Analyzing the Wild-West of Interrater Agreement in Affective Content Analysis on Text

A SYSTEMATIC LITERATURE REVIEW

1. BACKGROUND

- Human-computer interaction can peak if systems consider that humans' decisions are also influenced by their emotions
- *Affect* - umbrella term for all unconscious emotional experiences [1]
- *Text Affect Content Analysis* identifies the emotional state conveyed through written input [2]
- Affective models can **use manually labeled corpus** for training ground truth

VAGUE TERMS + HUMAN SUBJECTIVITY ≠ UNIFORMITY

- **No standard procedure** for conducting annotation
- **Interrater agreement (IRA)** is used to calculate consistency between labels
- Method of **computing IRA at researchers' discretion** from a large variety: *Scott's π*, *Krippendorff's α*, *Cohen's κ*, *Fleiss' κ*, % of full agreement, *ANOVA*, etc.

2. RESEARCH QUESTIONS

"How does interrater agreement influence the performance of text affect prediction models?"

- Targeted affective states
- Affect representation schemes
- Annotation process
- Trends in computing IRA
- Link between representation scheme and agreement
- Link between IRA computation method and performance

3. METHODOLOGY

- Systematic literature review following **PRISMA 2020** [3]
- **Literature databases**: *Scopus*, *Web of Science*, *IEEE Xplore*, *ACM Digital Library*
- Mainly focus on retrieving data about text corpus, not observing them put to use in learning models
- **Included papers**: corpus designed for text affect prediction, manually-labeled records
- **Excluded papers**: sentiment analysis datasets, multimodal affect prediction, non-English literature
- Relevant literature found at intersection of topics *Affect prediction*, *Text*, *Dataset* and *Manual labeling*
- **Feasibility constraints applied before manual screening:** by ease of scanning (removed *ACM Digital Library*), by keywords (only for *Scopus*), by field of expertise (only *Computer Science* publications)
- Data extraction performed during full-text filtering if paper isn't excluded
- **Results**: 41 papers included & 10 data papers manually-added as some literature only specified using a published dataset (Figure 1)

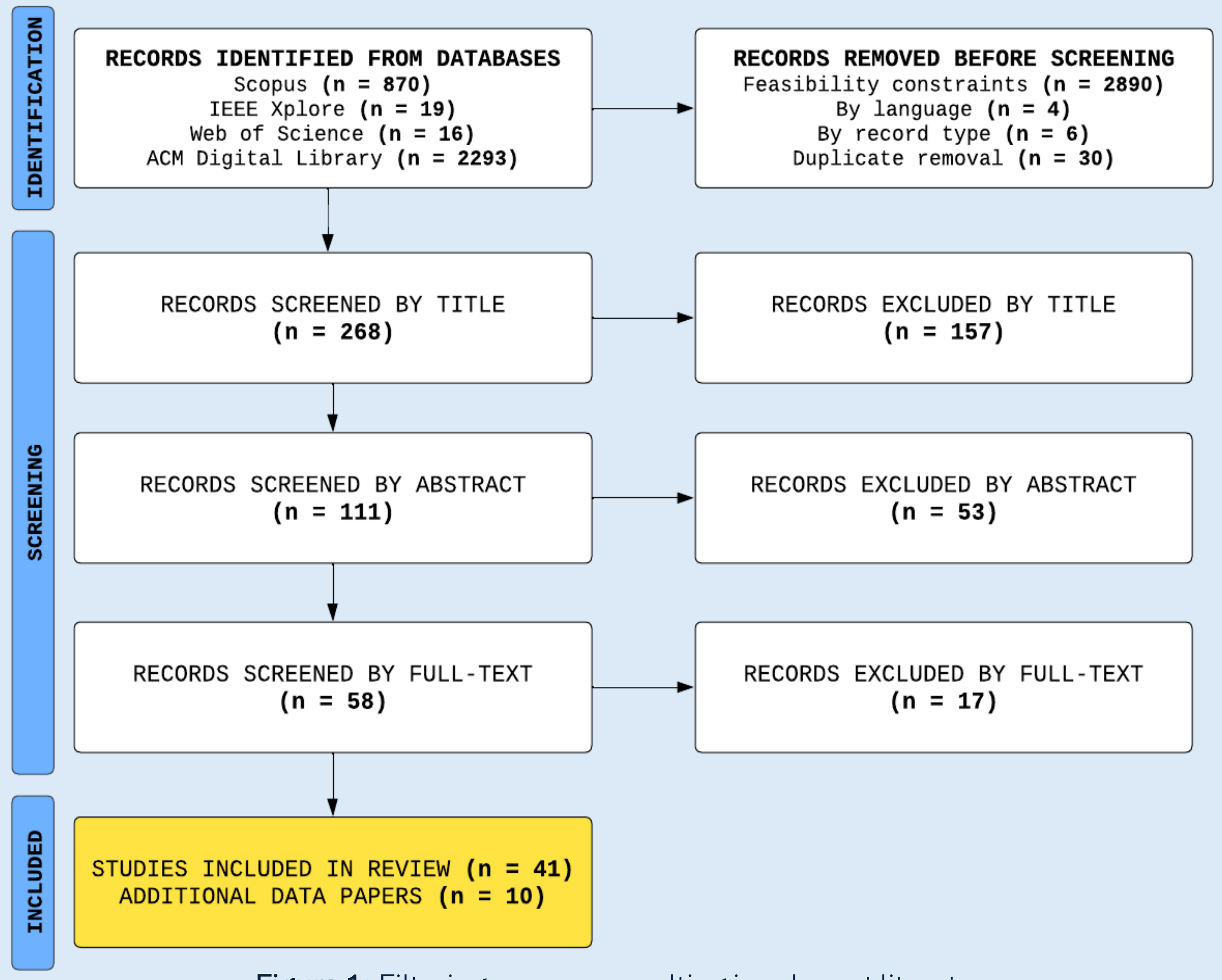


Figure 1: Filtering process resulting in relevant literature

4. RESULTS

1. Targeted affective states

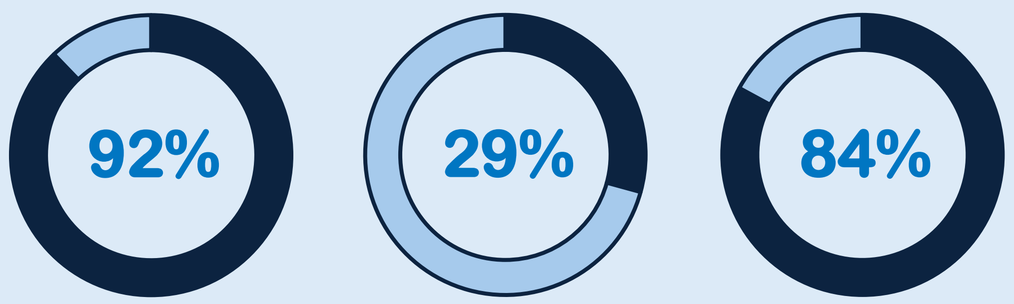
- 92% datasets portray emotions, 5% mood and 3% opinions
- 7% of datasets explain the perspective of emotions the annotators label (i.e. the general public's)

2. Affect representation schemes

- Categorical representation most used, even for dimensional approaches that augment discrete labels
- Large variety of sets of labels that convey similar emotions
- Most common labels represent negative emotions
- Variations of Ekman's Basic Emotions are most observable
- Justification behind ARS only for non-expressive labels (No emotion, Neutral, Other)

3. Annotation process

- 3 annotators most common (31%), only 7% are self-reports
- Actions towards facilitating agreement always done before annotating, most commonly to avoid random labeling



Compute IRA | Record high agreement | Facilitate agreement

4. Popularity of IRA calculation

- Miscellaneous statistical methods not commonly used for agreement were more prevalent until 2017 (Figure 2)
- From 2018 onwards, Fleiss' κ is steadily the most preferred, despite literature stating otherwise [4, 5]

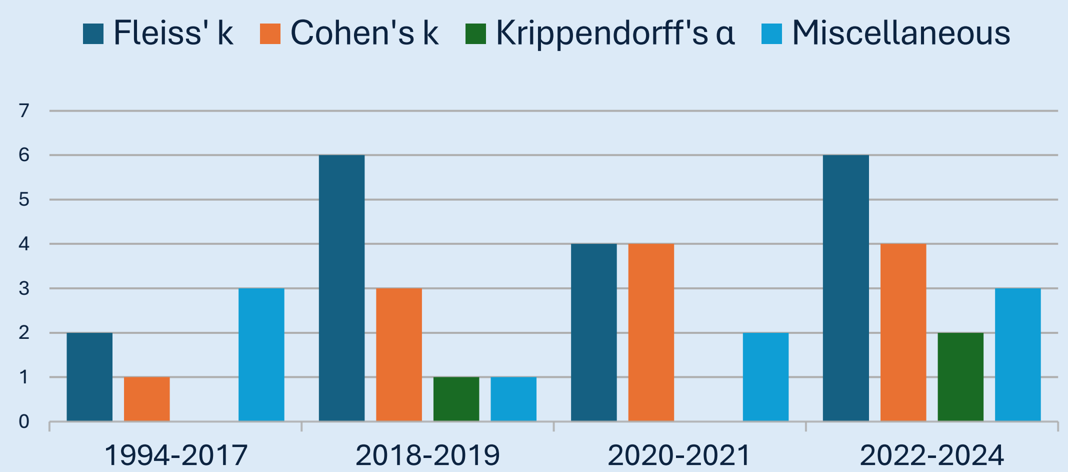


Figure 2: Popularity of IRA computation methods spanning 30 years

5. Relationship between ARS & agreement

- Ekman's basic emotions tend to lead to high agreement
- Neutral option doesn't necessarily increase agreement

6. Relationship between performance & schema

- Benchmarks are not presented by all data papers
- F-1 score varies between implementation of models trained on the same dataset
- No definitive conclusion due to poor data representation

5. DISCUSSIONS

- Usually, not more than 7 annotators employed for labeling
- Corrective actions not taken when annotation is concluded with low agreement
- Reporting agreement level might be misleading when on a subjective quality label is provided without any numerical metric computed

Limitations

- Time constraints led to removal of 2293 papers & didn't allow for investigating relationship between IRA computation and model performance
- Text has various degrees of expressivity when transmitting information
- Researchers have different interpretations of a "good" agreement level

6. CONCLUSIONS & FUTURE WORK

- Agreement enhancing techniques are considered & IRA is computed
- Methods of computing IRA have become more uniform, mostly using *Cohen's κ*, *Fleiss' κ and Krippendorff's α* and not generic statistical heuristics
- Annotating is still a chaotic procedure performed with a variety of settings

Future work

- Propose standard procedure for annotation
- Complete study with model performance analysis and possibly compare with influence of manual labeling for multimodal affect prediction systems

RESULTS DATA

REFERENCES

[1] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, 2014;

[2] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020;

[3] Page M J, McKenzie J E, Bossuyt P M, Boutron I, Hoffmann T C, Mulrow C D et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews *BMJ* 2021;

[4] Dewey M. E. (1983). Coefficients of agreement. *The British journal of psychiatry : the journal of mental science*, *143*, 487–489. https://doi.org/10.1192/bjp.143.5.487

[5] Warrens, M. J. (2013). A comparison of Cohen's kappa and agreement coefficients by Corrado Gini. *International Journal of Research and Reviews in Applie d Sciences*, *16*, 345-351.

AUTHOR: Macsim Violeta Mara (vmacsim@student.tudelft.nl)

TUDelft

SUPERVISOR & RESPONSIBLE PROFESSOR: Bernd Dudzik