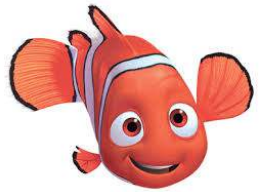


Computational Affective Science:

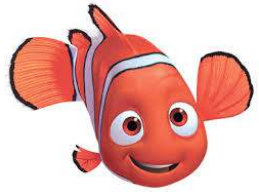
Exploring Fundamental Questions on Emotions through Language and Computation

Krishnapriya Vishnubhotla and Saif M. Mohammad

National Research Council Canada

vk22priya@gmail.com, uvgotsaif@gmail.com

@krishnapriyaVi5, @SaifMMohammad

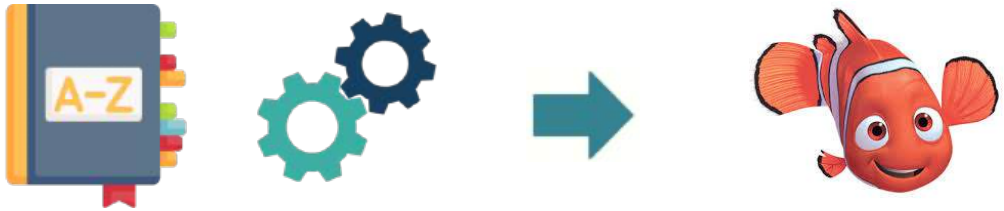


NLP for Affective Science:

Exploring Fundamental Questions on Emotions through Language and Computation

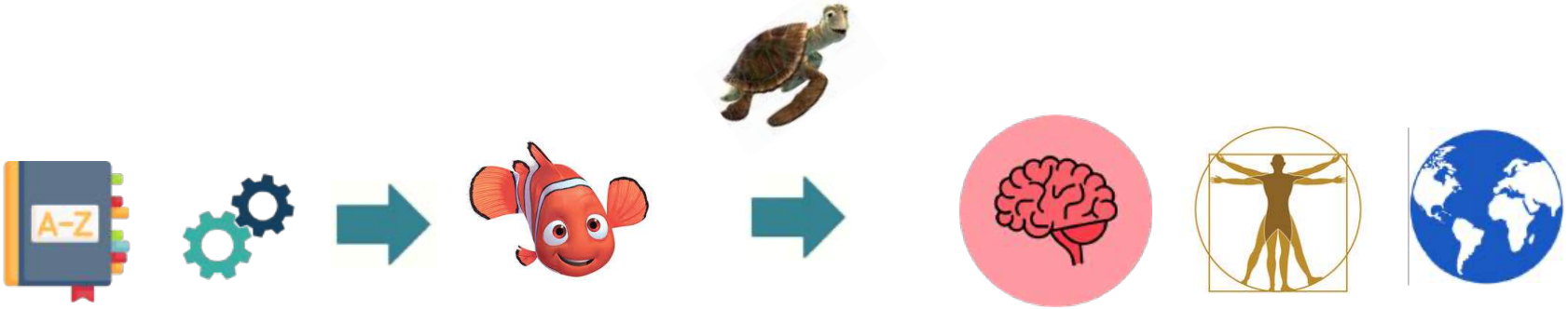
National Research Council Canada | Conseil national de recherches Canada

Canada



NLP for Affective Science:

A window into emotions through language and computation



NLP for Affective Science:

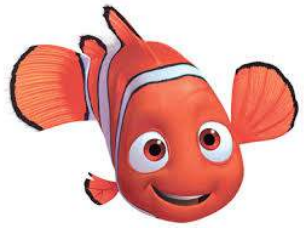
A window into emotions (mind, body, health, and behavior) through language and computation

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
  4. affective ethics

- Case Studies Exploring CAS Research



What are Affect and Emotions?

to start with…

Affect: the basic sense of feeling

Emotions: joy, sadness, fear, anger, etc.

Psychological Theories of Emotions

ON

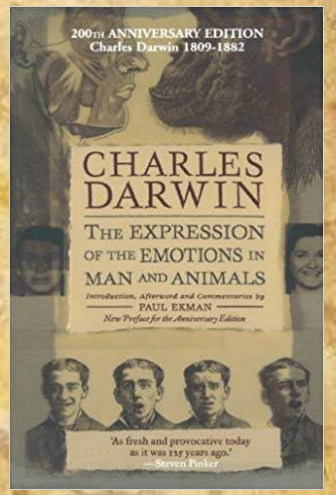
THE ORIGIN OF SPECIES

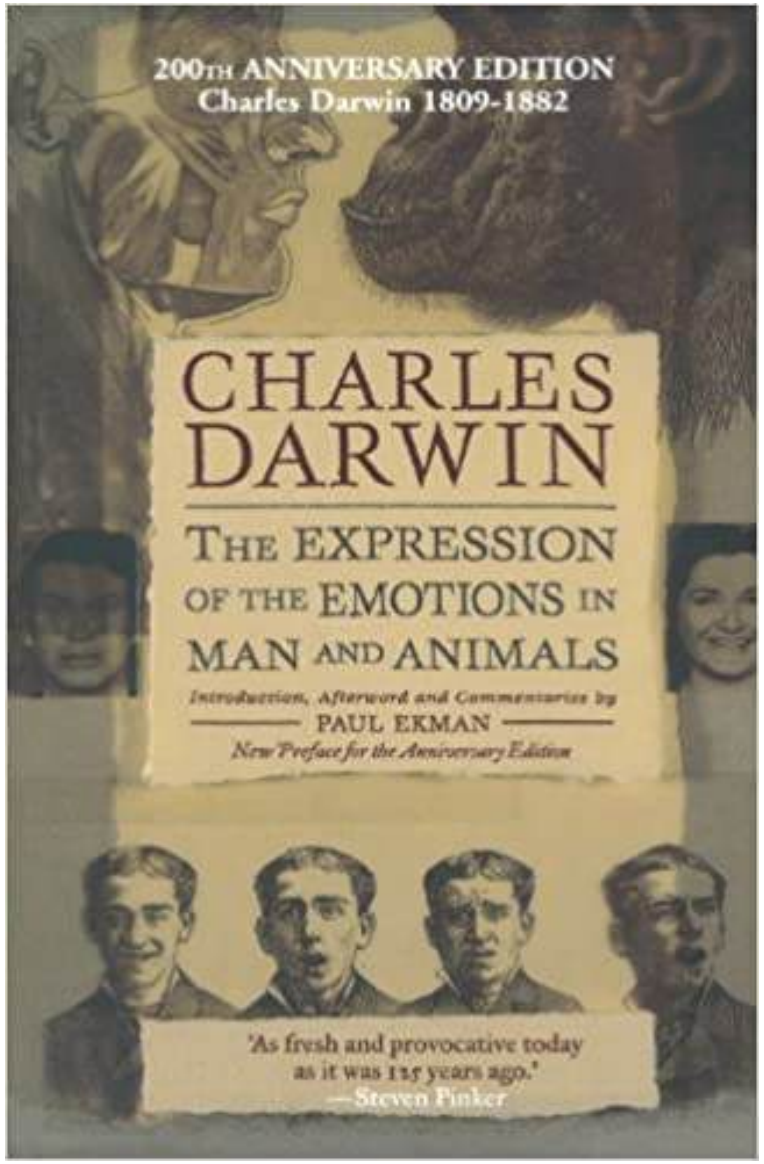
BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE

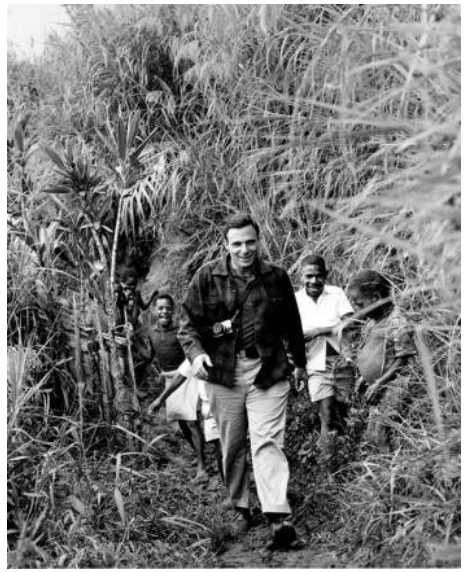
By CHARLES DARWIN, M.A.




200TH ANNIVERSARY EDITION
Charles Darwin 1809-1882
CHARLES DARWIN
THE EXPRESSION OF THE EMOTIONS IN MAN AND ANIMALS
Introduction, Afterword and Commentaries by
PAUL EKMAN
New Preface for the Anniversary Edition
'As fresh and provocative today as it was 125 years ago.'
—Steven Pinker

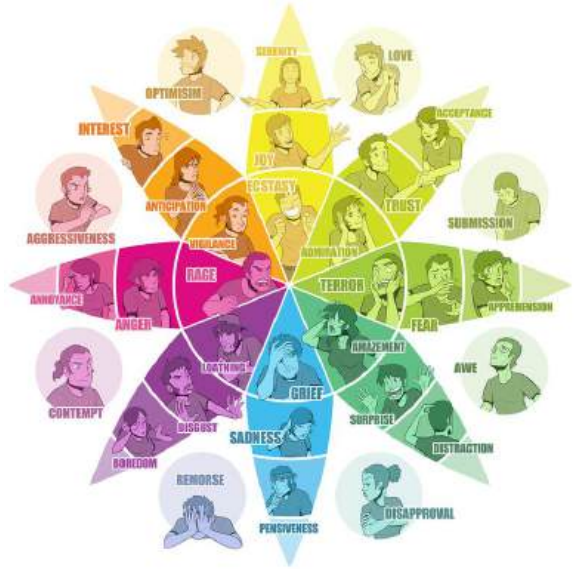
Theories of Emotion

Margaret Mead
Cultural anthropologist



Paul Ekman, Psychologist

- Mead, 1950s: culture determines emotion
- Paul Ekman, 1971: Six Universal Basic Emotions
  - Plutchik, 1980: Eight Basic Emotions
  - And many others



Plutchik's Emotion Wheel
Image credit: Julia Belyanevych

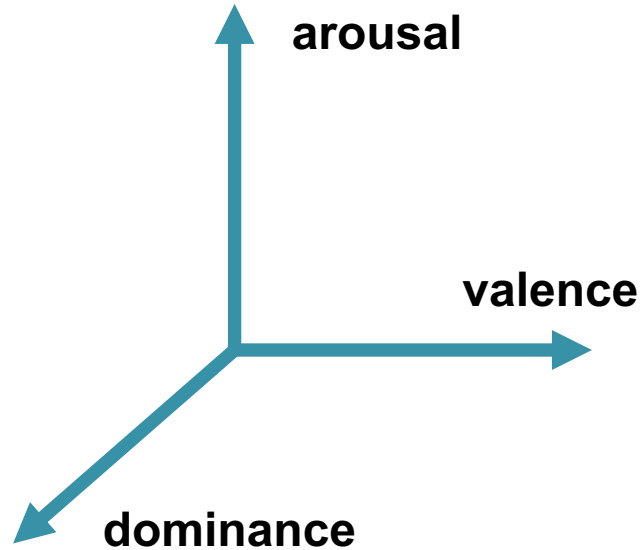
Core Dimensions of Connotative Meaning


Osgood


Russell

Influential factor analysis studies (Osgood et al., 1957; Russell, 1980, 2003) have shown that the three most important, largely independent, dimensions of word meaning:

- valence (V): positive/pleasure – negative/displeasure
- arousal (A): active/stimulated – sluggish/bored
- dominance (D): powerful/strong – powerless/weak



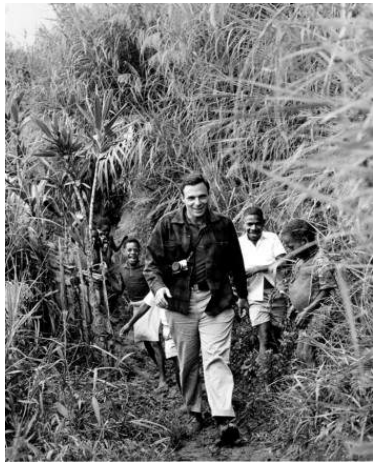
Thus, when comparing the meanings of two words, we can compare their V, A, D scores. For example:

- *banquet* indicates more positiveness than *funeral*
- *nervous* indicates more arousal than *lazy*
- *queen* indicates more dominance than *delicate*

Theories of Emotion

Margaret Mead
Cultural anthropologist

Paul Ekman
Psychologist and discoverer of micro expressions.



Lisa Barrett
University Distinguished Professor of Psychology, Northeastern University

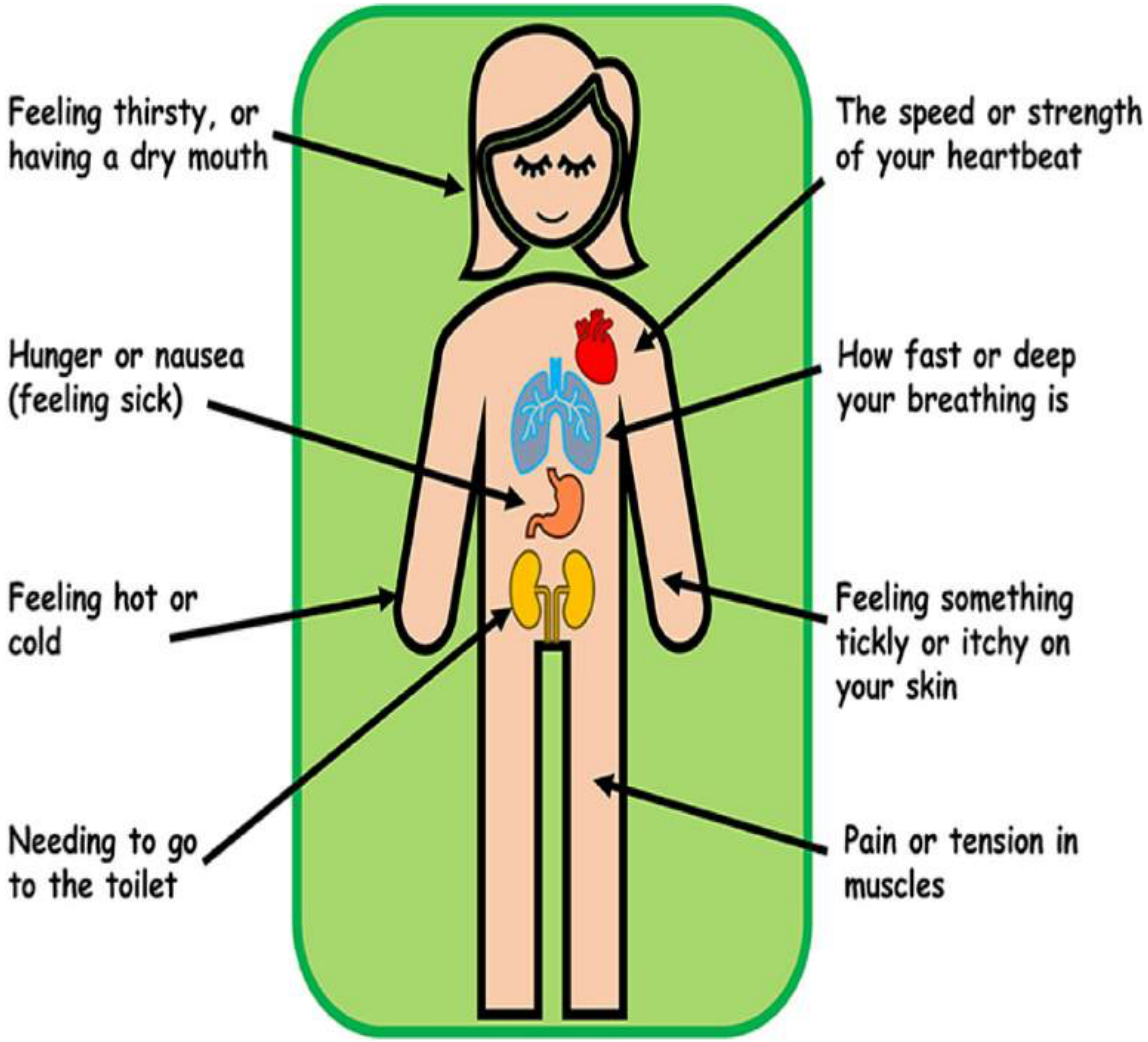
Theory of Constructed Emotion (Barrett, 2017)

- the brain constructs emotions
- important tenets of BET discredited ("basic" emotions)
- stress on variability

Interoception

the ability to sense signals about your bodily state

Interoceptive signals
sensory representations of the interior of the body (viscera)



Constructed Emotion

External Situation	Bodily Signal	Possible Emotions
 Public speaking in front of an audience	Heart races Palms sweat Stomach churns	Anxiety Excitement
 Walking alone at night and hearing footsteps behind you	Muscles tense Heart rate spikes Breathing shallow	Fear Apprehension Hypervigilance
 Sitting alone at café	Slow breathing Relaxed muscles	Loneliness Contentment

Affect

The basic sense of feeling:

- Transduced and summarized from interoceptive signals
  - sensory representations of the interior of the body (viscera)
- A feature of consciousness
  - occurs in every moment (whether you're aware of it or not)
- Key dimensions
  - Valence: displeasure to pleasure
  - Arousal: idle/sluggish to activated
  - Dominance: weak/loss of control to strong/having a sense of control

Emotions

Constructed by the brain using:

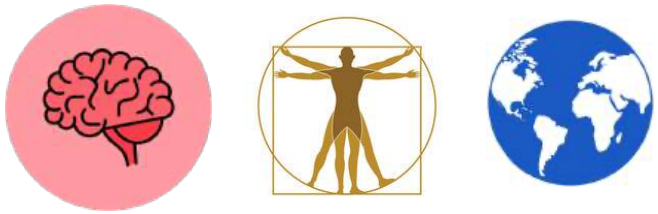
- affective and interoceptive signals
- "emotion concepts" from one's culture learned through socialization
- predictive coding

The brain categorizes the continuous affect into discrete categories (analogous to color perception)

- joy, sadness, fear, anger, etc.

Why Emotions and Affect Matter

- Determine human experience and behavior
- Condition our actions
- Central in organizing meaning
  - No cognition without emotion
- A window into
  - understanding our mind (cognition), body (health, well-being), how we make sense and interact with the world
  - the evolutionary forces that shaped us



Affective Science

Interdisciplinary field focused on understanding emotions and affect

- How do affect and emotions work?
  - affective and emotional processes
  - affective neuroscience
  - emotion regulation
- How do they impact our mental health, physical health, and behaviour?
  - wellbeing, emotion-related disorders
- What agency do we have in managing our emotions?

The Language of Emotions

Language is a powerful way of expressing emotions

- can express numerous emotional shades
  - terms with fuzzy boundaries, overlapping meanings, socio-cultural influences, etc.
- usually conveyed by connotation (and not denotation)
  - can be subtle, direct, ambiguous, deceptive
  - can be creative
  - can be conscious expression or subconscious manifestations

Computational Affective Science (CAS) aka NLP for Affective Science

Computational Analysis of Emotions Through Language

- Challenging
  - see previous slide on language
- Powerful
  - makes use of large amounts of text
  - simple to complex NLP techniques
  - language impacts thought and how we construct emotions
    linguistic relativity aka Sapir–Whorf hypothesis
- Complementary view to traditional affective science approaches in psychology
  - makes use of, complementary, ecologically valid data

Computational Analysis of Emotions Through Language

Affective Science Questions

- How do emotions work?
- What impacts our emotions?
- How do emotions relate to our bodies and health?
- How do we regulate emotions?

Linguistics

- How do we use language to convey emotions?
- How does language impact emotions?

Social Science

- How do emotions impact social cognition, morality, stereotypes, and behavior?

AI

- What tools can we build to help people, clinicians, social scientists?

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
  4. affective ethics

- Case Studies Exploring CAS Research

CAS: Areas of Research

1. The Nature of Affect

   The Relationship of Affect with:
   a. the mind, cognition
   b. the body
   c. the world we interact with

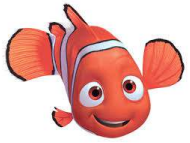


2. Affective Data and Resources

3. Affective Tasks

4. Affective Ethics

CAS 1: The Nature of Affect



Computational experiments that add to our understanding of affect and emotions.

- findings relevant to theories of emotion
- the biology of emotions
- the neuroscience of emotions
- models of emotions
  - appraisal models, dimensional models (valence / arousal / dominance), models of constructed emotion, cognitive-affective architectures, emotion dynamics (how emotions emerge, intensify, decay, or transition over time), emotion granularity, emotion regulation, affective embodiment, evolutionary affect development, developmental affect (how emotions and affect change over a life span), emotion and cognition, etc.

Note that many of these can be applied to:

- human beings, animals, and even text by artificial agents

CAS 1a: Affect and the Mind

Includes work on:

- Cognition
- Constructing meaning (sense making)
- Theory of mind
- Beliefs, Stereotypes, Opinions, Stance, Narratives, Stories
- Mental Health, Psychopathology, Mental Disorders

Overlaps with:

- Psycholinguistics
- Digital Humanities (DH)
- Computational Social Science (CSS)
- Hate Speech, Stereotypes, Bias work

CAS 1b: Affect and the Body

Embodiment

the idea that the body plays a central role in how we think, feel, and understand the world—not just the brain alone.

cognition, emotion, and meaning are shaped by our bodily experiences, including perception, movement, physiology, and interaction with the environment.

Example:

- Understanding the word *"grasp"* activates motor areas involved in grasping

Emotions are not just mental labels; they involve bodily patterns:

- Anxiety: increased heart rate, muscle tension
- Calmness: slower breathing, relaxed posture

How we interpret bodily signals shapes emotional experience (e.g., interoception)

CAS 1b: Affect and the Body (continued)

Language and Meaning

Many abstract concepts are grounded in physical experience:

- Valence: good/bad as approach/avoid
- Dominance: up/down, big/small
- Time: moving forward/backward

Language reflects these embodied mappings

Overlaps with:

- Cog Sci
- Psychology
- Health Science
- Linguistics
- CS

AI and Robotics

- Embodiment suggests intelligence emerges from acting in the world
- Robots learn better when perception and action are tightly coupled

Health

- mental, physical

CAS 1c: Affect and the World

Modeling:

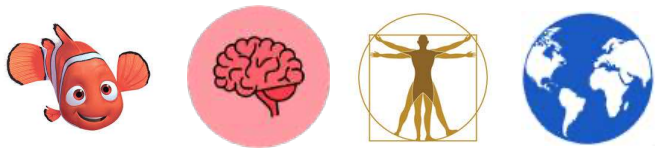
- Interpersonal affect
- Empathy
- Group-level affect modeling
- Affect contagion
- Polarization
- In-group and Out-group bias, stereotypes
- Hate speech
- How we react to misinformation
- Computational models of emotion regulation

etc.

Overlaps with:

- Social Sciences
- DH
- Information Sciences
- Psychology
- Communication Sciences
- Health Science

Affective states, cognitive states, bodily states, and the state of the world all interact in complex ways.



CAS 2: Affective Data and Resources

Work on compiling and annotating affect-related information:

- **Modality:** text, speech, images, physiological signal processing (ECG, EEG, GSR, multimodal biosensing), multimodal
- **Granularity:** Words, MWEs, Sentences, Posts, Streams of Text, Chat/Conversations/Dialogues, Documents, images, memes, videos
- **Genre:** Social media (blogs, microblogs), news, essays
- **Domain:** climate change, health, commerce, surveillance

CAS 3: Affective Tasks

Emotion Recognition, Generation

- Sentence/Post/Instance level: classification/regression/generation
  - emotion classification (discrete emotions, dimensional ratings)
  - emotion intensity estimation
  - emotion-cause detection (what triggers an emotion in text, video, or interaction)
  - context-aware affect inference (considering culture, situation, social setting)
  - generating text conveying appropriate affect

- At an Aggregate Level: determine trends across a large number of instances
  - group affect over time, across locations, etc. (emotion arcs)
  - groups affect towards an entity/issue (e.g., climate chance, vaccines) over time (emotion arcs)
  - group--group affect differences
  - document-level and cross-document emotion analysis

CAS 4: Affective Ethics, Fairness, Theory Integration, Philosophical Implications

- Bias and generalizability of affective systems across demographics
- Privacy and ethics in affective data collection
- Critically examining whether automatic NLP systems are relying on the current and valid theories of affect and emotion
- What it means for machines to model or simulate affect
  - Broader societal implications of affective artificial agents
- Explainability and interpretability in computational affective models

CAS Applications

- **Mental Health and Well-Being**
  - Early detection of anxiety, depression, stress, and burnout
  - Emotion-aware interventions, therapy support, and self-tracking tools
- **Human–Computer Interaction (HCI)**
  - Emotion-adaptive interfaces and personalized user experiences
  - Affective computing for more natural, empathetic AI systems
- **Natural Language Processing and AI**
  - Emotion, sentiment, and mental-state analysis of text and speech
  - Emotion-aware chatbots, moderation systems, and recommender systems

CAS Applications (continued)

- **Education and Learning Analytics**
  - Detecting engagement, confusion, frustration, and motivation
  - Adaptive tutoring systems and emotionally responsive feedback
- **Health Communication and Public Health**
  - Analyzing emotional framing in health messaging
  - Improving risk communication and behavior-change campaigns
- **Media, Marketing and Persuasion**
  - Understanding emotional impact of narratives, ads, and campaigns
  - Measuring audience response and affective engagement

How is CAS different from traditional sentiment analysis and affective computing?

Lets take a step back…

What are the Research Question Types in NLP Research?

A. About Computational Modeling, ML
B. About Language
C. About Affect and Emotions
D. About People, social behavior, evolution, psychology/cognitive, health, etc.

Some combination of the above

What is more common?

The center of gravity is closer to A in NLP, Sentiment Analysis, and Affective Computing.

In CAS it is closer to C.

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
  4. affective ethics

- Case Studies Exploring CAS Research

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
  4. affective ethics

- **Case Studies** exploring CAS Research

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
     a. word-emotion datasets
     b. sentence/post-emotion datasets
  3. affective tasks
  4. affective ethics

- **Case Studies** Exploring CAS Research

word --

2a: Word-Emotion Association Lexicons

Affect Datasets for Words

Lexicons for both categorical emotions as well as for valence, arousal, and dominance

- Lists of words associated with joy, sadness, fear, etc.
- Lists of words and their valence, arousal, and dominance scores

Affect Lexicons: VAD

- Affective Norms of English Words (ANEW) (Bradley and Lang, 1999)
  - ~1,000 words, 9-point rating scale
  - The ratings from annotators averaged to obtain a final score for the word
- Other than English
  - Redondo et al. (2007): Spanish
  - Vo et al. (2009): German

Affect Lexicons: Categorical Emotions

- NRC Word–Emotion Association Lexicon aka NRC Emotion Lexicon or EmoLex (2010)
  - provides associations for ~14k English words with
  - eight (Plutchik) emotions: anger, fear, joy, sadness, anticipation, disgust, surprise, trust
  - translations in over 100 languages

Widely used.

Use of The Emotion Lexicons

- For research by the scientific community
  - Computational linguistics, psychology, digital humanities, robotics, public health research, social science, etc.
- To analyze text
  - Brexit tweets, Radiohead songs, Trump tweets, election debates,…
  - **Wishing Wall**, uses the NRC Emotion lexicon to visualize wishes. Displayed in:
    - **Barbican Centre**, London, England, 2014
    - **Tekniska Museet**, Stockholm, Sweden, 2014
    - **Onassis Cultural Centre**, Athens, Greece, 2015
    - **Zorlu Centre**, Istanbul, Turkey, 2016

creativity

- In commercial applications

Affect Lexicons: VAD

- Affective Norms of English Words (ANEW) (Bradley and Lang, 1999)
  - ~1,000 words, 9-point rating scale
  - The ratings from annotators averaged to obtain a final score for the word
- Other than English
  - Redondo et al. (2007): Spanish
  - Vo et al. (2009): German
  - Moors et al. (2013): for Dutch
- Warriner et al. (2013): 13k English words
- The NRC VAD lexicon (Mohammad, 2018, 2024): comparative annotations (instead of rating scales); English with translations; high split-half reliability (>0.9)
  - v1: 20k words
  - v2: 45k words, 10k MWEs

Reliability vs. Inter-Annotator Agreement

Inter-Annotator Agreement (IAA)

- Measures how much annotators agree with each other
- Reflects agreement on labels across annotators
- Common metrics: Cohen's κ, Fleiss' κ, Krippendorff's α
- High IAA implies: annotations are applied similarly

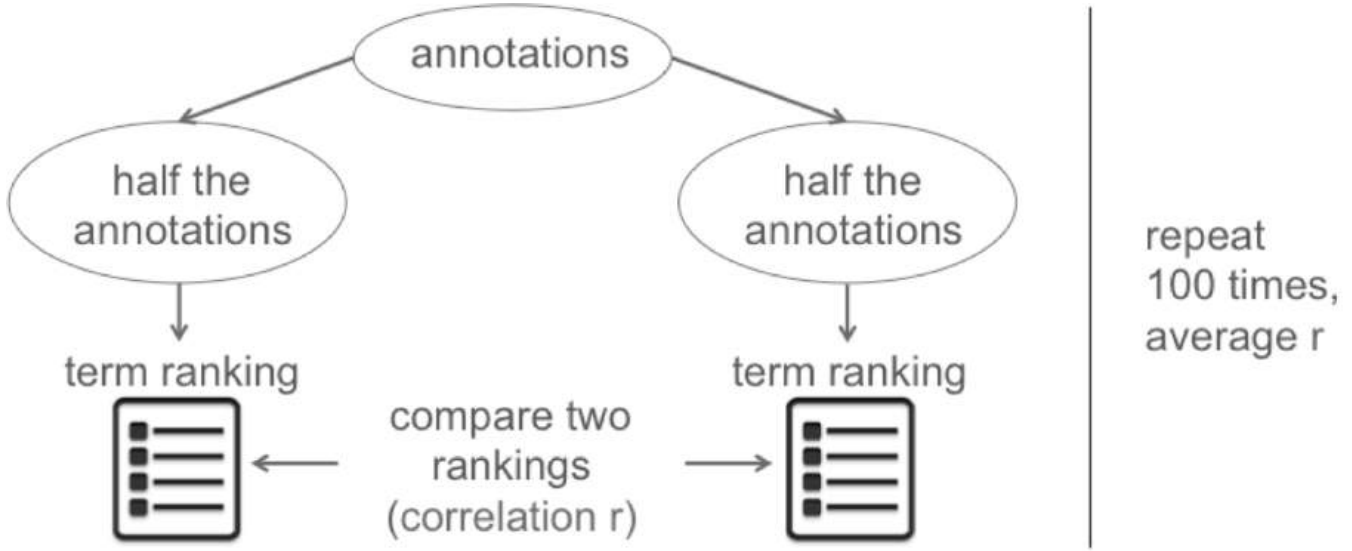
Reliability

- Measures how dependable and reproducible the annotations are
- Reflects stability/trustworthiness of the aggregated scores/labels
- Common metrics: split-half reliability, split-half class match reliability
- High reliability implies: repeat annotations will lead to similar aggregated scores/labels

Reliability (Reproducibility) of Annotations

Average split-half reliability (SHR): approach to determine consistency

(Kuder and Richardson, 1937; Cronbach, 1946)



Some Things We Learned About Emotions

- Variability
  - cultural
  - context
  - person-to-person
  - what/how you ask about emotions
    - more agreement for valence; less for fear, sadness, etc.
    - associated vs. evoked
    - greater agreement with comparative annotations

- Substantial common ground

Draw inferences at aggregate level

- determine broad trends

Affect Lexicons: Categorical Emotions

- NRC Word–Emotion Association Lexicon aka NRC Emotion Lexicon or EmoLex (2010)
  - provides associations for ~14k English words with
  - eight (Plutchik) emotions: anger, fear, joy, sadness, anticipation, disgust, surprise, trust
  - translations in over 100 languages
- The NRC Emotion Intensity Lexicon aka Affect Intensity Lexicon (2018-19)
  - provides intensity scores for ~6000 words found to be associated with the 8 emotions
- Word-Anxiety-Calmness Lexicon (WorryWords) (2024)
- Warmth and Competence (and Sociability, Trust) (2025)

Anxiety

the anticipatory unease about a potential (future) negative outcome

- common and beneficial human emotion
- can sometimes manifest into mental disorders
  - mismatch: current environment and
    what anxiety response slowly evolved to address

Why create language resources for anxiety?

- Understanding anxiety and the underlying mechanisms (Psych, Health)
  - how it relates to other emotions and affect
  - how it relates to our body
  - how anxiety changes with age, socio-economic status, weather, green spaces, etc.
  - identifying coping mechanisms, clinical interventions to manage anxiety
- Study how anxiety manifests in language (Ling.)
  - how language shapes anxiety
  - how culture shapes the language of anxiety
- Tracking the degree of anxiety towards targets of interest such as climate change, government policies, biological vectors, etc. (Health, Policy)
- Developing automatic systems for detecting anxiety (NLP)
- Studying how anxiety impacts behaviour in physical and virtual environments (SS)
- Studying anxiety in stories, character development, etc. (DH)

WorryWords

Repository of manually derived word–anxiety associations

- Scale: maximum calmness (-3) to maximum anxiety (3)
  - real-valued scores and also coarse categorical labels (e.g, low anxiety, high anxiety)
- Size
  - 44K English words
  - 10K English MWEs
- Quality
  - interspersed gold (control) questions
  - show that the anxiety associations are highly reliable
    - split-half reliability of 0.82

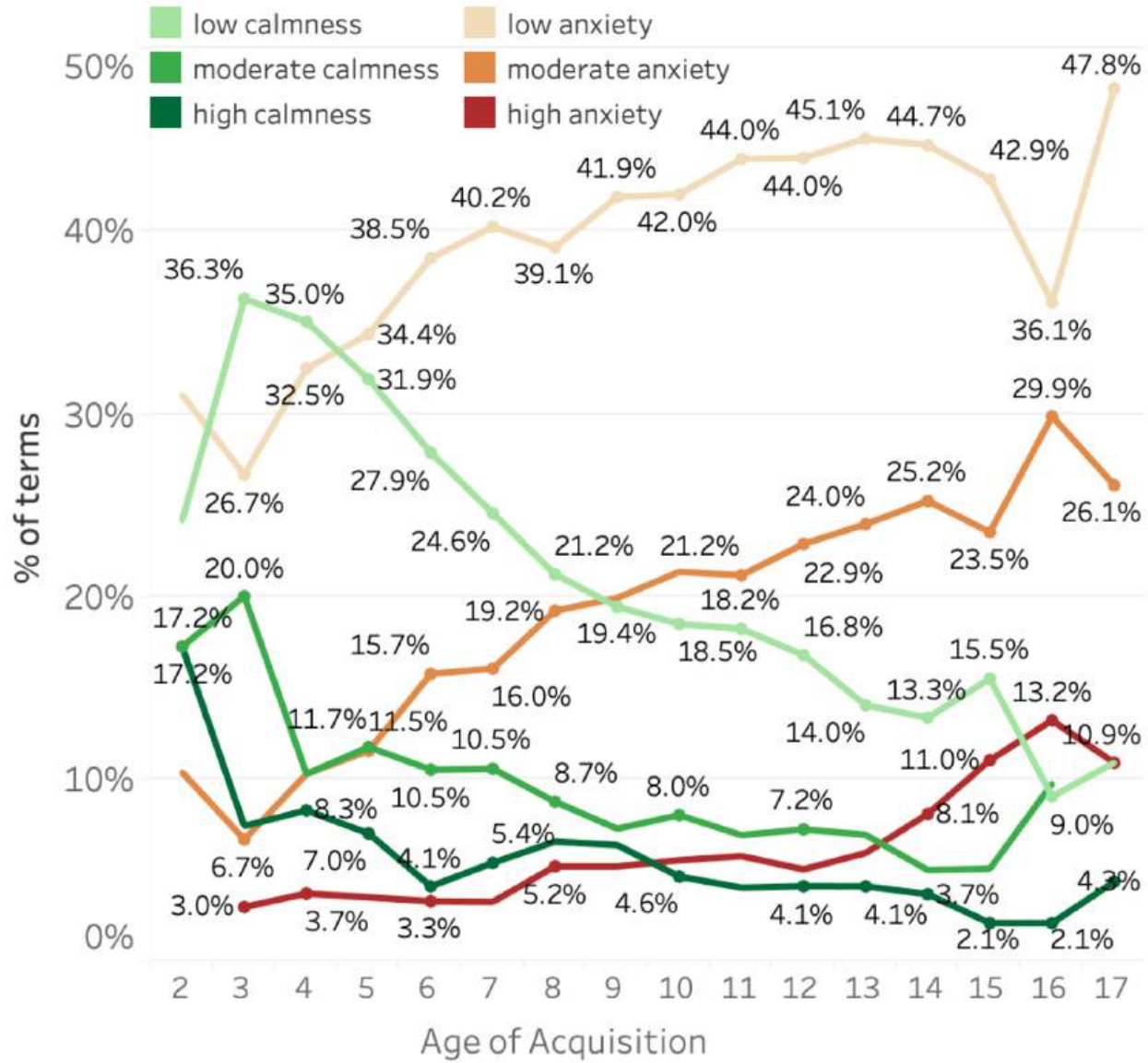
Term	Score
suffocative	3.00
manic	2.41
riskily	1.72
ceramic	0.12
conformed	-1.71
lullaby	-2.79

EMNLP 2024:

WorryWords: Norms of Anxiety Association for over 44K English Words. Saif M. Mohammad.

WorryWords

study the rate at which children acquire anxiety words with age



Used WorryWords to

Track the change of anxiety in streams of text

Words of Warmth:

Trust and Sociability Norms for over 26k English Words

Saif M. Mohammad, ACL 2025

National Research Council Canada | Conseil national de recherches Canada

Canada

Warmth And Competence

Warmth And Competence

Primary dimensions we assess people and groups on:

- Warmth (W) 
  - Sociability (S): friendliness, gregariousness, and conviviality 
  - Trust (T): morality, goodness, sincerity, and integrity 
- Competence (C) 
  - ability, power, dominance, and assertiveness

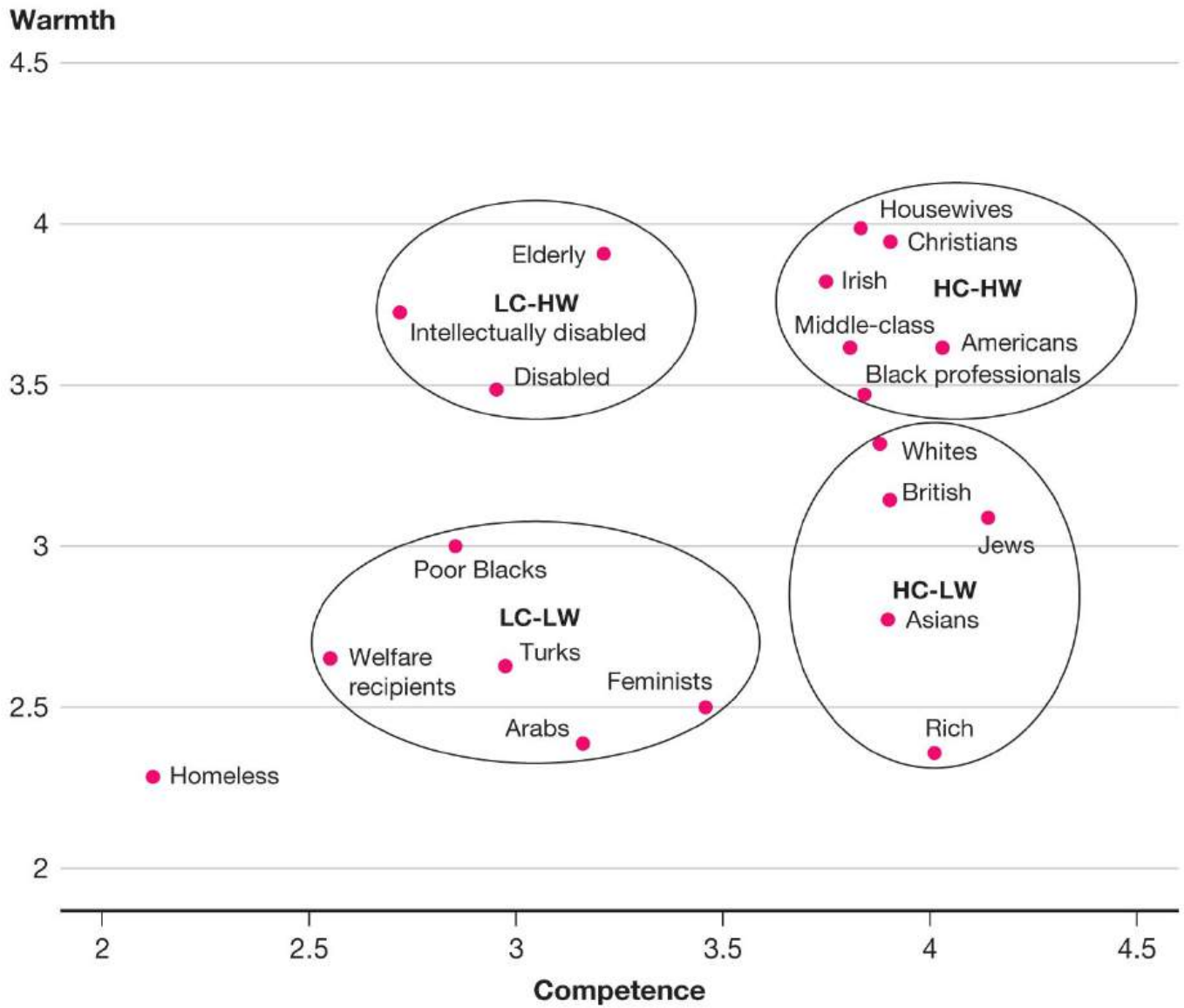
[Decades of social cognition and stereotype research. Notably by Susan Fiske and colleagues. Evolutionary benefits.]

American Stereotypes

Cuddy, Fiske, and Glick 2008

Tied to who is considered to be in their:

- ingroup
- outgroup



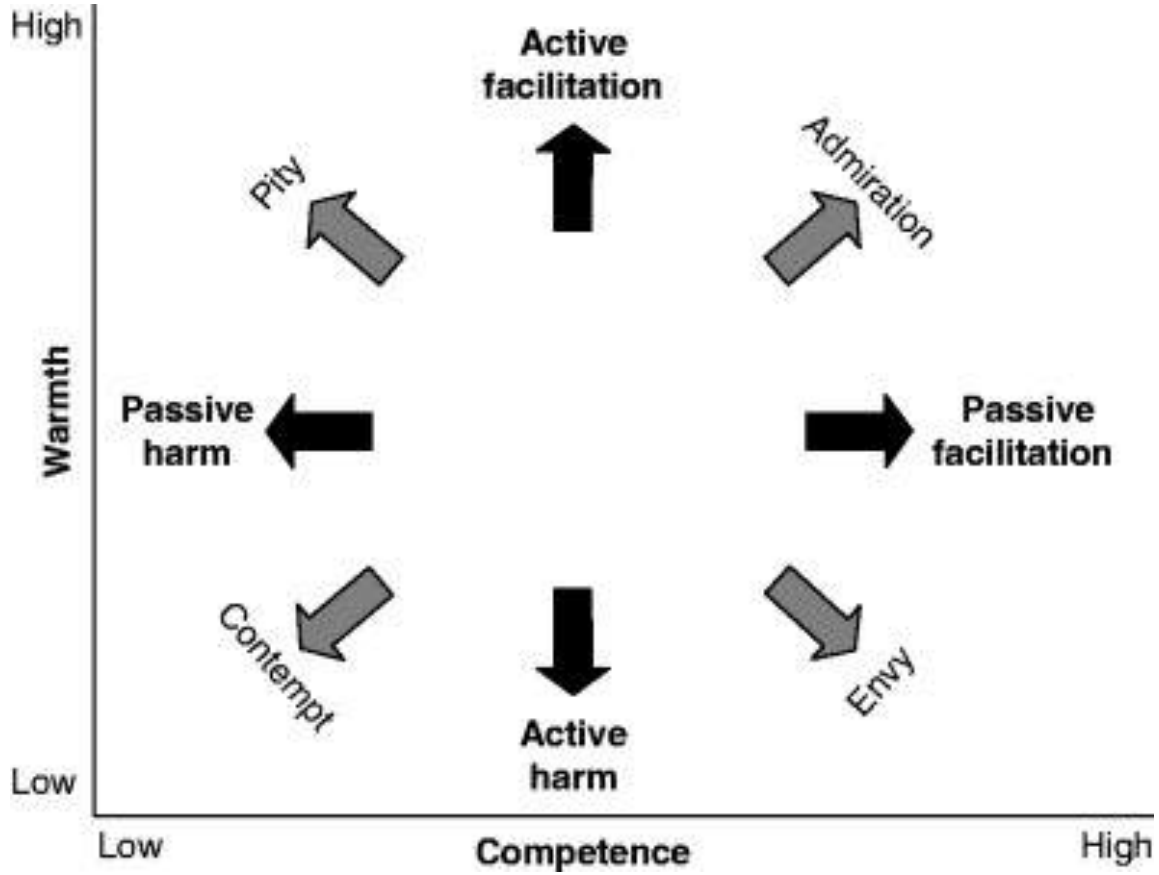
American Stereotypes

Cuddy, Fiske, and Glick 2008

Tied to who is considered to be in their:

- ingroup
- outgroup

Different quadrants associated with different emotions



Why create language resources for WCTS?

In Psychology and Social Cognition

- What kind of WCTS assessments do children develop first?

In Computational Social Science, NLP

- Levels of WCTS in public discourse (climate change, vaccines, etc.)

In HCI and NLP

- Perceptions of WCST of people towards artificial agents

In Digital Humanities and NLP

- Role do WCTS in developing compelling characters and story arcs

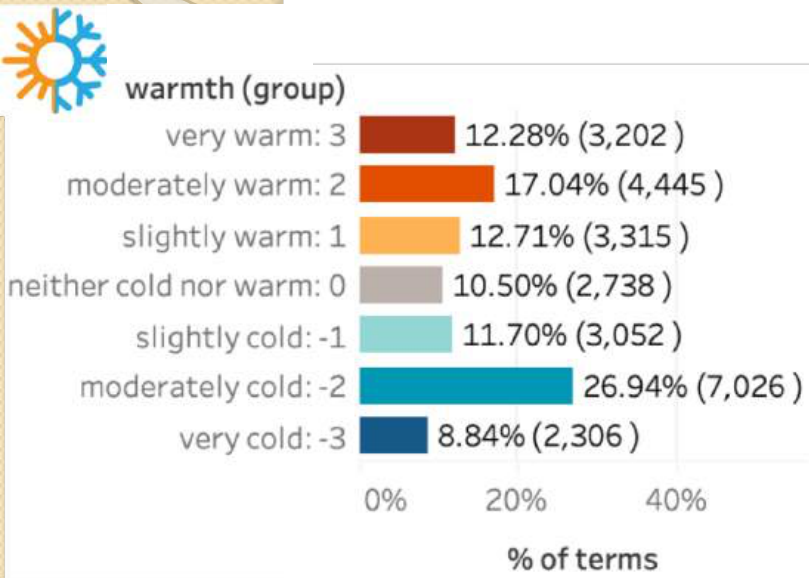
In Commerce

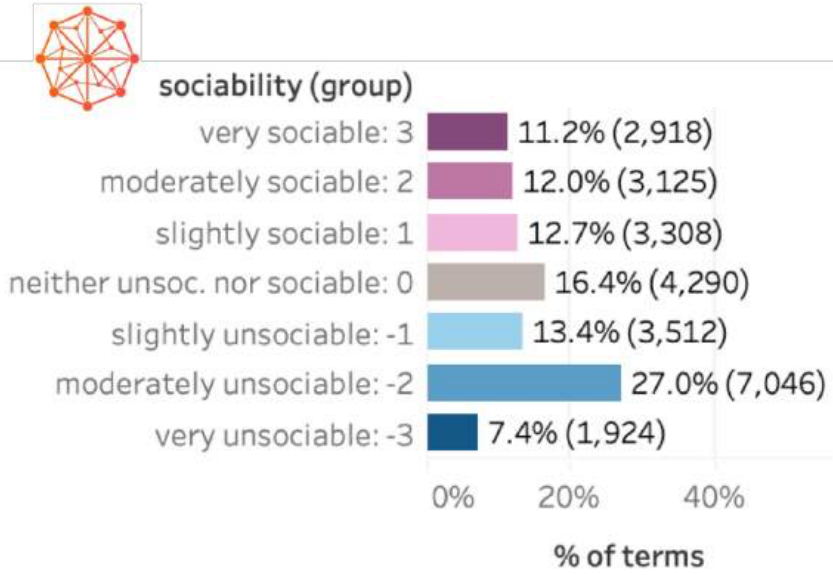
- Warmth, competence, trust, and sociability towards one's product on social media

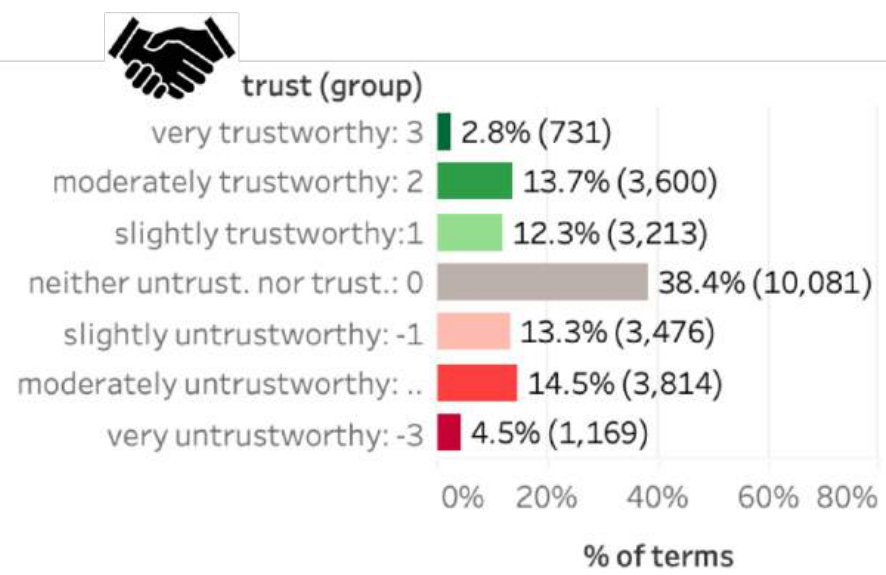
Words Of Warmth Lexicon

- Manually derived fine-grained scores of association for 26K English words
  - maximum coldness / unsociability / untrustworthiness (-3) to maximum warmth /sociability / trustworthiness (3)
    - W scores taken as union of S and T scores
- The annotations are reliable (high split-half reliability scores)
- Warmth analyses are often done along with competence (aka dominance) analyses
  - we include in the lexicon the competence scores taken from the NRC VAD Lexicon v2 (Mohammad, 2025)

Class Distributions





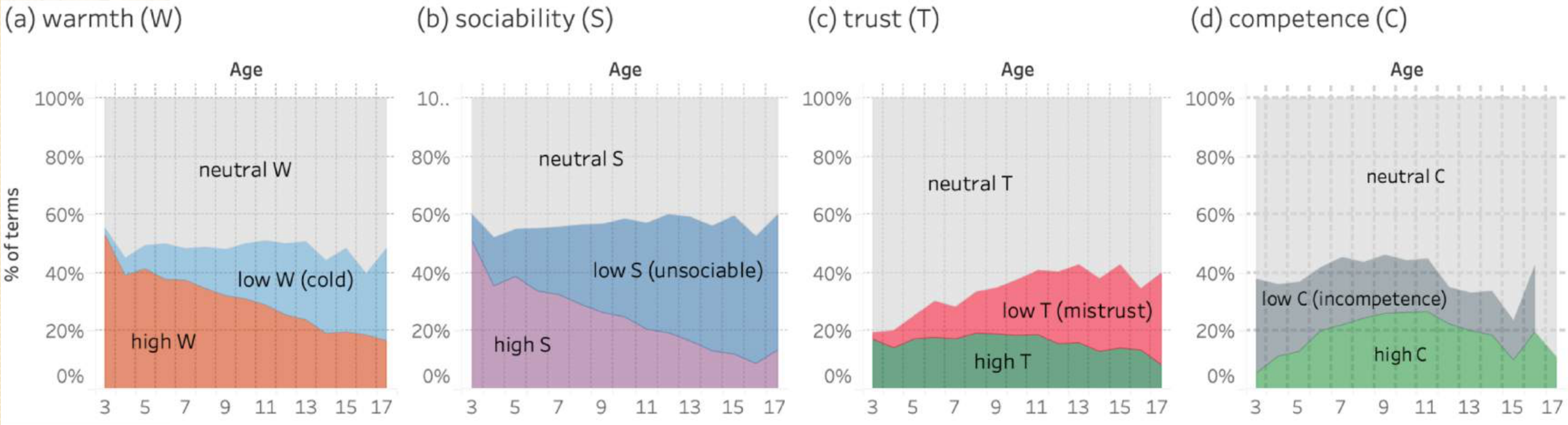


Experiments

1. At what rate do children acquire WCST words?

Rate at which children acquire WCST words

Graphs created using: Words Of Warmth, NRC VAD, Age of Acquisition lexicons.



W/S/T scores −1.5 to 1.5: neutral; <= −1.5: low; >= 1.5: high. C scores −0.33 to 0.33: neutral; −1 to −0.33: low C; 0.33 to 1: high C.

- Higher % for polar W words vs. polar C words: consistent with the primacy of valence hypothesis (not primacy of C).
- Higher % for polar S words as opposed to polar T words in early years: S is more important than T (and morality).
- Among the polar words, the early years are marked with a greater % of high-WST words, as well as low-C words.

Experiments

2. Case Studies of W and C Stereotypes

Measuring Stereotypes

Windows into stereotype towards various targets: Two methods

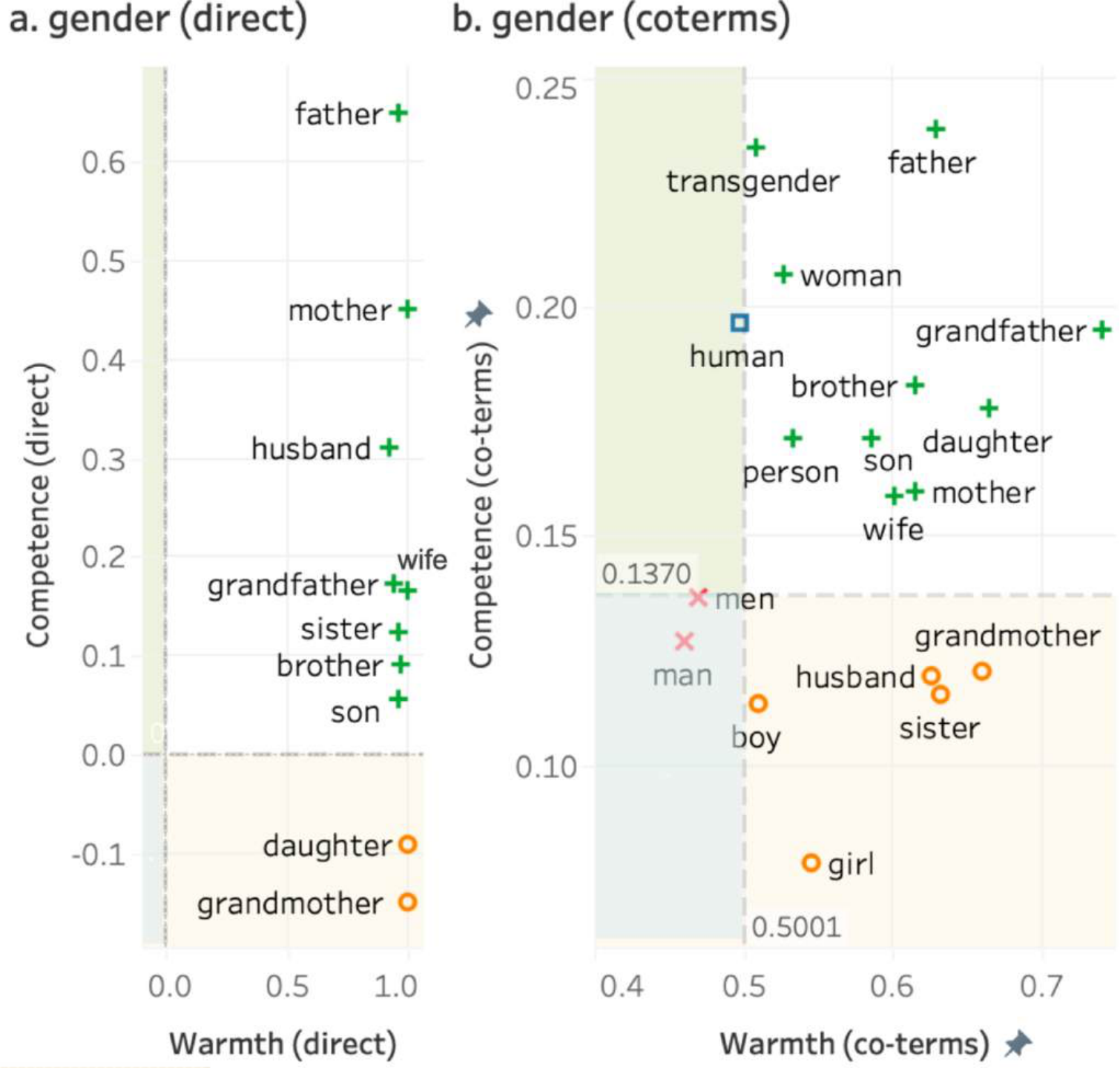
- Direct: Direct Target Lookup in the WCTS lexicon
- Co-terms: Examining WCTS of terms co-occurring with the target terms in the TUSC dataset

  American and Canadian geo-located posts on X from 2015 to 2021 (Vishnubhotla and Mohammad, 2022).

  Average W and C scores of all 3.1 million posts in the corpus: 0.5001, 0.1370 (in a range from -1 to 1).

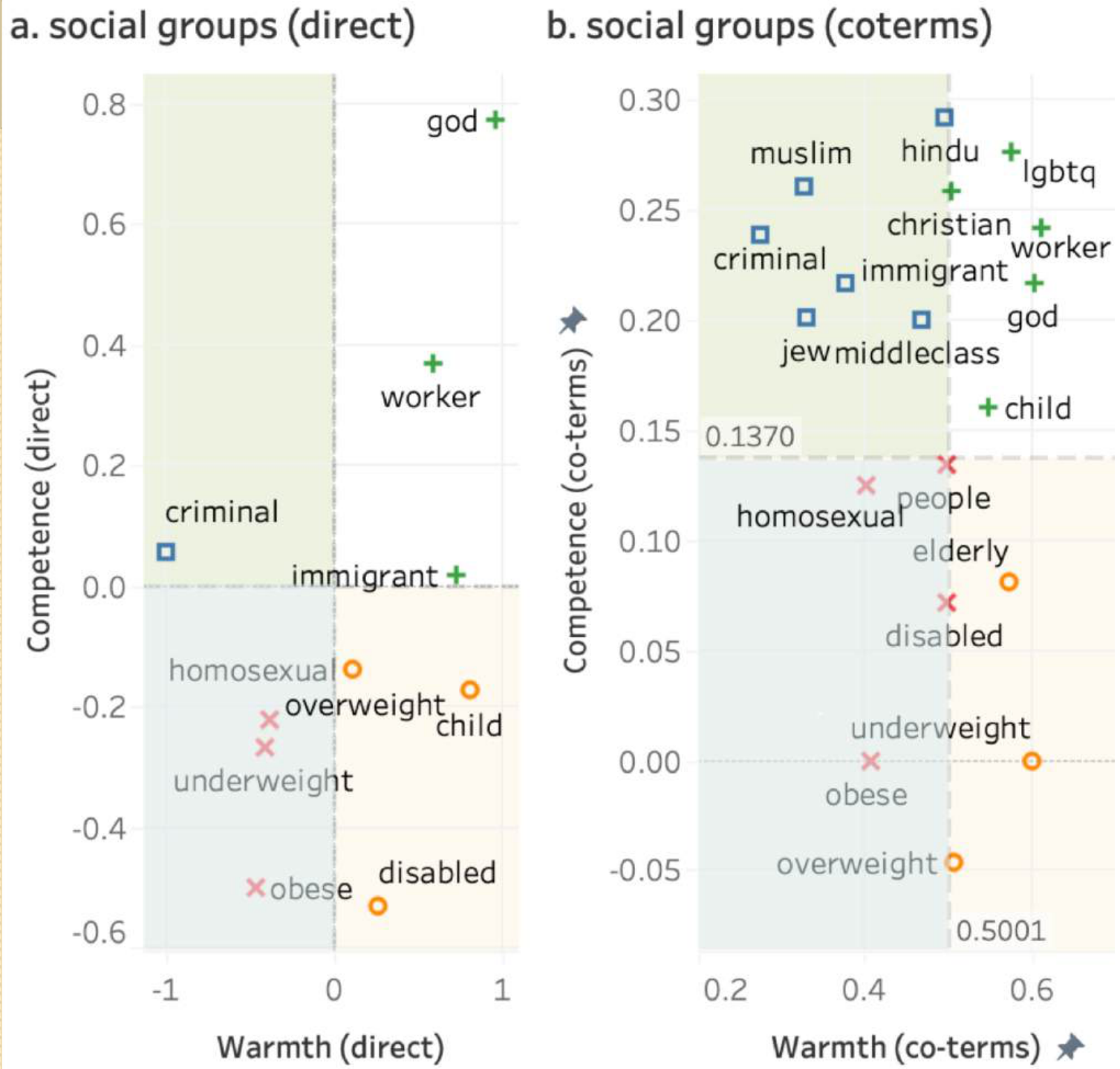
Ethical Considerations

- Associations and stereotypes; not inherent properties
- Consider coverage, domain, ambiguity, socio-cultural effects, etc.
- Ethics Sheet for Emotion Recognition (Mohammad, 2022) [CL Journal]
- Best Practices in the Use of Emotion Lexicons (Mohammad, 2020)



Relations

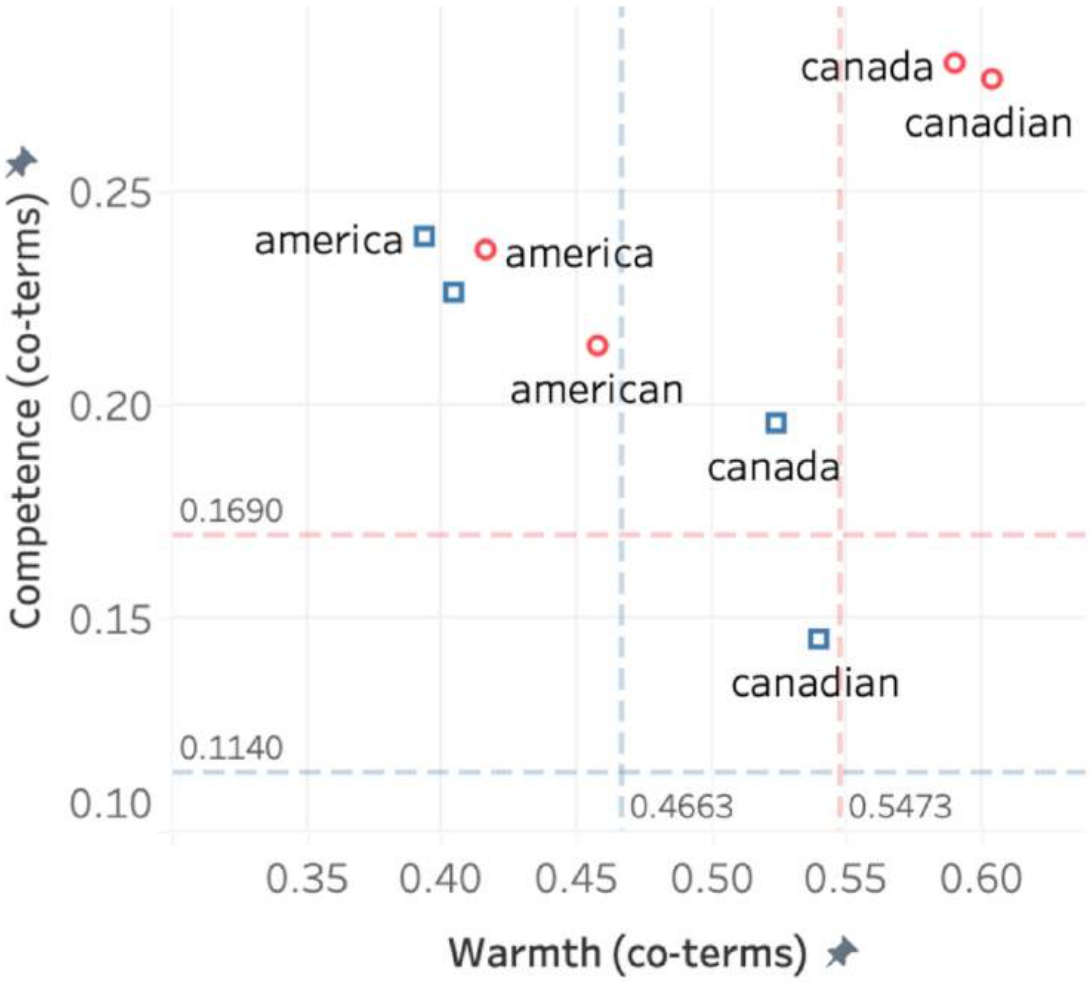
- Direct WC: people consider all these terms as high W; substantial variations in their C.
- In contrast, the co-term plots show that our language has marked differences for these terms for C and W.
- Clear gender stereotypes reflected in these scores



Social Groups

- *muslim, jew, immigrant:* low-W scores (consistent with known negative stereotypes in US, CA)
- *elderly, underweight:* low-C, high-W; *overweight:* even lower C score *obese:* low-W, low-C scores
- *god:* high direct W and C scores; but the discourse around *god* on X is such that the term gets lower co-terms-based C score than many other terms

In-group–Out-group Dynamics



The blue dashed lines indicate the average W and C scores of all posts by Americans and red dashed lines indicate the averages for Canadians.

- posts by Canadians in general have higher W and C scores than posts by Americans
- Canadians view themselves as more competent and much warmer than their neighbour (consistent with ingroup and out-group stereotypes)
- while Americans view themselves as more competent than Canadians, they too perceive Canadians as warmer (the *Canadians are nicer* stereotype overrides the outgroup stereotype)

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
     a. word-emotion datasets
     b. sentence/post-emotion datasets
  3. affective tasks
  4. affective ethics

- **Case Studies** Exploring CAS Research

post --

2b: Affect Datasets for Sentences/Posts (Annotated Corpora)

What can be annotated?

Text (and Context):

- News articles and headlines, essays
- Blogs/micro-blogs
- Narratives/stories
- Conversational data
- …

Annotators:

- Crowdsourced
- Expert or self (writer) annotation
- Self-supervised labels/proxies

Annotation:

- Choose between class labels
- Open-ended coding
- Comparative annotations (BWS)

- Emotion category labels (expressed by, towards, evoked in, ..)
- Emotion intensity
- V/A/D intensities
- Appraisal evaluations
- Semantic role labels (experiencer, agent, trigger, ..)

Emotions in Tweets

[1] label tweets for emotion categories using the associated hashtags.

- Six emotion classes (Ekman): *#anger, #disgust, #fear, #happy, #sadness*, and *#surprise*.

1. *Feeling left out... #sadness*
2. *My amazing memory saves the day again! #joy*
3. *Some jerk stole my photo on tumblr. #anger*
4. *Mika used my photo on tumblr. #anger*
5. *School is very boring today :/ #joy*
6. *to me.... YOU are ur only #fear*

Table 2: Example tweets with emotion-words hashtags.

hashtag	# of instances	% of instances
#anger	1,555	7.4
#disgust	761	3.6
#fear	2,816	13.4
#joy	8,240	39.1
#sadness	3,830	18.2
#surprise	3,849	18.3
Total tweets	21,051	100.0
# of tweeters	19,059	

Table 3: Details of the Twitter Emotion Corpus.

Twitter Emotion Corpus (TEC):

- >20k tweets self-labelled for emotion category.
- Class imbalance: certain emotions are expressed more on social media.

Are hashtags good labels?

- Validate with a classification setup: identify emotion in tweet without the hashtag (F1 49.9).
- Test transferability of features across domains.

[1] Saif Mohammad. 2012. #Emotional Tweets. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Emotion and Semantic Role Labels in Tweets

Table 1
The FrameNet frame for emotions. The roles examined in this paper are in bold.

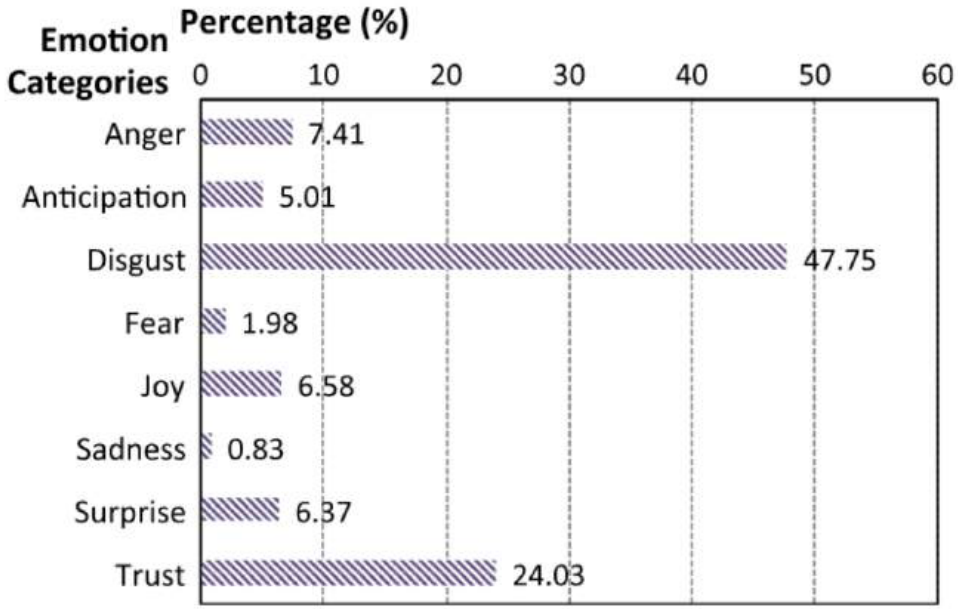
Role	Description
Core:	
Event	The Event is the occasion or happening that Experiencers in a certain emotional state participate in
Experiencer	The Experiencer is the person or sentient entity that experiences or feels the emotions
Expressor	The body part, gesture, or other expression of the Experiencer that reflects his or her emotional state
State	The State is the abstract noun that describes a more lasting experience by the Experiencer
Stimulus	The Stimulus is the person, event, or state of affairs that evokes the emotional response in the Experiencer
Topic	The Topic is the general area in which the emotion occurs It indicates a range of possible Stimulus
Non-Core:	
Circumstances	The Circumstances is the condition(s) under which the Stimulus evokes its response
Degree	The extent to which the Experiencer's emotion deviates from the norm for the emotion
Empathy_target	The Empathy_target is the individual or individuals with which the Experiencer identifies emotionally
Parameter	The Parameter is a domain in which the Experiencer experiences the Stimulus
Reason	The Reason is the explanation for why the Stimulus evokes a certain emotional response

- FrameNet defines a set of roles applicable for Emotions that go beyond just the State or Degree of emotion experienced/expressed.
- [2] annotate electoral tweets for information about "**who** feels **what** towards **whom**".
  - Annotations for a questionnaire are crowdsourced.

[2] Mohammad, Saif M., Xiao-Dan Zhu, Svetlana Kiritchenko and Joel D. Martin. "Sentiment, emotion, purpose, and style in electoral tweets." Inf. Process. Manag. 51 (2015): 480-499.

Aspects annotated:

- Emotion category: choose among 19 predefined label categories, or define own.
  - Mapped back to Plutchik set of 8.
- Sentiment: positive or negative.
- Degree: High, medium, low.
- Stimulus/target: towards whom or what?
- Trigger: words in tweet that express the emotion.
- Cause: reason for emotion being expressed.
- Topic: what is the tweet about?



Annotations:

- High agreement for whether Emotion is expressed, and what the topic is.
- Moderate agreement for category, stimulus, and trigger.
- Additional annotations: purpose, style, stance

Emotion Intensities in Tweets

[3] create the Tweet Emotion Intensity dataset of tweets annotated for *anger*, *fear*, *joy*, and *sadness* intensities.

- Best-worst scaling annotation scheme: present annotators with *n* items, and ask them to choose the item with the highest (best) and lowest (worst) intensity of emotion.
- Use annotations to rank items acc. to emotion intensity; obtain normalized scores in a range (0 to 1).

Emotion	Train	Dev.	Test	All
anger	857	84	760	1701
fear	1147	110	995	2252
joy	823	74	714	1611
sadness	786	74	673	1533
All	3613	342	3142	7097

How much do hashtags matter for conveying emotion intensity?

- tweets without hashtags have lower intensity scores on average.
- i.e, hashtags are *not* redundant.

[3] Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion Intensities in Tweets. In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017), pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Scaling Up Emotion Annotation

GoEmotions Dataset: going beyond common Ekman/Plutchik categories, [4] gather 58k Reddit comments labelled for 27 emotion categories (plus Neutral).

- Emotion category taxonomy derived from "the semantic space of emotion" theory in Psychology.
- Crowdsourced annotations: choose label(s) from predefined list.
- Inter-annotator agreement varies based on emotion category.
  - Significantly associated with linguistic simplicity of expression ("thanks" for gratitude → high agreement)
- Hierarchical clustering of labels based on correlation as a distance metric shows groupings based on sentiment.
- Classification model trained on GoEmotions transfers well to other domains.
- Criticisms: data and annotator bias; quality issues due to *mislabeling.

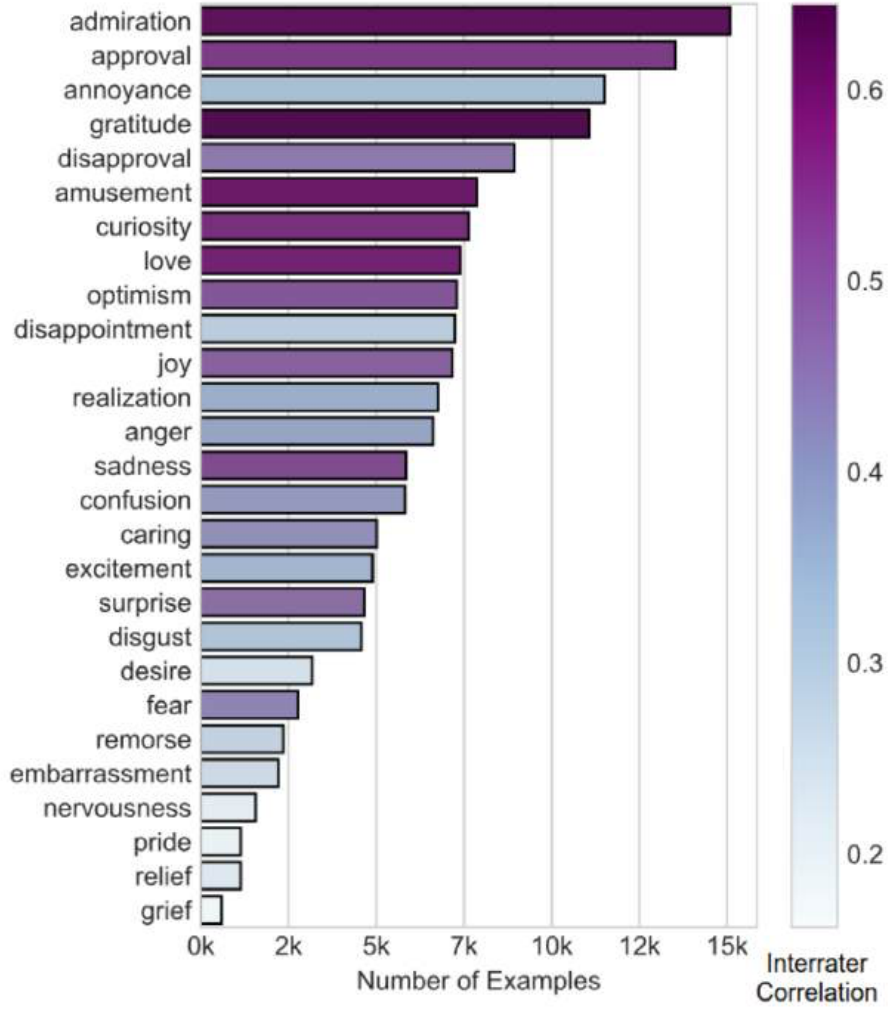


Figure 1: Our emotion categories, ordered by the number of examples where at least one rater uses a particular label. The color indicates the interrater correlation.

[4] Demszky, Dorottya, Movshovitz-Attias, Ko, Cowen, Nemade and Ravi. "GoEmotions: A Dataset of Fine-Grained Emotions." (2020).
* https://surgehq.ai/blog/30-percent-of-googles-reddit-emotions-dataset-is-mislabeled

Scaling Up Emotion Annotation

[5] collect posts from a social media platform *Vent,* self-annotated by posters for the associated emotion category.

- Idea: self-reports offer a more reliable ground-truth label.
- Collected data includes:
  - >33 million “vents”, posted by 934,095 users.
  - 705 user-annotated emotion labels, organized into 63 emotion categories.
  - Social network features: time of posting, reactions to a post, social network of users.

The Vent Dataset:

- 60% of users posted less than 10 vents; a small group of users posted >10^4 vents.
- 50% of users use only 5 labels; a few use more than 200.

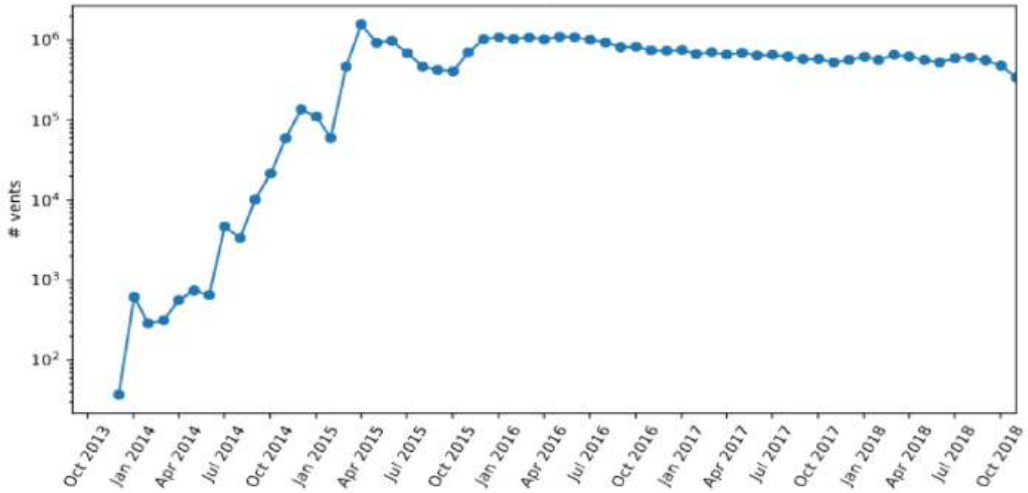
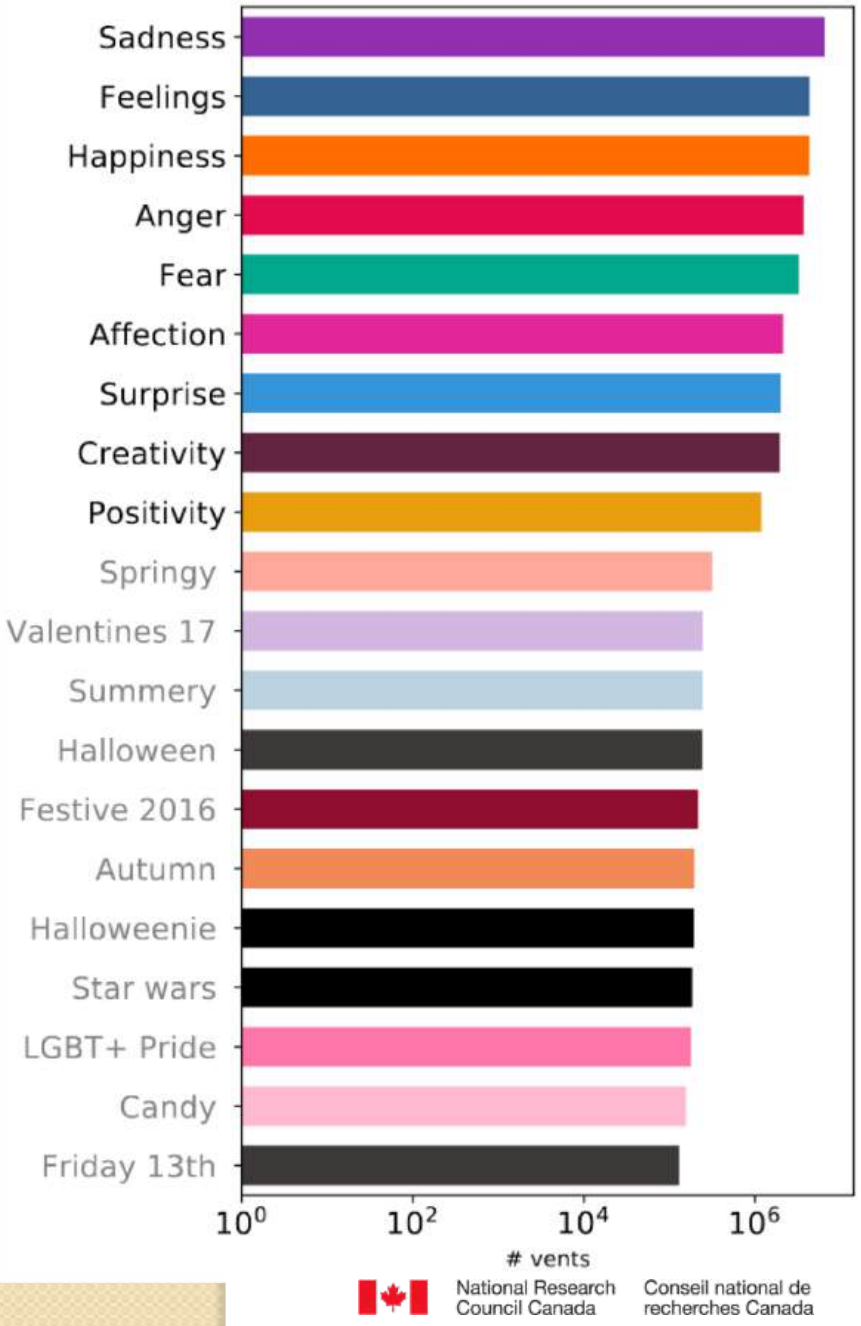
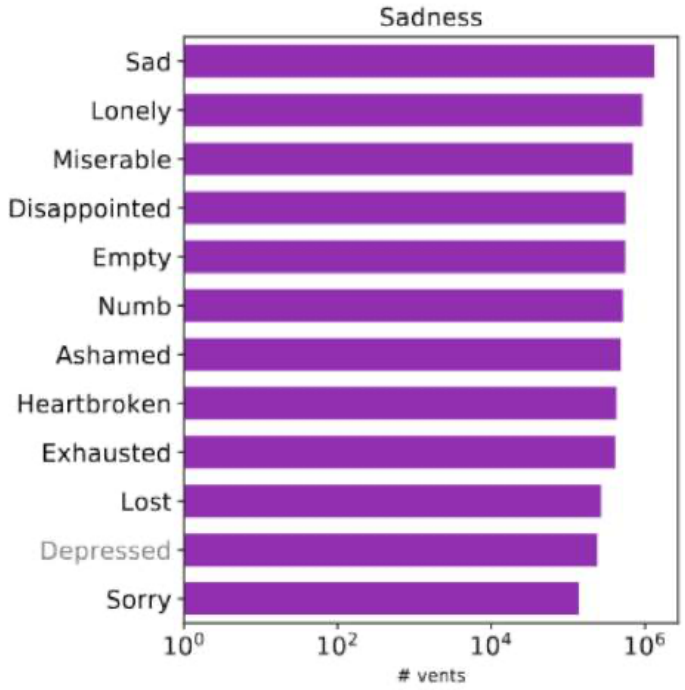


Figure 6: Aggregated monthly activity shows an increase that reached nearly a million of vents .

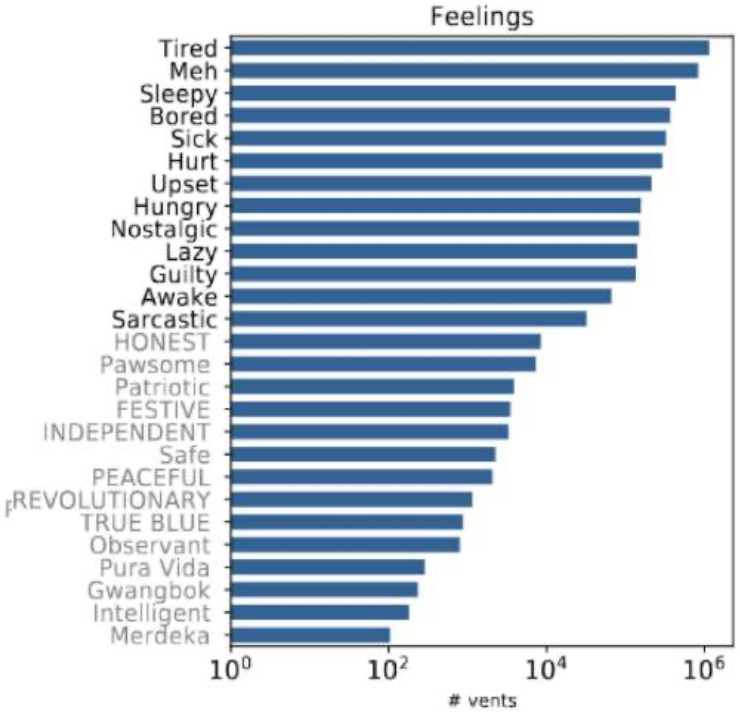
[5] Lykousas, Nikolaos, Constantinos Patsakis, Andreas Kaltenbrunner, and Vicenç Gómez. 2019. “Sharing Emotions at Scale: The Vent Dataset”. Proceedings of the International AAAI Conference on Web and Social Media 13 (01):611-19. https://doi.org/10.1609/icwsm.v13i01.3361.

Emotion Labels in the Vent dataset





(a) Category *Sadness*



(b) Category *Feelings*

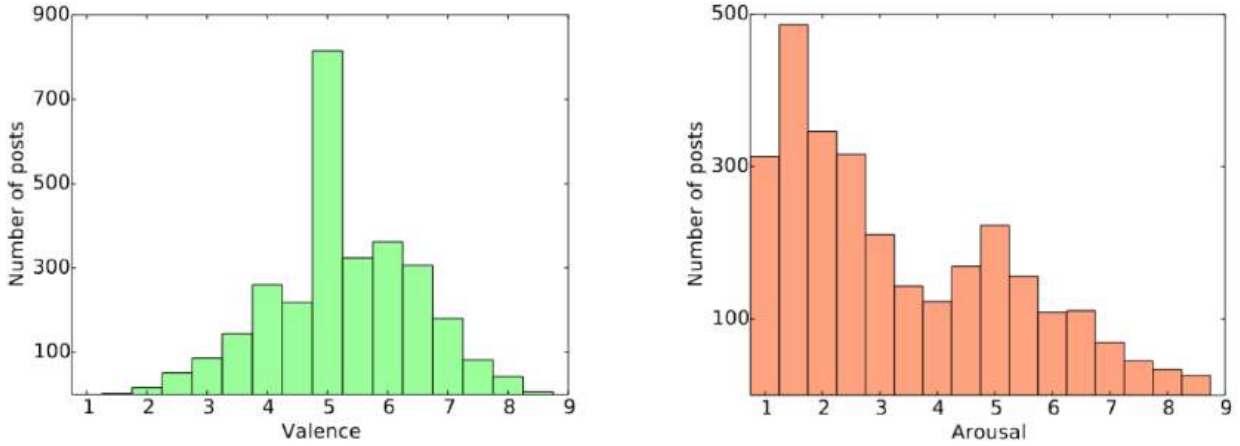
Datasets for Dimensional Models of Affect

- Dimensional models of emotion are an alternative to modelling affect as discrete emotion categories:
  - Circumplex model: affect is represented in a two-dimensional space of valence and arousal.
  - A linear combination of V and A scores can be mapped to an emotion.

- [6] collect and annotate Facebook posts for Valence and Arousal intensity:
  - Valence: polarity or sentiment
  - Arousal: proxy for intensity
  - 9-point Likert scale for both dimensions.
  - Achieve high inter-annotator correlation (~0.8).

Valence of posts	1–9	1–3.5	1–4	6–9	6.5–9
Correlation to arousal	.222	-.047	-.201	.226	.085
Mean arousal	3.35	3.85	3.47	4.31	4.68

Table 3: Correlation with arousal and mean arousal values for different posts grouped by valence.



(a) Valence. (b) Arousal.

Figure 1: Histrograms of average rating scores.

[6] Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling Valence and Arousal in Facebook posts. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 9–15, San Diego, California. Association for Computational Linguistics.

Writer and Reader Emotions

Emotion expressed by the writer and emotions felt by different readers can differ:

- "*Italy defeats France in World Cup final.*" is (probably) neutral for the writer, negative for a fan of the France team, and positive for an Italy supporter.
- [7] crowdsource annotations for:
  - V, A,D, and emotion (Ekman category) expressed by writer (5-point scale)
  - V,A,D and emotion evoked in an *average* reader.
- IAA for READER is significantly higher than WRITER.
  - READER annotations also have higher average emotionality.

Corpus	Domain	Raw	Filtered
SE07	news headlines	1,250	1,192
MASC	blogs	1,378	1,336
	essays	1,196	1,135
	fiction	2,893	2,753
	letters	1,479	1,413
	newspapers	1,381	1,314
	travel guides	971	919
Sum		**10,548**	**10,062**

Table 1: Genre distribution of the raw and filtered EMOBANK corpus.

Mapping between Dimensional and Categorical annotations:

- kNN classification of emotion category given VAD values match human performance (of emotion category annotations) [avg 0.52; Joy 0.78; Surprise 0.17).
- i.e, mapping from VAD to Categorical datasets is feasible.

[7] Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Emotion Triggers and Appraisals

Long texts and narratives often express multiple emotions, with distinct triggers and subjective evaluations:

- Appraisal theory views emotions as resulting from a subjective *evaluation* of an event:
  - "I lost my job and I can't pay my rent anymore" (fear)
  - "I lost my job and it was totally unfair" (anger)
  - "I lost my job and it was toxic workplace" (relief).

- Psychology theories describe different evaluative dimensions of an appraisal:
  - Ellsworth and Smith (1988): pleasantness, effort, certainty, attention, responsibility, control.
  - Scherer and Fontaine (2013): relevance, implication, coping potential, normative significance.
  - OCC model: a more complex logical flow of interacting component-wise evaluations.

- In NLP: identifying emotion cause: "why is this emotion experienced in this situation?"

- Cause identification be framed as an extractive or abstractive NLP task:
  - Extract the span of text or clause that describes the cause.
  - Synthesize the trigger as a natural language description.

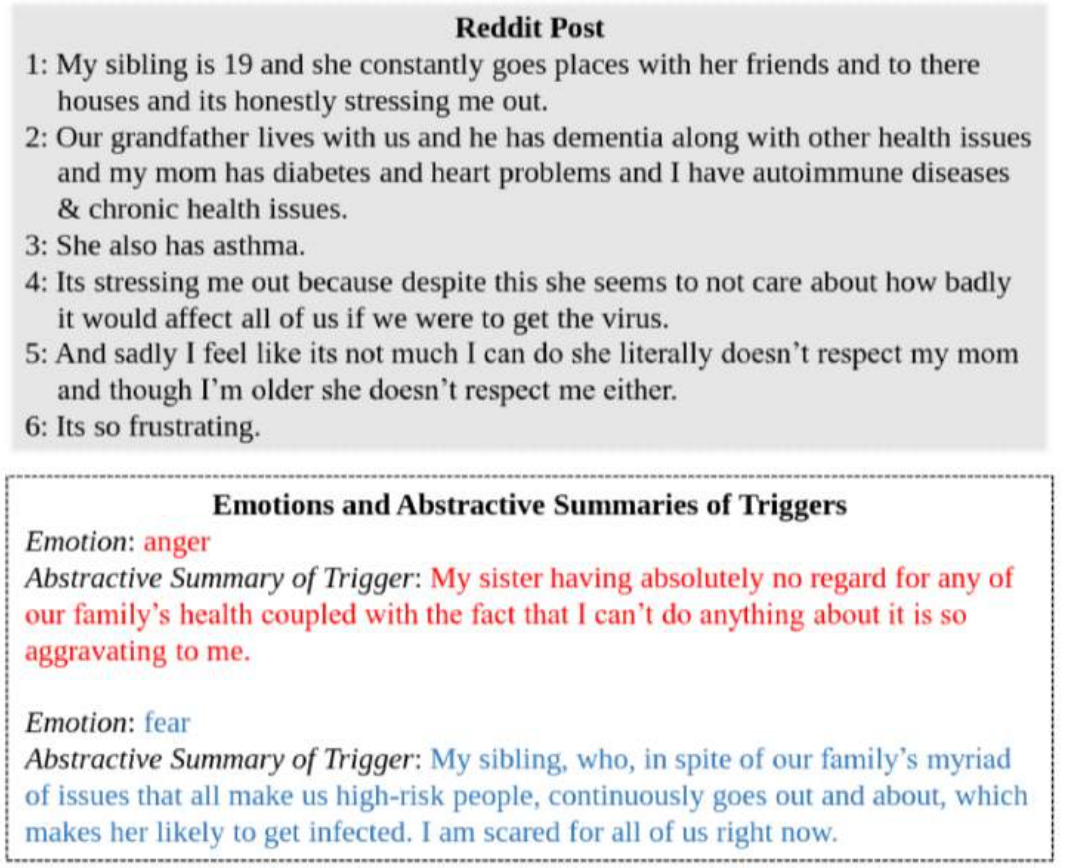
Emotions and Triggers: Reddit Posts

[8] collect 1833 Reddit posts about the COVID-19 pandemic:

- Annotated for 7 emotion labels (Plutchik without *surprise*)
- and a descriptive trigger for the emotion.

CovidET Dataset:

- 2.46 emotions per post, on average.
- Higher proportion of negative emotions (effect of domain).
- 26.9 words per trigger, on average.



Reddit Post

1: My sibling is 19 and she constantly goes places with her friends and to there houses and its honestly stressing me out.
2: Our grandfather lives with us and he has dementia along with other health issues and my mom has diabetes and heart problems and I have autoimmune diseases & chronic health issues.
3: She also has asthma.
4: Its stressing me out because despite this she seems to not care about how badly it would affect all of us if we were to get the virus.
5: And sadly I feel like its not much I can do she literally doesn't respect my mom and though I'm older she doesn't respect me either.
6: Its so frustrating.

Emotions and Abstractive Summaries of Triggers

Emotion: anger
Abstractive Summary of Trigger: My sister having absolutely no regard for any of our family's health coupled with the fact that I can't do anything about it is so aggravating to me.

Emotion: fear
Abstractive Summary of Trigger: My sibling, who, in spite of our family's myriad of issues that all make us high-risk people, continuously goes out and about, which makes her likely to get infected. I am scared for all of us right now.

Figure 1: An example from COVIDET, with perceived emotion(s) identified and their trigger(s) summarized.

[8] Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022. Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9436–9453, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[9] expand 241 posts from the CovidET dataset with evaluations along 24 dimensions:

- Dimensions are drawn from psychology taxonomy (Yeo and Ong 2023).
- Some dimensions cannot be clearly ascertained from the text alone:
  - "consistency with internal values"
  - "fairness"
  - "consistency with social norms"

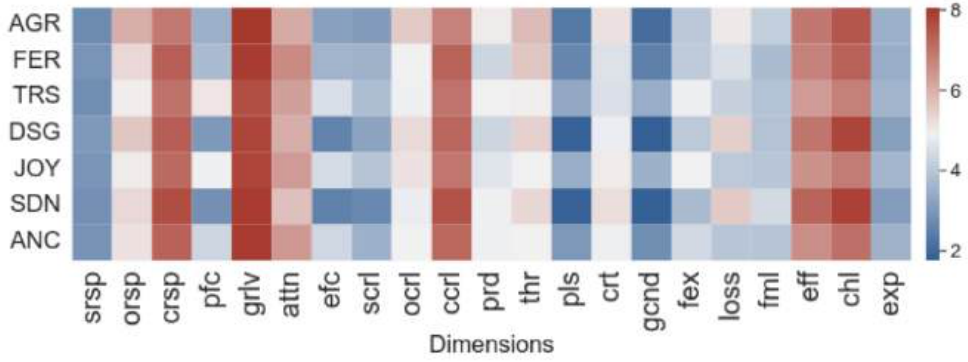


Figure 4: Mean Likert-scale ratings for each dimension in each emotion.

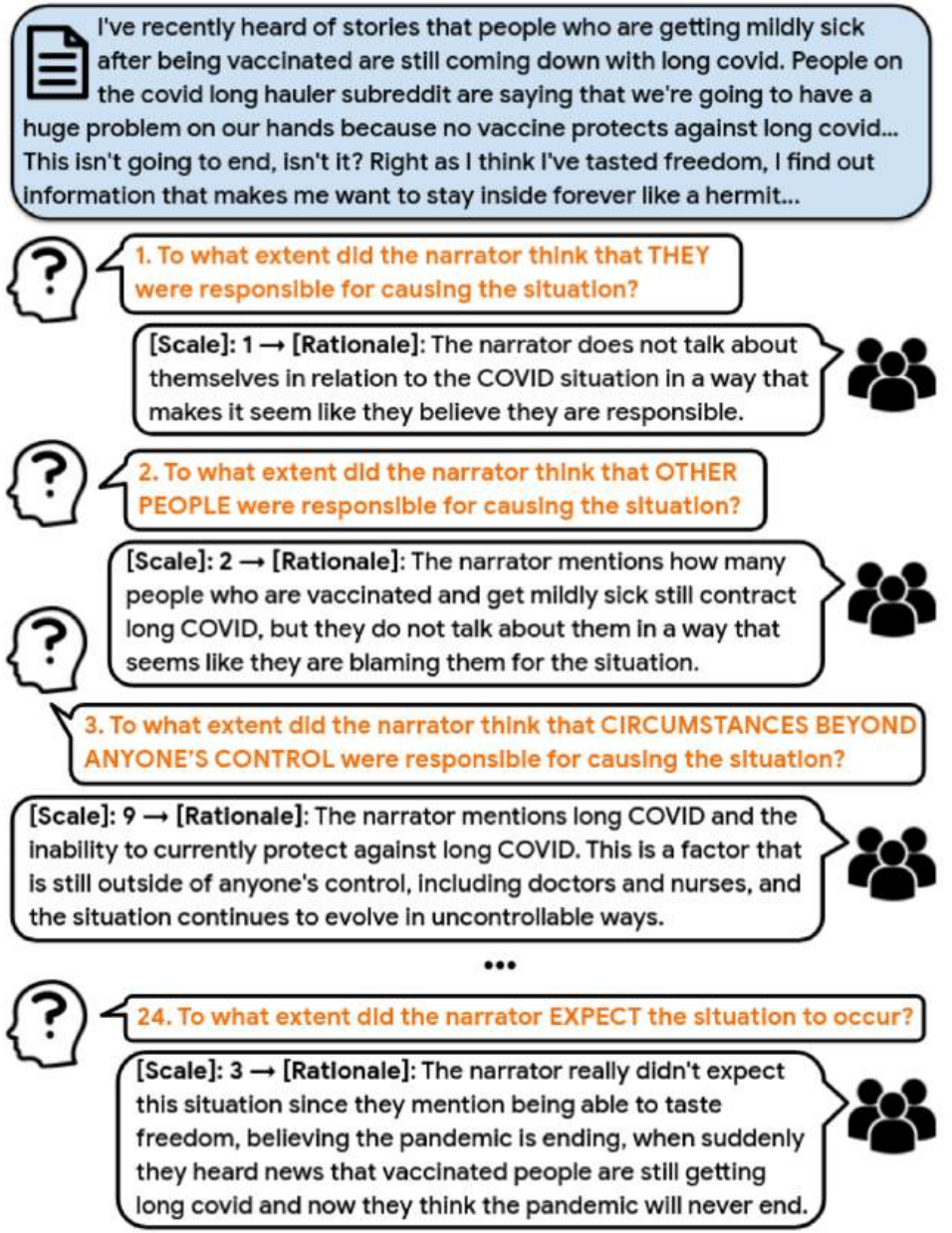


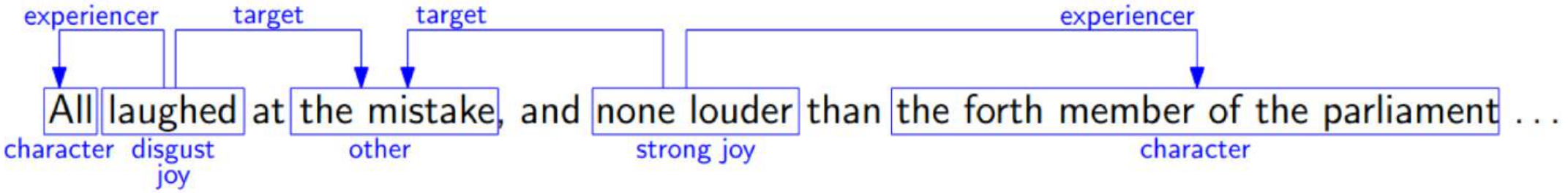
Figure 1: An example from COVIDET-APPRAISALS.

[9] Zhan, Hongli, Desmond Ong, and Junyi Jessy Li. "Evaluating subjective cognitive appraisals of emotions from large language models." In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 14418-14446. 2023.

Emotions in Narratives

[10] annotate excerpts from literary fictional texts for semantic role labels associated with emotion: experiencer, cause, target.

- All aspects are annotated as extractive spans.
- Data: 1720 sentence triples sampled from 200 fictional works from Project Gutenberg.
  - Annotations are obtained for the middle sentence.
- Difficulties in annotation:
  - Literature in particular is highly subjective, particularly for “cause” and “target”.
  - Example: “*they had never seen . . . what was really hateful in his face; . . . they could only express it by saying that the arched brows and the long emphatic chin gave it always a look of being lit from below . . .* ”

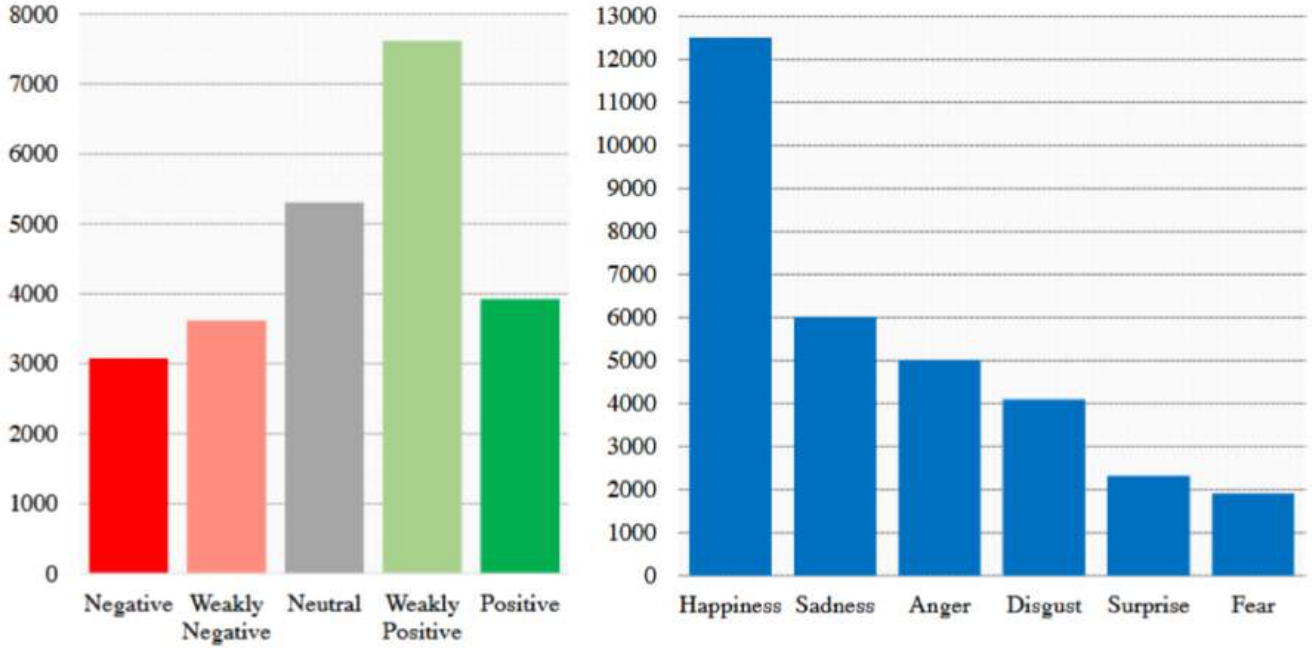


[10] Evgeny Kim and Roman Klinger. 2018. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Multimodal Datasets

Human communications utilize a combination of verbal and non-verbal signals.

- CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset annotates 23,453 video segments (corresponding to transcript sentences) for:
  - Sentiment (Likert scale) | Ekman emotion intensity (Likert scale)



[11] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Multilingual Datasets

Non-English languages are underserved in the availability of annotated datasets; automatic translation methods are often utilized to transfer annotations:

- [12] create XED, an annotated dataset of Finnish and English sentences annotated for Plutchik emotion categories.
  - English annotations are projected to an additional 30 languages by using a dataset of parallel sentences (OPUS movie subtitles).
- [13] create SAMSEMO, a multimodal and multilingual emotion recognition dataset of video scenes in 5 languages (English, Dutch, Spanish, Polish, Korean).
  - Manual annotations of Ekman emotions (plus Neutral and Other)
- [14] review text emotion datasets and categorize ~30 datasets based on annotation scheme, data source, and language.

[12] Öhman, Emily, Marc Pàmies, Kaisla Kajava and Jörg Tiedemann. "XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection." International Conference on Computational Linguistics (2020).
[13] Bujnowski, P., Kuzma, B., Paziewski, B., Rutkowski, J., Marhula, J., Bordzicka, Z., Andruszkiewicz, P. (2024) SAMSEMO: New dataset for multilingual and multimodal emotion recognition. Proc. Interspeech 2024, 2925-2929, doi: 10.21437/Interspeech.2024-212
[14] Koufakou, A., Nieves, E. Review of recent emotion-annotated text corpora and resources. *Lang Resources & Evaluation* **59**, 4313–4347 (2025). https://doi.org/10.1007/s10579-025-09828-1

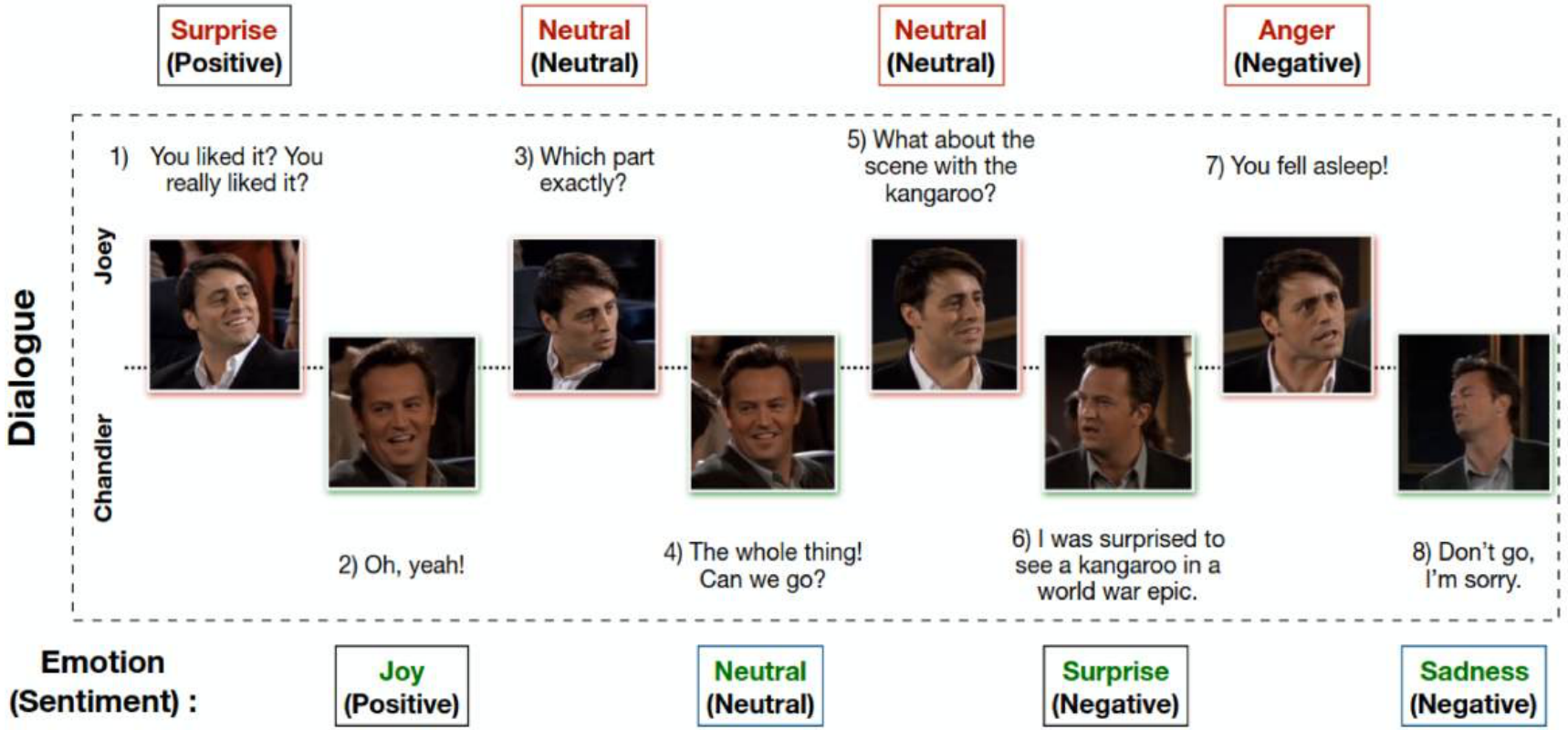
Table 6 Language and related text corpora

Language	#	Dataset
English	20	Affect in Tweets; CARER; Covid-worry; enISEAR; EmoEvent; EmoContext; EmotionLines; Github-love; GoEmotions; GoodNews; RECCON; RU-EN; SenWave; StackOv-GS; TweetEval; Universal Joy; Us vs. Them; WASSA-21, WASSA-23; XED
Spanish	4	Affect in Tweets; EmoEvent; FB-Emo-SP; Universal Joy
Italian	3	FEEL-IT; MultiEmo-IT; Universal Joy
German	2	deISEAR; Universal Joy
Arabic	4	Affect in Tweets; AraEmoCorpus; IAEC; SenWave
Chinese	2	SenWave; Universal Joy
Other	7	CEDR (Russian); Palo-Emo-GR (Greek); RU-EN (Roman Urdu); ShEMO (Persian); SenWave (translated); UIT-VSMEC (Vietnamese); Universal Joy (18 languages total); XED (Finnish and 30 more languages by translation)

Multilingual corpora: Universal Joy is the most diverse multi-lingual corpus with 18 languages (based on Facebook posts and associated “feeling” tags) [14].

Emotions in Conversations

Conversational data is a natural way to understand how emotions are shaped dynamically in social interactions.



Example from MELD: a multi-modal, multi-party, emotion dataset.

Emotions in Conversations

- EmotionLines ([15]) is a textual dataset of dialogues from TV show scripts (and private conversations from a messaging app), where each utterance is annotated with an emotion label (Ekman + neutral).
- MELD ([16]) augments EmotionLines with corresponding audio-visual clips for each utterance.
  - They also re-annotate the data for emotion labels, with additional context available to annotators.

Utterance	Speaker	MELD	EmotionLines
I'm so sorry!	Chandler	sadness	sadness
Look!	Chandler	surprise	surprise
This guy fell asleep!	Chandler	anger	non-neutral

Table 5: Difference in annotation between EmotionLines and MELD.

[15] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).
[16] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset fo Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536, Florence, Italy. Association for Computational Linguistics.

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
     a. sentence/post-level classification/regression/generation
     b. aggregate-level (group comparisons, emotion arcs, etc.)
  4. affective ethics

- **Case Studies** Exploring CAS Research

3a: Affective Tasks

sentence/post-level classification/regression/generation

Emotion Recognition and Estimation

Recognizing emotions expressed in a text is generally framed as a classification or regression task:

- Current dominant approaches involve prompting LLMs to generate class labels or intensity scores:
  - Does not require significant training data (other than few-shot examples).
  - Allows for more flexibility in choosing task-relevant class labels.
  - Ease of use and evaluation in multilingual settings.

- Finetuning can out-perform in-context learning, with the added cost of requiring labelled data:
  - Smaller finetuned models can also offer improved latency without a drop in performance for domain-specific tasks.

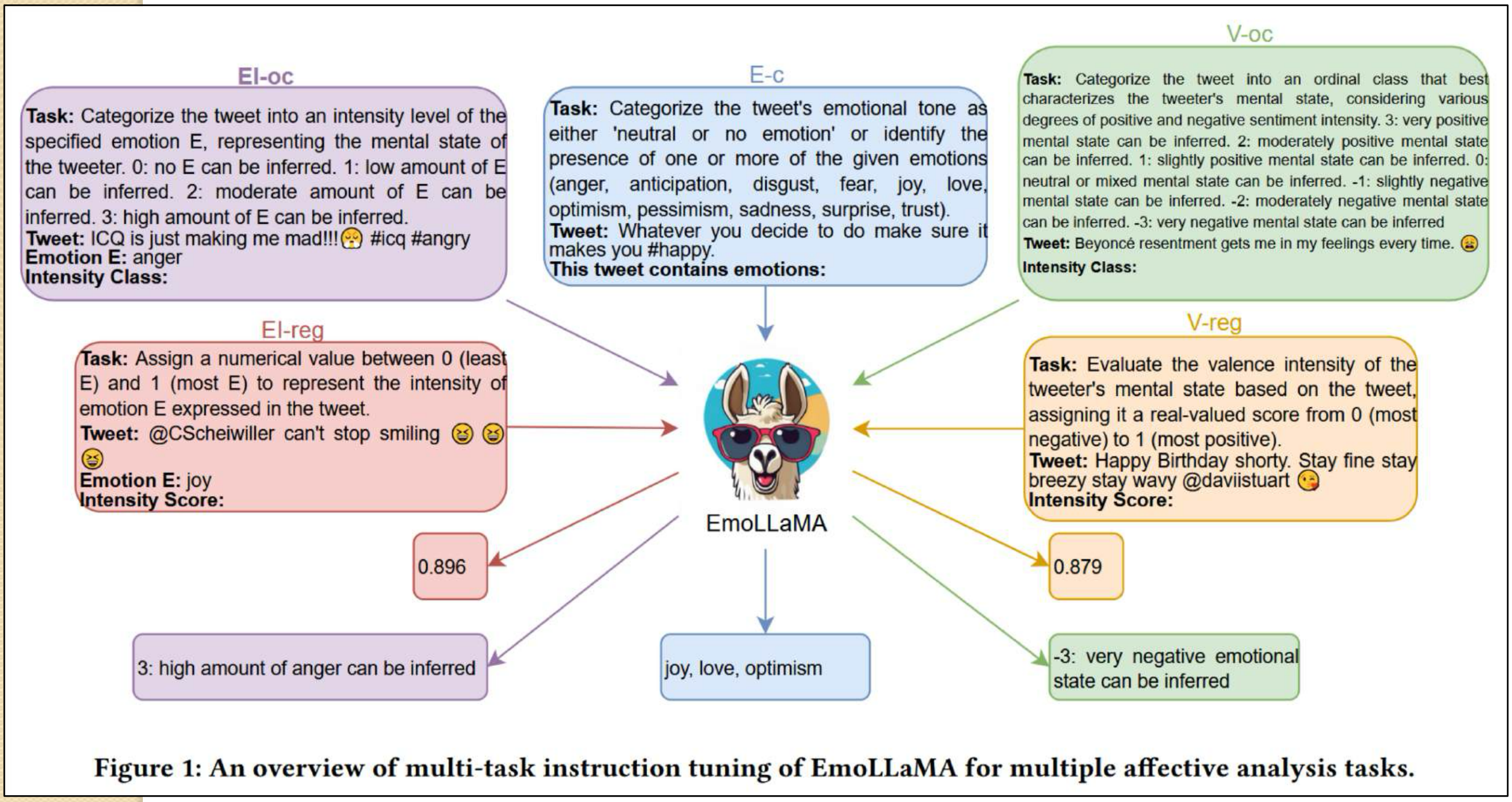


Figure 1: An overview of multi-task instruction tuning of EmoLLaMA for multiple affective analysis tasks.

[17] Liu, Zhiwei et al. "EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis." Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2024): n. pag.

model	EI-reg					EI-oc					V-reg	V-oc	E-c		
	ave	anger	fear	joy	sadness	ave	anger	fear	joy	sadness	valence	valence	acc	mi-F1	ma-F1
Leaderboard(1)	0.799	0.827	0.779	0.792	0.798	0.695	0.706	0.637	0.720	0.717	0.873	0.856	0.609	0.724	0.592
PLMs															
BERT-base	0.785	0.800	0.781	0.783	0.742	0.683	0.698	0.656	0.712	0.665	0.840	0.805	0.567	0.718	0.568
RoBERTa-base	0.717	0.670	0.736	0.769	0.694	0.664	-	-	-	-	0.845	0.772	0.563	0.721	0.536
SentiBERT	0.722	0.724	0.740	0.731	0.691	0.665	-	-	-	-	0.835	0.763	0.535	0.700	0.522
Zero-shot/few-shot methods															
Falcon	0.114	0.147	0.082	0.095	0.131	0.033	0.022	0.017	0.031	0.061	0.135	0.189	0.190	0.318	0.253
Vicuna	0.281	0.307	0.257	0.260	0.299	0.214	0.238	0.193	0.186	0.241	0.298	0.579	0.220	0.359	0.253
LLaMA2-7B-chat	0.194	0.176	0.257	0.097	0.247	0.120	0.112	0.138	0.115	0.114	0.094	0.497	0.257	0.414	0.286
LLaMA2-13B-chat	0.488	0.524	0.506	0.398	0.526	0.194	0.262	0.178	0.119	0.216	0.312	0.568	0.274	0.424	0.302
ChatGPT	0.599	0.637	0.573	0.569	0.618	0.455	0.500	0.428	0.363	0.529	0.637	0.748	0.382	0.546	0.429
ChatGPT-FS	0.550	0.572	0.482	0.587	0.560	0.473	0.502	0.410	0.407	0.573	0.739	0.791	0.413	0.563	0.466
GPT-4	0.656	0.699	0.575	**0.686**	0.667	**0.620**	**0.656**	**0.579**	**0.618**	**0.629**	0.811	0.788	0.444	0.572	0.497
GPT-4-FS	**0.679**	**0.704**	**0.654**	0.679	**0.678**	0.562	0.623	0.523	0.515	0.585	**0.825**	**0.793**	**0.460**	**0.582**	**0.515**
Emotion-based instruction-tuning methods															
EmoBART	0.795	0.798	0.803	0.795	0.782	0.725	0.705	0.742	0.723	0.729	0.851	0.835	0.528	0.686	0.548
EmoT5	0.783	0.785	0.797	0.798	0.751	0.717	0.703	0.733	0.726	0.707	0.852	0.836	**0.559**	**0.712**	**0.568**
EmoOPT	0.825	**0.827**	0.830	0.837	0.805	0.753	0.739	0.751	0.762	**0.759**	**0.887**	0.843	0.532	0.680	0.550
EmoBLOOM	0.791	0.802	0.797	0.790	0.776	0.732	0.725	0.717	0.746	0.740	0.857	0.822	0.528	0.683	0.552
EmoLLaMA-7B	0.822	0.819	0.821	0.837	0.809	0.743	0.738	0.722	0.768	0.745	0.879	0.843	0.545	0.695	0.563
EmoLLaMA-chat-7B	0.824	0.825	0.830	0.832	0.810	0.751	0.748	0.754	0.764	0.739	0.876	0.827	0.534	0.693	0.540
EmoLLaMA-chat-13B	**0.831**	**0.827**	**0.835**	**0.843**	**0.817**	**0.763**	**0.755**	**0.764**	**0.777**	0.755	0.886	**0.860**	0.537	0.696	0.545

Evaluation Results on a series of English-language Emotion Recognition and Intensity estimation tasks with finetuned PLMs and prompted LLMs. [17]

[17] Liu, Zhiwei et al. "EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis." Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2024): n. pag.

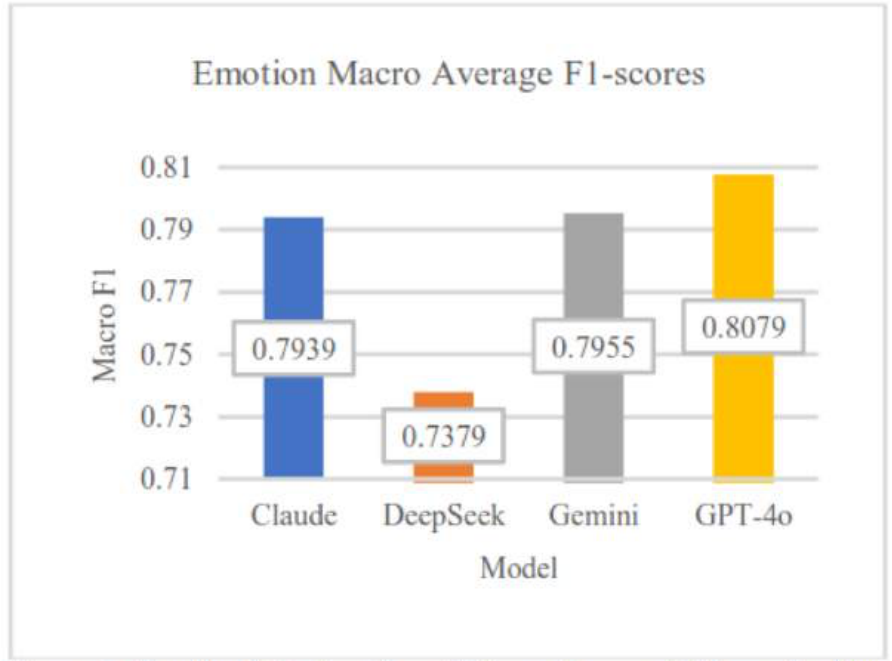


Figure 2 - Emotion Detection Overall Macro Average F1 Score Rankings

Performance of SoTA LLMs on Emotion Detection in Persian Tweets [18]

Model	Love	Happiness	Sadness	Anger	Fear	Macro-Precision
Human Annotation	0.777	0.900	0.890	**1.000**	0.617	0.837
ChatGPT-4	0.790	**0.910**	**0.900**	0.980	0.630	**0.842**
llama3_8b	0.766	0.446	0.577	0.834	0.786	0.682
llama3_70b	0.807	0.636	0.698	0.735	0.953	0.766
llama3.1_8b	0.658	0.757	0.468	0.837	0.955	0.735
llama3.1_70b	0.761	0.809	0.668	0.833	0.947	0.804
llama3.2_3b	0.612	0.862	0.506	0.613	0.862	0.691
llama3.3_70b	0.713	0.825	0.686	0.819	0.965	0.798
gemma2_2b	0.556	0.804	0.461	0.946	0.495	0.652
gemma2_9b	**0.846**	0.662	0.688	0.804	0.646	0.729
gemma2_27b	0.791	0.773	0.744	0.862	0.936	0.801
qwen_7b	0.705	0.704	0.614	0.778	0.610	0.682
qwen2_7b	0.798	0.725	0.701	0.809	0.739	0.754
qwen2.5_7b	0.758	0.760	0.622	0.791	0.666	0.719
phi3_14b	0.817	0.544	0.524	0.691	0.911	0.697
phi4_14b	0.705	0.788	0.524	0.816	0.916	0.750
qwq_32b	0.777	0.239	0.730	0.531	**1.000**	0.655

Performance of SoTA LLMs on Emotion Detection in Indonesian Tweets [19]

[18] Tohidi, Kian, Kia Dashtipour, Simone Rebora, and Sevda Pourfaramarz. "A Comparative Evaluation of Large Language Models for Persian Sentiment Analysis and Emotion Detection in Social Media Texts." arXiv preprint arXiv:2509.14922 (2025).

[19] A. H. Nasution, A. Onan, Y. Murakami, W. Monika and A. Hanafiah, "Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets," in IEEE Access, vol. 13, pp. 94009-94025, 2025

Emotion Cause/Trigger Detection

Large Language Models can be prompted to generate triggers corresponding to expressed emotions:

- How good are they?
- To what extent do emotion prediction models rely on features that reflect emotion triggers?

[20] annotate subsets of three social media emotion for *words/phrases* that contribute to the emotion label (extractive triggers):

- GPT-4, Llama2-13b, and Alpaca-13B are evaluated for emotion detection and trigger detection using prompting.
- BERT-based PLMs are finetuned on the above datasets for emotion detection; SHAP values are used to identify most salient words.

[20] Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2024. Language Models (Mostly) Do Not Consider Emotion Triggers When Predicting Emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 603–614, Mexico City, Mexico. Association for Computational Linguistics.

Example 1. Unless he has actually *threatened* to or used them against you I think you're overreacting [annoyance]. Ask him to please delete them [caring].
GPT-4: overreacting [annoyance] Ask him to please delete them [caring]
Llama2Chat: delete them [annoyance]
Alpaca: threatened, delete them [annoyance]
EmoBERTa-SHAP: threatened, used, against [anger]

Example 2: Everyone loved the *Snack Trade* [love]
GPT-4: loved [love]
Llama2Chat: loved [love]
Alpaca: love [love]
EmoBERTa-SHAP: love [joy]

Example 3: Between the two cancers, [sadness] I feel *confused* and *alone*
GPT-4: alone [sadness]
Llama2Chat: alone [sadness]
Alpaca: alone [sadness]
EmoBERTa-SHAP: confused, alone, cancers [sadness]

Example 4: All we can do now is pray [anticipation] that mother nature [fear] shows them *mercy*
GPT-4: pray [sadness], mother nature [fear]
Llama2Chat: pray [anticipation], show them mercy [fear]
Alpaca: show them mercy [fear]
EmoBERTa-SHAP: mother, nature, mercy [fear]

Highlighted: gold labels | Underlined: keyphrases
Italics: EmoLex works

Dataset	GPT4	Llama2	Alpaca	Emo-BERTa
HurricaneE.	0.920	0.851	0.756	0.483
CancerEmo	0.914	0.842	0.723	0.378
GoEmotion	0.907	0.820	0.707	0.341

Table 2: Macro F1 score for emotion detection. All scores are based on results of few-shot prompting.

- LLMs (GPT-4) perform best on both emotion detection and trigger identification.
  - Some emotions are harder than others (“anticipation”)
- Salient features for models largely do not align with annotated triggers (except GPT-4; ~0.28 overlap)
- Comparing model-extracted triggers with emotion lexicons:
  - Low overlap with explicit emotion terms.
  - Keyphrases extracted using TopicRank show higher alignment with triggers.
  - Emotion-dependant: Anger, Joy, Sadness are more explicit compared to Anticipation, Fear, Disgust.

LLMs as Emotion Annotators

[21] test how well LLM-assigned emotion labels align with human-annotated datasets:

- Zero-shot prompting with GPT-4 is comparable to the performance of a finetuned BERT model.
  - For both classification labels and Likert-scale ratings.
  - Confusion matrix analysis: mis-classification between similarly-valenced emotion groups; preference for “shame” over “guilt”.
- Human evaluation study on samples with disagreement between GPT-4 and human annotations:
  - GPT-4 annotations are preferred.
  - Open-ended generations by GPT-4 are preferred over those restricted to class labels.
- LLM annotations can be used to identify “low-quality” (?) human annotations.

Table 2: *GPT-4 zero-shot vs. BERT finetuned performance across four dataset. Better performances are in bold.*

	Macro-F1 ↑		UAR ↑	
	GPT-4	*BERT*	*GPT-4*	*BERT*
ISEAR	**0.739**	0.726	**0.747**	0.727
SemEval	0.511	**0.548**	0.476	**0.495**
GoEmotions	0.375	**0.521**	**0.485**	0.469
	PCC ↑		**MAE ↓**	
	GPT-4	*BERT*	*GPT-4*	*BERT*
Emobank	**0.764**	0.321	0.645	**0.442**

[21] Niu, Minxue, Mimansa Jaiswal, and Emily Mower Provost. "From Text to Emotion: Unveiling the Emotion Annotation Capabilities of LLMs." In Proc. Interspeech 2024, pp. 2650-2654. 2024.

Affective Understanding as Social Intelligence

- Emerging benchmarks for evaluating LLMs include affect and emotion-related tasks under the broad category of “subjective language understanding”.
  - sentiment, emotion, stance,
  - sarcasm, humour, metaphor, intent understanding.

- Theory-of-mind (ToM) is defined as the “capacity to attribute mental states to others (and oneself), in order to explain and anticipate behaviour” [22]
  - a key human (and animal?) ability to navigate social situations, previously studied in Cognitive Science.
  - Several ToM benchmarks for evaluating LLM capabilities have emerged in the last couple of years.
  - Abilities in Theory of Mind Space (ATOMS) framework: beliefs, intentions, desires, emotions, knowledge, percepts, non-literal communications.

[22] Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of Mind in Large Language Models: Assessment and Enhancement. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31539–31558, Vienna, Austria. Association for Computational Linguistics.

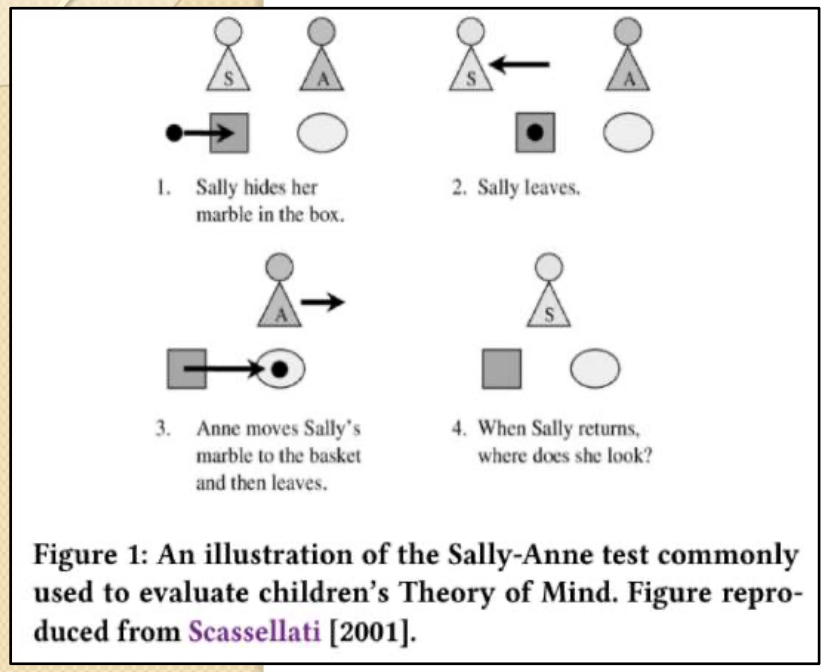


Figure 1: An illustration of the Sally-Anne test commonly used to evaluate children's Theory of Mind. Figure reproduced from Scassellati [2001].

A basic ToM task used in Cognitive Science research. [23]

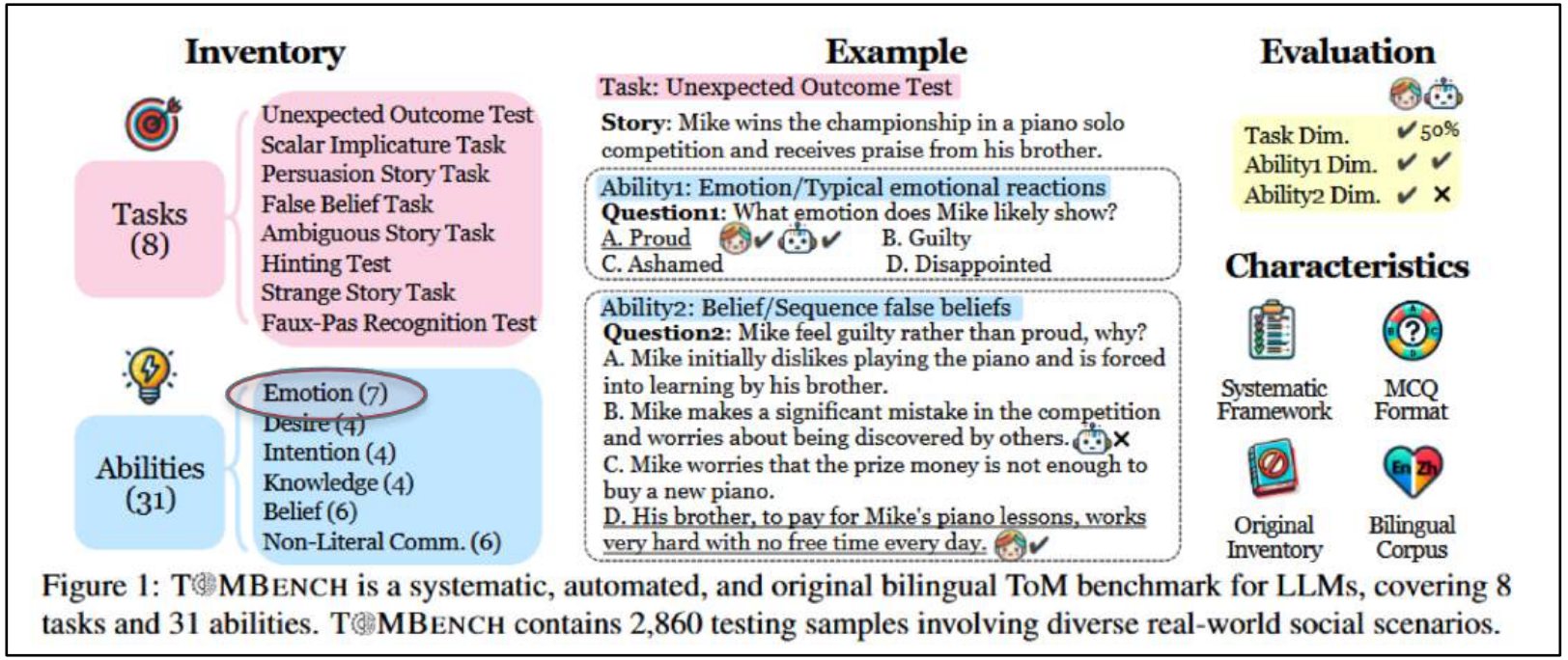


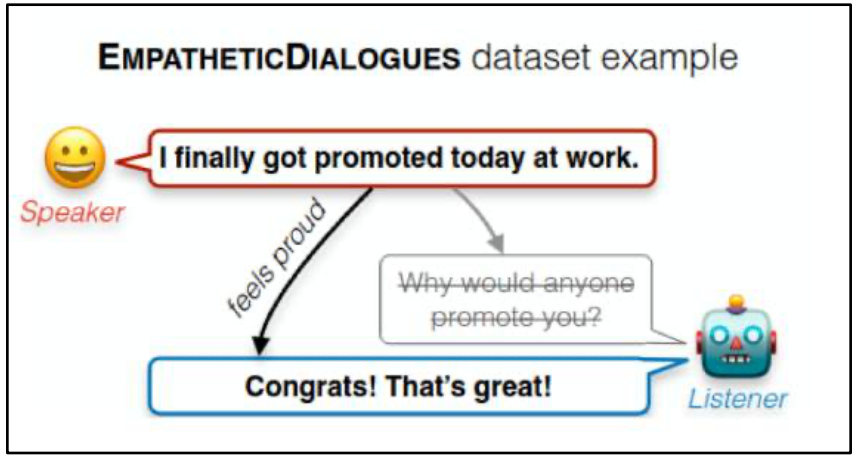
Figure 1: T MBENCH is a systematic, automated, and original bilingual ToM benchmark for LLMs, covering 8 tasks and 31 abilities. T MBENCH contains 2,860 testing samples involving diverse real-world social scenarios.

Complex ToM benchmarks for LLMs involving emotional reasoning. [24]

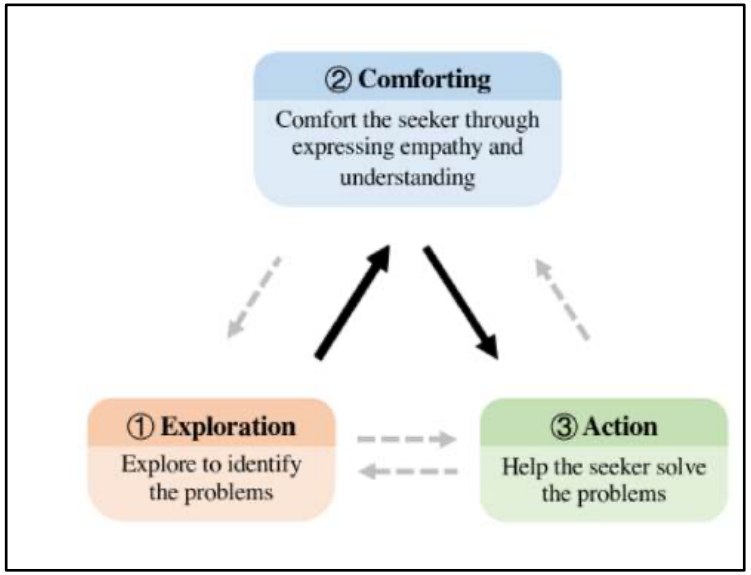
[23] Wang, Qiaosi, Xuhui Zhou, Maarten Sap, Jodi Forlizzi, and Hong Shen. "Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective." arXiv preprint arXiv:2504.10839 (2025).
[24] Chen, Zhuang, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao et al. "ToMBench: Benchmarking Theory of Mind in Large Language Models." In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15959-15983. 2024.

Affective Generation

- Generating text that conforms to certain affective constraints or pursues affective goals:
  - General-purpose chatbots need to understand the emotions expressed by the user, as well as generate an appropriate response.
  - Popular tasks: Empathetic Response Generation (ERG) and Emotional Support Conversations (ESC).



Acknowledging the user in an appropriate way. [25]



Conversational goals aimed at providing emotional support. [26]

[25] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
[26] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3469–3483, Online. Association for Computational Linguistics.

Emotional Support Conversation (ESC) Framework

Framework for evaluating dialogue systems as Emotional Support agents:

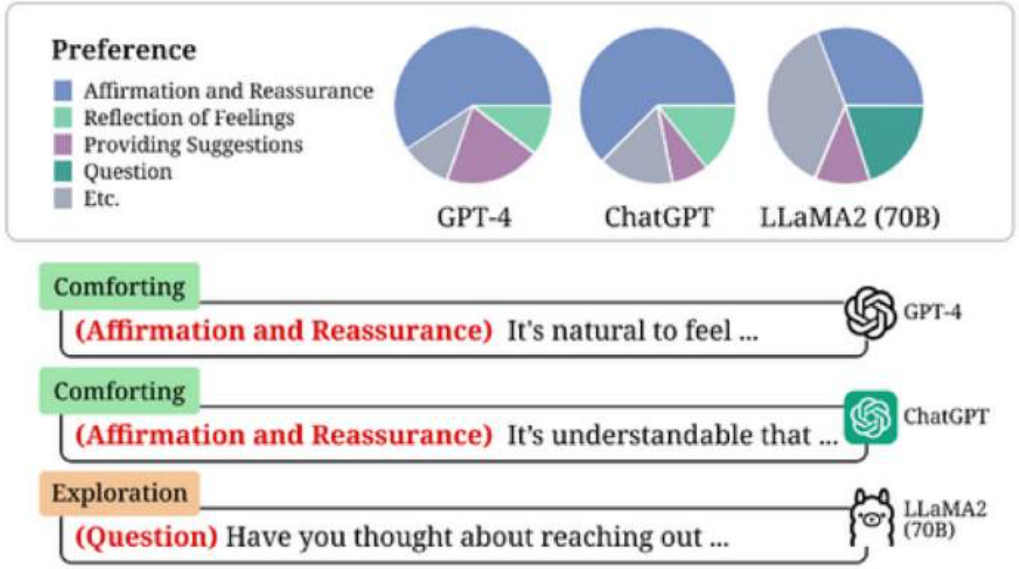
- Grounded in the Helping Skills theory*, intended for support through social conversations rather than professional counselling.
- Three stage process:
  - Exploration, Comforting, Action
  - each involving strategy selection and strategy-constrained generation.
- The ESConv dataset gathers sample conversations following the above framework from trained crowdworkers.

*Hill, Clara E.. "Helping Skills: Facilitating Exploration, Insight, and Action." (1999).

LLMs for ESC

[27] evaluate how LLMs perform on the ESConv dataset.

- LLMs show a preference for certain strategies, limiting performance.
- They propose methods to mitigate preference bias:
  - response refinement with feedback and CoT.
  - delegate the strategy selection step to a separate finetuned model.
- Findings:
  - incorporating an external strategy planner is highly effective.
  - Deeper reasoning and thinking strategies deepen the preference bias.
  - Mitigating preference bias results in improved performance on the ESC task.



[27] Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Safeguarding LLM Chatbots for Mental Health

General-purpose LLMs are not explicitly designed for therapeutic use, leading to inappropriate and harmful behaviours and outcomes:

- [28] present EmoAgent, a system that evaluates persona-based conversational AI systems for risks of inducing psychological distress:
  - incorporates clinically-validated tools to evaluate risks: Patient Health Questionnaire (PHQ-9) for depression, Peters et al. Delusions Inventory (PDI) for delusion, Positive and Negative Syndrome Scale (PANSS) for psychosis.
  - inserts safeguard agents as an intermediary between the user and the AI chatbot: monitor the conversation, predict potential harm, intervene by delivering corrective feedback to the AI model.
  - Empirical studies with simulated human-AI conversations:
    - simulate user profiles based on a cognitive model.
    - iteratively generate dialogues between simulated user and the chatbot model.
    - Findings: “delusion” exhibits the highest overall deterioration rates post-conversations.
- Complex affect and emotion tasks emerge with downstream use-cases.

[28] Qiu, Jiahao, Yinghui He, Xinzhe Juan, Yiming Wang, Yuhan Liu, Zixin Yao, Yue Wu, Xun Jiang, Ling Yang and Mengdi Wang. “EmoAgent: Assessing and Safeguarding Human-AI Interaction for Mental Health Safety.” ArXiv abs/2504.09689 (2025): n. pag.

Multilinguality: Cultural Variation in Emotion Word Usage

Conceptualization and usage of emotion words in different languages/cultures is varied:

- influenced by linguistic, historical, social contexts, and lived experiences of individuals and populations.
- [29] investigate this variation, and how well LLMs capture or reflect these variations, with empirical studies using multi-lingual corpora and multi-lingual LLMs.
- Languages: English, Spanish, Chinese, and Japanese.

RQ1: Does implicit and explicit alignment (of models) inappropriately anchor emotion embeddings to English?

- Obtain emotion word embeddings from (parallel) monolingual and a multi-lingual RoBERTA models.
- Compute “similarity of joy in language X and Y” as “correlation between vectors of distance between joy-anger, joy-sadness, joy-elation, joy-happiness for RoBERTA-X and RoBERTA-Y”.
- Multilingual model aligns more closely with English.
  - Non-english language embeddings become more similar to English in the multilingual model.

[29] Havaldar, Shreya, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. "Multilingual language models are not multicultural: A case study in emotion." In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pp. 202-214. 2023.

RQ2: Do emotion embeddings reflect known psychological cultural differences?

- Consider differences in Pride and Shame between the US and Japan:
  - Shame has a more desirable affect in Eastern cultures; Pride is inhibited.
- Project embeddings for “Shame” and “Pride” from language-specific models to the V-A space.
  - English-Pride and Japanese-Shame have slightly higher valence.
  - But findings are not conclusive.

RQ3/4 (Generative): Do LLMs reflect social norms of emotion in generative settings?

- Consider culturally-specific scenarios: “How would you feel if your guests chose to keep their shoes on when entering your home?”
- Prompt the LLM to generate responses, priming with cultural cues (“You live in Japan.”) or with translations.
- LM responses in non-English languages are rated as being less culturally-appropriate.
  - Q: Does this reflect the lack of cultural knowledge of emotions in LLMs, or a lack of multilingual capability?
- Other studies [30]: “prompting in English with explicit country context often outperforms in-language prompts for culture-aware emotion and sentiment understanding.”

[30] Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural Lenses on Emotion - Benchmarking LLMs for Cross-Cultural Emotion Understanding. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.

Cannot find the PDF but seems like a very interesting study…

Tak, Ala Nekouvaght, Jonathan Gratch and Klaus R Scherer. “Aware Yet Biased: Investigating Emotional Reasoning and Appraisal Bias in Large Language Models.” IEEE Transactions on Affective Computing 16 (2025): 2871-2880.

“This paper reports two studies investigating the emotional reasoning of Large Language Models (LLM). Previous research has suggested that LLMs are surprisingly accurate at predicting human emotions from text descriptions of situations and reason in a way that is consistent with appraisal theory—a leading theory of emotion. Study 1 tests this claim with a large multilingual corpus (English, French, and German) of autobiographical descriptions of emotionally charged events. We confirm that GPT-4, one of the most advanced and widely studied LLMs, shows a remarkable ability to predict emotion and appraisals. We further show this ability is language-independent, with accuracy being consistent across languages and unaffected by the language of the prompt. However, GPT-4 struggles to accurately predict certain emotions (shame, fear, and irritation) and fails to understand appraisal dimensions related to control and power. We repeat the experiments with Gemini-2.0-Flash and find a remarkably similar pattern of strengths and weaknesses, although it consistently outperforms GPT-4. Study 2 examines a possible mechanism for these failures based on the idea of cognitive appraisal bias. In psychological appraisal theory, appraisal bias is the idea that people evaluate situations in biased, often unrealistic ways. By testing both models on a set of situations designed to identify appraisal bias, we find they exhibit strong—but similar—appraisal bias; for example, evaluating situations as if they were a person high in agreeableness and low in power. We further offer evidence suggesting that LLMs could be debiased by incorporating a person's personality in the prompt. This research underscores LLMs’ capabilities and limitations in emotional reasoning, though highlights one mechanism underlying this limitation and suggests an approach for addressing these limits.”

This Tutorial

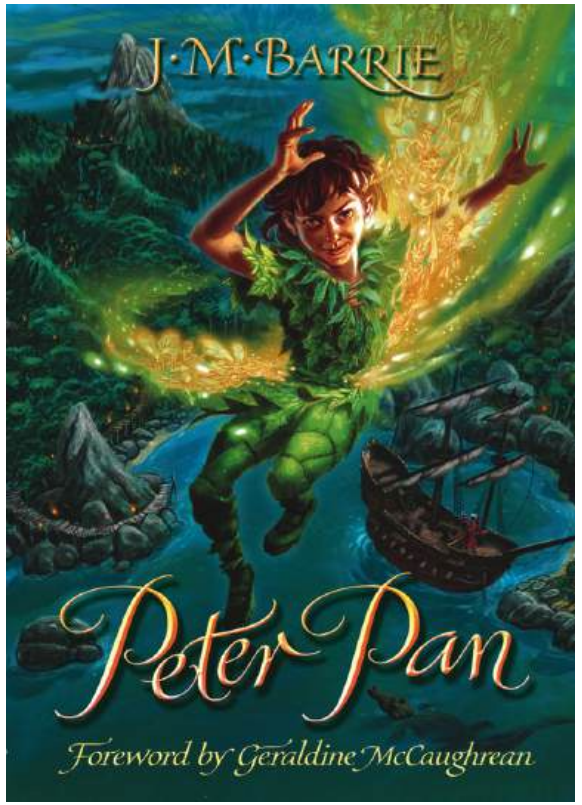
- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
     a. sentence/post-level classification/regression/generation
     b. aggregate-level (emotion arcs, group comparisons, etc.)
  4. affective ethics

- Case Studies Exploring CAS Research

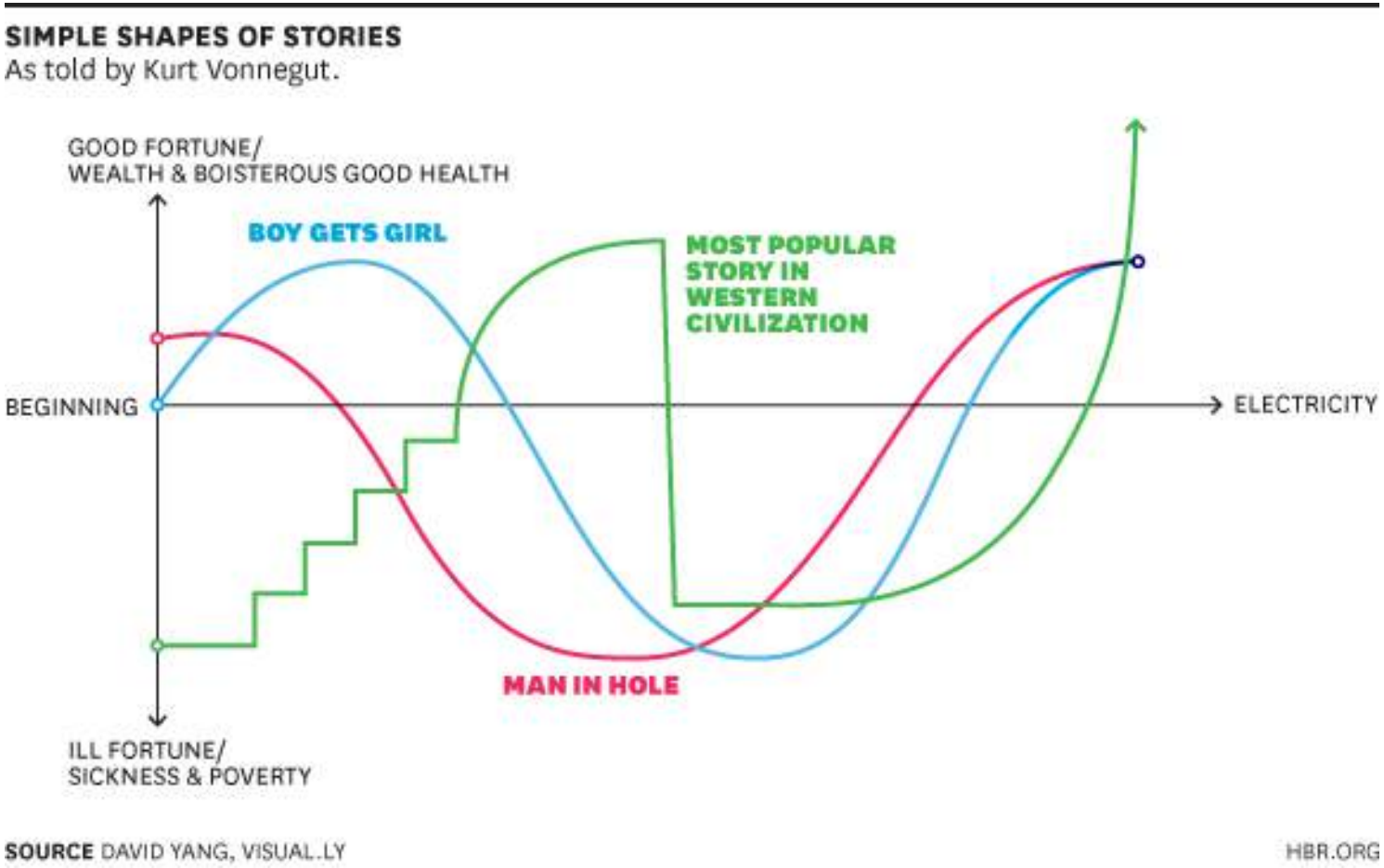
3b: Affective Tasks

aggregate-level (emotion arcs, group comparisons, etc.)



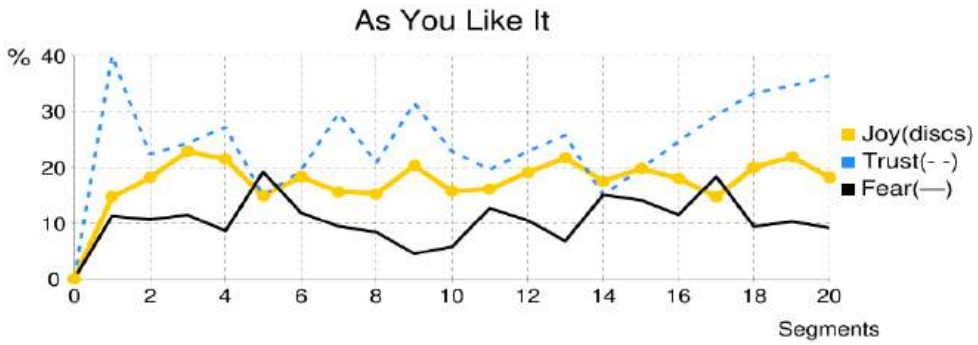
Emotions Arcs in CAS (mind, body, world), Commerce, Psychology, Health, and Humanities

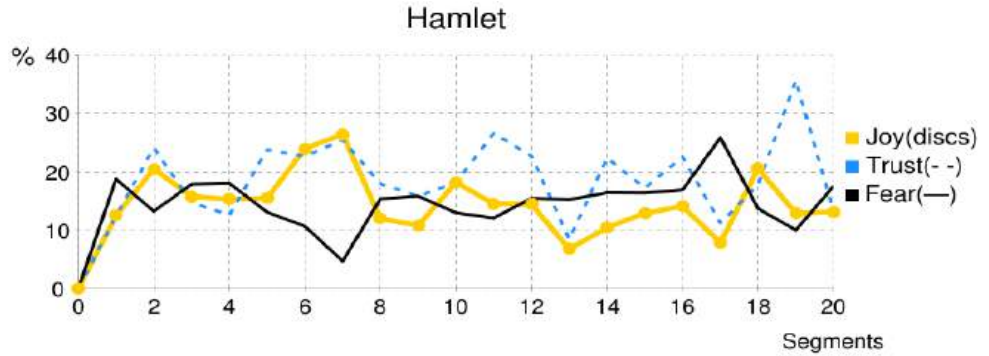
Tracking Emotions in Stories

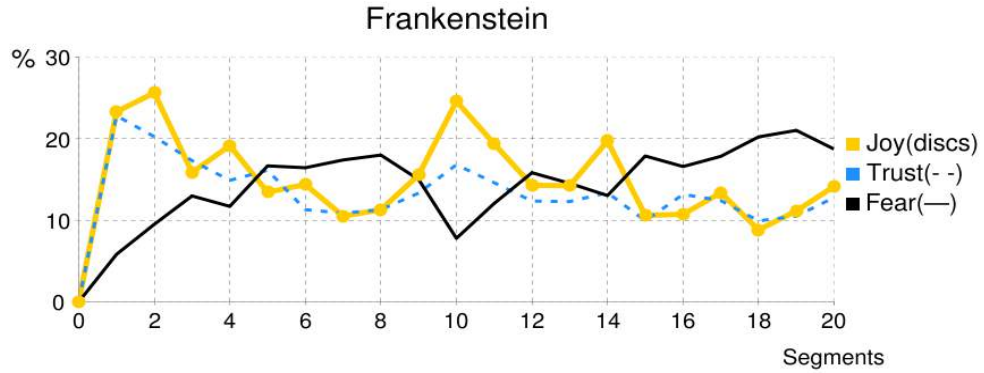


Back in 2011:

Tracking Emotions in Stories







From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales, Saif Mohammad, In Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), June 2011, Portland, OR.

Creating Emotion Arcs

- Lexicon-only approach
- ML approaches (sometimes making use of lexicons)

Lexicon-only approaches

- Pros
  - simple, accessible
  - interpretable
  - low-carbon
  - domain-free
- Cons
  - not highly accurate at instance level (context, long-distance dependencies)

Evaluating Emotion Arcs

Very little work!! No dataset of gold arcs.

Evaluating Emotion Arcs

Consider tracking anger in tweets associated with vaccines (week by week)

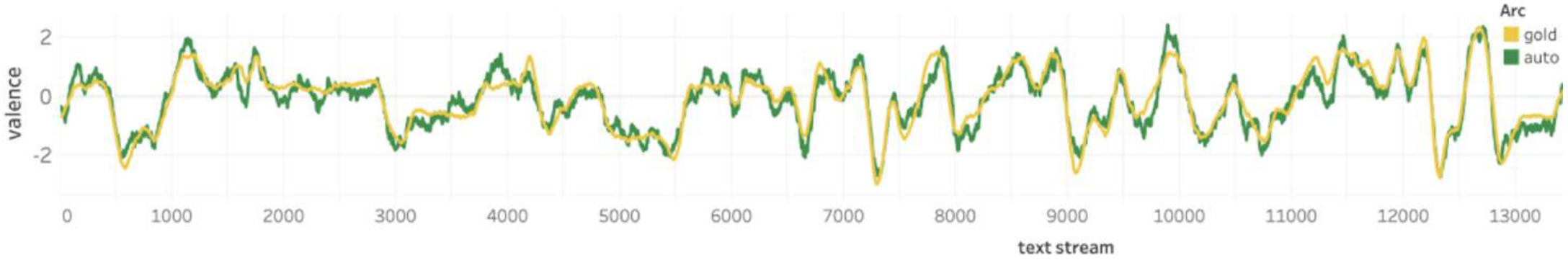
- Manually annotate 300,000 individual tweets from 2018 to 2024
- Take the percentage of tweets marked as joy in every week to create the emotion arc

Annotating data is a bottleneck

Daniela Teoderescu

2023 EMNLP: Evaluating Emotion Arcs Across Languages

- make use of existing emotion datasets (usually 2 to 5K instances)
- sample instances with replacement to generate random but non-trivial arcs
- create gold emotion arcs as usual



2023 EMNLP: Generating High-Quality Emotion Arcs Using Emotion Lexicons

- Used 36 datasets that had emotion-labeled sentences/tweets to create gold arcs
- For various affect categories, multiple languages, and other characteristics

Key Conclusions:

- lexicon-only based methods are extremely accurate
- aggregating information from hundreds of tweets/instances to create points of the emotion arcs very powerful

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
     a. sentence/post-level classification/regression/generation
     b. aggregate-level (group comparisons, emotion arcs, etc.)
  4. affective ethics

- Case Studies Exploring CAS Research

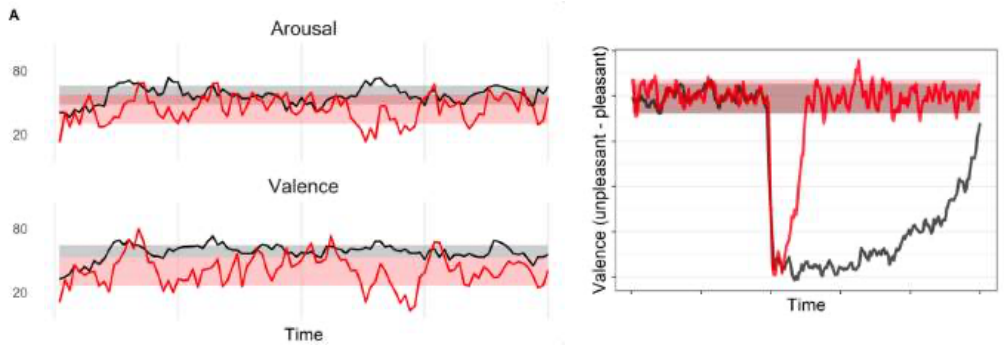
1: Affect and the Mind

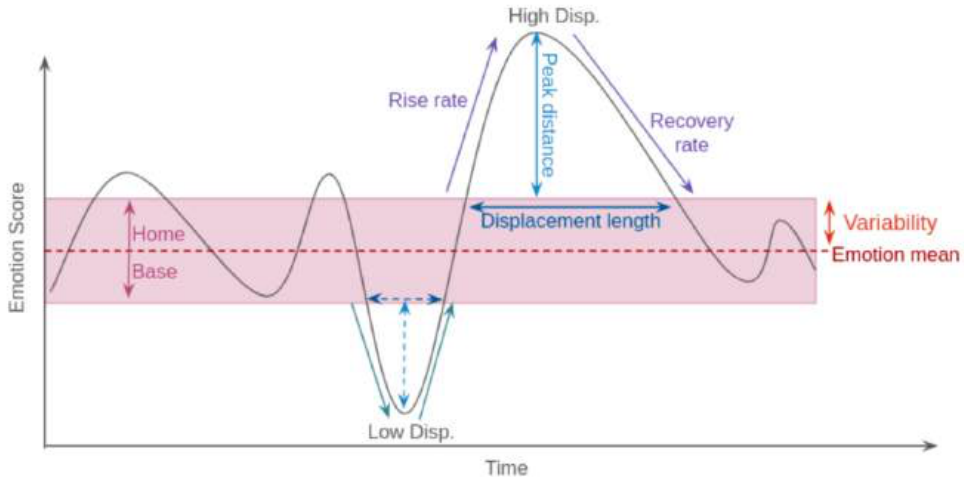
aggregate-level (emotion arcs)

Emotion Dynamics (from Psychology)

Study of change in emotional state with time

- intensive longitudinal data (repeated self-reports of emotional state)
- quite difficult to obtain such data



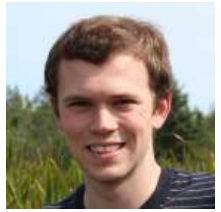


Another window into emotions is through our words:

- E.g., if happier, we are likely to utter more happiness-associated words

Utterance Emotion Dynamics: study of change in emotion words over time
(Hipson and Mohammad, 2021)

PLOS One



Will Hipson

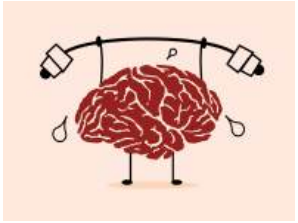
2021: Emotion Dynamics of Fictional Characters

LREC

Krishnapriya (KP) Vishnubhotla

2022: Tweet Emotion Dynamics

Emotion Word Usage in Tweets from US and Canada



EMNLP

Daniela Teodorescu

Tiffany Cheng

Alona Fyshe

2023: Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers.


1: Affect and the Mind–World

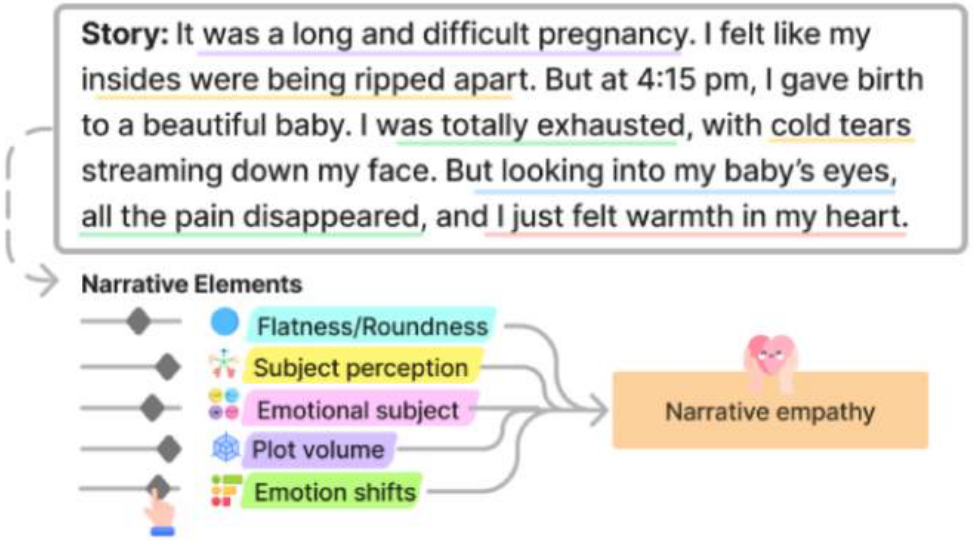
Emotions in Stories

- stories and narratives are fundamentally about the mind—but they are structured through the world
- narratives act as interfaces between internal affective–cognitive states and external world (social–ecological realities)

Affect in Narratives

- Storytelling is everywhere, and is deeply entwined with Affect:
  - Fiction
  - Societal/historical
  - Personal

- The affective dynamics of a narrative have downstream correlations with:
  - Genre, plot archetypes, quality/success
  - Cognitive reframing, mental health, identity formation.
  - Cultural and social perception
  - Reader preferences



Emotions in a personal narrative of a birthing story.

Emotions in Movies

The emotion arcs of stories have been talked about for a long time:

- Fundamental questions: what do we mean by “the emotional arc of a story”?
  - Emotions expressed by writer
  - Emotions invoked in reader
  - Emotions associated with character(s)? protagonists?

- Computational operationalization: emotions expressed in a text segment (sentence, chunk, paragraph) or audio-visual scene.

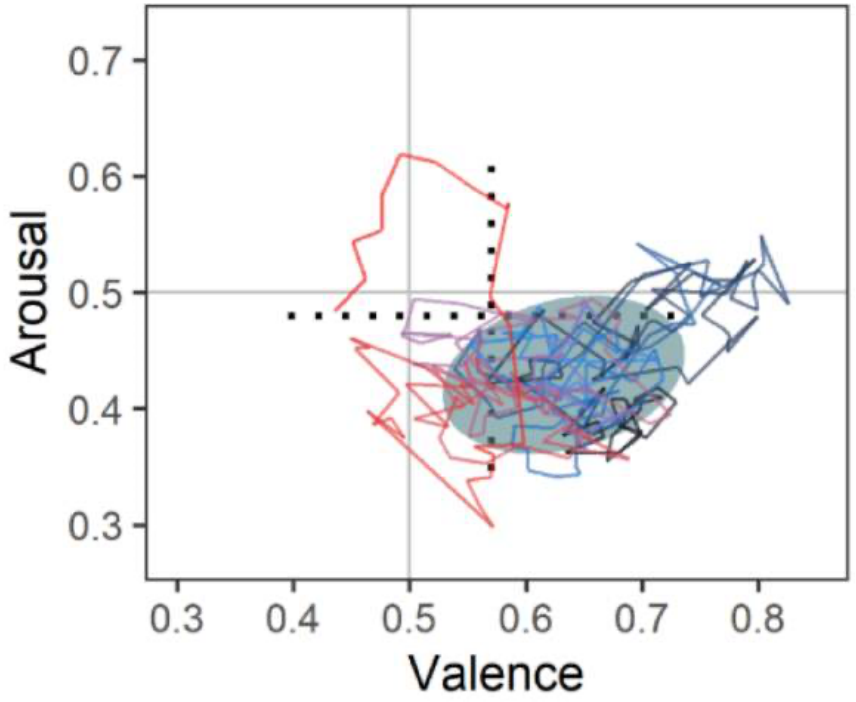
- [34] study the emotions of individual characters in movies using dialogue text with the Emotion Dynamics framework.
- [35] study the emotions of dialogues in movies using a combination of audio and text data.

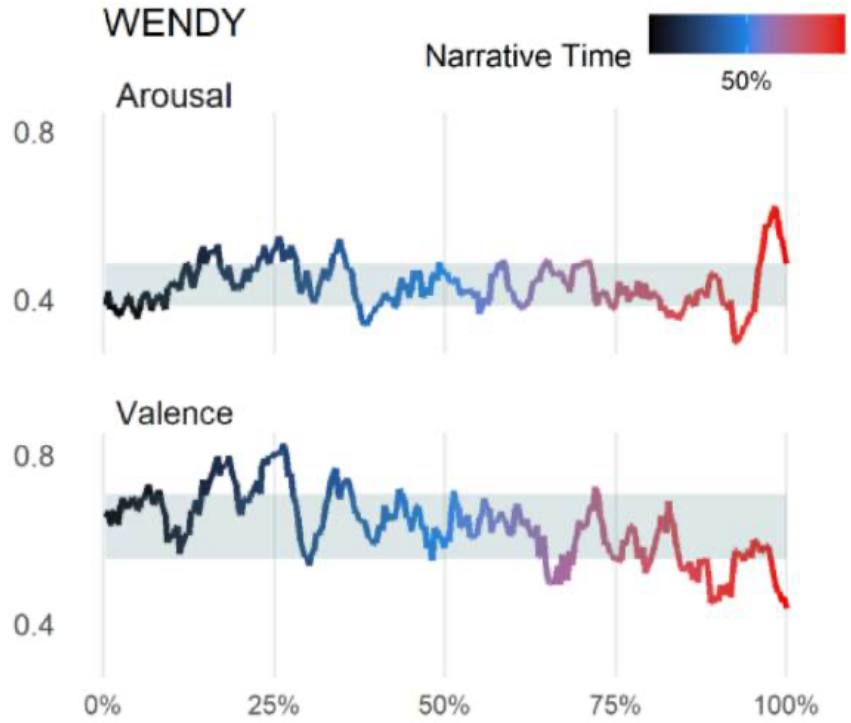
[34] Hipson, Will E., and Saif M. Mohammad. "Emotion dynamics in movie dialogues." PloS one 16.9 (2021): e0256153.
[35] Zhou, Naitian, and David Bamman. "Once More, With Feeling: Measuring Emotion of Acting Performances in Contemporary American Film." arXiv preprint arXiv:2411.10018 (2024).

Character Utterance Emotion Dynamics

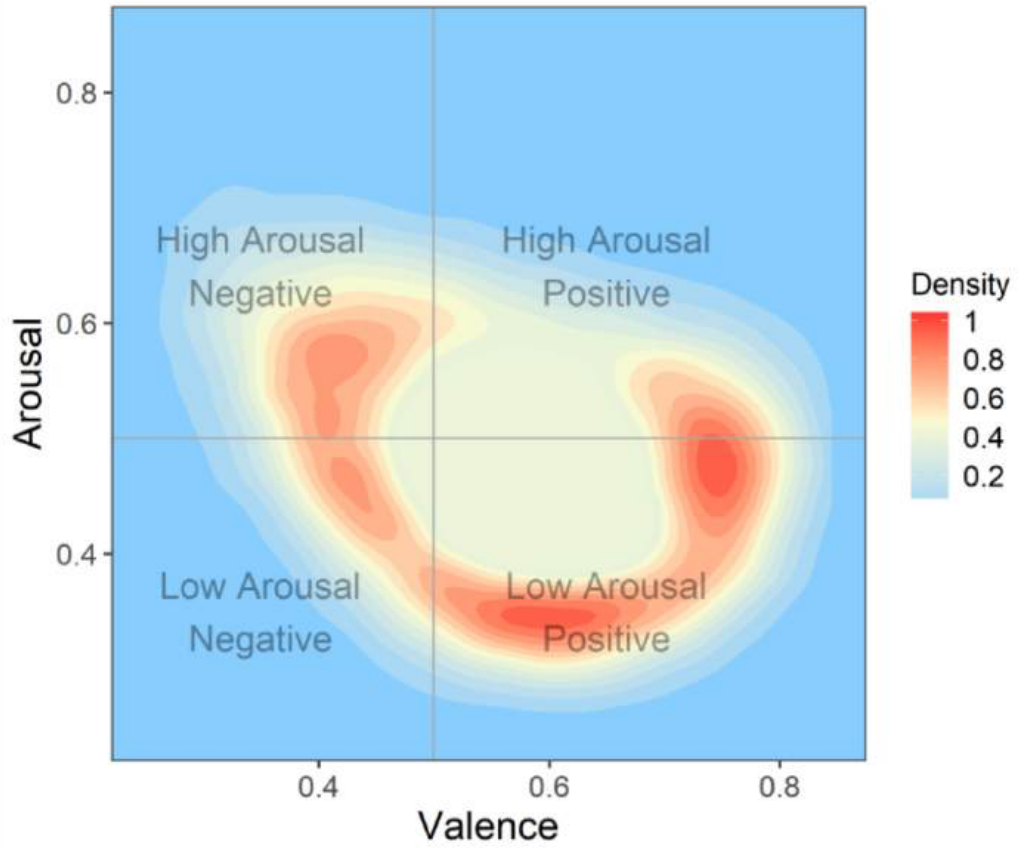
Data: the IMSDb corpus of movie scripts:

- 1,123 movies, 2687 main characters.
- Model the emotion dynamics for each character along the valence and arousal axes.





A sample emotion trajectory in 1D and 2D state spaces: Wendy from The Shining. [34]



Rank	Character	Movie Title	Rec.
1	Jacob	Nightmare on Elm Street	0.107
2	Chad	Burn After Reading	0.106
3	Andrew	The Breakfast Club	0.105
4	Rennie	Friday the 13th	0.101
5	Jimmy	Magnolia	0.101
2610	Jonson	Anonymous	0.019
2611	Agnis	Shipping News, The	0.018
2612	Paul	Manhattan Murder Mystery	0.017
2613	Jack	Burlesque	0.017
2614	Chigurh	No Country for Old Men	0.015

Where do characters tend to have the highest displacements? [34]

- low arousal positive (content)
- moderate arousal positive (happy)
- high arousal negative (agitated)

Which characters recover the fastest (and slowest) from emotional displacements? [34]

Emotions in Film

[35] Data: 2307 contemporary American films from 1980-2022.

- Align single-speaker audio segments with text from transcripts (sentence-level boundaries).
- Infer emotions expressed by speech (audio) and those expressed by the text.
- Cluster text segments by semantic similarity to obtain dialogue phrase groups.

Examples of utterances which are clustered into dialogue phrase groups.

Phrase Groups
"Let's go, let's go, let's go!", "Let's go, let's go!", "Let's go right now go go", "Go, let's go, let's go.", "Okay guys, let's go."
"Oh, pleasure to meet you.", "It's so nice to finally meet you.", "It is a pleasure to finally meet you.", "Oh, it's nice to meet you.", "It's so nice to meet you!"

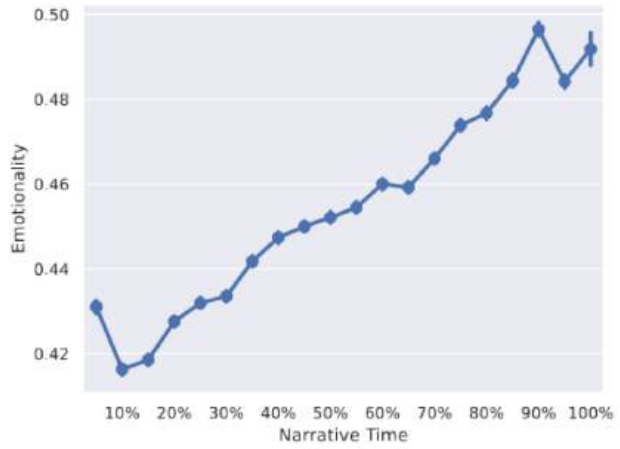
Emotional Variation within Phrase Groups

Dialogue phrase groups with the highest and lowest emotional range scores. The table shows a representative phrase from each group.

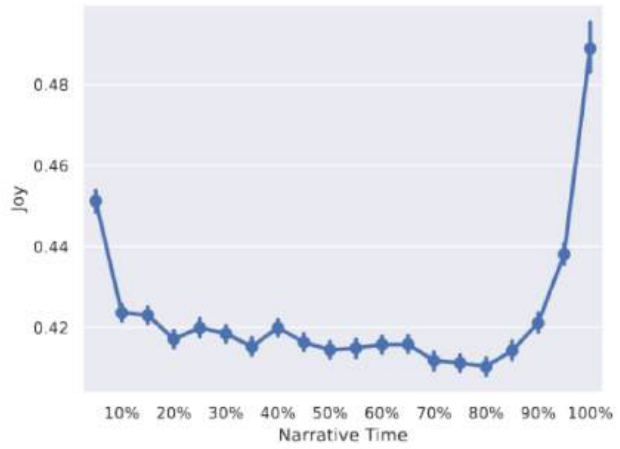
Low Emotional Range		High Emotional Range	
Phrase	**Entropy**	**Phrase**	**Entropy**
"Could I ask you something?"	-17.02	"All rise."	-7.85
"This is your captain speaking."	-16.56	"Are you out of your mind?"	-7.88
"Is that okay?"	-16.53	"What the fuck wrong with you?"	-7.99
"Can I get something for you?"	-16.17	"You're alive."	-8.32
"Can I get something to drink?"	-16.16	"You saved my life."	-8.34
"Hey, what can I get you?"	-15.91	"Don't you understand?"	-8.34
"You wanna come?"	-15.65	"Don't be so afraid."	-8.39
"Yeah, that's good."	-15.36	"You son of a bitch."	-8.40
"Any questions?"	-15.26	"You scared the shit out of me!"	-8.44
"That's correct."	-15.25	"Ow."	-8.44

Entropy is modelled as the entropy of the Dirichlet distribution that maximizes the likelihood of the observed emotion vectors. [35]

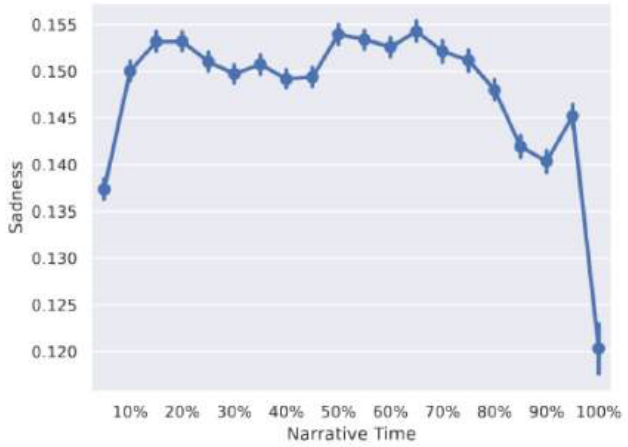
Emotionality over Narrative Time



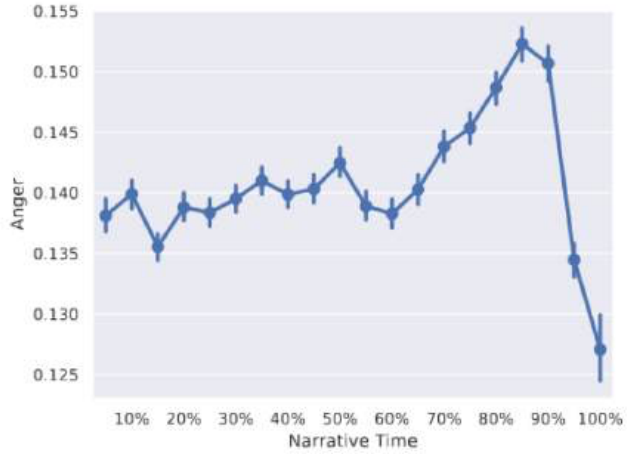
(a) Emotionality



(b) Joy



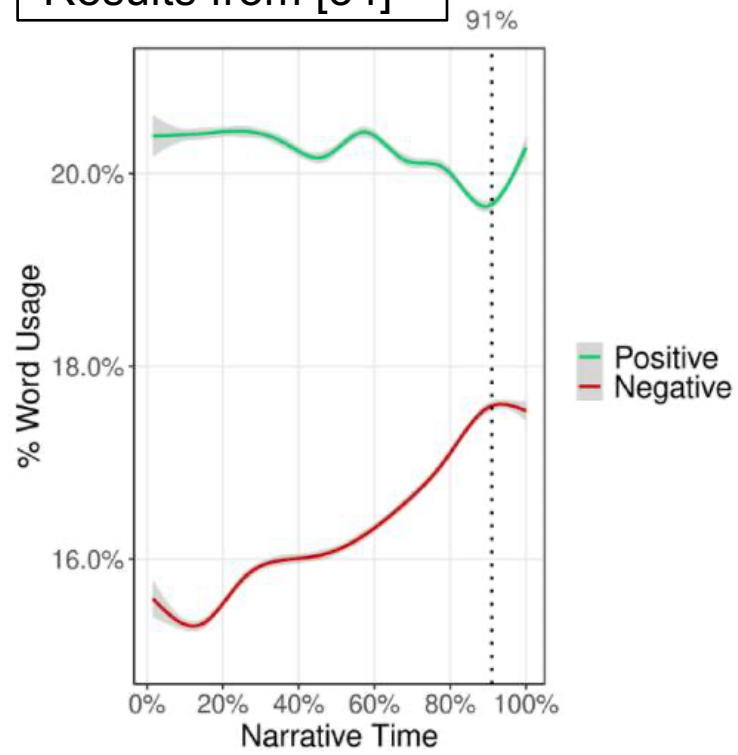
(c) Sadness

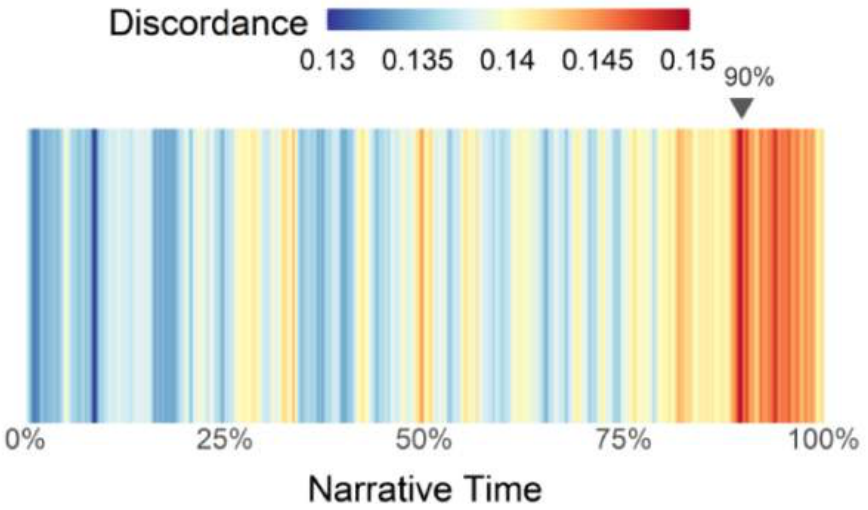


(d) Anger

Results from [35]

Results from [34]

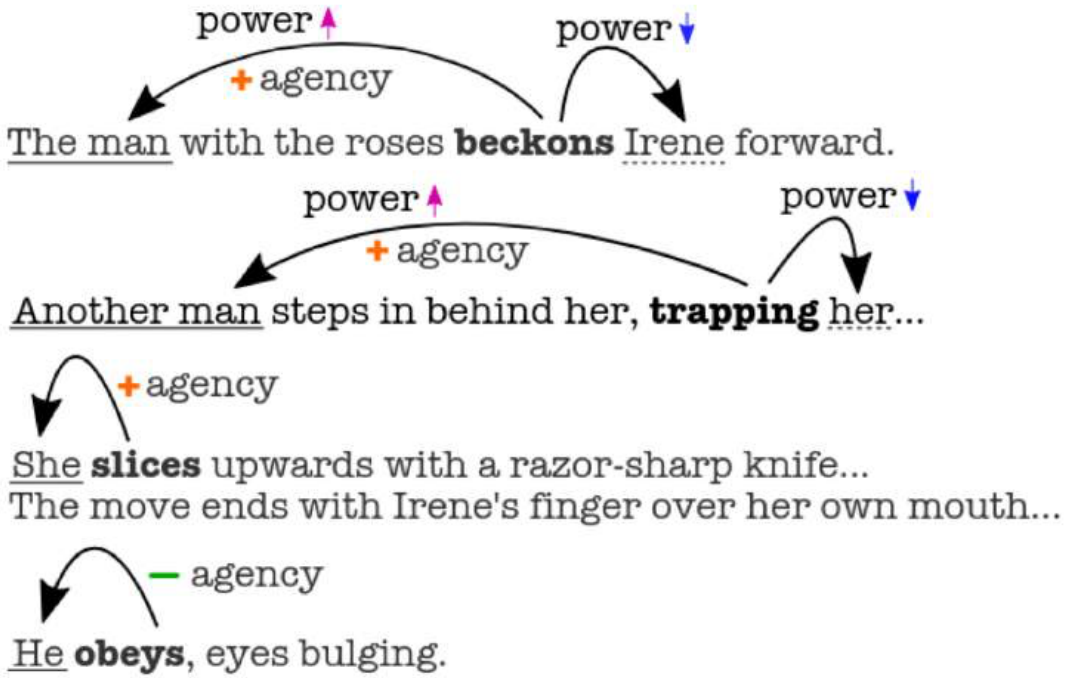




Biases in Stories

Biases in how social groups are portrayed in narratives are pervasive: by gender, race, social class, sexuality, religion, …

- Common character “tropes”: the dumb blonde, the miserly banker, the strong male rescuer.
- [36] formalize dimensions of *power* and *agency* afforded to characters.
- [37] look at biases in *generated* stories from a GPT-3 model, for associations with *appearance*, *intellect,* and *power.*



[36] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.
[37] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In Proceedings of the Third Workshop on Narrative Understanding, pages 48–55, Virtual. Association for Computational Linguistics.

Gender Biases in Stories

Findings from movie scripts [36]:

- Narration: male characters are portrayed with higher agency and higher power than female characters.
- Dialogue: male characters display more power and agency than female counterparts.
- LIWC features:
  - Women use more Hedges and Agreement words
  - Men more imperative sentences and inhibitory language.

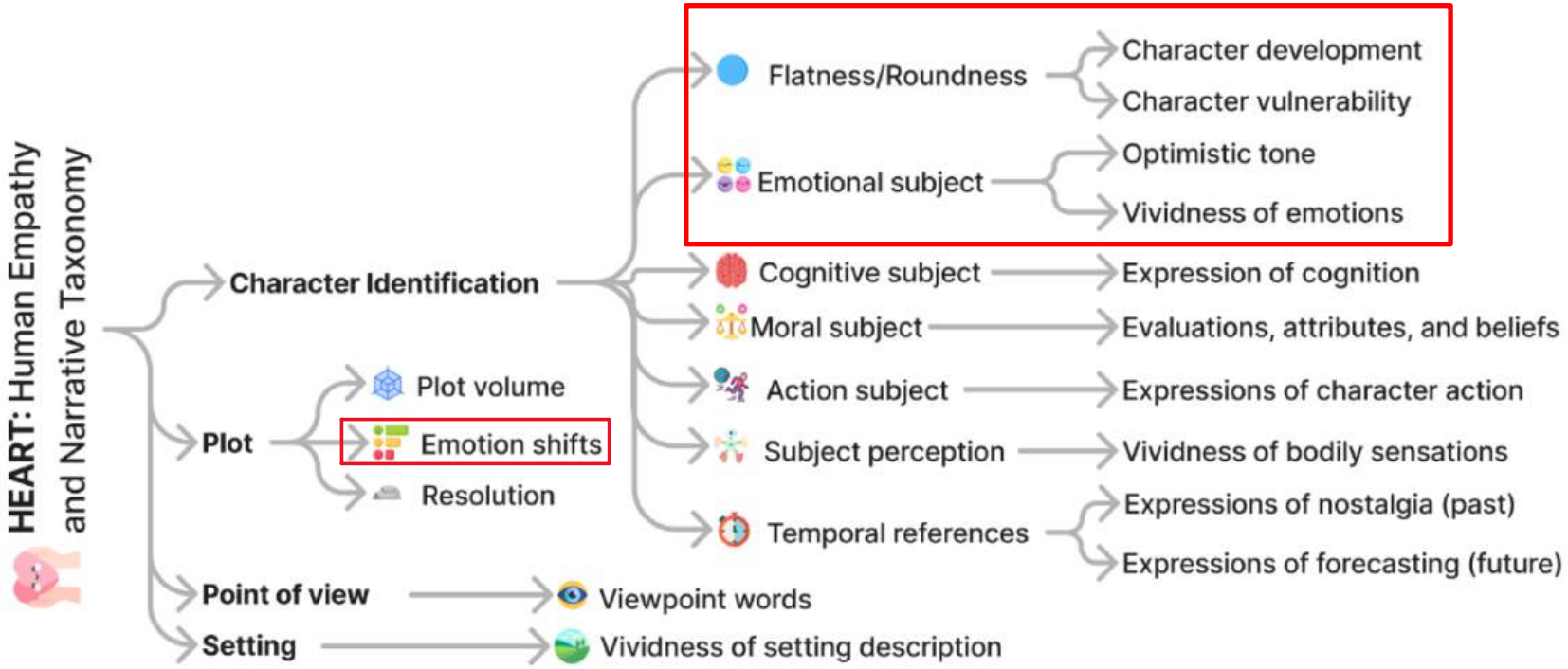
Findings from generated stories [37]:

- Topic distributions: feminine characters are more likely to be discussed in topics associated with *family*, *emotions*, and *body parts*; masculine characters with *politics, war, sports, crime*.
- Feminine characters are more likely to be described based on appearance.
- Masculine characters have higher power associations.

Empathy in Personal Narratives

We often share personal narratives in the hope of evoking empathy:

- Can we computationally model the “emotional resonance of a narrative”?
- [38] develop a theory-driven taxonomy of narrative style elements that relate to empathy.



[38] Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024. HEART-felt Narratives: Tracing Empathy and Narrative Style in Personal Stories with LLMs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1026–1046, Miami, Florida, USA. Association for Computational Linguistics.

Data:

- 874 personal narratives sourced from social media sites, podcasts, and storytelling interactions.
- Narrative features are measured using LLMs:
  - Prompted GPT-4 model.
  - Human validation shows this outperforms lexica-based features.
- Crowdworkers rate empathy towards stories.
  - Annotation study collected answers to several open-ended questions measuring reader characteristics and reading experience.

Findings:

- Vividness of emotions significantly impacts downstream empathy.
- Other significant effects: character development (flatness/roundness, depth), and plot volume (flow of events).
- The same narrative can evoke different levels of empathy in readers:
  - significant effect of demographics: age, sex, ethnicity, and trait empathy (baseline level of empathy).
- Other studies ([39]): complex interactions between mood and emotions of readers prior to reading the story, and those expressed in the story.

[39] M. Roshanaei, C. Tran, S. Morelli, C. Caragea and E. Zheleva, "Paths to Empathy: Heterogeneous Effects of Reading Personal Stories Online," 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Washington, DC, USA, 2019, pp. 570-579, doi: 10.1109/DSAA.2019.00072.

1: Affect and the Mind–Body

Emotion Granularity

We saw…

Emotion Dynamics: Individual Emotion Arcs

next…

Emotion Granularity: Pairs of Emotion Arcs

Emotion Granularity/Differentiation (from Psychology)

Some people:

- recognize, identify, describe their feelings using precise terms
  - like guilt, anger, frustration, or helplessness
- can reliably describe these concepts using language
  - distinguishing between angry and sad, elated and content, etc.

Others:

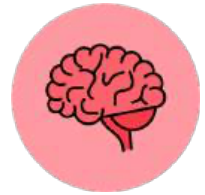
- tend to use more broad terms to convey emotions
  - a general sense of feeling bad or feeling low
- co-endorsing multiple emotions

Emotion granularity (Barrett et al., 2001)

- this ability to experience and categorize emotions in very specific terms
- the degree of not co-endorsing multiple emotions

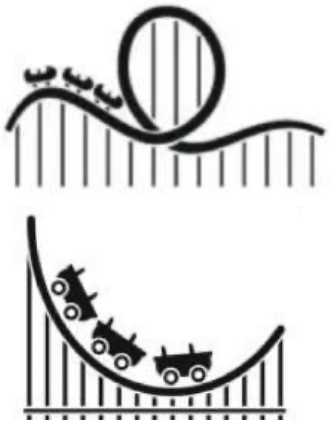
Lisa Barrett

Emotion Granularity and Health



- Mental health (Erbas et al., 2014, 2018)
  - depression (Starr et al., 2017)
  - anxiety (Seah et al., 2020)
  - borderline personality disorder (Dixon-Gordon et al., 2014, Suvak et al., 2011)
  - show less neural reactivity to rejection (Kashdan et al., 2014)
- Physical health (Hoemann et al., 2021)
  - cardiovascular physiological activity and stress (Bonar et al., 2023)
- Behavior
  - maladaptive behaviours such as binge drinking, aggression, and self-injurious behavior (Dixon-Gordon et al., 2014, Kashdan et al., 2015)
  - school behaviour (Brackett, Rivers, Reyes, & Salovey, 2012)
  - eating disorders (Selby et al., 2013)
  - less likely to retaliate aggressively (i.e., verbally or physically assault) against someone who has hurt them (Pond et al., 2012)

Emotion Granularity from Text (our work)



- To what extent are we co-endorsing multiple emotions **in text**?
  - through connotations and not necessarily denotations
- Compute emotion arcs for various emotions
- Compute emotion granularity (EG): correlation between pairs of emotion arcs
- Show that text by those who have self-disclosed to have certain mental health conditions (depression, PTSD, ADHD, etc.) have significantly lower EG than text by control group

EMNLP 2024:
Emotion Granularity from Text: An Aggregate-Level Indicator of Mental Health


Krishnapriya (KP) Vishnubhotla


Daniela Teodorescu


Mallory Fedman


Kristen Lindquist

Sophie Wu

Jan Philip Wahle

The Language of Interoception:

Examining Embodiment and Emotion Through a Corpus of Body Part Mentions

Impact of the body on our

- brain
- how we conceptualize and perceive things
- emotions

Work from

- Medicine
- Psychology
- Affective Science
- Cognitive Science
- Philosophy

Key Theories

- Embodiment Theory
  - our physical body and its interactions with the world are fundamental to our cognitive processes, emotions, and experiences (not separate)

- The Theory of Conceptual Metaphor (Lakoff & Johnson, 1980)
  - abstract concepts (like time, emotion, power, and relationships) map to concrete source domains (like body, movement, or space)'

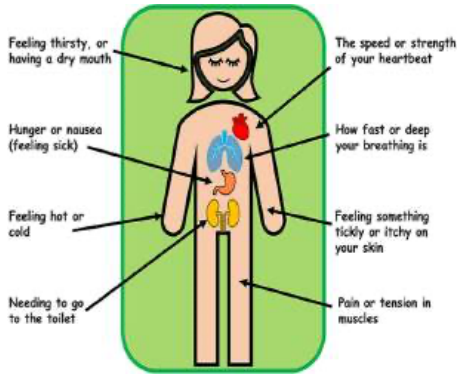
    e.g., *grasp* (a physical act) to mean *understand*
  - abstract domains are understood via source domains like the body, movement, or space.

Body Part Mentions (BPMs) offer a rich entry point into conceptual metaphor.

Studying them can help reveal how language encodes embodied thought.

Key Theories (continued)

- Mindfulness Theory
  - developing an awareness of the present-moment
    > positive outcomes (reduced stress, improved emotional regulation, well-being)
- Theory of Constructed Emotion
  - Emotions are actively constructed by the brain in each moment based on past experiences, context, and sensory information
- Theories on Interoception
  - Better interoception positively correlated with better physical and mental health
    - Emotional regulation (Zamariola et al., 2019)
    - Emotional decision-making (Dunn et al., 2010)
    - Emotional granularity (ability to distinguish emotions) (Zamariola et al., 2024)



The Language of Interoception:

Examining Embodiment and Emotion Through a Corpus of Body Part Mentions

Can we use language to shed light on this connection of the body with the mind, emotions, health, and behavior?

Body Part Mentions

Instances of language where words referring to parts of the body are used.

Examples:

My head hurt after the smog from last week.

Her chest tightened after hearing the news.

His back ached after the long flight.

I could feel my heart racing.

She rolled her eyes at the comment.

The hair on my arms stood up.

Ambiguity with BPMs

BPMs that abstractly refer to a person's body.

Thank you for always being by my side.

That went over my head.

BPMs that use body parts as active metaphor, but do not refer to a person's body.

Hands down, the best film of the year.

The heart of the matter is (...)

BPMs with fossilized/weak metaphorical connections.

I will be right back.

Let's head out.

Everyday use of BPMs can tell us about:

- the relationship between the body and mind
- how natural language is connected to the physical world from which it originates

Body Research Questions

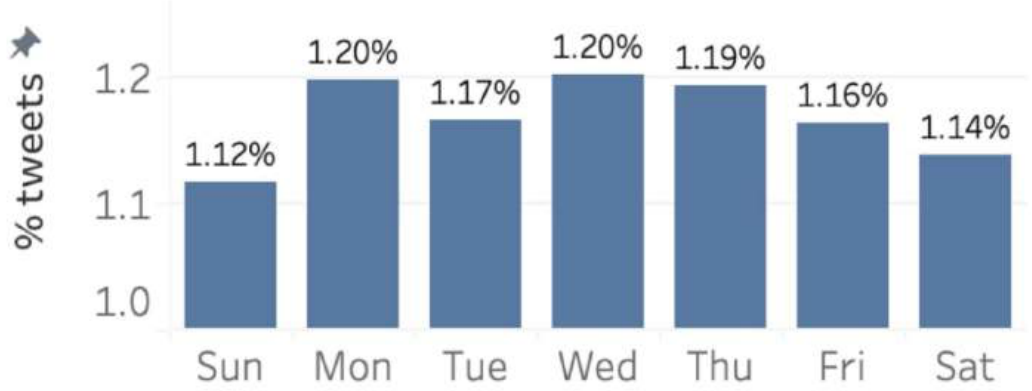
1. To what extent do we use body-related words? 5 to 10%
2. To what extent do we talk about our own body parts versus others' body parts?
3. Which of our body parts do we refer to most often? Do we refer to our body differently in different online contexts?
4. Does the time of day/week/year impact whether we refer to our body?
5. Do individuals in different regions refer to their bodies at different frequencies?

B4.a. Mentions of “my <BPM>” by Month

Mentions of “my <BPM>” peak in spring and summer, decline in fall, and stay lowest in winter.

- May reflect increased bodily awareness in warmer, more active months (sunlight, time outside).

B4.b. Mentions of "my <BPM>" by Day of Week

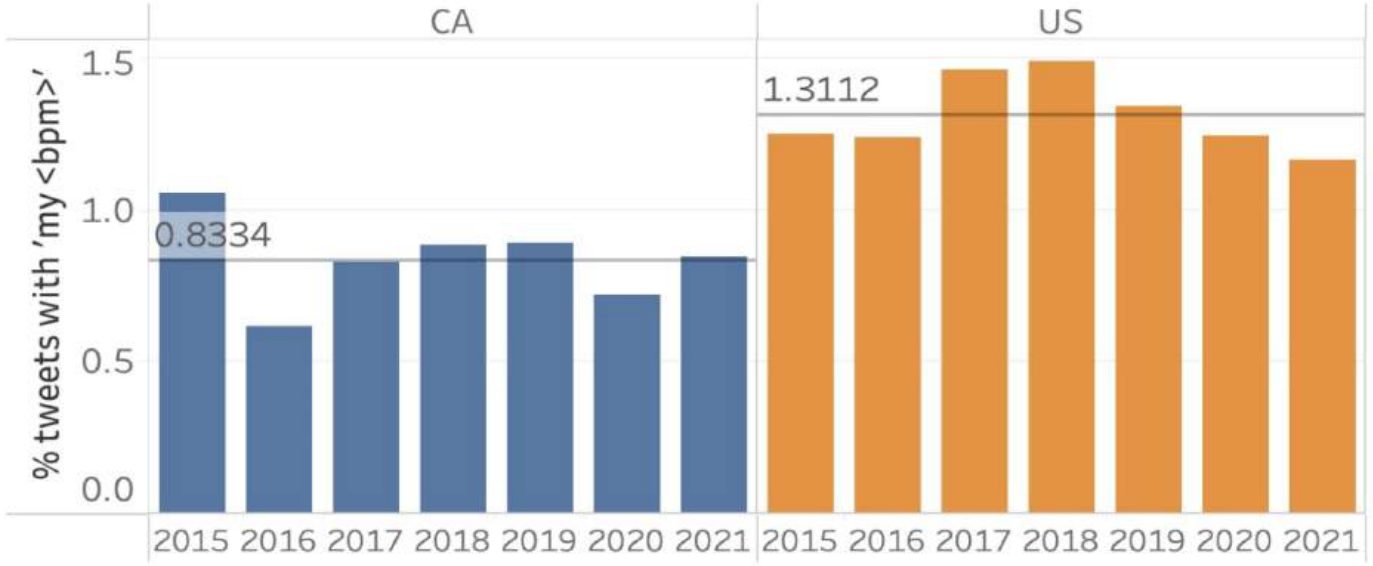


- Usage rises from Weekends to mid-week, then declines
- Suggests a connection to the structure and fatigue of the work week

References to our bodies are not static—they reflect seasonal, environmental, and social rhythms in our lives.

B5.a. Do individuals in different regions refer to their bodies at different frequencies?

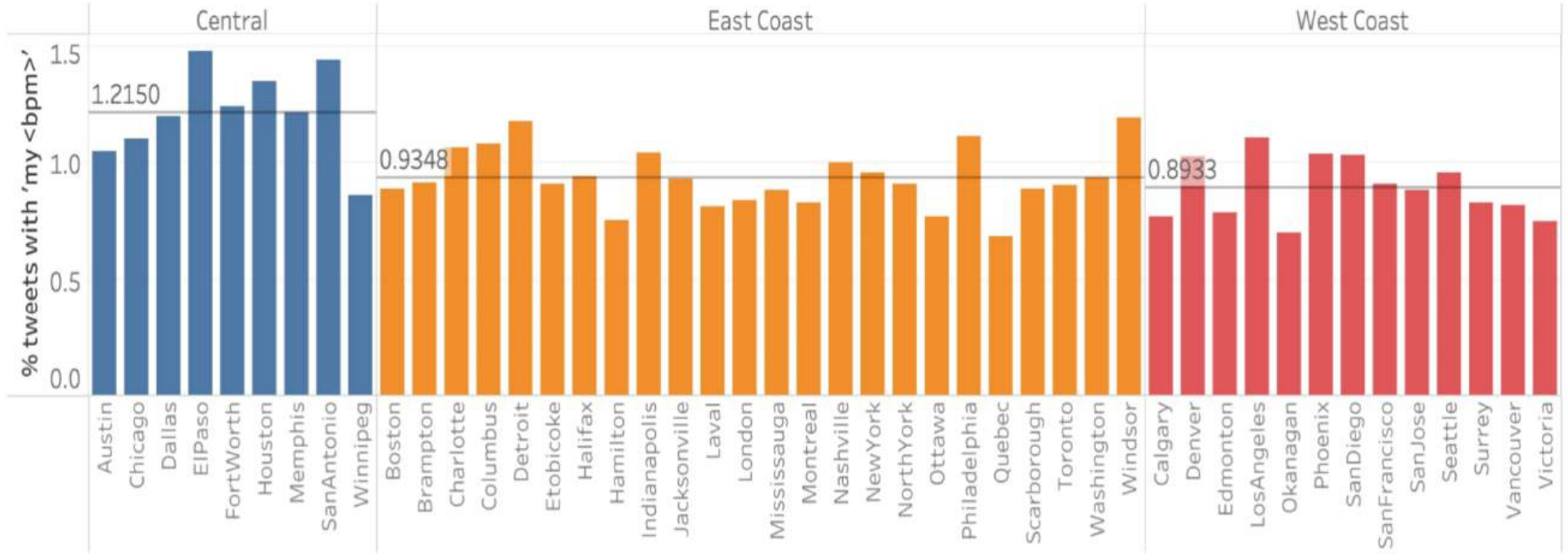
Canada vs. US across Years



Americans refer to their bodies more than Canadians in tweets.

B5.b. Do individuals in different regions refer to their bodies at different frequencies?

US cities: Central, East Coast, West Coast



Central cities show higher BPM usage than coastal cities.

Body and Affect Research Questions

1. Do posts with body part mentions have markedly different emotional associations?
2. What is the impact of explicitly embodied emotion on the emotions expressed through body part mentions?
3. Do individual body part mentions co-occur with markedly different emotion distributions?

Instances	Metric	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
my <any BPM>	Mean	27.6	37.7	21.2	29.4	32.2	28.2	21.7	37.8
	Std. Dev.	5.3	6.9	5.3	6.6	5.0	7.0	3.6	7.4
		0 20 40	0 20 40 60	0 10 20 30	0 20 40	0 20 40	0 20 40	0 10 20 30	0 20 40 60
		% posts	% posts	% posts	% posts	% posts	% posts	% posts	% posts

Instances	Metric	ΔAnger	ΔAnticipation	ΔDisgust	ΔFear	ΔJoy	ΔSadness	ΔSurprise	ΔTrust
arm	Mean	12.1	-4.9	2.8	11.5	-3.8	14.0	-0.2	-3.3
back	Mean	3.1	-5.2	-0.8	4.6	-1.1	3.2	-2.3	-3.7
belly	Mean	-6.9	-6.4	4.5	-3.4	-1.6	-1.5	-6.9	-8.8
blood	Mean	1.8	0.1	9.8	3.7	3.1	2.3	-0.3	3.2
body	Mean	3.9	0.5	6.6	6.5	2.0	8.7	-0.4	-1.8
brain	Mean	1.3	-1.3	1.0	1.9	-4.6	3.5	-2.6	-3.5
chest	Mean	9.1	-3.5	6.2	5.6	-1.6	6.5	-2.8	-5.3
ear	Mean	1.0	-11.7	2.3	-2.6	-5.0	-2.4	-5.6	-10.8
eye	Mean	-2.3	-3.7	-2.1	-1.0	-3.3	-1.6	-0.7	-4.1
face	Mean	1.5	-5.5	3.6	-1.9	3.0	-0.8	4.2	-0.5
feet	Mean	3.0	-3.3	1.9	1.2	-1.2	2.0	-1.5	-2.4
finger	Mean	3.3	-3.3	2.3	4.0	0.0	1.4	2.5	-1.9
foot	Mean	5.9	-4.4	3.6	9.4	-4.9	10.2	0.1	-1.5
forehead	Mean	3.8	-7.9	6.6	-2.9	-6.4	-4.0	-3.8	-13.9
hair	Mean	-6.8	-4.7	-1.0	-9.5	-5.8	-6.4	-7.4	-10.5
hand	Mean	2.5	-2.1	0.1	4.0	0.4	1.0	-0.8	-0.5
head	Mean	1.7	-4.8	0.3	0.2	-5.4	0.2	-3.4	-4.6
heart	Mean	0.8	2.9	-1.1	7.2	13.6	7.6	1.2	6.4
knee	Mean	13.4	-4.2	2.1	15.9	-8.0	18.7	-3.6	-5.8
leg	Mean	9.5	-2.7	7.2	15.9	-3.8	17.8	-0.1	-4.5
lungs	Mean	18.3	28.5	18.8	19.6	6.7	11.3	10.7	26.2
mouth	Mean	8.3	-3.4	8.8	2.5	-5.3	0.3	-2.9	-3.0
neck	Mean	1.7	-6.6	1.3	0.1	-6.8	2.8	0.9	-8.4
nerves	Mean	5.5	-13.9	6.0	-0.4	-12.2	1.7	-7.4	-16.4
nose	Mean	1.0	-8.6	4.6	-2.3	-7.9	0.4	-4.3	-8.6
shoulder	Mean	8.4	-3.2	6.4	10.7	-2.7	10.4	1.5	-1.5
skin	Mean	4.1	-2.5	12.7	2.4	2.5	5.0	-0.8	2.8
stomach	Mean	12.3	-6.7	21.1	11.4	-5.1	21.6	-5.9	-8.9
teeth	Mean	0.3	-6.5	2.8	-0.2	-2.7	3.2	-4.2	-4.9
throat	Mean	12.9	-10.1	7.8	10.1	-6.6	10.9	-2.9	-10.4
tongue	Mean	-0.7	-9.6	2.0	-5.6	-5.9	-5.9	-4.7	-8.8
		-10 0 10 20	-10 0 10 20	-10 0 10 20	-10 0 10 20	-10 0 10 20	-10 0 10 20	-10 0 10 20	-10 0 10 20

BPMs and Emotions

- Different body parts are associated different emotions

  my stomach → sadness

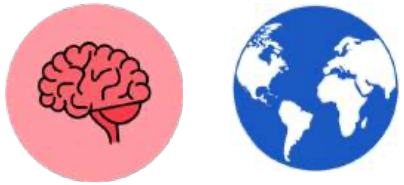
  my chest → anger

- Negative emotions (anger, fear, sadness) dominate many high-frequency BPMs.

Are BPMs correlated with health outcomes?

	Freq. Mental Distress		Freq. Phys. Distress		Life Expectancy		Physical Inactivity	
	Spearman's *r*	*p*-value	Spearman's *r*	*p*-value	Spearman's *r*	*p*-value	Spearman's *r*	*p*-value
a. Number of tweets	-0.170	0.418	-0.167	0.425	0.290	0.160	-0.243	0.242
b. Prop. of <*Fear word*> tweets	-0.230	0.231	-0.370	0.054	0.160	0.403	**-0.460**	**0.014**
c. Prop. of "*my* <*BPM*>" tweets	**0.497**	**0.012**	**0.721**	**0.000**	**-0.409**	**0.043**	**0.704**	**0.000**
d. Prop. of "<*BPM*>" tweets	**0.527**	**0.007**	**0.553**	**0.004**	**-0.613**	**0.001**	**0.539**	**0.006**

- Proportion of BPM tweets is a strong predictor of negative health outcomes
- Emotion word use weakly correlated (at best)



1: Affect and the Mind–World

…

Computational Social Science: Intergroup Bias and Empathy

- People are susceptible to in-group bias: favourable and preferential attitudes towards those who are perceived to belong to the same group (in-groups), compared to out-group members.
  - Groups can be defined by various social factors: race, gender, nationality, economic class, political party, or even favourite sports team.

- Intergroup biases can reduce empathy towards out-groups:
  - brain scan studies show “participants tend to have much stronger empathy-related neural responses when watching in-group members in pain than out-group members.”*
  - A major component of prejudice, stereotyping, and discrimination.

- Intergroup biases are encoded in language:
  - How do people talk about their in-groups vs out-groups?
  - Linguistic Intergroup Bias: positive in-group and negative out-group behaviours are described more abstractly than negative in-group and positive out-group behaviours.

* Sheng, F., Liu, Y., Zhou, B., Zhou, W., & Han, S. (2013). Oxytocin modulates the racial bias in neural responses to others’ suffering. Biological Psychology, 92(2), 380–386. https://doi.org/10.1016/j.biopsycho.2012.11.018

Intergroup Bias and Emotion in Political Tweets

[31] annotate a dataset of tweets from members of the US Congress that mention another US Congress member:

- Political party affiliation is a proxy for in-group vs out-group relationship
- Annotated for "emotion expressed by speaker towards target" (Plutchik + none)

Emotion	All	In-Group	Out-Group
Admiration	15.5	22.2	9.1
Anger	8.2	1.0	15.1
Disgust	7.4	0.3	14.2
Interest	22.9	27.2	18.6
Joy	26.7	32.2	21.4
Sadness	2.5	2.6	2.4
No Emotion	16.8	14.5	19.1

Table 3: Proportion of emotions in different interpersonal contexts

a. **In-group**: We stand w @Doe, who has seen a lot worse than cheap insults from an insecure bully. #MLKDAY weekend.
b. **Out-group**: Parents and families live in constant fear for their children with food allergies. A worthy bi-partisan cause - thank you @Doe for your leadership on this issue.

- Negative emotions are almost always expressed in out-group settings.
- Positive emotions are more evenly distributed.
- Classification experiments indicate that emotion features help prediction of IGR – intuitive because of imbalance distribution.
  - but they are not sufficient by themselves.

[31] Govindarajan, Venkata S, Katherine Atwell, Barea M. Sinno, Malihe Alikhani, David Ian Beaver and Junyi Jessy Li. "How people talk about each other: Modeling Generalized Intergroup Bias and Emotion." Conference of the European Chapter of the Association for Computational Linguistics (2022).

The Language of Dehumanization

Drawing on social psychology research, [32] identify and operationalize four components that manifest in language: negative evaluations, denial of agency, moral disgust, and metaphors invoking “vermin”.

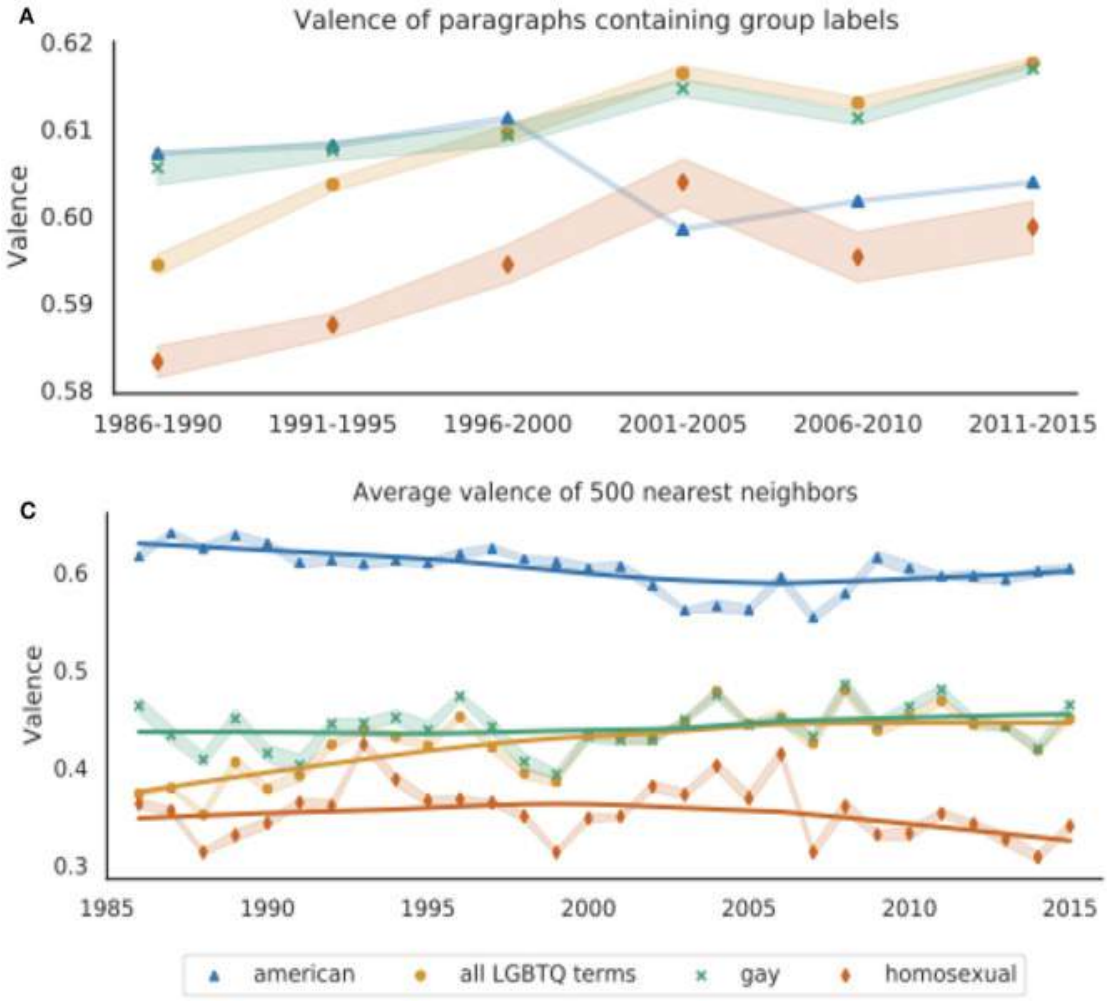
Overview of linguistic correlates and our operationalizations for four elements of dehumanization.

Dehumanization element	Operationalization
Negative evaluation of target group	Paragraph-level sentiment analysis
	Connotation frames of perspective
	Word embedding neighbor valence
Denial of agency	Connotation frames of agency
	Word embedding neighbor dominance
Moral disgust	Vector similarity to disgust
Vermin metaphor	Vector similarity to vermin

[32] Mendelsohn J, Tsvetkov Y, Jurafsky D. A Framework for the Computational Linguistic Analysis of Dehumanization. Front Artif Intell. 2020 Aug 7;3:55. doi: 10.3389/frai.2020.00055. PMID: 33733172; PMCID: PMC7861242.

Data (focus group: LGBTQ Communities)

- *New York Times* articles from 1986-2015.
  - Containing terms from a list of LGBTQ terms.
  - Control: articles mentioning “Americans”



Findings:

- Generally, LGBTQ groups have become more positively evaluated over time:
  - Except for the label “homosexual”.
  - “American” is used in more positive contexts.
- LGBTQ groups experience greater denial of agency.
- Associations with moral disgust and “vermin” generally reduce over time; again, there are connotative differences between the terms “gay” and “homosexual”.

Biases in Large Language Models

If human text encodes biases in attitudes and affect towards in-group and out-group members, is the same behaviour reflected in Large Language Models?

- [33] model this via an emotion intensity prediction task.
- Consider a narrative described by an *experiencer* belonging to a social group *s1* involving an *emotion*.
  - The LLM is assigned to be the *perceiver* of the emotion, with a persona belonging to social group *s2.*
- Intergroup bias is measured by the (average) difference in LLM-perceived emotion intensity for in-groups (s1==s2) vs out-groups (s1!=s2).
- Social groups are defined based on race, nationality, and religion.

[33] Hou, Yu, Hal Daum'e and Rachel Rudinger. "Language Models Predict Empathy Gaps Between Social In-groups and Out-groups." North American Chapter of the Association for Computational Linguistics (2025).

Findings:

- Race, nationality, and religion all show significantly higher predicted intensities for in-group pairs.
- Results are robust across groups, prompt variations, and LLMs.
- Biases based on social groups:
  - White perceivers are the most empathetic; Black perceivers are the least empathetic.
  - Asians receive the least amount of empathy; Hispanics the most.
  - Similar cultural effects observed for Nationality and Religion.
  - Replicates certain historical realities:
    - Israel vs Palestine; Ukraine vs Russia – lowest average intensities.

Where does this matter?

- LLMs are increasingly becoming a part of social conversations: participants, mediators, impersonators.
- Desired behaviours and biases should be measured when LLMs are used in these settings.

This Tutorial

- Core Theories of Affect and Emotion
  - What is Computational Affective Science?
  - Why it matters?

- Overview of the Areas of Work in CAS
  1. nature of affect; affect and the mind, body, world
  2. affective data
  3. affective tasks
     a. sentence/post-level classification/regression/generation
     b. aggregate-level (group comparisons, emotion arcs, etc.)
  4. affective ethics

- Case Studies Exploring CAS Research

4: Affective Ethics

…

Emotion Recognition: Task

1. Inferring emotions felt by the speaker

Emotion Recognition: Task

1. Inferring emotions felt by the speaker
2. Inferring emotions of the speaker as perceived by the reader/listener
3. Inferring emotions that the speaker is attempting to convey
4. Inferring emotions evoked in the reader/listener
5. Inferring emotions of people mentioned in the text
6. Inferring whether what is described is good for pre-determined target of interest

Emotion Recognition: Task

1. Inferring emotions felt by the speaker
2. Inferring emotions of the speaker as perceived by the reader/listener
3. Inferring emotions that the speaker is attempting to convey
4. Inferring emotions evoked in the reader/listener
5. Inferring emotions of people mentioned in the text
6. Inferring whether what is described is good for pre-determined target of interest
7. Inferring the intensity of the emotions discussed above
8. Inferring patterns of speaker's emotions over long periods of time, across many utterances; including the inference of moods, emotion dynamics, and emotional arcs
9. Inferring speaker's emotions/attitudes/sentiment towards a target product, movie, person, idea, policy, entity, etc.
10. Inferring emotionality of language used in text (regardless of whose emotions)
11. Inferring how language is used to convey emotions such as joy, sadness, loneliness, hate, etc.
12. …

Emotion Recognition: Task

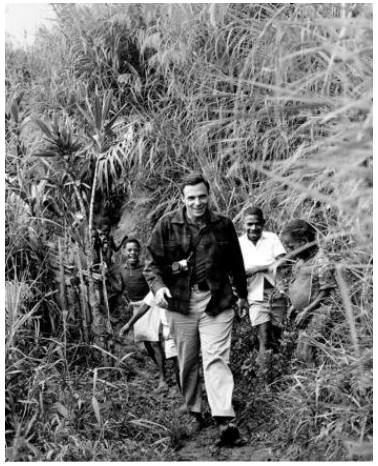
1. Inferring emotions felt by the speaker
2. Inferring emotions of the speaker as perceived by the reader/listener
3. Inferring emotions that the speaker is attempting to convey
4. Inferring emotions evoked in the reader/listener
5. Inferring emotions of people mentioned in the text
6. Inferring whether what is described is good for pre-determined target of interest
7. Inferring the intensity of the emotions discussed above
8. Inferring patterns of speaker's emotions over long periods of time, across many utterances; including the inference of moods, emotion dynamics, and emotional arcs
9. Inferring speaker's emotions/attitudes/sentiment towards a target product, movie, person, idea, policy, entity, etc.
10. Inferring emotionality of language used in text (regardless of whose emotions)
11. Inferring how language is used to convey emotions such as joy, sadness, loneliness, hate, etc.
12. …

All of these come with…
Benefits, Potential Harms, Ethical Considerations

Theories of Emotion

Margaret Mead
Cultural anthropologist

Paul Ekman
Psychologist and discoverer of micro expressions.



Lisa Barrett
University Distinguished Professor of Psychology, Northeastern University

Theory of Constructed Emotion (Barrett, 2017)

- the brain constructs emotions
- important tenets of BET discredited ("basic" emotions)
- stress on variability

Computational Analysis of Emotions and Automatic Emotion Recognition (AER)

A force that helps unlock:

- how emotions work
- how they relate to our health, language, behavior, social interactions,…
- numerous commercial applications that benefit society

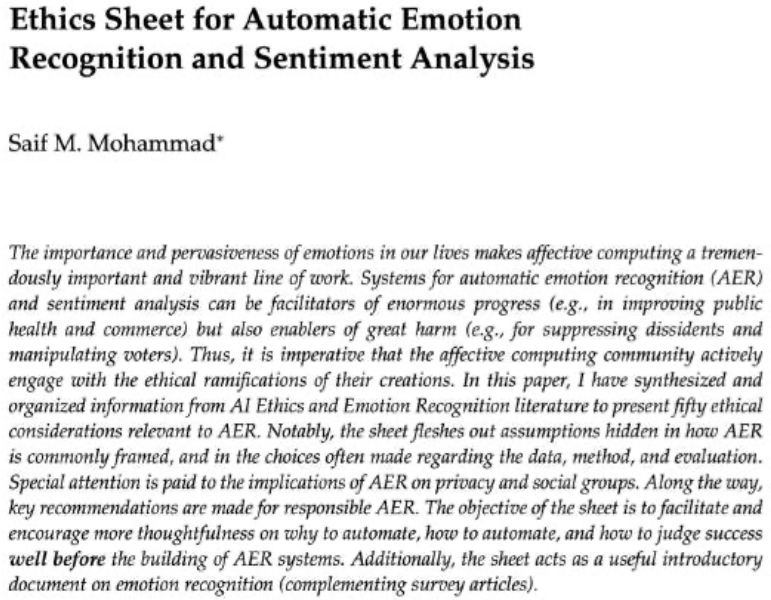
A tool for substantial harm, e.g.:

- mass application on vulnerable populations
- unreliable approaches
- privacy concerns
- perpetuation of physiognomy

Ethics Sheet for
Automatic Emotion Recognition and Sentiment Analysis

Medium Blog Post in summer of 2021:
https://medium.com/@nlpscholar/ethics-sheet-aer-b8d671286682

CL Journal June 2022



Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis

Saif M. Mohammad*

The importance and pervasiveness of emotions in our lives makes affective computing a tremendously important and vibrant line of work. Systems for automatic emotion recognition (AER) and sentiment analysis can be facilitators of enormous progress (e.g., in improving public health and commerce) but also enablers of great harm (e.g., for suppressing dissidents and manipulating voters). Thus, it is imperative that the affective computing community actively engage with the ethical ramifications of their creations. In this paper, I have synthesized and organized information from AI Ethics and Emotion Recognition literature to present fifty ethical considerations relevant to AER. Notably, the sheet fleshes out assumptions hidden in how AER is commonly framed, and in the choices often made regarding the data, method, and evaluation. Special attention is paid to the implications of AER on privacy and social groups. Along the way, key recommendations are made for responsible AER. The objective of the sheet is to facilitate and encourage more thoughtfulness on why to automate, how to automate, and how to judge success ***well before*** *the building of AER systems. Additionally, the sheet acts as a useful introductory document on emotion recognition (complementing survey articles).*

Template

50 considerations grouped under:

- *Task Design*
- *Data*
- *Method*
- *Impact and Evaluation*

} common phases in system development

- *Implications for Privacy and Social Groups*

TASK DESIGN

A. Theoretical Foundations

1. Task Design and Framing
2. Theoretical Models and their Implications
3. Meaning and Extra-Linguistic Information
4. Wellness and Health Implications
5. Aggregate Level vs. Individual Level Prediction

B. Implications of Automation

6. Why Automate
7. Embracing Diversity
8. Participatory/Emancipatory Design
9. Applications, Dual Use, Misuse
10. Disclosure of Automation

DATA

C. Why This Data

11. Types of data
12. Dimensions of data

D. Human Variability–Machine Normativeness

13. Variability of Expression, Conceptualization
14. Norms of Emotions Expression
15. Norms of Attitudes

… 50!

First SemEval Shared Task on African Languages:

SemEval 2023: AfriSenti: Detecting Sentiment in African Languages

Labeled sentiment datasets for **14** languages from 3 language families

Led by African researchers

Shared Task on Emotions

SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

Labeled emotion datasets for **35** languages

-- most from Africa and Asia

Most popular task in CodaLab for the year 2024

SemEval 2025 Best Shared Task Award

ACL 2025 Best Resource Paper Award

Affect in GenAI Development and its Ethics

Ethics of AI Assistants

- Training Large Language (and Vision) models for language modelling has led to a surprising emergence of “intelligence”:
  - Artificial systems can understand and interface with human communication.

- The release and public availability of systems like ChatGPT means that everyday people have everyday conversations with these models:
  - One study* found Therapy/Companionship to be #1 use of Generative AI in 2025.
  - Other top use-cases include Personal Advice, Life Organization, and Brainstorming.

- Given the increasing use of LLMs for personal problem-solving, emotional support, and companionship:
  - What affective capabilities should these systems posses?
  - What guardrails should we have in place?
  - What insights do Affective Science, Cognitive Science, and Psychology offer?

*Marc Zao-Sanders. How People Are Really Using Gen AI in 2025 — hbr.org. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025, 2025. [Accessed 02-05-2025]

Incorporating Psychology Theories in LLM Training

[40] illustrate how theories from cognitive, developmental, behavioural, social, psycholinguistic, and personality theories integrate into four key stages of LLM development:

- Key among these are learning methods from cognitive psychology influencing training mechanisms for LLMs: sequential and continuous learning; reward formulations in reinforcement learning.

- Personality psychology theories influence studies on improving collaborative multi-agent LLM frameworks.
  - Does assigning certain “personas” to LLMs improve collaborative performance? (easy-going, confident, reflective, …) [41]
  - Human cognitive biases are reflected in LLM behaviours: conformity in collaborative scenarios, impacting their accuracy [42].

[40] Liu, Zizhou, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R. Greene and Julia Hirschberg. “The Mind in the Machine: A Survey of Incorporating Psychological Theories in LLMs.” ArXiv abs/2505.00003 (2025): n. pag.
[41] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.
[42] Weng, Zhiyuan, Guikun Chen, and Wenguan Wang. "Do as We Do, Not as You Think: the Conformity of Large Language Models." In The Thirteenth International Conference on Learning Representations.

Anthropomorphization of LLMs

- Multiple studies show that human-like qualities are increasingly being attributed to LLMs:
  - “Third-party evaluators perceive AI as more compassionate than expert humans” [43]
  - “Large Language Models Produce Responses Perceived to be Empathic” [44]
  - “Exploring relationship development with social chatbots: A mixed-method study of replika.” [45]
  - “It happened to be the perfect thing”: experiences of generative AI chatbots for mental health” [46]

- Should LLMs be used in sensitive domains like healthcare and personal advice?
  - Empirical evidence shows benefits as well as safety risks [47,48]:
    - LLMs can be steered towards helpful traits like displaying empathy, cultural awareness and sensitivity, factual grounding, and balancing diverse perspectives.
    - As all computational systems, they can fail in predictable or surprising ways: bias, sycophancy, hallucinations, and security vulnerabilities.
  - No straight answer – just like human relationships, human-AI relationships can be healthy or unhealthy.
  - Alignment of AI systems with human values needs to be bi-directional: humans are susceptible to emotional control, manipulation, deception, cognitive dependence, social isolation, etc.[49]

[43] Ovsyannikova, Dariya, Victoria Oldemburgo De Mello and Michael Inzlicht. “Third-party evaluators perceive AI as more compassionate than expert humans.” Communications Psychology 3 (2025): n. pag.

[44] Lee, Yoon Kyung, Jina Suh, Hongli Zhan, Junyi Jessy Li and Desmond C. Ong. “Large Language Models Produce Responses Perceived to be Empathic.” 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII) (2024): 63-71.

[45] Pentina, Iryna, Tyler Hancock, and Tianling Xie. "Exploring relationship development with social chatbots: A mixed-method study of replika." Computers in Human Behavior 140 (2023): 107600.

[46] Siddals, Steve, John B Torous and Astrid Coxon. ““It happened to be the perfect thing”: experiences of generative AI chatbots for mental health.” NPJ Mental Health Research 3 (2024): n. pag.

[47] Zhang, Yutong, Dora Zhao, Jeffrey T. Hancock, Robert Kraut and Diyi Yang. “The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being.” ArXiv abs/2506.12605 (2025): n. pag.

[48] Zhang, Renwen, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan and Yi-Chieh Lee. “The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships.” Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (2024): n. pag.

[49] Gabriel, Iason et al. “The Ethics of Advanced AI Assistants.” ArXiv abs/2404.16244 (2024): n. pag.

Human-Machine Relationships

Published in 2005, [50] discusses the psychological aspects of human-machine relationships:

- "Maintaining relationships involves managing expectations, attitudes and intentions, ….Relationships are also fundamentally social and emotional;"
- Relational agents: "We define relational agents as computational artifacts designed to build long-term, social-emotional relationships with their users."
- The idea of *relational agents* is not limited to current AI systems; humans can form attachments with " … a number of embodiments: jewelry, clothing, handheld, robotic, and other nonhumanoid physical or nonphysical forms."

- Social psychology has a rich literature of study on the psychology of social relationships, and their trajectories over time:
  - Dimensional model: axes of power, social distance | equal vs unequal, superficial vs intense, ….| trust

- Empirical study of a long-term relational agent for changing health behaviours:
  - longitudinal self-report data collected over six-weeks for relational, non-relational, and control groups.
  - Relational agent was perceived to be more helpful, but did not translate to differences in exercise behaviour.

[50] Bickmore, Timothy W. and Rosalind W. Picard. "Establishing and maintaining long-term human-computer relationships." ACM Trans. Comput. Hum. Interact. 12 (2005): 293-327.

Socio-affective Alignment for AI Systems

[51] discuss human-machine relationships in the context of current AI agents:

- AI Alignment: the process of formally encoding values or principles in AI systems.
  - Technical challenge: how to encode values
  - Social challenge: what values to encode
    - Macro-level: socio-technical challenges (influence of and on social structures)
    - Micro-level: socio-affective challenges (influence of and on individuals)

- Human values and preferences are influenced by social rewards and interactions:
  - AI systems are now becoming part of this social landscape, perceived as independent agents rather than mediators (unlike other technologies like the internet or The Algorithm)
  - Which relationships will we prioritize?
  - How will our moral perceptions and judgments change?
  - How will changing human preferences feed back into AI systems?

- Identify three key dilemmas: short-term vs long-term goals; boundaries of autonomy; balancing human and AI relationships.

P.S: Who determines alignment goals of AI systems? Corporations, stakeholders, governments, regulatory bodies, …

[51] Kirk, Hannah Rose, Iason Gabriel, Chris Summerfield, Bertie Vidgen and Scott A. Hale. "Why human–AI relationships need socioaffective alignment." Humanities and Social Sciences Communications 12 (2025): n. pag.

Wrapping Up

…

Summary

- Affect is a fundamental aspect of the human experience, and along with Cognition, shapes how we perceive and interact with the world around us.
- Our language use is a window into our affective states: word usage itself can tell us about how our emotional states and feelings towards others.
- Computational models of affective language use offer insights into the fundamental nature of affect and its interaction with the world around us:
  - mind-body: Emotion dynamics, granularity, relation to mind and body.
  - mind-world: Narratives, social dynamics, AI technologies.
- AI systems do not have internal affective states, but:
  - Models rely on patterns of human language use, which reflect affective biases and behaviour.
  - AI systems are becoming a part of our communicative landscape: models need to understand and shape affective dynamics effectively and ethically.

The 1st Workshop on Computational Affective Science

Co-located with LREC 2026 in Palma de Mallorca, 11-16 May.

https://casworkshop.github.io

Computational Affective Science

A Window into emotions, mind, body, health, and behavior through language and computation

Slides, Papers, Datasets, Lexicons, Code
Available at: www.saifmohammad.com/WebPages/CAS-tutorial.html

vk22priya@gmail.com, saif.mohammad@nrc-cnrc.gc.ca, uvgotsaif@gmail.com

@krishnapriyaVi5, @SaifMMohammad