Emotion-Annotated Data in NLP: Perspectives on Recent Resources and Practices

Anna Koufakou **Elijah Nieves**
Department of Computing & Software Engineering, Florida Gulf Coast University, USA
akoufakou@fgcu.edu

Abstract

Automated Emotion Recognition in language is a challenging task that has attracted considerable attention especially in recent years. We present a summary of our findings and observations based on a thorough analysis of recent resources and related practices.

1 Introduction

The task of automated emotion recognition (AER) in language as an application of NLP has substantial real-world application potential from analyzing large corpora of literary texts to enhancing intelligent chatbots. AER deals with recognizing specific emotions in the text, such as anger, sadness, or joy. This is a challenging task, as opposed to simpler tasks such as sentiment analysis or polarity detection. For example, consider the statement "An old friend called out of the blue": this could convey surprise and joy, or nostalgia and sadness, or even anxiety, depending on the individual and their relationship with the friend.

When corpora are collected for AER, they must be annotated with emotions. There is great variability, for example, in annotation schemes (e.g., single vs. multi-label) or even which emotions are used. Annotating the data with emotions is inherently not a simple task: in the example we used above, we could consider the speaker's versus the reader's perspective, and the latter might differ from one person to another (Mohammad, 2022).

Despite these challenges, there has been a surge of AER research since 2018 (Plaza-del Arco et al., 2024). In this work, we summarize our findings and discuss our observations based on our earlier systematic study of recent emotion-annotated resources and related practices (2018 to now).[1]

[1]The in-depth study and analysis is part of a larger (journal) article. This extended abstract summarizes our findings citing example articles: they should be interpreted as representative examples and not as a comprehensive review.

2 Resources and related practices

Overview: Many of the datasets we found in our review came from social media such as Twitter/X (Mohammad et al., 2018; Barbieri et al., 2020) or Reddit (Demszky et al., 2020). There were a few corpora containing self-written statements or essays (Kleinberg et al., 2020; Troiano et al., 2019). Other sources include TV scripts (Hsu et al., 2018), news headlines (Oberländer et al., 2020), movie subtitles (Öhman et al., 2020) and literary narratives (Liu et al., 2019). Related to topics, certain corpora explore reactions to news (Tafreshi et al., 2021; Huguet Cabot et al., 2021) or COVID-19 (Yang et al., 2020). Regarding emotions, most datasets feature a few basic emotions, with distributions varying across datasets.

Lack of Resources/Awareness: Based on our study, there is currently no comprehensive resource (repository) that encompasses all available emotion-annotated corpora. Existing well-known ML/NLP data repositories such as Hugging Face contain only a handful of emotion-annotated text corpora, most of which are lacking documentation or even citation/author. Regarding studies and surveys, in 2018, Oberländer and Klinger (2018) conducted a unified framework and analysis of 14 corpora. Their work was shared online to allow comparisons of the corpora. Recent surveys we reviewed, e.g. (Deng and Ren, 2021; Kusal et al., 2023), do not cover many of the datasets we discovered in our study. Recently, Plaza-del Arco et al. (2024) reviewed over 150 ACL papers and offered a detailed overview of trends and gaps, aligning with many of our findings.

It should be noted that certain datasets are well-known in the NLP community: e.g. *GoEmotions* (Demszky et al., 2020) and *TweetEval* (Barbieri et al., 2020) based on *Affect in Tweets* (Mohammad et al., 2018), each cited more than 700 times (per Google scholar, Aug. 2024). These are also examples of the sweeping trend we observed: most data were collected from social media or online forums, with each record containing a short post or comment. On the other hand, we identified very different corpora that are not as well-known, e.g. *Real World Worry Waves Dataset (RW3D)* (van der Vegt and Kleinberg, 2023) with UK survey responses (essays and emotion self-ratings) related to COVID-19 over 3 years, and *EmotionArcs* (Öhman et al., 2024) with emotional arcs from over 9,000 English novels. Sharing available resources in a centralized repository would improve resource awareness and standardization, with great potential to advance research in this field, for example benchmarking efforts.

Issues with existing resources: As we mentioned earlier, most data come from social media or online forums. The language on these platforms includes misspellings, emojis, abbreviations, etc., which makes it difficult to parse or follow. The collected datasets do not represent linguistic patterns or human communication outside of these platforms. Additionally, it is known that many of the social media platforms are biased: for example, platforms such as Reddit are not representative of a diverse population, and are instead biased towards young male users (Demszky et al., 2020). The vast majority of data is in English, a common issue in NLP. There are efforts to present datasets in other languages, and we also found some datasets in multiple languages, e.g. *Universal Joy* (Lamprinidis et al., 2021) with anonymized Facebook posts in 18 languages. A recent work explored emotion detection in low/moderate-resource languages with transfer learning (Tafreshi et al., 2024). Specifically for emotion detection, one should consider linguistic and cultural differences (De Bruyne, 2023).

There is great variability in how data has been annotated with emotions. The collectors of the data have followed different emotion taxonomies borrowed from Psychology (Plaza-del Arco et al., 2024). Most datasets use basic emotions, usually relying on Ekman taxonomy from the 1970s (Ekman, 1992), and to a lesser extent Plutchik (1984). The Ekman representation links emotions to facial expressions or similar: this has been challenged by Barrett (2017), who also highlighted the need to consider context in interpreting emotions. De Bruyne (2023) discussed how these basic emotions are too broad to be realistic and thus useful. At the same time, using a large number of detailed emotions might increase the overlap of the emotions, which is harder (more confusing) for annotators (Öhman, 2020). A few works used fine-grained emotion labeling, e.g. (Demszky et al., 2020; Imran et al., 2022).

Annotation practices also reveal challenges. For example, there is a lack of detailed or uniform reporting on annotator demographics and training (Plaza-del Arco et al., 2024), and great variability in number of annotators, metrics for inter-annotator agreement etc., as also observed by Stajner (2021). Only a couple of works followed data statements proposed by Bender and Friedman (2018). Given the subjectivity of the AER task, it could be beneficial to consider disagreements in annotations (Basile et al., 2021). Finally, ethical considerations, as presented by Mohammad (2022), emphasize the importance of having tasks that are clearly defined and also considering how emotions are expressed and perceived by different individuals.

Limited Interdisciplinary Work: Research integrating NLP with Humanities, Psychology and Social Sciences remains limited. McGillivray et al. (2020) and Öhman et al. (2023) focused on Digital Humanities, Behnke et al. (2023) on Psychology/CS perspectives, and Demszky et al. (2023) discussed LLMs in Psychology. Interdisciplinary collaboration is vital for refining emotion models and developing culturally and contextually relevant applications.

3 Conclusions

Based on our in-depth exploration of recent AER resources and related practices, we summarized our findings including issues and challenges. By exploring strategies for overcoming them, we can promote a more integrated approach that enhances the effectiveness and applicability of AER techniques. Towards that goal, we recently presented a unified framework built from several emotion-annotated corpora, with which we conducted initial benchmarking experiments (Koufakou et al., 2024). We shared our code and data information online.[2] Building on our in-depth analysis of related datasets, we are currently curating a repository that includes details and comparisons, while also seeking ways to engage with researchers beyond the NLP community.

[2]https://github.com/a-koufakou/EmoDetect-Unify

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online.

Lisa Feldman Barrett. 2017. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 1:23.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Maciej Behnke, Stanislaw Saganowski, Łukasz D Kaczmarek, and Przernysław Kazienko. 2023. Emotions studied by computer scientists and psychologistsa complementary perspective. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 206–211. IEEE.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Jiawen Deng and Fuji Ren. 2021. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. them: A dataset of populist attitudes, news bias and emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945, Online. ACL.

Mia Mohammad Imran, Yashasvi Jain, Preetha Chatterjee, and Kostadin Damevski. 2022. Data augmentation for improving emotion recognition in software engineering communication. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring emotions in the covid-19 real world worry dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Anna Koufakou, Elijah Nieves, and John Peler. 2024. Towards a new benchmark for emotion detection in NLP: A unifying framework of recent corpora. In *Proceedings of the 2nd GenBench workshop on generalisation (benchmarking) in NLP*.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, pages 1–87.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. Dens: A dataset for multi-class emotion analysis. 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 6293–6298.

Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz. 2020. Digital humanities and natural language processing:je taime... moi non plus. *Digital Humanities Quarterly*, 14(2).

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

Laura Ana Maria Oberländer 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.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *DHN post-proceedings*, pages 134–144.

Emily Öhman, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer Nielbo. 2024. Emotionarcs: Emotion arcs for 9,000 literary texts. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 51–66.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552.

Emily Öhman, Michael Piotrowski, and Mika Hämäläinen. 2023. The great digital humanities disconnect: The failure of dh publishing. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 132–137.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Sanja Stajner. 2021. Exploring reliability of gold labels for emotion detection in Twitter. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1350–1359, Held Online. INCOMA Ltd.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online. ACL.

Shabnam Tafreshi, Shubham Vatsal, and Mona Diab. 2024. Emotion classification in low and moderate resource languages. *arXiv preprint arXiv:2402.18424*.

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.

Isabelle van der Vegt and Bennett Kleinberg. 2023. A multi-modal panel dataset to understand the psychological impact of the pandemic. *Scientific Data*, 10(1):537.

Qiang Yang, Hind Alamro, Somayah Albaradei, Adil Salhi, Xiaoting Lv, Changsheng Ma, Manal Alshehri, Inji Jaber, Faroug Tifratene, Wei Wang, et al. 2020. Senwave: Monitoring the global sentiments under the covid-19 pandemic. *arXiv preprint arXiv:2006.10842*.