

Age Dataset Documentation

This dataset contains linguistic and demographic information extracted from Reddit posts and Twitter/X posts (via [TUSC](#)), focusing on users who self-identify their age in their posts.

Dataset Statistics

Users with Self-Identified Age

Dataset	Time Period	Number of Users
Reddit	2010-2022	2,405,284
TUSC-city	2020-2021	26,916
TUSC-country	2015-2021	1,142

Posts from Users with Self-Identified Age

Dataset	Time Period	Number of Posts
Reddit	2010-2022	36,444,262
TUSC-city	2020-2021	4,207,298
TUSC-country	2015-2021	27,940

Dataset Files

Reddit Dataset

- **reddit_users.tsv**: Contains Reddit users who self-identified their age
- **reddit_user_posts.tsv**: Contains all posts from self-identified users with linguistic features

TUSC (Twitter/X) Datasets

- **tusc_country_users.tsv**: Contains Twitter/X users who self-identified their age (country-level location)
- **tusc_country_user_posts.tsv**: Contains all posts from self-identified users with linguistic features (country-level)
- **tusc_city_users.tsv**: Contains Twitter/X users who self-identified their age (city-level location)
- **tusc_city_user_posts.tsv**: Contains all posts from self-identified users with linguistic features (city-level)

Dataset Construction Process

1. Data Sources

- **Reddit**: JSON Lines files containing Reddit posts from 2010-2022 from [Pushshift](#)
- **TUSC**: Parquet files containing geolocated Twitter/X posts from [TUSC](#)

2. Processing Pipeline

The dataset was constructed using a two-stage pipeline:

Stage 1: Self-Identification Detection

- Scans posts/tweets to find users who self-identify their age using regex patterns to detect age mentions
- Resolves multiple age mentions to determine birth year
- Outputs user files with demographic information

Stage 2: Feature Extraction

- Collects all posts from self-identified users
- Applies feature extraction using various lexicons
- Computes age at post time based on birth year
- Outputs post files with all features

3. Filtering Criteria

- **Text length:** 5-1000 words
- **Age range:** 13-100 years old
- **Excluded authors:** [deleted], AutoModerator, Bot (Reddit only)
- **Valid self-identification:** Must match one of the regex patterns
- **Remove posts marked as adult material** (over_18 flag, Reddit only)
- **Remove posts with title but no body text** (Reddit only)
- **Remove promoted/advertised posts** (Reddit only)

Age Extraction

Regex Patterns Used

The system uses 6 regex patterns to detect age self-identification:

1. **Direct age statement:** `\bI(?:\s+am|'m)\s+(\d{1,2})\s+years?\s+old\b`
 - Example: "I am 25 years old", "I'm 30 year old"
2. **Age with contextual boundaries:** `\bI(?:\s+am|'m)\s+(\d{1,2})(?=\s*(?:$|[,.!?;:\-]|(?:and|but|so|yet)\s))`
 - Example: "I am 25.", "I'm 30, and...", "I am 25 but..."
3. **Birth year (4-digit):** `\bI(?:\s+was|\s+am|'m)\s+born\s+in\s+(\d{2}|20(?:0\d|1\d|2[0-4]))\b`
 - Example: "I was born in 1998", "I am born in 2005"
4. **Birth year (2-digit with apostrophe):** `\bI(?:\s+was|\s+am|'m)\s+born\s+in\s+'(\d{2})\b`

- Example: "I was born in '98", "I'm born in '05"

5. **Birth date (full format):** `\bI\s+was\s+born\s+on\s+(?:\d{1,2}(?:st|nd|rd|th)?\s+)?(?:January|February|March|April|May|June|July|August|September|October|November|December|Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Sept|Oct|Nov|Dec)\s+(?:\d{1,2}(?:st|nd|rd|th)?,?\s+)?(19\d{2}|20(?:0\d|1\d|2[0-4]))\b`

- Example: "I was born on 15 March 1998", "I was born on March 15th, 1998"

6. **Birth date (numeric format):** `\bI\s+was\s+born\s+on\s+\d{1,2}[/\-]\d{1,2}[/\-](19\d{2}|20(?:0\d|1\d|2[0-4]))\b`

- Example: "I was born on 03/15/1998", "I was born on 15-03-1998"

False Positive Prevention

- Word boundaries ensure complete word matches
- Contextual boundaries for pattern 2 (punctuation or conjunctions)
- Year ranges limited to 1900-2024
- Age filtering: only 13-100 years old accepted
- First-person requirement ("I") ensures self-identification

Age Resolution Algorithm

1. Extract all age/birthyear mentions from text
2. Convert ages to birth years (post year - age)
3. Filter out ages below 13 during conversion
4. Cluster similar birth years (within 2 years)
5. Weight birth years (1.0) higher than ages (0.8)
6. Select cluster with highest score (weight sum + count × 0.1)
7. Compute weighted average as final birth year
8. Calculate resolved age and filter if not between 13-100

Lexicons Used

NRC Lexicons

- **NRC VAD Lexicon** (Version 1, July 2018)
 - Contains valence, arousal, and dominance scores (0-1) for words
 - Source: [NRC Word-Emotion Association Lexicon](#)
- **NRC Emotion Lexicon** (Version 0.92, July 2011)
 - Maps words to 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and 2 sentiments (positive, negative)
 - Source: [NRC Emotion Lexicon](#)
- **NRC WorryWords Lexicon** (Anxiety/Calmness)
 - Contains anxiety scores from -3 (very calm) to +3 (very anxious)

- Source: [NRC Word-Worry Association Lexicon](#)
- **NRC MoralTrust Lexicon** (Version: Jan 5, 2025)
 - Contains moral trustworthiness scores
 - Source: [NRC Lexicons](#)
- **NRC SocialWarmth Lexicon** (Version: Jan 5, 2025)
 - Contains social warmth scores
 - Source: [NRC Lexicons](#)
- **NRC CombinedWarmth Lexicon** (Version: Jan 5, 2025)
 - Contains combined warmth scores
 - Source: [NRC Lexicons](#)

Other Lexicons

- **ENG Tenses Lexicon** (Version 3, April 2022)
 - Maps words to their grammatical forms (past, present, etc.)
 - Source: [UniMorph English](#)
- **Body Part Words**: Union of two sources:
 - [Collins Dictionary Body Parts List](#)
 - [Enchanted Learning Body Parts List](#)

Feature Descriptions

Demographic Features

- **Author**: User ID/username
- **DMGMajorityBirthyear**: Resolved birth year from self-identification
- **DMGRawBirthyearExtractions**: Raw extracted age/year values
- **DMGAgeAtPost**: Age when the post was created

Post Metadata

- **PostID**: Unique post identifier
- **PostCreatedUtc** (Reddit) / **PostCreatedAt** (TUSC): Timestamp
- **PostSubreddit** (Reddit only): Subreddit name
- **PostTitle** (Reddit only): Post title
- **PostSelftext** (Reddit) / **PostText** (TUSC): Post content
- **PostScore** (Reddit only): Post score
- **PostNumComments** (Reddit only): Number of comments
- **PostCountry/PostCity** (TUSC only): Location information

Body Part Mentions (BPMs)

- **HasBPM**: Any body part found in text

- **MyBPM:** Body parts after "my"
- **YourBPM:** Body parts after "your"
- **HerBPM:** Body parts after "her"
- **HisBPM:** Body parts after "his"
- **TheirBPM:** Body parts after "their"

Pronoun Features

Binary flags for presence of pronouns:

- **PRNHasI:** Contains "I"
- **PRNHasMe:** Contains "me"
- **PRNHasMy:** Contains "my"
- **PRNHasMine:** Contains "mine"
- **PRNHasWe:** Contains "we"
- **PRNHasOur:** Contains "our"
- **PRNHasOurs:** Contains "ours"
- **PRNHasYou:** Contains "you"
- **PRNHasYour:** Contains "your"
- **PRNHasYours:** Contains "yours"
- **PRNHasShe:** Contains "she"
- **PRNHasHer:** Contains "her"
- **PRNHasHers:** Contains "hers"
- **PRNHasHe:** Contains "he"
- **PRNHasHim:** Contains "him"
- **PRNHasHis:** Contains "his"
- **PRNHasThey:** Contains "they"
- **PRNHasThem:** Contains "them"
- **PRNHasTheir:** Contains "their"
- **PRNHasTheirs:** Contains "theirs"

NRC VAD Features

- **NRCAvgValence/Arousal/Dominance:** Average scores (0-1)
- **NRCHasHigh/LowValenceWord:** Presence of extreme values
- **NRCCountHigh/LowValenceWords:** Count of extreme values
- (Similar patterns for arousal and dominance)

NRC Emotion Features

For each emotion (anger, anticipation, disgust, fear, joy, sadness, surprise, trust):

- **NRCHas[Emotion]Word:** Binary presence
- **NRCCount[Emotion]Words:** Word count

For sentiments (positive, negative):

- **NRCHasPositive/NegativeWord:** Binary presence
- **NRCCountPositive/NegativeWords:** Word count

NRC WorryWords Features

- **NRCHasAnxietyWord**: Presence of anxious words
- **NRCHasCalmnessWord**: Presence of calm words
- **NRCAvgAnxiety/Calmness**: Average scores
- **NRCHasHighAnxiety/CalmnessWord**: Presence of extreme scores
- **NRCCountHighAnxiety/CalmnessWords**: Count of extreme scores

NRC Moral/Social/Warmth Features

Moral Trust Features

- **NRCHasHighMoralTrustWord**: Presence of high moral trust words (OrdinalClass=3)
- **NRCCountHighMoralTrustWord**: Count of high moral trust words
- **NRCHasLowMoralTrustWord**: Presence of low moral trust words (OrdinalClass=-3)
- **NRCCountLowMoralTrustWord**: Count of low moral trust words
- **NRCAvgMoralTrustWord**: Average moral trust score

Social Warmth Features

- **NRCHasHighSocialWarmthWord**: Presence of high social warmth words (OrdinalClass=3)
- **NRCCountHighSocialWarmthWord**: Count of high social warmth words
- **NRCHasLowSocialWarmthWord**: Presence of low social warmth words (OrdinalClass=-3)
- **NRCCountLowSocialWarmthWord**: Count of low social warmth words
- **NRCAvgSocialWarmthWord**: Average social warmth score

Combined Warmth Features

- **NRCHasHighWarmthWord**: Presence of high warmth words (OrdinalClass=3)
- **NRCCountHighWarmthWord**: Count of high warmth words
- **NRCHasLowWarmthWord**: Presence of low warmth words (OrdinalClass=-3)
- **NRCCountLowWarmthWord**: Count of low warmth words
- **NRCAvgWarmthWord**: Average warmth score

Additional Features

- **WordCount**: Total word count in the text