The SuperEmotion dataset

Enric Junqué de Fortuny
Managerial Decision Making
IESE, Barcelona, Spain
ejunque@iese.edu

Abstract

Despite the wide-scale usage and development of emotion classification datasets in NLP, the field lacks a standardized, large-scale resource that follows a psychologically grounded taxonomy. Existing datasets either use inconsistent emotion categories, suffer from limited sample size, or focus on specific domains. The SuperEmotion dataset addresses this gap by harmonizing diverse text sources into a unified framework based on Shaver's empirically validated emotion taxonomy, enabling more consistent cross-domain emotion recognition research.

1 Introduction

This report describes the SuperEmotion dataset, the world's largest Shaver compliant emotion dataset for natural language processing. We developed the SuperEmotion dataset by aggregating multiple existing emotion datasets and remapping categories into Shaver's primary emotion classes. The dataset encompasses 552,821 samples labeled across the primary emotions: *joy, sadness, anger, fear, love*, and *surprise* as well as a *neutral* category. The source datasets integrated into this collection include:

- **MELD** (Poria et al., 2019): A multimodal dataset for emotion recognition in conversations. The text comes from scripts and transcribed dialogues from the TV show *Friends*, capturing multi-party spoken interactions with speaker context.
- **GoEmotions** (Demszky et al., 2020): A large-scale dataset of carefully filtered English Reddit comments annotated with 27 fine-grained emotions plus neutrality. It is well-suited for studying subtle affect in user-generated social media content.
- **TwitterEmotion** (Saravia et al., 2018): A Twitter-based dataset developed for context-aware emotion recognition, containing tweets labeled across multiple emotion categories. The short, informal text reflects real-world online communication.
- **ISEAR** (Scherer, 1997): The International Survey on Emotion Antecedents and Reactions consists of structured questionnaire responses. Participants described personal experiences that triggered one of seven basic emotions, yielding formal first-person narratives.
- **SemEval** (Mohammad et al., 2018): A dataset from the Semantic Evaluation series, particularly Task 1 on Affect in Tweets. It contains tweets annotated for emotional intensity and valence, providing rich insight into affective language in short-form, real-time social media.
- **CrowdFlower** (Van Pelt and Sorokin, 2012): A crowdsourced dataset of tweets labeled with emotion categories such as sadness, joy, and anger. The data was used to evaluate how well laypeople agree on emotion labels in noisy, informal text.

By consolidating these resources, the SuperEmotion dataset addresses class imbalances and provides a more diverse and extensive foundation for training robust emotion classification models following

a well-understood taxonomy. Detailed statistics, data distribution, and guidelines for accessing the dataset are provided in this report.

2 Dataset Construction

2.1 Dataset Composition

The SuperEmotion dataset is constructed by aggregating multiple publicly available emotion classification datasets. Table 1 summarizes the primary datasets integrated into this collection.

Dataset	Train	Validation	Test	Classes
MELD	9,989	1,109	2,610	7
TwitterEmotion	16,000	2,000	2,000	6
ISEAR	416,809	-	-	6
GoEmotions	43,410	5,426	5,427	28
CrowdFlower	39,998	-	-	13
SemEval	6,634	872	3,184	11
Total	532,840	9,407	13,221	
SuperEmotions	439,361	54,835	58,625	7

Table 1: Overview of datasets aggregated in the SuperEmotion dataset, including the number of emotion classes in each source.

2.2 Preprocessing and Quality Control

To ensure consistency across diverse source formats, we applied several preprocessing steps:

1. **Text Normalization**: We standardized text formatting by removing excessive whitespace, normalizing unicode characters, and ensuring consistent punctuation placement.
2. **Deduplication**: We identified and removed exact duplicate texts to prevent test set leakage and avoid biasing the dataset toward redundant patterns.
3. **Data Splits**: For datasets without predefined splits, we created stratified train/validation/test partitions (80%/10%/10%) to preserve label distribution across splits.
4. **Metadata Preservation**: While harmonizing the emotion taxonomy, we retained source information for each example, enabling analysis of domain-specific patterns and potential biases.

2.3 Shaver's Taxonomy

Emotion classification has been approached in a variety of ways in psychology, linguistics, and affective computing. Some models, such as Eckman's framework (Eckman, 1972), define a small set of discrete universal emotions (e.g., anger, fear, joy, sadness, surprise, disgust), primarily grounded in facial expressions. Others, like Russell's circumplex model (Russell and Barrett, 1999), represent emotions in a continuous space defined by valence and arousal dimensions. Plutchik's wheel of emotions (Plutchik, 1980, 2001) combines discrete emotions with intensity scaling and mixing of different primary emotions.

After careful consideration, we adopt the taxonomy proposed by Shaver et al. (1987), which provides a psychologically grounded, hierarchical classification of emotions based on empirical clustering of 135 emotion terms. Through free-listing, sorting, and similarity judgments, Shaver et al. identified six basic-level emotions (*love, joy, anger, sadness, fear,* and *surprise*[1]) which serve as cognitively salient and linguistically frequent prototypes in English.

[1]Note: We include surprise even though Shaver gave it less important as it appears in all datasets and is clearly of interest to the NLP community.

We selected Shaver's taxonomy because the taxonomy is lexically grounded—built from natural language emotion terms - which aligns well with the input modality of most NLP systems and simplifies annotation. Second, it balances granularity and coverage, capturing sufficient emotional nuance for robust modeling while remaining tractable for supervised learning and being robust for potential expansions of the dataset.

2.4 Label Harmonization

To align heterogeneous label sets from different source datasets, we mapped related labels into these six core categories using Shaver's original definitions and keywords (Shaver et al., 1987).

When handling ambiguous cases during label harmonization, we prioritized semantic similarity to Shaver's prototypical emotion terms. For instance, labels like 'optimism' could reasonably map to either *joy* or be excluded as a distinct anticipatory state; we assigned it to *joy* based on its positive valence and association with pleasant feelings. Similarly, *confusion* was mapped to *surprise* rather than *fear* based on its cognitive rather than threat-oriented nature. Where a source dataset used idiosyncratic labels without clear mapping to Shaver's categories (e.g., *empty*, *curiosity*), we excluded these examples rather than forcing them into potentially inappropriate categories. A complete mapping is provided in Table 2.

Emotion	Source Labels
Anger	anger, anger, anger, anger, anger, anger, annoyance, disapproval, disgust, disgust, disgust, hate
Fear	fear, fear, fear, fear, fear, nervousness, worry
Joy	amusement, enthusiasm, excitement, fun, happiness, joy, joy, joy, joy, joy, optimism, optimism, pride, relief, relief
Love	admiration, approval, caring, desire, gratitude, love, love, love, love, love, trust
Sadness	disappointment, embarrassment, grief, pessimism, remorse, sadness, sadness, sadness, sadness, sadness, sadness
Surprise	confusion, realization, surprise, surprise, surprise, surprise, surprise, surprise
Neutral	boredom, neutral, neutral, neutral
Dropped	anticipation, curiosity, empty

Table 2: Emotion label breakdown mapped to Shaver's categories. Dataset sources are color-coded as follows: GoEmotions, ISEAR, MELD, Crowdflower, SemEval, TwitterEmotion.

We also introduce a **Neutral** category to capture instances with low or absent emotional valence. Labels that lacked a clear conceptual or empirical correspondence with Shaver's taxonomy such as *anticipation*, *curiosity*, and *empty*—were excluded from the final label set, and their associated examples were removed from the dataset.

Source	Neutr.	Surp.	Fear	Sad.	Joy	Anger	Love	Dropped	Total
MELD	6,436	1,636	358	1,002	2,308	1,968	0	0	13,708
TwitterEmotion	0	719	2,373	5,797	6,761	2,709	1,641	0	20,000
ISEAR	0	14,972	47,712	121,187	141,067	57,317	34,554	0	416,809
GoEmotions	17,772	4,295	929	4,032	7,646	7,838	15,557	2,723	60,792
Crowdflower	8,817	2,187	8,457	5,165	9,270	1,433	3,842	827	39,998
SemEval	0	566	1,848	3,607	5,065	4,780	1,757	1,527	19,150
Total	33,025	24,375	61,677	140,790	172,117	76,045	57,351	5,077	570,457

Table 3: Distribution of samples across emotion categories and datasets in the SuperEmotion dataset. All values reflect merged train/validation/test splits, including samples that were mapped to `Dropped`.

Table 3 summarizes the distribution of emotion labels across all source datasets after mapping to the unified taxonomy. Note that counts may exceed those in Table 1 due to the multi-label nature of the task, where a single text sample can be annotated with multiple emotions. Column **X** indicates

the number of observations removed due to conceptual incongruence with Shaver's taxonomy (e.g., labels like *anticipation* or *curiosity*).

Figure 1 visualizes the overlap between emotion categories, showing how frequently different emotions co-occur within the same instance. More precisely, it visualizes the conditional probability of observing emotion Y given emotion X is present, expressed as a percentage. The asymmetric nature of the matrix reflects that $P(Y|X) \neq P(X|Y)$ for most emotion pairs. For example, 5.5% of texts annotated as *joy* also contain *love*, while only 1.8% of texts labeled as *love* also contain *joy*.

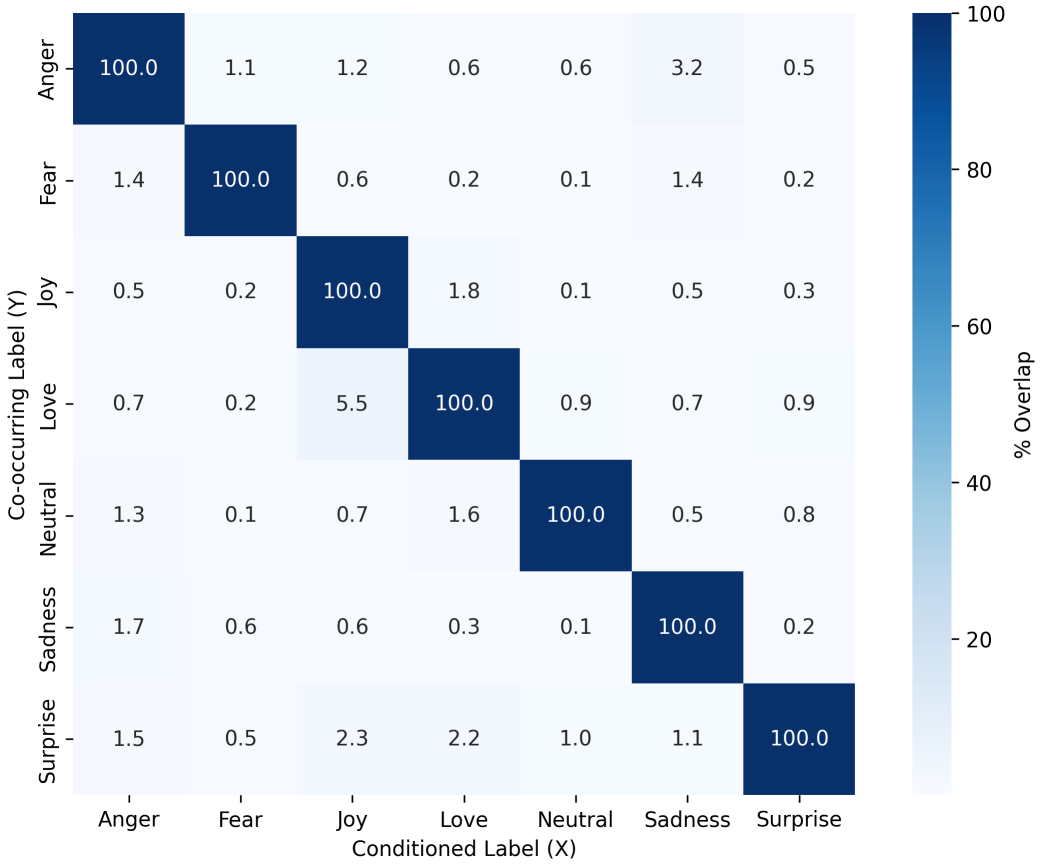

Figure 1: Label co-occurrence heatmap showing the percentage of samples annotated with emotion X (X-axis) that are also annotated with emotion Y (Y-axis), denoted as $P(Y \mid X) = \frac{\#(X \cap Y)}{\#(X)}$. Diagonal values are always 100%, as each annotation trivially co-occurs with itself.

3 Final Considerations

3.1 Data Accessibility

The dataset is publicly available on Hugging Face at the following URL: `https://huggingface.co/datasets/cirimus/super-emotions`. Alternatively, users can download the dataset using the Hugging Face `datasets` library in python:

```
from datasets import load_dataset
dataset = load_dataset("cirimus/super-emotion")
```

When the dataset is updated, we will update this versioned repository while keeping a copy of the old data for archival purposes. The dataset described in this document is version 1.

3.2 Limitations and Ethical Considerations

While the SuperEmotion dataset offers advantages in scale and emotional coverage, several limitations should be acknowledged:

1. **Dataset Biases**: The aggregation inherits biases from source datasets, including potential cultural and demographic skews in emotion expression and annotation. Moreover, the dataset introduces a new sampling bias due to ISEAR's dominance in size. Researchers may therefore want to stratify across datasets to mitigate this problem.
2. **Contextual Limitations**: Many samples lack conversational context that might influence emotion interpretation.
3. **Annotation Quality**: Source datasets employed different annotation methodologies and annotator populations, potentially introducing inconsistencies in label quality.
4. **Privacy Considerations**: Though all datasets excluded personally identifiable information, users should remain cautious when deploying models trained on this data in applications involving sensitive contexts.

We encourage researchers to consider these limitations when developing emotion recognition systems and to supplement with domain-specific data when appropriate.

4 Conclusion

By creating the SuperEmotion dataset, we contribute to emotion recognition research by harmonizing multiple existing datasets into a consistent taxonomy based on Shaver's psychological framework. By addressing class imbalances and providing a diverse text collection, and reducing taxonomical inconsistencies, this resource enables more robust emotion classification models.

The harmonized labels, clear documentation, and easy accessibility through Hugging Face facilitate immediate application in natural language processing tasks. We anticipate this dataset will support advances in affective computing, human-computer interaction, and sentiment analysis.

Future work may expand this collection with additional data, especially in secondary dimensions of Shaver's aspects.

References

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. arXiv:2005.00547 [cs].

Eckman, P. (1972). Universal and cultural differences in facial expression of emotion. In *Nebraska symposium on motivation*, volume 19, pages 207–284. University of Nebraska Press Lincoln.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350. Publisher: JSTOR.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. arXiv:1810.02508 [cs].

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819. Place: US Publisher: American Psychological Association.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, 73(5):902–922.

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086.

Van Pelt, C. and Sorokin, A. (2012). Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 765–766, Scottsdale Arizona USA. ACM.