# Affect, Body, Cognition, Demographics, and Emotion: The ABCDE of Text Features for Computational Affective Science

**Jan Philip Wahle**[*†], **Krishnapriya Vishnubhotla**[*‡], **Bela Gipp**[†], **Saif M. Mohammad**[‡]
[†]University of Göttingen, Germany; [‡]National Research Council Canada
wahle@uni-goettingen.de; vkpriya@cs.toronto.edu, saif.mohammad@nrc-cnrc.gc.ca
[*]equal contribution

Dataset hf.co/datasets/jpwahle/abcde
Code github.com/jpwahle/abcde

**Abstract**

Work in Computational Affective Science and Computational Social Science explores a wide variety of research questions about people, emotions, behavior, and health. Often they make use of language data that is first labeled with relevant information such as the use of emotion words and age of the speaker. Even though many resources and algorithms exist to enable such labeling, finding and using them is still a substantial impediment, especially to practitioners in fields outside of computer science. Here, we present the ABCDE dataset ("Affect, Body, Cognition, Demographics, and Emotion"), a large-scale collection of over 400 million released text instances from social media, blogs, books, and AI-generated sources, annotated for a number of features relevant to computational affective and social science. ABCDE facilitates inter-disciplinary research in wide range of fields, including affective science, cognitive science, the digital humanities, sociology, political science, and computational linguistics.

**Keywords:** computational social science, computational affective science, scientometrics

## 1. Introduction

Language is a rich medium of self-expression and communication. It is a product of historical, cultural, embodied, and cognitive processes. Consequently, computational approaches to the analysis of large text datasets have become a powerful tool for studying feelings, thought, body/health, and behavior of individuals and populations.

In the domain of Computational Affective Science (CAS), research has shown that language use is strongly associated with, and indicative of, mental states and cognitive processes in the human brain (Guntuku et al., 2017). When we are sad, we express ourselves using words associated with low valence (negative sentiment), or even subtle linguistic signals such as shorter, more terse utterances. When we lack confidence, we tend use more qualifiers such as *"I think"*, *"probably"*, and *"right?"*. And so on. Linguistic properties of utterances are thus immensely useful signals of our mental models and processes, as well as how we as humans perceive ourselves and the world around us. At the same time, language is only an indirect trace of the underlying states we care about. Many mental states and cognitive processes are latent and must be inferred from behavior and context. Thus while it is difficult to determine the exact mental state of an individual from an utterance (something we do not recommend pursuing automatically anyway), the power of language analysis lies in bringing forth trends and patterns of mental states of whole populations (aggregate-level analysis). Language use remains a natural and scalable behavioral trace that can be used to infer patterns of relative differences and change at an aggregate level. For example, how has the use of emotion words by a population

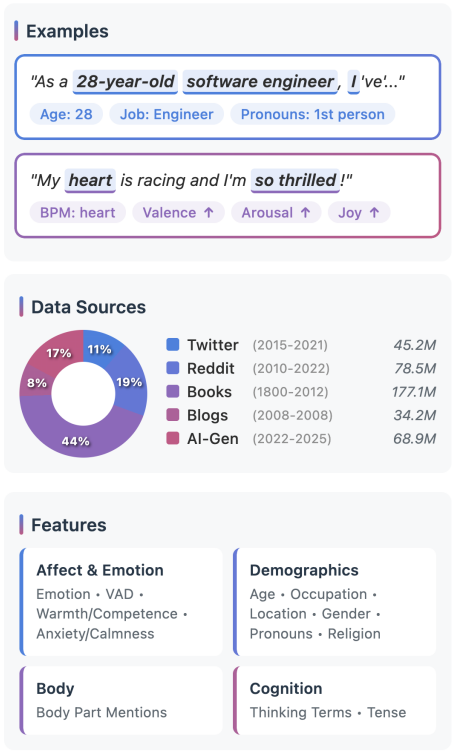


Figure 1: Overview of the ABCDE dataset.



*Proceedings of Computational Affective Science (CAS) @ LREC 2026*, pages 217–234
16 May 2026. 

changed with time and what psychological changes that indicates; how a cohort of people uses warmth and competence words when posting about different social groups; what moral frames are invoked when talking about controversial issues (such as immigration), etc.

Similarly, in the broader domain of social sciences, insights derived from textual data can complement those derived from qualitative methodologies like surveys and self-reports. Computational methods have the added advantage of enabling research using a much larger and more diverse pool of data when compared to lab studies with smaller groups of participants, thereby increasing the reliability of measurements. For example, Computational Social Science (CSS) researchers have used social media data to study mechanisms of influence and information diffusion in social networks (Goel et al., 2016), monitor public sentiment and stance towards specific entities and topics (Dodds et al., 2011), identify changes in the contextualized semantic meanings of word types in different communities (Lucy and Bamman, 2021), etc.

Another line of research that is of increasing interest to linguists, cognitive scientists, and computer scientists alike is the studying of AI-generated text. Machine learning based training of multi-billion parameter neural networks has resulted in Large Language Models that interface with and communicate in natural language with surprising fidelity (Achiam et al., 2023). Consequently, researchers are looking at the outputs of these models to understand statistical associations and biases that are encoded in them, particularly as LLMs take on a larger role in the data science pipeline as de-facto models for classification, labeling, and synthetic data generation (Ziems et al., 2024; Farrell et al., 2025). Researchers are also analyzing the "reasoning traces" or chains-of-thought (CoTs) of these models to better understand the artificial "cognitive" reasoning processes in these models, and how they compare with human processes (Korbak et al., 2025).

All of this research that makes use of "language as data" relies on three key tasks as a pre-requisite to subsequent statistical analyses: collecting the textual data, tagging it with associated metadata (extra-linguistic context), and the measurement of features of interest (quantifying, for example, sentiment from tweets). The dynamic nature of pipelines for data collection on the web, as well as the rapid pace of development of computational models of measurement for various constructs that are of interest in the social sciences, make this a technically challenging process for most researchers, especially those outside of computer science. Further, even for those comfortable with natural language processing, compiling and annotating a large number of diverse datasets and features is time and labor intensive.

In this work, we compile and release **ABCDE**, a collection of text datasets automatically tagged with metadata associated with speaker demographics, and labeled for 136 lexical features that are of broad utility for researchers in the affective and social sciences. The selected datasets (aggregated from various primary sources) span multiple domains, genres, and time-periods. We group our text features into five thematic sets: **A**ffect, **B**ody, **C**ognition, **D**emographics, and **E**motion.

Our primary motivation in creating this dataset is to unify textual features that encode aspects of behavior and cognition that are of interest to researchers in the affective sciences. While emotion and affect have been prominent topics of study here, recent advances in embodied cognition emphasize how bodily awareness and interactions with the environment shape many of these cognitive functions, and have downstream correlations with mental, emotional, and physical health (Wu et al., 2025). We therefore include body part mentions as a core feature of our dataset, allowing researchers to quantify patterns at the intersection of bodily mentions and emotions/affect. With Demographic information, we also consider features that encode *social context*, in addition to linguistic and lexical features of the text itself. This is a crucial variable for most research in the affective and social sciences, where concepts like emotion and cognition are not constant, stable traits, but demonstrate significant variation with demographic features like age, within and across population sub-groups — in fact, it is this variation that is often the phenomenon of interest (Hoemann et al., 2025; Gutchess and Rajaram, 2023). Age-related differences have been reported in the arousal structure of emotion concepts (Trnka et al., 2022), and age and culture can jointly shift cognitive performance and style (Na et al., 2017).

We make our resource publicly available for open access and use by researchers. We hope the **ABCDE** dataset will facilitate inter-disciplinary research in wide range of fields, including affective science, cognitive science, the digital humanities, sociology, political science, and NLP.

## 2. Related Work

The past several decades of research in computational linguistics and natural language processing has led to the development of tools that can be used to extract rich linguistic information from textual data, such as StanfordCoreNLP (Manning et al., 2014), NLTK (Bird et al., 2009), and spaCy[1]. However, these are aimed largely at researchers

[1] https://spacy.io/

with familiarity and expertise in these domains of research, with rich programming experience, and target features that are useful for linguistic analysis, like part-of-speech tags and dependency relations. In the social sciences, by contrast, researchers are generally interested in quantifying more abstract, higher-level features of text, like sentiment or emotion intensity, political leaning, satire, literary quality, etc. (Licht et al., 2025)

There exists a small but growing set of accessible text analysis toolkits that have a narrower focus on specific domains of research. The Cornell Conversational Analysis Toolkit (Chang et al., 2020) is a collection of datasets and computational scripts that enable research into aspects of social networks, conversational dynamics, and the sociolinguistics of online interactions. GutenTag is an NLP-driven tool for Digital Humanities research in particular, with a web-based interface that automatically extracts several features of interest from literary texts (Brooke et al., 2015). In the Affective Sciences domain, packages like VADER (Hutto and Gilbert, 2014) and Textblob[2] are widely used for sentiment analysis. The Emotion Dynamics toolkit (Vishnubhotla and Mohammad, 2022) provides scripts to quantify patterns of change in emotional expression in textual utterances over time.[3]

A prime example of a text processing toolkit that has seen broader use in the social, affective, and cognitive sciences, and the digital humanities, is LIWC (Linguistic Inquiry and Word Count), a proprietary software that quantifies several psychometric properties of words to facilitate studying the links between language, cognition, and psychology (Boyd et al., 2022).

We position **ABCDE** as not just a feature extraction tool, but a much more comprehensive repository of data mapped to metadata and pre-computed features. We compile, clean, and annotate multiple text datasets from diverse domains that are of broad interest to affective and cognitive science researchers, including social media, blogs, and AI-generated text. By pre-computing features for these datasets, our resource lowers the technical barrier of entry for many researchers, and also encourages reproducibility of empirical research in the field. The **ABCDE** dataset therefore functions as a much more thorough and standardized starting point, on top of which researchers can directly apply statistical methods of analysis in order to answer various research questions.

## 3. The ABCDE Dataset

We compiled a large-scale, longitudinal collection of textual datasets from various primary sources, spanning data from social media platforms, books, and blogs. Additionally, we compiled a selection of AI-generated text, including human–LLM conversational data, LLM reasoning traces, preference datasets, and datasets from AI-generated text detection tasks. We then computed and recorded the text features enumerated in Section 3.2 for every text *instance* in each of these datasets. Note that a text instance can be defined at different levels of granularity – features can be recorded at the sentence-level (with each sentence in turn linked to the original Reddit post, blog post, AI-generated story, etc.), or for equal-sized chunks of tokens from each dataset, among others. Our definition of what constitutes a text instance differs from one dataset to another, and is enumerated in Section 3.1 alongside the dataset descriptions. In Section 3.2, we expand on our annotated features and the methods used to quantify them given a text instance.

### 3.1. Data Sources

We annotated datasets spanning millions of records from Twitter, Reddit, blogs, books, and AI-generated content:

**Twitter (TUSC, 2015–2021)**: 45.2M geolocated tweets sourced from the TUSC dataset (Vishnubhotla and Mohammad, 2022). Each tweet is considered an instance.

**Reddit (Pushshift, 2010–2022)**: 78.6M posts crawled from Reddit via Pushshift archives hosted by the Internet Archive (Baumgartner et al., 2020). We remove posts that are marked as adult (over_18), promoted, containing images or videos, and having fewer than 5 or more than 1,000 words; each post constitutes a text instance.

**Books (Google, 1800–2012)**: 177.1M 5-gram occurrences (1.74M unique 5-grams) from the English Fiction subset of the Google Books Ngram Corpus (version 20120701) (Google Inc., 2012). Each 5-gram is treated as an instance. While these instances are short, they remain useful for large-scale frequency and longitudinal analyses.

**Blogs (Spinn3r, 2008)**: 34.2M personal blog entries from the ICWSM 2009 conference (Burton and Soboroff, 2009). Each blog post is an instance.

**AI-Generated Texts (Various, 2022–2025)**: 68.9M AI-completions from 15 distinct datasets, including conversational texts (WildChat-1M (Zhao et al., 2024), LMSYS-Chat-1M (Zheng et al., 2024), PIPPA (Gosling et al., 2023), HH-RLHF (Bai et al., 2022), Prism (Kirk et al., 2024), and APT (Wahle et al., 2022); persuasive essays (Anthropic-Persuasiveness (Durmus et al., 2024)); AI text detection datasets (M4 (Wang et al., 2024), MAGE (Li et al., 2024), LUAR (Soto et al., 2024)); reasoning traces (General Thoughts 430k (Reasoning, 2024), Reasoning Shield (Li et al., 2025), SafeChain (Jiang et al., 2025), STAR-1 (Wang et al.,

[2] https://textblob.readthedocs.io/

[3] https://github.com/Priya22/EmotionDynamics

2025)); and narratives (TinyStories (Eldan and Li, 2023)). For conversational datasets, we define the LLM response from a single user–chatbot interaction turn as a text instance (multi-turn conversations are split into multiple instances). Preference datasets generally comprise of a pair of LLM generations (in response to a user prompt) along with an indicator of the preferred generation; we take each generation separately to be an instance of AI-generated text. The AI-generated text detection datasets consist of LLM generations of varying lengths, intended to simulate human text in specific domains like news articles and social media posts; each generation is considered an instance. For reasoning traces, we disregard the final model output and only use the chain-of-thought text for each query as the instance. Instances are multi-feature. One text can activate several feature families, and some short texts may not activate any feature in a given family.

### 3.2. Annotated Features

We extract linguistic features through a set of lexicons, word lists, and regular expressions, organized along five dimensions relevant to computational affective science: Affect, Emotion, Body, Cognition, and Demographics. Word lists and lexicons as a measurement method offer several advantages, particularly in inter-disciplinary studies: ease-of-use, generalization power, reliability, adaptability, and interpretability, among others. Variation in word choices across populations and data subsets is also intertwined with multiple cognitive and social processes relating to language use (such as power dynamics, cultural and demographic factors, medium of communication, etc.). Appendix A.1 summarizes the released resources, and Appendix A.2 reports per-source text statistics.

Affect and Emotion features are characterized using word association lexicons; Body and Cognition features through a curated list of terms, compiled from multiple sources, that indicate a strong association with body and cognition-related activities; and Demographic features through hand-constructed regular expressions and heuristic rules. For the demographic feature "Occupation", for example, we use a regex to parse self-disclosure statements of the type "I am/work at/employed as *[occupation]*." from a user, and apply the extracted term to all text instances from that user. The demographic feature "Age", which refers to the age of the user at the time of posting, is computed by first using a similar regex for age or date-of-birth, and then applying a heuristic rule that combines this with the timestamp of the post.

For each feature dimension (say, valence), and each text instance (say, a tweet), we use the associated word-level lexicon to compute an aggregate instance-level score in multiple ways: the **average** valence score of the constituent terms, a binary **flag** indicating if a word from the valence lexicon is present in the instance, and the **count** of the number of lexicon terms present in the instance. The appropriateness of a particular aggregate measure for a feature will depend on the specifics of the research question being answered, and we leave this decision to the users. For example, length-normalized count features are more appropriate as a comparative indicator for texts of different lengths, rather than the average intensity score.

**Affect:** Affect refers to the fundamental neural processes that broadly determine and regulate internal experiences of emotion, mood, and feelings. Affective states are generally characterized along three principal dimensions: valence (scale of positive–negative), arousal (scale of active–passive), and dominance (scale of competent–incompetent / powerful–weak), together referred to as VAD, which form our feature dimensions for Affect. We match words against the NRC VAD lexicon (Mohammad, 2018a), which maps words to a real-valued intensity score between 0 (lowest) and 1 (highest). We further define 'High' and 'Low' VAD features, where only lexicon entries with scores $\geq 0.66$ and $\leq 0.33$ respectively are considered.

**Emotion:** We compute average, count, and binary presence flag features for multiple discrete emotions: the eight basic emotions of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, computed using the NRC Emotion Intensity lexicon (Mohammad, 2018b); warmth, competence, sociability, and trust using the WCST Lexicon (Mohammad, 2025a); and anxiety and calmness with the NRC WorryWords lexicon (Mohammad, 2024).

**Body**: We identify Body Part Mentions (BPMs) by matching text tokens against curated lists of 292 anatomical unigrams, bigrams, and trigrams from Zhuang et al. (2024); Wu et al. (2025), capturing mentions with possessive pronouns (e.g., "my heart", "her hand"). Features are recorded as binary flags for the presence of a body part mention, and for each possessive pronoun, as a list of BPMs used with that pronoun (i.e, `MyBPM` is a list containing BPMs used with the pronoun 'my'). Appendix A.3 lists the body part words used.

**Cognition**: We classify cognitive processes by identifying "thinking words" derived from categories outlined in Bloom's Taxonomy and Queensland's Glossary of Cognitive Verbs.[4,5] We categorize 98 unigrams into 11 categories (e.g., analyzing, learning, decision-making). A binary flag feature indicating presence is recorded for each category. Ap-

[4] https://adp.uni.edu/documents/bloomverbscognitiveaffectivepsychomotor.pdf

[5] https://www.qcaa.qld.edu.au/downloads/senior-qce/common/snr_glossary_cognitive_verbs.pdf

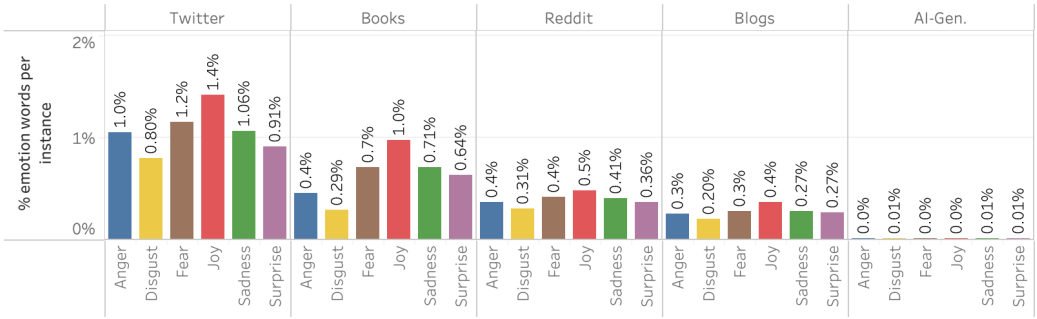


Figure 2: **Q1. Emotion:** Percent of words per instance from the six Ekman emotion categories.

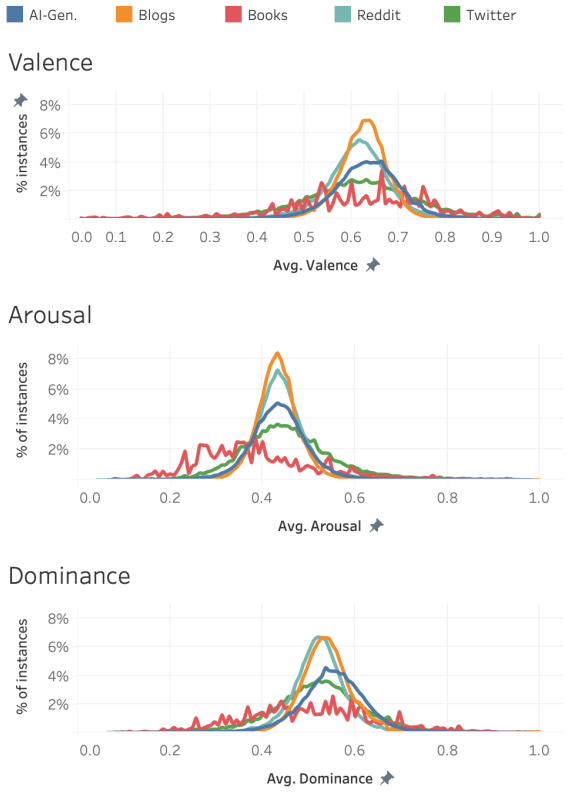


Figure 3: **Q1. Affect:** Distributions of average Valence, Arousal, and Dominance (VAD) scores of all words per instance, by source.

pendix A.4 lists the full grouped inventory.

**Demographics:** We extract demographic attributes, including age, occupation, gender, country, city, and religion, using regular expression matching and structured dictionary lookups. We map occupations according to the U.S. Bureau of Labor Statistics Standard Occupational Classification (SOC) system, and match gender, country, city, and religion references against controlled vocabularies from Wikipedia and Geonames datasets.[6] These regexes are not exhaustive but target high-precision first-person self-disclosure patterns and public controlled vocabularies. Appendix A.5 shows the full list of patterns used.

**Focus features:** We also identify pronouns (possessive and non-possessive) and verb tenses (past, present, future) through direct lexical matching and morphological tagging using the English UniMorph dataset.[7] This enables quantifying linguistic usage patterns related to personal references and temporal framing.

### 3.3. Lexical vs. ML Features

The measurement of a construct such as "anger" or "cognitive activity" given a text can be operationalized as a feature in many ways, and is dictated by the requirements of downstream use-case(s). If the goal is to predict the level of anger expressed in a single utterance by a person in real-time, trained machine learning models or pre-trained large language models will be more accurate than word-level lexicon aggregates. If, on the other hand, our goal is to measure the changes in anger levels expressed on Twitter in the USA over the last decade, plotting the (normalized) density of usage of anger-associated words in tweets from each year is a valid measurement method.

In other words, while lexicon-based methods are not as accurate as neural models at instance-level estimation (i.e, for estimating the sentiment of a particular sentence), they are comparable and sufficient when used as tools of aggregate-level analysis. This means that relative patterns of change, for comparing emotional intensity across different sources, or quantifying the patterns of changes in sentiment over a temporal period (longitudinal analysis), can be estimated with high accuracy using lexicon-based methods. The source resources also

[6] https://en.wikipedia.org/wiki/List_of_gender_identities, List_of_religions_and_spiritual_traditions,List_of_countries_and_dependencies_by_population,www.bls.gov/soc/2018, https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000/

[7] https://github.com/unimorph/eng

provide validation evidence for aggregate use (Mohammad, 2018a; Mohammad and Turney, 2013; Zhuang et al., 2024).

In Teodorescu and Mohammad (2023), the authors empirically demonstrate that emotion arcs (i.e, temporal fluctuations of valence intensity) with word-level lexicons of valence highly correlate with the ground-truth arcs (with correlation scores $> 0.9$) for social media texts, provided certain hyperparameters (like the width of the temporal window) are appropriately set. In the domain of digital humanities, Öhman et al. (2024) measure the agreement of lexicon-generated arcs for classic fiction novels with human annotation, and find high consensus. In their work, the NRC emotion intensity lexicon was customized to the literary domain with a word-similarity measure, which is one of the recommended practices for the use of emotion lexicons (Mohammad, 2023).

## 4. Key Research Questions

We use ABCDE to shed light on nine research questions, grouped by feature family.

**Q1. Affect–Emotion (AE):** *To what extent are affect- and emotion-associated words used in different media?*
*Motivation.* Affect and emotion lexicons provide aggregate signals about how people express emotion and affect across domains. These are central to population-level work in affective and social science to trace well-being, emotional contagion, or cultural mood shifts. For NLP, these signals offer interpretable, domain-portable emotion features for sentiment modeling and model alignment.
*Results.* Figure 2 shows the length-normalized counts (i.e, instance-level word density) of lexicon terms for six core emotion categories. Observe that Twitter has the highest emotion word usage rates (e.g., ∼1.4% joy tokens per instance), followed by Books and Reddit. AI-generated text uses explicit emotion words rarely (≈0.01%). Figure 3 shows the distributions of average intensity scores for Valence, Arousal, and Dominance (VAD): human-authored sources center around neutral-to-positive valence (roughly 0.60–0.65), moderate dominance, and mid-to-low arousal. Books and AI-generated text display slightly lower arousal.
**Q2. Body (B):** *How often do people refer to body parts in text? Does it vary in different media?*
*Motivation.* Body parts mentioned (BPMs) can reveal how people linguistically connect to their bodies and thus how embodied experience manifests in text. In CSS, BPMs allow measurement of health, stress, or embodied metaphors in everyday discourse. In NLP, they can serve as features for affective and health-related modeling.

| | |
|---|---|
| Books | 0.29% |
| Twitter | 1.19% |
| AI-Gen. | 1.02% |
| Blogs | 4.47% |
| Reddit | 8.16% |

Figure 4: **Q2. Body:** Percent of instances with a positive flag for possessive body part mentions (e.g., *my head/heart/*etc.) by source.

*Results.* Figure 4 shows that Reddit (8.16%) and Blogs (4.47%) contain the most possessive BPMs (references to '*my* ‹*BPM*›'); Books and AI-generated text contain very few. Figure 5 highlights the most frequent first-person BPMs: *my head* is most common (especially on Reddit with 0.91%), while *my heart* appears more in Blogs (0.55%). Mentions like *my body, my face,* and my *hands* vary systematically with domain.
**Q3. Cognition (D):** *How prevalent are cognition ("thinking") words across domains, and which subprocesses are most/least common?*
*Motivation.* Cognition vocabulary (understanding, remembering, deciding, etc.) offers a linguistic window into reasoning and mental-state discourse. For CSS, this enables cross-platform comparison of analytical concepts (and one can draw relations to emotional language). For NLP, such features open a window into the interpretability of models of deliberation, argumentation, and education.
*Results.* Figure 6 shows that human-authored sources use these terms far more than AI-generated text or Books. Blogs and Reddit lead overall: words of *understanding* (5.68%/5.57%), *general cognition* (3.94%/3.81%), and *memory recall* (3.55%/3.31%). Twitter shows the same ranking of categories, but at lower overall rates (e.g., understanding ∼1.09%). AI-generated text underexpresses cognition terms across categories (e.g., problem solving ∼0.09%, explanation ∼0.20%). Across corpora, *analyzing* is the least frequent cognition term in our set.
**Q4. Demographics-Age (D-Age):** *What are the self-disclosed age distributions in different media?*
*Motivation.* Age is a key demographic in CSS for understanding generational differences in discourse, politics, or emotion, and in NLP for bias control and stratified sampling. Self-disclosed age data helps characterize who participates in online discourse and calibrate downstream demographic analyses.
*Results.* Figure 7 shows that late teens to late 20s are most common in Reddit, whereas the 30s are most common in Twitter. Tails decline steadily there-

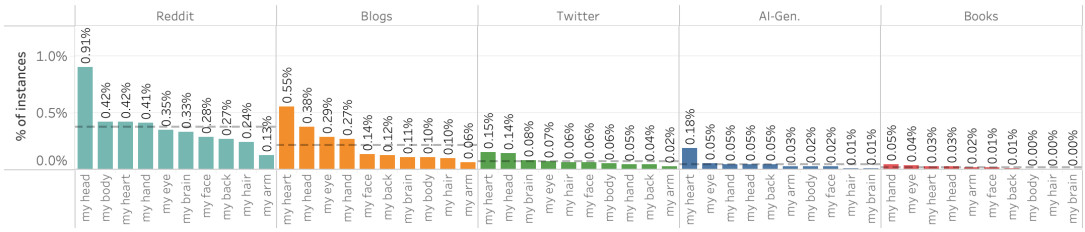


Figure 5: **Q2. Body:** % of possessive BPMs for the top ten most common BPMs, across sources. Dashed lines represent the average occurence of BPMs per dataset.

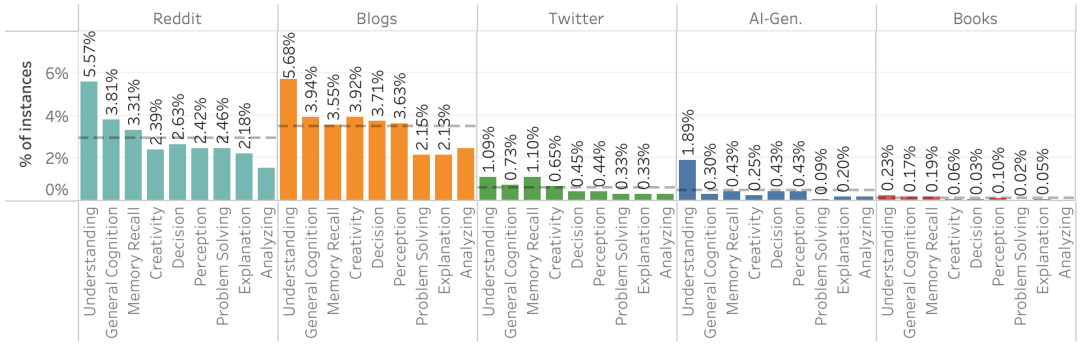


Figure 6: **Q3. Cognition:** % of instances with a positive flag for cognitive terms from different categories. Dashed lines represent the average occurence of cognitive terms per dataset.

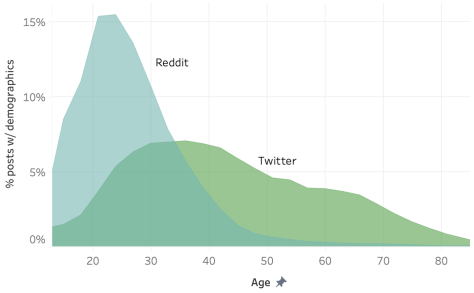


Figure 7: **Q4. Demographics:** Age distribution of posts with self-disclosed age.

after. This suggests that Reddit users are markedly younger than Twitter users.

**Q5. D-Occupation:** *Which occupations are most frequently self-disclosed?*

*Motivation.* In most modern societies, occupations are key parts of people's identity, values, and social interaction. In CSS, they enable studies of professional discourse, social stratification, and belief-linked behavior. In NLP, such self-labels offer interpretable subgroup features for fairness and personalization.

*Results.* Figure 8 shows Reddit users often self-identify as creative or technical professionals (e.g., musicians, software engineers), whereas Twitter skews toward public-facing roles (e.g., journalists, authors, executives).

**Q6. D-Gender:** *Which genders do users most often self-disclose?*

*Motivation.* Gender is a central dimension of social identity (and inequality). Self-identified gender in text provides a precise, ethically grounded way to study participation patterns, self-representation, and language variation without resorting to gender inference or heuristics. In CSS, this enables analysis of representation and discourse norms; in NLP, it supports fairness auditing in generation, pronoun resolution, and toxicity detection.

*Results.* Figure 9 shows the majority of people self-disclose as cis men and women, with smaller but notable transgender (5–8%) and non-binary (1–2%) self-descriptions. These values highlight the presence of gender-diverse voices in online spaces and illustrate the potential for intersectional analyses when combined with other attributes (e.g., affect, occupation). Importantly, these reflect rates among disclosed posts, not population estimates. For downstream use, such disclosures serve as positive-only labels—absence of gender mention should be treated as missing, not negative.

**Q7. D-Religion:** *With which religions do users most often identify?*

*Motivation.* Religious self-identification is an important dimension of cultural identity and social behavior. In CSS, religion enables studies of moral language, group cohesion, and polarization; in NLP, it is critical for evaluating fairness, bias, or sentiment models concerning faith-related content.

*Results.* Figure 10 shows that users mostly iden-

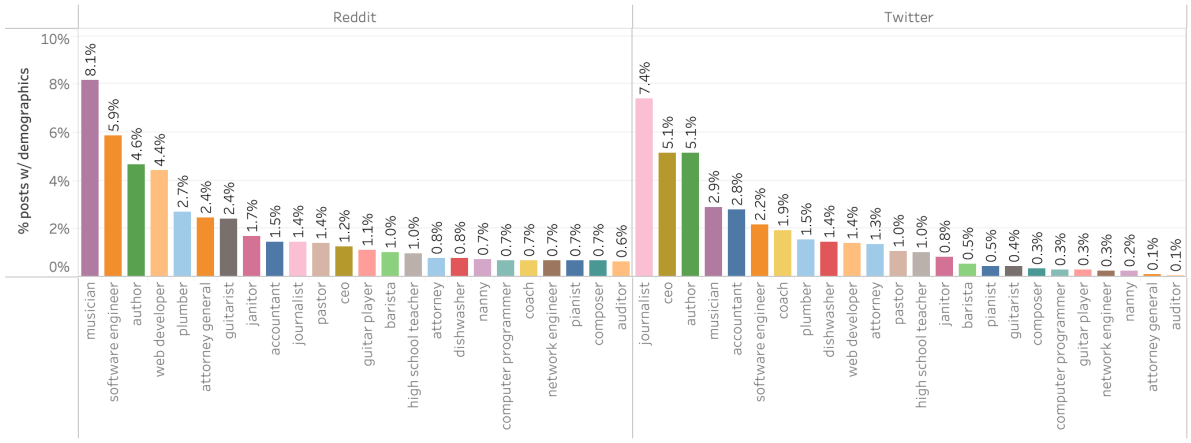


Figure 8: **Q5. D:** Occupation distribution of posts with self-disclosed occupation.

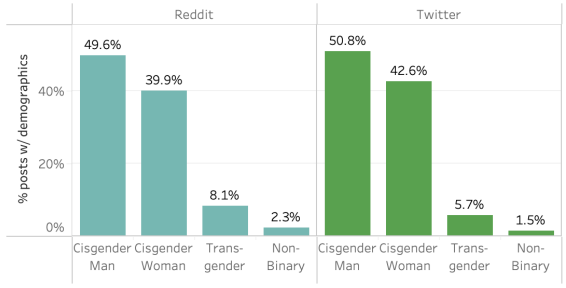


Figure 9: **Q6. D:** Gender distribution of posts with self-disclosed gender.

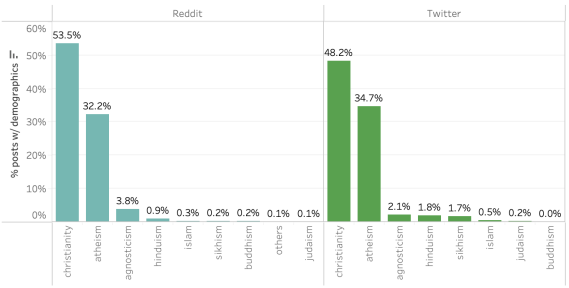


Figure 10: **Q7. D:** Religion distribution for posts with self-disclosed faith.

tify with Christianity and atheism on both Reddit and Twitter. On Reddit, Christianity accounts for 53.5% of religion-disclosing posts and atheism for 32.2%; on Twitter, Christianity (48.2%) and atheism (34.7%) again lead. Agnosticism follows (Reddit 3.8%, Twitter 2.1%), with smaller communities representing Hinduism, Sikhism, Islam, Judaism, and Buddhism (each under 2%). Twitter exhibits relatively greater religious diversity (notably Hinduism and Sikhism) than Reddit.

**Q8. D-Tense-Pronouns:** *How do tense and pronoun usage differ by dataset?*

*Motivation.* Tense and person reference capture

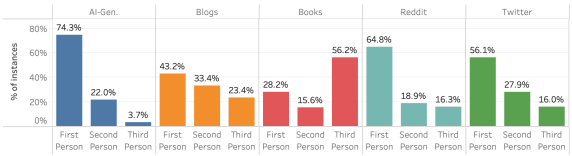


Figure 11: **Q8. D:** % instances with a positive flag for a first-, second-, or third-person pronoun.

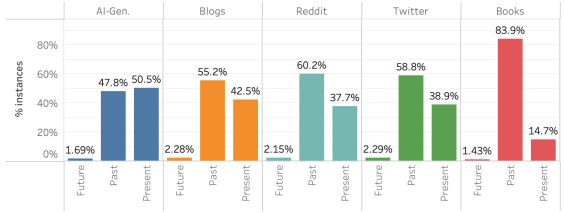


Figure 12: **Q8 Cognition:** % instances with a positive flag for a specific tense verb or modal.

narrative stance and interactional focus, which are core constructs in discourse and stylistics. For CSS, they reveal how people recount vs. instruct, self-narrate vs. address others. For NLP, they support interpretable style transfer and alignment of narrative voice in generation.

*Results.* Figure 12 shows that future tense is rare (1.4–2.3%) across all media. Past tense dominates Books (83.9%) and remains high on Reddit (60.2%), Twitter (58.8%), and Blogs (55.2%). AI-generated text slightly favors present (50.5%) over past (47.8%). Pronoun distributions (Figure 11) indicate that first-person usage is most common on Reddit and Blogs (self-narration), whereas second-person occurs more on Twitter and in AI dialogues (addressivity).

**Q9. ABCDE:** *How can ABCDE support cross-feature analyses across Affect, Body, Cognition,*

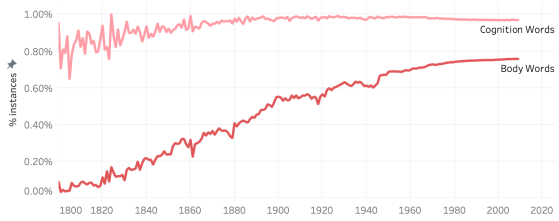


Figure 13: **Q9. B,C:** % instances with body and cognition terms over 200 years in fiction books.

*Demographics, and Emotion?*
*Motivation.* Many questions in affective and social science require connecting multiple linguistic signals at once; for example, how much embodied words are being used compared to cognitive words, and how that has changed over time. Cross-feature views are a primary use case for ABCDE, which provides ready access to many features on the same instances and across instances over time.
*Results.* As a demonstration, let us say we are interested in the rate at which body and cognition terms have been used over the last 200 years in English-fiction books (Google Books 5-grams, 1800—2010). Figure 13 plots the results. Observe that the share of instances with at least one *body* term rises more than an order of magnitude — from about **0.05%** circa 1800 to roughly **0.50%** by **1900**. It then still continues to rise, albeit at a slower rate of increase, to **0.62%** in the **1930s–1940s**. There is a dip in the years of World War 2 (1938 to 1945), but it recovers back up to about **0.66%** by **1950**. Mentions of body parts plateaus to around **0.73–0.75%** after **1980**. In contrast, the share of *cognition* terms has always been higher than that of body terms and perhaps more importantly, it has remained fairly steady over time (hovering around **0.95–0.99%** from the late 19th to mid-20th century, with a slight softening to about **0.97–0.98%** in the 1990s and 2000s). Consequently, the body-to-cognition ratio increase is stark from roughly **5%** in **1800** to about **77%** by the **2000s** (numbers not shown here) suggesting a long-term rise in embodied language. Researchers can exploit various combinations of features in ABCDE to similarly explore research questions of interest.

## 5. Conclusion

In this work, we presented ABCDE, a collection of text datasets and linguistic annotations that capture features of interest at the intersection of language, affective science, cognition, and social science. We make all artifacts of ABCDE available, along with the associated code. Notably, several recent works already use ABCDE, including sentence-level annotations of social perception dimensions (Ayesh et al., 2026), analyses of emotion expression across the lifespan in social media (Teodorescu et al., 2026), and analyses for large-scale lexical norms for warmth and trust (Mohammad, 2025b), and anxiety (Mohammad, 2026).

Additionally, we presented several representative analyses that reflect the potential of this dataset to answer fundamental research questions on the nature of affect, cognition, and behavior, using the rich signals encoded in language as data. Affective Scientists, for example, can use this resource to map the evolution of emotion word usage with age and social culture, study which emotions are most commonly associated with a focus on the self vs the other, the strength of the connection between bodily awareness and affective states, and the downstream impacts of such patterns on the mental and physical health of populations. The social media datasets in ABCDE can be used to answer central questions in computational social science on how social concepts are perceived and talked about in different sub-networks, and the change in these perceptions over time and in relation to real-world events. The inclusion of AI-generated texts enables a comparison of the ways in which large language models imitate or differ from human usage of natural language at an aggregate level, along dimensions such as emotion and cogitation, as well more fine-grained analyses of the dynamics of human-LLM conversations.

Future work will add information for more affective, social perception, and moral dimensions: notably the care–harm, authority–subversion, loyalty–betrayal, fairness–unfairness, and purity–degradation dimensions from the NRC Moral Foundations Lexicon (based on the Moral Foundation Theory of Haidt and Graham (2007)). We also hope that future work led by native speakers of various languages will include datasets in various languages.

## Limitations and Ethical Considerations

The scope and focus of our work was on English corpora. Thus, much of the research that the corpora and features enable are relevant to English and North America. We discuss some of the relevant limitations and ethical considerations below. We hope this work will pave the way for creating similar resources in various other languages, led by speakers of those languages.

1. English is the only language represented in our dataset. Many of the features of language use that we enumerate and operationalize here vary with language, and are reflective of underlying differences in cognition, affect, and behavior across population groups.

2. A large proportion of the data in the datasets that we consolidate originate from North American users, and therefore is not representative of English-speaking populations from other countries, regions, and cultures.

3. The lexicons we use to quantify various features are also reflective of the usage patterns and biases of the annotators involved in the process, who are largely from the English-speaking populations of US and Europe.

4. The broad use of static lexicons in our resource also may not account for the variations seen in language use with time, or across different domains of data.

5. Our feature set does not model syntax, discourse structure, or compositional effects. Even so, prior lexicon-based studies show that large-scale comparative trends can remain informative at aggregate scale (Teodorescu and Mohammad, 2023; Öhman et al., 2024).

We refer readers to Mohammad (2023) for an overview of the recommended practices in the usage of emotion lexicons, and Mohammad (2022) for the ethical considerations that are relevant in the use of automatic emotion recognition tools.

We also enumerate below some of the many ethical considerations that apply to research in computational affective science and social science:

1. Conclusions about language use, mental health, emotionality, etc. should be considered aggregate, population-level indicators of trends, rather than as tools for estimating or predicting the behavior of individuals.

2. Conclusions drawn from our resource should be validated by other sources of data and measurement methods, such as controlled user studies.

3. The creation of language datasets has several socio-cultural biases encoded in the process. What data, and whose data, is recorded, digitized, preserved, and published, either on the internet and in historical archives, is determined by social systems of power and influence. Language datasets, however large, should be critically interrogated to understand whose worldview is being represented, and to avoid over-claiming the generalizability of conclusions.

4. The released instance-level annotations include demographic attributes such as age, gender, religion, occupation, and location. We strictly prohibit any commercial use of our dataset, or its use for profiling, targeting, or individual-level inference.

## Acknowledgments

This work was supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation. Thanks to Terry Ruas and Lars Kaesberg for early feedback on this work.

## 6. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mutaz Ayesh, Saif M. Mohammad, and Nedjma Ousidhoum. 2026. Annotating dimensions of social perception in text: The first sentence-level dataset of warmth and competence.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 10:1–47.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47, Denver, Colorado, USA. Association for Computational Linguistics.

Keith Burton and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Z. Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. *CoRR*, abs/2005.04246.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, 6(12):e26752.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still generate coherent text? arXiv preprint arXiv:2305.07759.

Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. 2025. Large AI models are cultural and social technologies. *Science*, 387(6739):1153–1156.

Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2016. The structural virality of online diffusion. *Management Science*, pages 180–196.

Google Inc. 2012. Google books ngram corpus, version 20120701. `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`. English Fiction subset, 5-grams.

Tear Gosling, Alpin Dale, and Yinhe Zheng. 2023. PIPPA: A partially synthetic conversational dataset.

Sharath Chandra Guntuku, David Bryce Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Angela Gutchess and Suparna Rajaram. 2023. Consideration of culture in cognition: How we can enrich methodology and theory. *Psychonomic Bulletin & Review*, 30(3):914–931.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1):98–116.

Katie Hoemann, Yeasle Lee, Èvelyne Dussault, Simon Devylder, Lyle H Ungar, Dirk Geeraerts, and Batja Mesquita. 2025. The construction of emotional meaning in language. *Communications Psychology*, 3(1):99.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. SafeChain: Safety of language models with long chain-of-thought reasoning capabilities.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM alignment dataset.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. 2025. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*.

Changyi Li, Jiayi Wang, Xudong Pan, Geng Hong, and Min Yang. 2025. ReasoningShield: Content safety detection over reasoning traces of large reasoning models. *arXiv preprint arXiv:2505.17244*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Hauke Licht, Rupak Sarkar, Patrick Y Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. Measuring scalar constructs in social science with LLMs. *arXiv preprint arXiv:2509.03116*.

Li Lucy and David Bamman. 2021. Characterizing english variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of*

*52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Saif M. Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

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

Saif M. Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif M. Mohammad. 2024. Worrywords: Norms of anxiety association for 44,450 english words. In *Proceedings of The Annual Conference of the Empirical Methods on Natural Language Processing (EMNLP 2024, main)*, Miami, FL.

Saif M Mohammad. 2025a. Words of warmth: Trust and sociability norms for over 26k english words. *arXiv preprint arXiv:2506.03993*.

Saif M. Mohammad. 2025b. Words of warmth: Trust and sociability norms for over 26k English words. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18830–18850, Vienna, Austria. Association for Computational Linguistics.

Saif M. Mohammad. 2026. From composure to catastrophe: Norms of calmness–anxiety associations for 54,000 English words and multiword expressions. PsyArXiv preprint.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Jinkyung Na, Chih-Mao Huang, and Denise C. Park. 2017. When age and culture interact in an easy and yet cognitively demanding task: Older adults, but not younger adults, showed the expected cultural differences. *Frontiers in Psychology*, 8:457.

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.

General Reasoning. 2024. General thoughts 430k dataset. https://huggingface.co/datasets/GeneralReasoning/GeneralThought-430K.

Rafael Alberto Rivera Soto, Kailin Koch, Aleem Khan, Barry Y Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations*.

Daniela Teodorescu and Saif Mohammad. 2023. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4124–4137.

Daniela Teodorescu, Jan Philip Wahle, and Saif M. Mohammad. 2026. Age and affect in language: How emotion expression on social media varies across adulthood. In *Proceedings of the 1st Workshop on Computational Affective Science (CAS 2026)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).

Radek Trnka, Josef Mana, and Martin Kuska. 2022. Age-related differences in valence and arousal of emotion concepts. *Ageing & Society*, 42(9):1991–2007.

Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France. European Language Resources Association.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022. How large language models are transforming machine-paraphrase plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 952–963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun,

Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Jieru Mei, Brian R Bartoldson, Bhavya Kailkhura, and Cihang Xie. 2025. Star-1: Safer alignment of reasoning llms with 1k data. *arXiv preprint arXiv:2504.01903*.

Sophie Wu, Jan Philip Wahle, and Saif M Mohammad. 2025. The language of interoception: Examining embodiment and emotion through a corpus of body part mentions. *arXiv preprint arXiv:2505.16189*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2024. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2024. My heart skipped a beat! recognizing expressions of embodied emotion in natural language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3525–3537, Mexico City, Mexico. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

# A. Appendix: Feature Resources and Descriptive Statistics

The following appendix lists details on the released features in **ABCDE**, including source resources and descriptive statistics.

## A.1. Feature Summary

Table 1 summarizes the released feature families, the label patterns that appear in the data files, the underlying resources, and the per-instance output types.

| Family | Labels in release | Source resources and scope | Outputs |
|---|---|---|---|
| VAD | `NRCAvgValence`, `NRCAvgArousal`, `NRCAvgDominance`<br>`NRCHasHigh/LowValenceWord`<br>`NRCHasHigh/LowArousalWord`<br>`NRCHasHigh/LowDominanceWord`<br>`NRCCountHigh/LowValenceWords`<br>`NRCCountHigh/LowArousalWords`<br>`NRCCountHigh/LowDominanceWords` | NRC VAD (Mohammad, 2018a); 19,971 entries with scalar Valence, Arousal, and Dominance scores | average, count, binary |
| Discrete Emotion | `NRCHas[Emotion]Word`<br>`NRCCount[Emotion]Words`<br>for anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive, and negative | NRC Emotion Lexicon (Mohammad and Turney, 2013); 14,154 words with binary emotion and sentiment associations | count, binary |
| Anxiety / Calmness | `NRCHasAnxietyWord`, `NRCHasCalmnessWord`<br>`NRCAvgAnxiety`, `NRCAvgCalmness`<br>`NRCHasHighAnxietyWord`<br>`NRCCountHighAnxietyWords`<br>`NRCHasHighCalmnessWord`<br>`NRCCountHighCalmnessWords` | NRC WorryWords (Mohammad, 2024); 44,447 entries with signed anxiety or calmness scores | average, count, binary |
| Warmth / Trust | `NRCAvgMoralTrustWord`<br>`NRCAvgSocialWarmthWord`<br>`NRCAvgWarmthWord`<br>`NRCHasHigh/LowMoralTrustWord`<br>`NRCHasHigh/LowSocialWarmthWord`<br>`NRCHasHigh/LowWarmthWord`<br>`NRCCountHigh/LowMoralTrustWord`<br>`NRCCountHigh/LowSocialWarmthWord`<br>`NRCCountHigh/LowWarmthWord` | NRC MoralTrust, SocialWarmth, and CombinedWarmth (Mohammad, 2025b); 31,456–31,573 entries across the three lexicons | average, count, binary |
| BPMs | `HasBPM`, `MyBPM`, `YourBPM`<br>`HerBPM`, `HisBPM`, `TheirBPM` | Released BPM inventory; 292 entries in file order, with duplicate *toe* retained; compiled from public body-part lists and linked body-language work (Zhuang et al., 2024; Wu et al., 2025) | binary, list |
| Cognition | one `COGHas...Word` binary per category, including<br>`COGHasAnalyzingEvaluatingWord`<br>`COGHasCreativityIdeationWord`<br>`COGHasGeneralCognitionWord`<br>`COGHasLearningUnderstandingWord`<br>`COGHasExplanationArticulationWord` | Released cognition inventory; 12 categories and 98 terms | binary |

| Family | Labels in release | Source resources and scope | Outputs |
|---|---|---|---|
| Demographics | `Author`, `DMGMajorityBirthyear` `DMGRawBirthyearExtractions` `DMGRawExtractedAge`, `DMGAgeAtPost` `DMGRawExtractedGender` `DMGRawExtractedCity` `DMGCountryMappedFrom` `ExtractedCity` `DMGRawExtractedCountry` `DMGRawExtractedReligion` `DMGMainReligionMappedFrom` `ExtractedReligion` `DMGMainCategoryMappedFrom` `ExtractedReligion` `DMGRawExtractedOccupation` `DMGSOCTitleMappedFrom` `ExtractedOccupation` | Regex extraction plus controlled vocabularies: 6 age regexes, SOC occupations, 69 gender terms, 231 countries, top 50k GeoNames cities, and 157 religion rows | extracted strings, mapped labels, numeric age |
| Pronouns | binary `PRNHas...` binarys from `PRNHasI` to `PRNHasTheirs` across first-, second-, and third-person forms | Released pronoun inventory covering first-person singular and plural, second person, and third-person feminine, masculine, and plural or neutral forms | binary |
| Time / Tense | `TIMEHasPastVerb`, `TIMECountPastVerbs` `TIMEHasPresentVerb` `TIMECountPresentVerbs` `TIMEHasFutureModal` `TIMECountFutureModals` `TIMEHasPresentNoFuture` `TIMEHasFutureReference` | UniMorph English plus future-modal rules; tense features derived from present, past, and future-oriented forms | count, binary |
| General | `WordCount` | tokenized text stream for each instance | count |

Table 1: Released annotation features in **ABCDE**

## A.2. Dataset Summary Statistics

Table 2 reports the instance definition and lexical statistics for each source. MTLD (Measure of Textual Lexical Diversity) estimates the average token span over which lexical variation remains above a fixed threshold (McCarthy and Jarvis, 2010). MATTR (Moving-Average Type-Token Ratio) averages the type-token ratio over a sliding window to reduce text-length sensitivity (Covington and McFall, 2010). For both metrics, higher values indicate greater lexical diversity. MATTR ranges from 0 to 1, while MTLD has no fixed upper bound. A few contrasts are notable. Twitter has the shortest instances on average (20.83 tokens) because of its limited posting length, but the highest lexical-diversity values (MTLD 252.31, MATTR 0.88), whereas Books are fixed-length 5-grams and therefore have the lowest diversity values (MTLD 6.23, MATTR 0.09). Blogs and AI-generated texts are much longer on average (199.57 and 258.65 tokens), but AI-generated text remains substantially less lexically diverse than either Blogs or Reddit.

| Source | Instances | Instance definition | Mean tokens | Median tokens | MTLD | MATTR |
|---|---|---|---|---|---|---|
| Twitter | 45.2M | one post | 20.83 | 16.00 | 252.31 | 0.88 |
| Reddit | 78.5M | one post | 117.78 | 72.00 | 91.52 | 0.81 |
| Books | 177.1M (1.74M unique) | one 5-gram occurence | 5.00 | 5.00 | 6.23 | 0.09 |
| Blogs | 34.2M | one sentence | 199.57 | 96.00 | 107.38 | 0.83 |
| AI-generated | 68.9M | one AI response | 258.65 | 143.00 | 30.60 | 0.67 |

Table 2: Per-source descriptive statistics. MTLD and MATTR are computed over the concatenated token stream for each source (MATTR window = 50; MTLD threshold = 0.72). Higher values on both metrics indicate greater lexical diversity; MATTR ranges from 0 to 1, while MTLD has no fixed upper bound.

Table 3 reports demographic coverage as counts in thousands and percentages of total source instances with age, occupation, gender, country, city, and religion annotations.

| Source | Age | Occupation | Gender | Country | City | Religion |
|---|---|---|---|---|---|---|
| Twitter | 2,150k (4.8%) | 525k (1.2%) | 502k (1.1%) | 4,900k (10.8%) | 3,496k (7.7%) | 334k (0.7%) |
| Reddit | 33,495k (42.7%) | 3,041k (3.9%) | 3,451k (4.4%) | 12,558k (16.0%) | 8,653k (11.0%) | 2,283k (2.9%) |

Table 3: Demographic coverage with self-disclosure annotations. Cells show the number of instances in thousands, together with the percentage of total source instances for that source.

## A.3. Full BPM Inventory

Table 4 shows the BPM words used in alphabetical order.

| First-Letter Group | Terms |
|---|---|
| A-C | adam's apple, abdomen, abdominal, adenoid, adenoids, adrenal gland, alvine, anal, anatomy, ankle, anus, appendicular, appendix, arch, arm, armpit, arterial, artery, axilla, axillary, back, ball of the foot, belly, belly button, toe, bladder, blood, blood vessels, body, bone, brachial, brachium, brain, breast, buttocks, caecum, calf, capillary, capital, caput, cardiac, carpal, carpus, cartilage, cell, cerebral, cervical, cervical vertebrae, cervix, cheek, chest, chin, ciliary, cilium, circulatory system, clavicle, clitoral, clitoris, coccyx, collar bone, colon, colonic, crural, crus, cubital, cutaneous, cutis |
| D-H | diaphragm, digestive system, digital, dorsal, duodenal, duodenum, ear, ear lobe, elbow, encephalon, endocrine system, epiglottis, erythrocyte, esophagus, eye, eyebrow, eyelash, eyelashes, eyelid, face, fallopian tubes, feet, femur, fibula, filling, finger, fingernail, fist, follicle, foot, forearm, forehead, foreskin, frontal, gallbladder, gastric, gena, genal, genicular, genu, gingiva, gingival, gland, glands, glottal, glottic, glottis, groin, gullet, gum, gums, hair, half-moon, hamstring, hand, head, heart, heel, hepatic, hip, humerus |
| I-M | ileum, immune system, index finger, inguinal, instep, intestine, intestines, iris, jaw, jejunum, kidney, knee, knuckle, labial, labyrinth, laryngeal, larynx, leg, leucocyte, ligament, lingua, lip, liver, lobe, loin, lumbar, lumbar vertebrae, lumbus, lung, lungs, lymph node, lymphocyte, mandible, manual, manus, metacarpal, metatarsal, midriff, molar, mouth, muscle |
| N-R | nail, nape, naris, nasal, nates, navel, neck, nerve, nerves, neural, nipple, nose, nostril, nucha, occipital, occiput, oesophagus, organs, ovarian, ovary, oviduct, palm, palpebral, pancreas, pancreatic, patella, pectoral, pedal, pelvis, penile, penis, phalanges, pharyngeal, pharynx, pinky, pituitary, plantar, pollex, popliteal, pore, prepuce, pubes, pubic, pulmonary, pupil, radius, rectal, rectum, red blood cells, respiratory system, ribcage, ribs |
| S-W | sacrum, scalp, scapula, scrotum, senses, shin, shoulder, shoulder blade, skeleton, skin, skull, sole, spinal column, spinal cord, spine, spleen, sternum, stomach, stomatic, superciliary, sural, talus, tarsal, teeth, temple, temporal, tendon, testes, testicle, testicular, thigh, thoracic, thorax, throat, thumb, thyroid, tibia, tissue, toe, toenail, tongue, tonsil, tonsils, tooth, torso, trachea, trunk, ulna, umbilical, umbilicus, ureter, urethra, urinary system, uterine, uterus, uvula, vagina, vaginal, vein, vena, venous, venter, ventral, vertebra, vesical, vulva, waist, white blood cells, windpipe, womb, wrist |

Table 4: Released BPM inventory in file order.

## A.4. Cognition Word Inventory

Table 5 shows the cognition word inventory.

| Category | Terms |
|---|---|
| Analyzing & Evaluating | analyze, appraise, assess, critique, diagnose, differentiate, discern, discriminate, distinguish, evaluate, investigate, self-evaluate, test |
| Creativity & Ideation | brainstorm, conceptualize, create, diverge, evolve, fantasize, ideate, imagine, innovate, invent, pretend, visualize |
| General Cognition | contemplate, deliberate, focus, introspect, reason, reflect, regulate, ruminate |
| Learning & Understanding | accommodate, assimilate, associate, comprehend, empathize, explore, grasp, internalize, learn, study, understand |
| Decision Making & Judging | calculate, choose, decide, deduce, determine, estimate, infer, judge, prioritize, resolve |
| Problem Solving | plan, revise, solve, strategize, troubleshoot |
| Higher-Order Thinking | abstract, categorize, classify, generalize, hypothesize, interpret, synthesize |
| Confused or Uncertain Thinking | doubt, self-question |
| Memory & Recall | consolidate, encode, forget, memorize, recall, rehearse, remember, retrieve |
| Perception & Observation | detect, identify, label, notice, observe, perceive, recognize, scan, spot, trace |
| Prediction & Forecasting | anticipate, forecast, forethink, predict, project |
| Explanation & Articulation | articulate, define, describe, discuss, elaborate, explain, verbalize |

Table 5: Cognition word inventory grouped by category.

## A.5. Demographic Regex Overview

Table 6 summarizes the self-disclosure regex patterns used by the demographic extractor.

| Attribute | Pattern(s) |
|---|---|
| age | (i) "I am/I'm **XX** years old" (explicit age statement)<br>(ii) "I am/I'm **XX**" (+ end of string, punct./conj.)<br>(iii) "I was/am born in **YYYY**" (four-digit year)<br>(iv) "I was/am born in '**YY**" (two-digit year with apostrophe)<br>(v) "I was born on DD Month **YYYY**" (full date format)<br>(vi) "I was born on **MM/DD/YYYY**" (and similar date formats) |
| occupation | (i) "I am/I'm (a/an/the) **[Occupation]** (at **[Company]**) (+ punt./conj.)"<br>(ii) "I work as (a/an/the) **[Occupation]** (at **[Company]**)"<br>(iii) "My job/occupation/role is (to be) (a/an/as) **[Occupation]**"<br>(iv) "I'm (currently) employed as (a/an/the) **[Occupation]**" |
| gender | (i) "I am/I'm (a/an) **[Gender]** (+ punctuation/conjunction)"<br>(ii) "I identify as (a/an) **[Gender]**"<br>(iii) "My gender (identity) is **[Gender]**"<br>(iv) "I'm/I am (a) (transgender/trans) **[Gender]**" |
| country | (i) "I am/I'm from (the) **[Country]**" (current primary location)<br>(ii) "I am/I'm (a/an/the) **[Nationality]** (+ punt./conj.)"<br>(iii) "I live in/at (the) **[Country]**"<br>(iv) "I come from (the) **[Country]**"<br>(v) "My nationality/citizenship is **[Nationality]**"<br>(vi) "I was/am born and raised in (the) **[Country]**"<br>(vii) "I am/I'm originally from (the) **[Country]**" |
| city | (i) "I am/I'm from **[City]** (+ punctuation/conjunction)"<br>(ii) "I live in/at **[City]**"<br>(iii) "I'm/I am (currently) residing/based in **[City]**"<br>(iv) "My (current/home) city/town is **[City]**"<br>(v) "I (grew up / was raised) in **[City]**" |
| religion | (i) "I am/I'm (a/an/a practicing) **[Religion]** (+ context)"<br>(ii) "My religion/faith is **[Religion]**"<br>(iii) "I (actively) practice **[Religion]**"<br>(iv) "I am/I'm a follower of **[Religion]**"<br>(v) "I converted to **[Religion]**"<br>(vi) "I was raised/born (as a/an) **[Religion]**"<br>(vii) "I identify as **[Religion]** / identify with **[Religion]**" |

Table 6: Extraction patterns for labeling social media users with demographic attributes. For a full list of **[Occupation]**, **[Genders]**, **[Country]**, **[City]**, and **[Religion]**, please refer to the GitHub repository.