JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Affect and Lexicons

Recap

- Word embeddings as methodology for corpus analysis
  - We often use embeddings to compute relations between sets of words:
    - {Woman, she, her, gal, girl}
    - {nurse, secretary, teacher}
  - Dimensions of beliefs
    - Gender, potency (power)
- Where do these words come from? What other types of word annotations are useful?

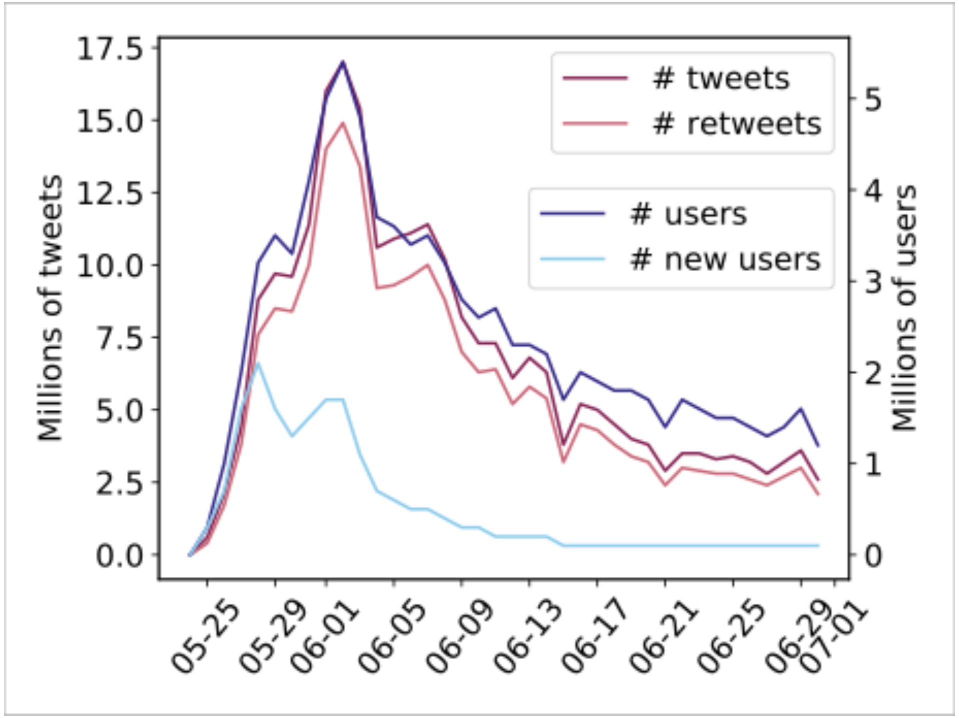
This class

- Psychology measures of affect and emotion
- Common lexicons, construction and uses
- [Data annotation]

Analysis Data: 34M tweets about the #BlackLivesMatter Movement

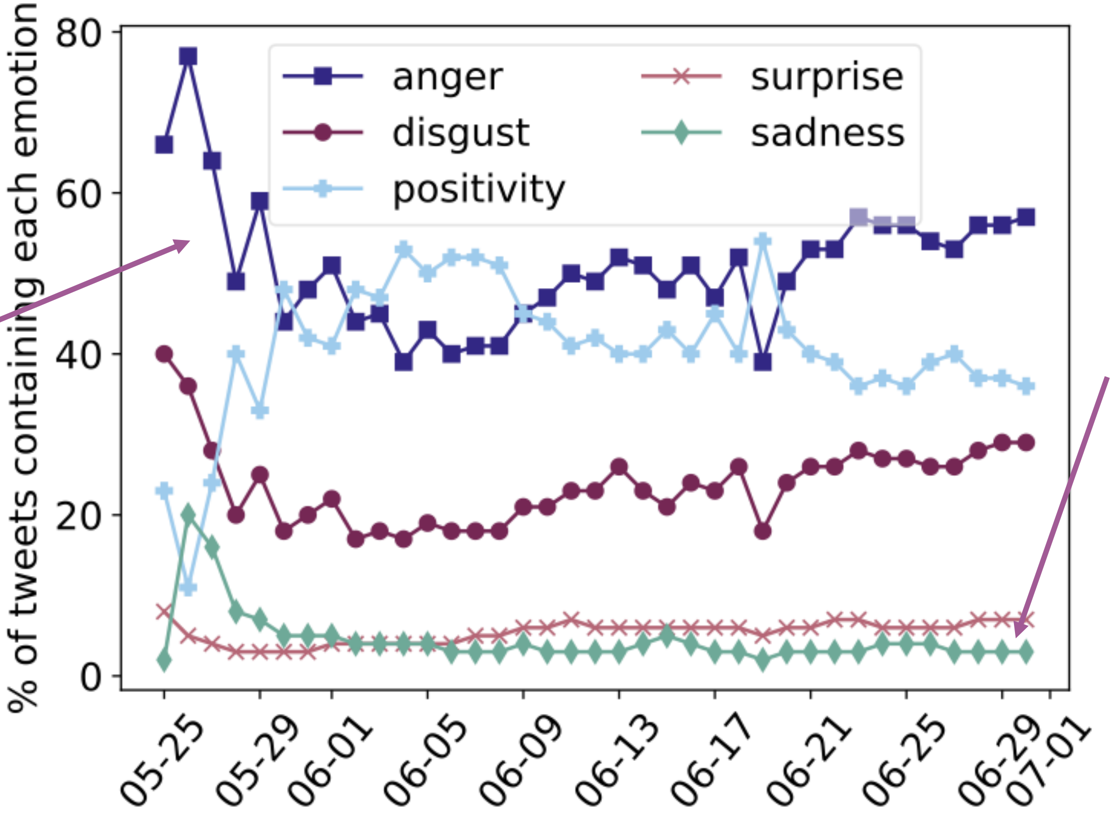
The term #BlackLivesMatter originated in posts made by activists Alicia Garza and Patrisse Cullors in 2013

#BlackLivesMatter
#JusticeForGeorgeFloyd
#ICantBreathe



Field et al. "An Analysis of Emotions and the Prominence of Positivity in #BlackLivesMatter Tweets" PNAS (2022)

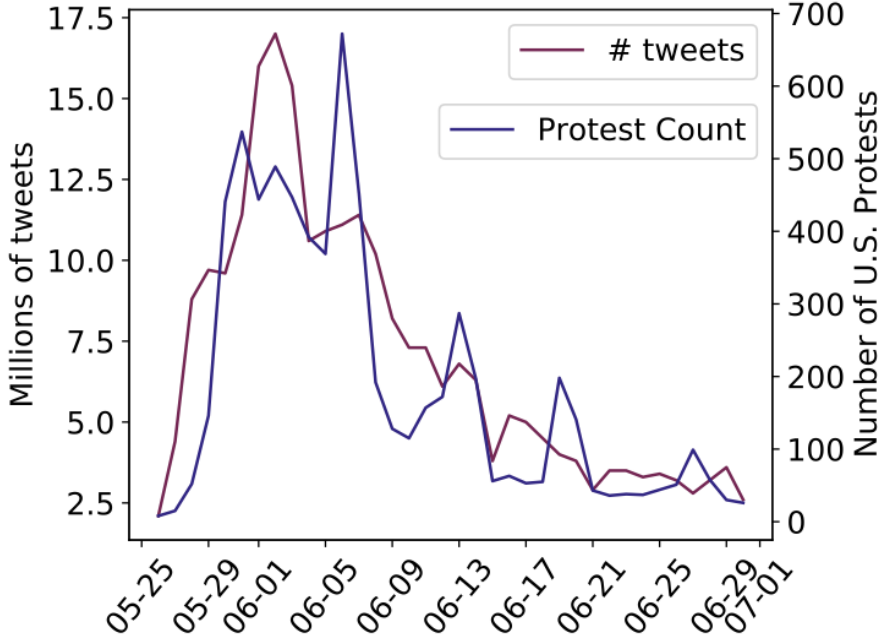
Emotions over time in tweets with pro-BLM hashtags



Initial peak in anger, eclipsed by more positive emotions

Low sadness and surprise?

Positivity is correlated with in-person protests



	Correlation with protest across states	Correlation with protests across cities
Anger	-0.43*	-0.16*
Disgust	-0.24	-0.21*
Positivity	0.48*	0.12*
Sadness	-0.38*	0.06
Surprise	-0.25	0.09

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

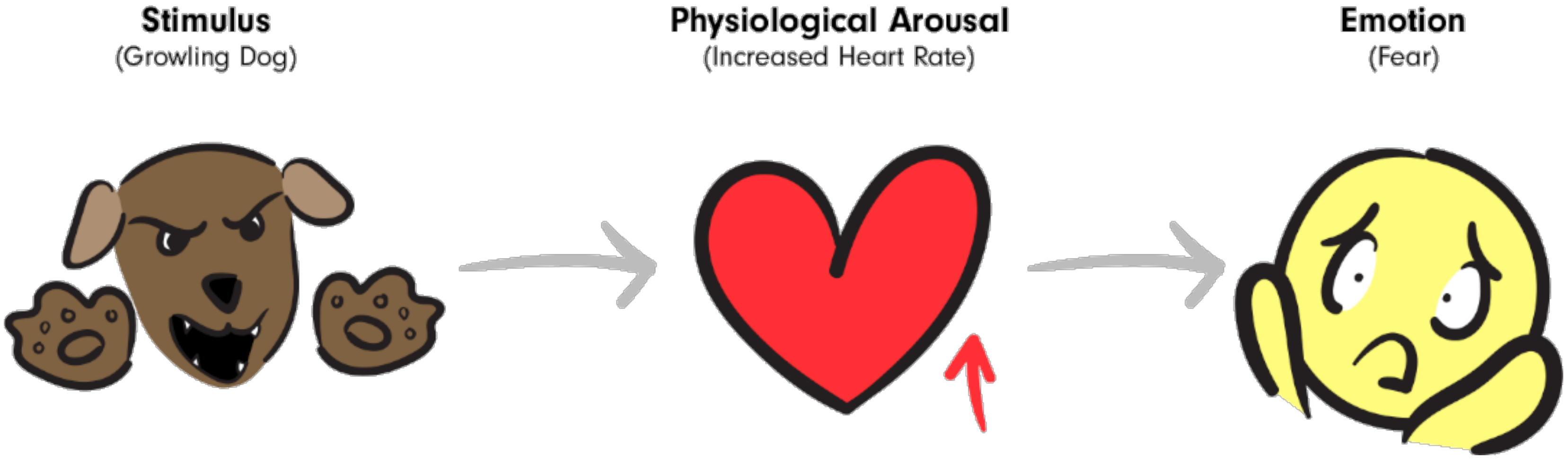
Emotion Taxonomies

What is an emotion?

- **Emotions** are a mix of
  - (1) physiological arousal (heart pounding)
  - (2) expressive behaviors (quickened pace),
  - (3) consciously experienced thoughts (is this a kidnapping?) and feelings (a sense of fear, and later joy)
- The puzzle for psychologists has been figuring out how these three pieces fit together

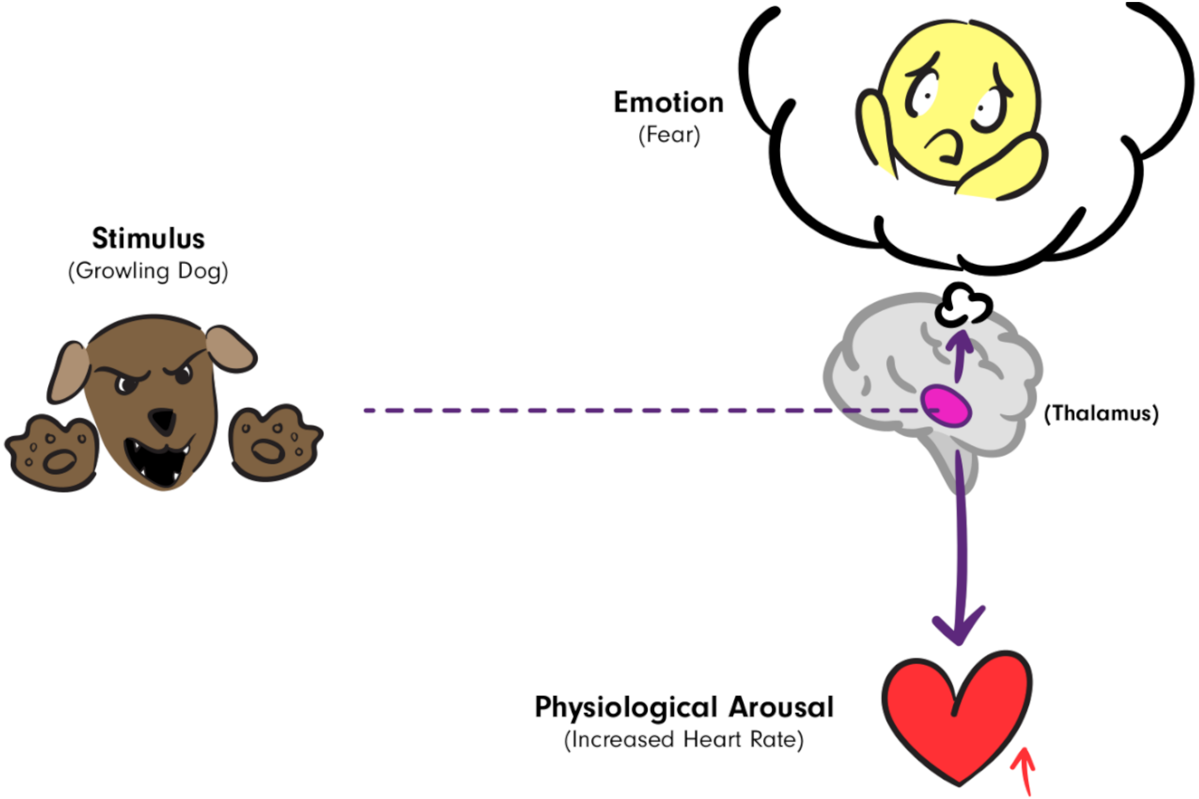
Myers, David G. *Psychology, in modules*. Macmillan, 2004

James-Lange Theory



https://pixorize.com/view/5126

Cannon-Bard Theory



https://pixorize.com/view/5133

Discrete Emotion Theory

- All humans have innate set of basic emotions that are cross-culturally recognizable
- “Discrete”: emotions are separate and distinct
- Distinguishable by neural, physiological, behavioral and expressive features
- A little historical context:
    - Darwin (1872) described “several facial, physiological and behavioral processes associated with different emotions in humans as well as animals”
    - Tomkins (1962, 1963) proposed 8 “pancultural affect programs”: surprise, interest, joy, rage, fear, disgust, shame and anguish

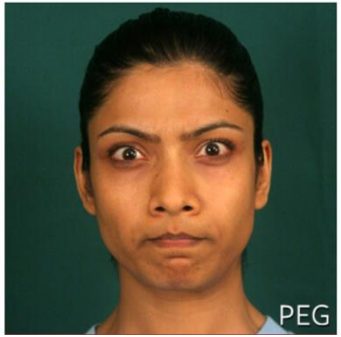
Giovanna Colombetti (2009) From affect programs to dynamical discrete emotions, Philosophical Psychology, 22:4, 407-425, DOI: 10.1080/09515080903153600

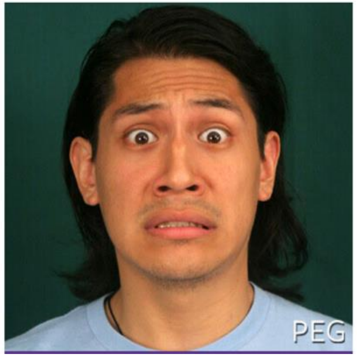
Paul Ekman and Carroll Izard Taxonomy

- “I and others found evidence…that certain facial expressions of emotion appeared to be universal”
- Example field experiments:
  - Show stress-inducing films to students in the US and Japan → Japanese and American students had virtually identical facial expressions
  - Show photographs of different emotion expressions to people in US, Japan, Chile, Argentina, and Brazil: people judged the same emotions in these countries
- Each basic emotion is a *family* of related states

Ekman, Paul, and Wallace V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Vol. 10. Ishk, 2003.

Paul Ekman’s Taxonomy

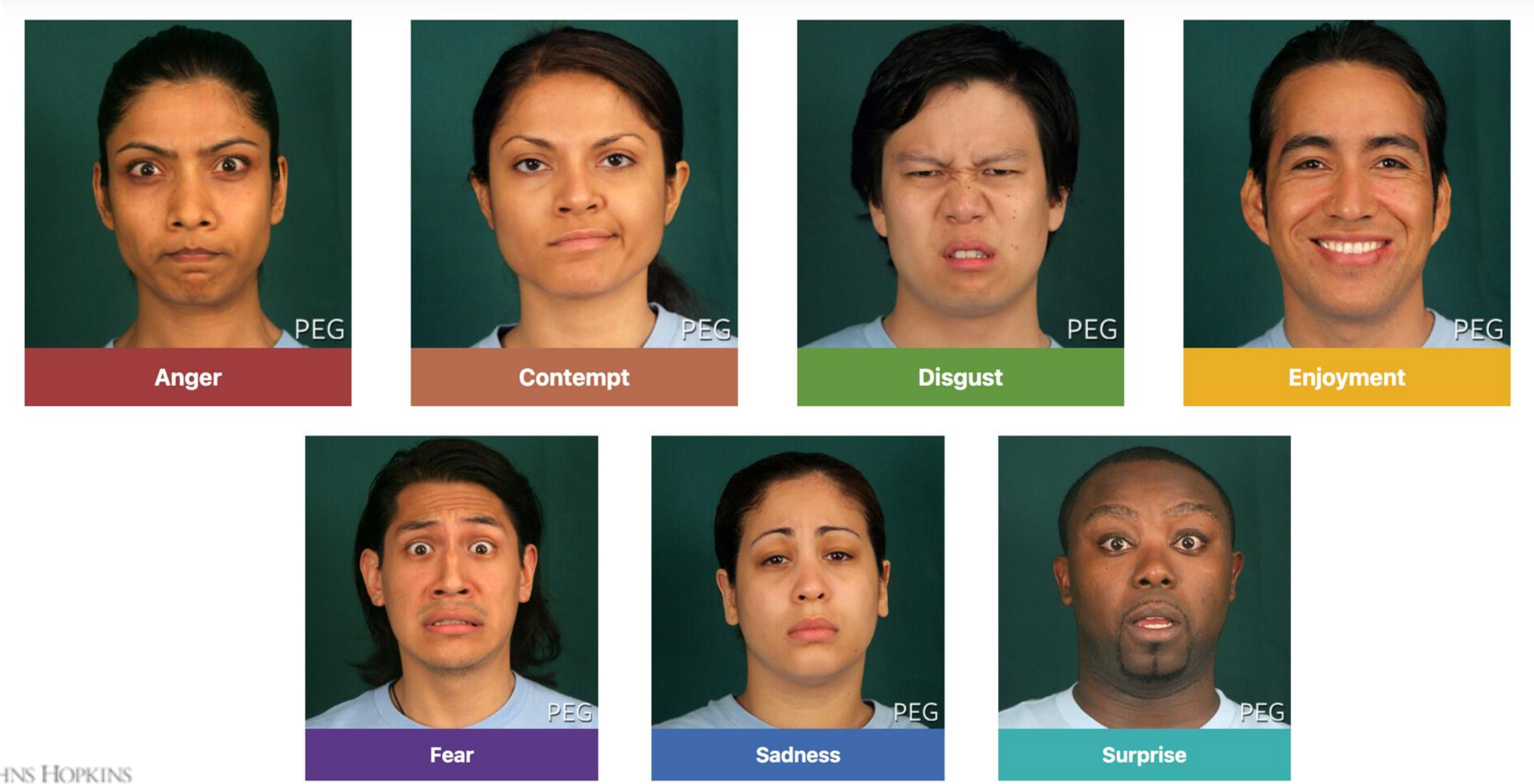




- Sadness
- Anger
- Enjoyment
- Disgust
- Surprise
- Fear
- Contempt

https://www.paulekman.com/universal-emotions/

Paul Ekman’s Taxonomy



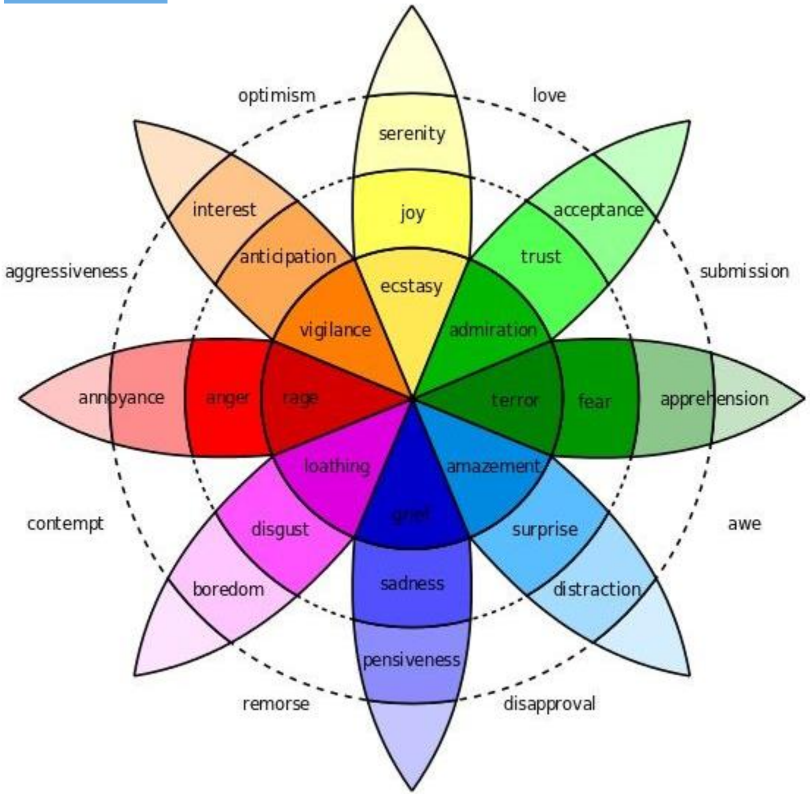
Critiques of Discrete Emotion Theory

- Failure to find correlations between neural and nervous system (ANS) activity and emotions

- Discrete Emotion Theory cannot account for rich variability and context-sensitivity of emotions (Russell and Barrett)
  - Factors other than immediate feeling can affect facial expressions (you may smile out of a desire to please others rather than happiness)
  - Emotions can elicit different responses: flight or fight response to fear
  - Expressions of emotions can differ across cultures

Barrett, 2006a, 2006b, 2006c; Barrett & Russell, 1999; Russell, 1980, 1991, 2003, 2005, 2006; Russell & Barrett, 1999

Plutchnik Emotion Taxonomy (Increasing continuity)



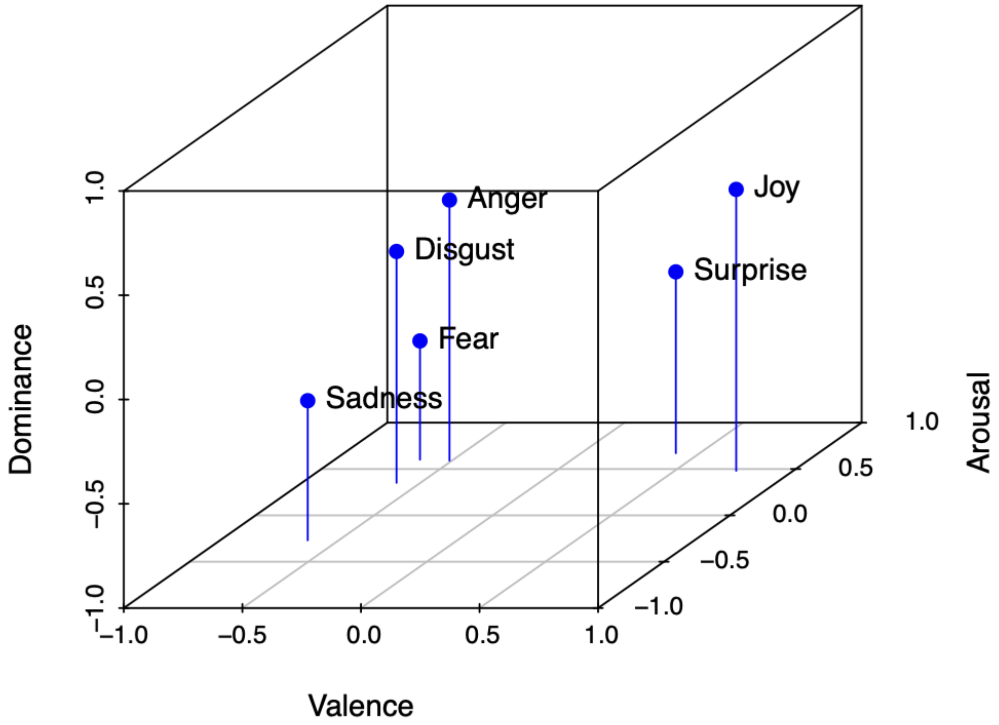
- Still have 8 basic emotions emotions in the center
- Different levels of intensity
- Some emotions are combinations of 8 core emotions
- Interactive demo: https://www.6seconds.org/2022/03/13/plutchik-wheel-emotions/

Alternate view: Continuous representation of affect

- Osgood et al. (1957) asked human participants to rate words along dimensions of opposites such as heavy– light, good–bad, strong–weak
- Factor analysis of these judgments revealed that the three most prominent dimensions of meaning:
  - **Valence**/Evaluation/Sentiment (good–bad)
  - **Dominance**/Power/potency (strong–weak)
  - **Activity**/Arousal/Agency (active–passive)

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11(3):273–294.

Alternate view: Continuous representation of affect



- Emotions can be mapped to these continuous dimensions, rather than being basic discrete categories
- Recall "gender subspace" idea: This seems well-suited to word embeddings? [Sort of works depending on embedding quality Field&Tsvetkov 2019]

Buechel, Sven, and Udo Hahn. "Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation." *Proceedings of the 11th linguistic annotation workshop*. 2017.

A note on ethics of AI for Emotion Detection

- Plethora of work on using AI for emotion detection and existence of commercial products that claim to be able to do so (e.g. based on facial recognition)
- Limited evidence that this possible
  - Distinction between true internal emotional state and outward expression
  - Lack of consensus on what emotions are among pyschologists
- High misuse potential
  - Faulty AI used to make impact decisions in domains like law, education, and employment
- Privacy
  - People generally find “emotion AI” invasive with little benefit for themselves

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1-68. https://doi.org/10.1177/1529100619832930

Pyle, K. Roemmich, and N. Andalibi, “US job-seekers’ organizational justice perceptions of emotion AI-enabled interviews,” Proceedings of the ACM on Human-Computer Interaction, vol. 8, 273 no. CSCW2, pp. 1–42, 2024.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Lexicons: Manual Construction

What are lexicons?

- A collection of words
- Words with labels
- Some popular lexicons:
  - Linguistic Inquiry and Word Count (LIWC): https://www.liwc.app/
  - NRC Emotion Lexicons: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
  - NRC-VAD Lexicon: https://saifmohammad.com/WebPages/nrc-vad.html
  - Connotations frames of power, agency, and sentiment https://github.com/maartensap/riveter-nlp

When are lexicons useful?

- Less ideal use case:
  - Simple classification model (text expresses “anger” if it has a word from an “anger” lexicon)
  - Classifier typically works much better but lexicons are extremely easy to implement (just have to count words) and very interpretable
- More common use cases:
  - Pre-filtering data
    - (e.g. hate speech has low prevalence in randomly sampled social media posts but we can use lexicons of offensive terms to identify what to annotate)
  - Data collection
    - Tweets or news articles that mention particular events
  - Testing robustness/bias, defining meaningful subsets or axes on a scale (think word embedding metrics)

LIWC

- Transparent text analysis program that counts words in “psychologically meaningful categories”
- Origins and motivation:
  - Words that people use are reflective of internal state, hidden intentions, psychological state
  - Walter Weintraub (1981, 1989) hand-counted people’s words in texts (political speeches, medical interviews, etc.) and noticed that first-person singular pronouns (e.g., I, me, my) were reliably linked to people’s levels of depression

LIWC Categories

- 80(+?) categories:
  - Straightforward language dimensions: articles, pronouns
  - More subjective dimensions: emotions, power,
  - Hierarchy of dictionaries:
    - “Anger” dictionary is a subset of “emotion” dictionary

https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf
Tausczik, Yla R., and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of language and social psychology* 29.1 (2010): 24-54.

LIWC Construction

1. Word Collection: ”judges brainstorm words for each category” (later versions of LIWC use earlier versions as starting point)
2. Judge Rating phase: 3-4 judges rate “goodness of fit” for each word for each category
3. Base Rate Analysis: Examine word frequency across corpora and remove infrequent words
4. Candidate Word Generation: Examine most frequent words in corpora and determine if they should be added to the Dictionary
5. Psychometric Evaluations: compute internal consistency statistics for each category and manually judge if words “detrimental to the internal consistency” should be omitted (judgements made by the 4 authors)
6. Refinement: Repeat steps 1-5 and check for mistakes
7. Addition of summary variables: add in categories that are summaries of others (e.g. emotional tone)

LIWC Takeaways

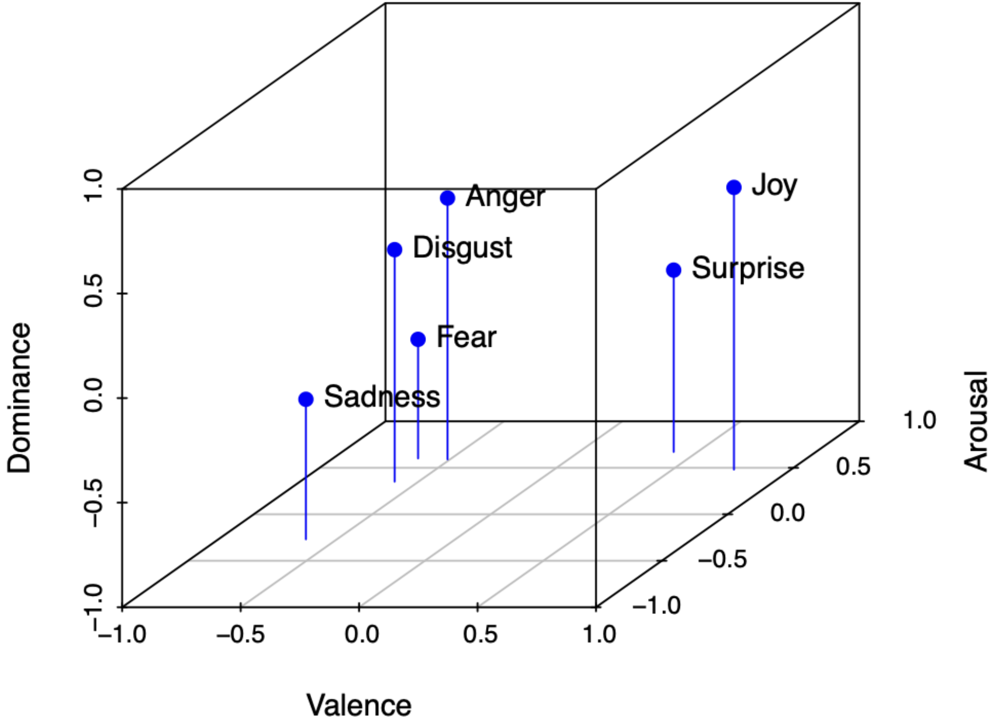
- Really popular resource:
  - Often preferred by social scientists because it was developed by psychologists
  - Commercial easy-to-use software where you can just upload texts and get scores

- Example of data set construction:
  - Relies on domain expertise, judgements of authors and domain experts (not just outsourcing to crowd workers)
  - Iterative process

- Often misused in scenarios it was not designed or evaluated for

Different Annotation Approach: VAD Lexicons

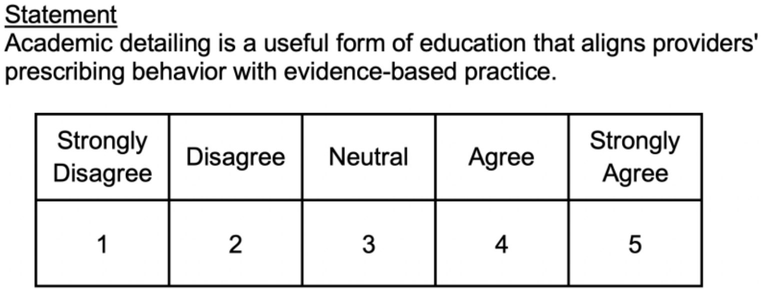
- LIWC defines discrete categories
- We might want more continuous ratings:
  - Is "annoyed" word associated with "anger"? {0, 1}
  - *How* associated is "annoyed" with "anger"? [0, 1]



Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Saif M. Mohammad. ACL 2018.

Different Annotation Approach: VAD Lexicons

- Likert Rating scale:



Statement

Academic detailing is a useful form of education that aligns providers' prescribing behavior with evidence-based practice.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

- Problems:
  - Fixed granularity
  - Difficult to maintain consistency across annotators
  - Difficult for an annotator to be self-consistent
  - Scale region bias

https://mbounthavong.com/blog/2019/5/16/communicating-data-effectively-with-data-visualization-part-15-divergent-stacked-bar-chart-for-likert-scales

Best-Worst Scaling

Out of these four words (A, B, C, and D):

Which word is associated with the most/highest valence?

Which word is associated with the least/lowest valence?

- By answering just these two questions, five out of the six inequalities are known:
  - Example: If A: highest valence and D: lowest valence
    - We know: A > B, A > C, A > D, B > D, C > D

Louviere & Woodworth, 1990

Best-Worst Scaling

$$score(w) = \frac{\#best(w) - \#worst(w)}{\#annotations(w)}$$

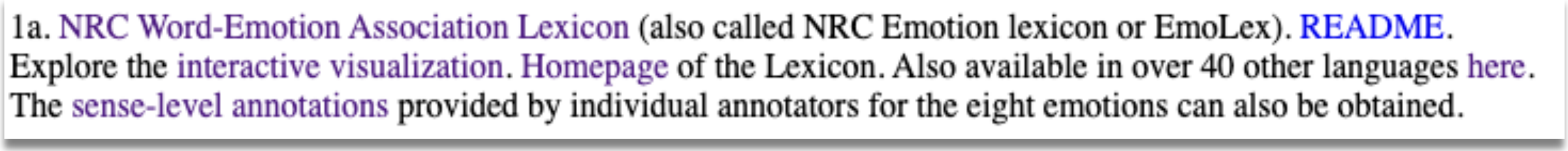
- Scores range from -1 to 1
- Empirically shown that three annotations each for 2N distinct 4-tuples is sufficient for obtaining reliable scores (where N is the number of items)

(Louviere, 1991; Kiritchenko and Mohammad, 2016; Kiritchenko and Mohammad 2017; Orme 2009)

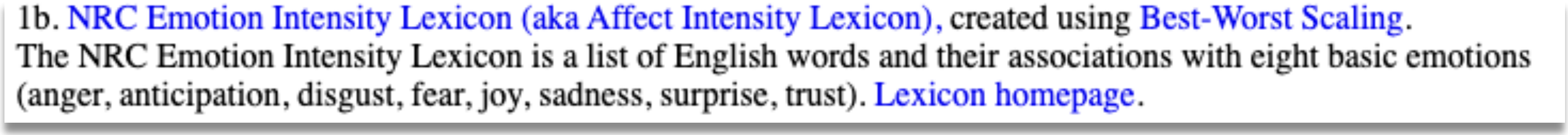
Score reliability: *split-half reliability (SHR)*

- Split all annotations (e.g. for each 4-tuple) into two halves
- Produce two sets of scores independently from the two halves
- Calculate correlation between the two sets of scores. If the annotations are of good quality, then the correlation between the two halves will be high.
- [Repeat for many, e.g. 100 trials]

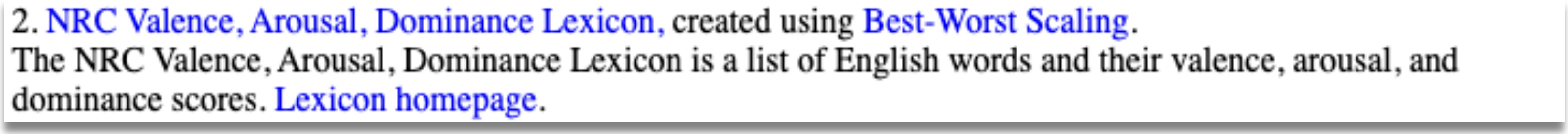
NRC lexicons



1a. NRC Word-Emotion Association Lexicon (also called NRC Emotion lexicon or EmoLex). README. Explore the interactive visualization. Homepage of the Lexicon. Also available in over 40 other languages here. The sense-level annotations provided by individual annotators for the eight emotions can also be obtained.



1b. NRC Emotion Intensity Lexicon (aka Affect Intensity Lexicon), created using Best-Worst Scaling. The NRC Emotion Intensity Lexicon is a list of English words and their associations with eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). Lexicon homepage.



2. NRC Valence, Arousal, Dominance Lexicon, created using Best-Worst Scaling. The NRC Valence, Arousal, Dominance Lexicon is a list of English words and their valence, arousal, and dominance scores. Lexicon homepage.

https://saifmohammad.com/WebPages/lexicons.html

Mohammad, Saif M. "NRC VAD Lexicon v2: Norms for valence, arousal, and dominance for over 55k English terms." *arXiv preprint arXiv:2503.23547* (2025).

Connotation Frames of Power, Agency, and Sentiment

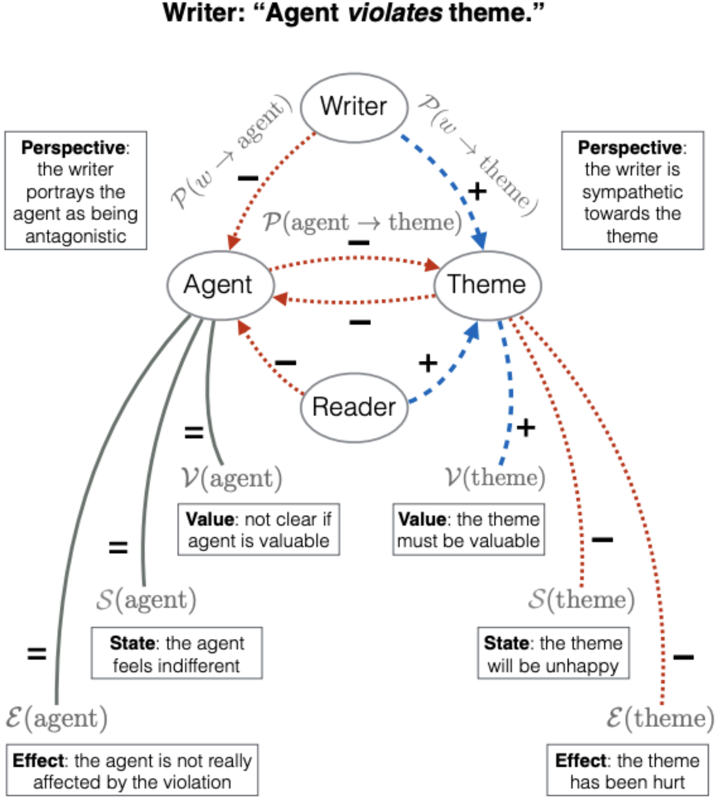
- Lexicon labels can be discrete or continuous, but they can also be directed
- Connotation frames are a formalism for analyzing subjective roles and relationships implied by a given predicate

"X dismisses Y"

  - Writer's Perspective: the writer treats Y more sympathetically but thinks of X as more of an antagonist
  - Reader's Perspective: the reader will likely feel sympathetic towards Y and think more poorly of X
  - X and Y's Mental State: X may feel indifferent. Y will feel distressed
  - X and Y's Perspective, X and Y's Value, Effect on X and Y

Rashkin, Hannah, Sameer Singh, and Yejin Choi. "Connotation Frames: A Data-Driven Investigation." *ACL*. 2016.
Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. EMNLP. 2017

Connotation Frames of Power, Agency, and Sentiment



Writer: "Agent *violates* theme."
Writer
$\mathcal{P}(w \rightarrow \text{agent})$
–
$\mathcal{P}(w \rightarrow \text{theme})$
+
Perspective: the writer portrays the agent as being antagonistic
Perspective: the writer is sympathetic towards the theme
$\mathcal{P}(\text{agent} \rightarrow \text{theme})$
–
Agent
Theme
–
–
+
Reader
=
+
$\mathcal{V}(\text{agent})$
$\mathcal{V}(\text{theme})$
Value: not clear if agent is valuable
Value: the theme must be valuable
=
–
$\mathcal{S}(\text{agent})$
$\mathcal{S}(\text{theme})$
State: the agent feels indifferent
State: the theme will be unhappy
=
–
$\mathcal{E}(\text{agent})$
$\mathcal{E}(\text{theme})$
Effect: the agent is not really affected by the violation
Effect: the theme has been hurt

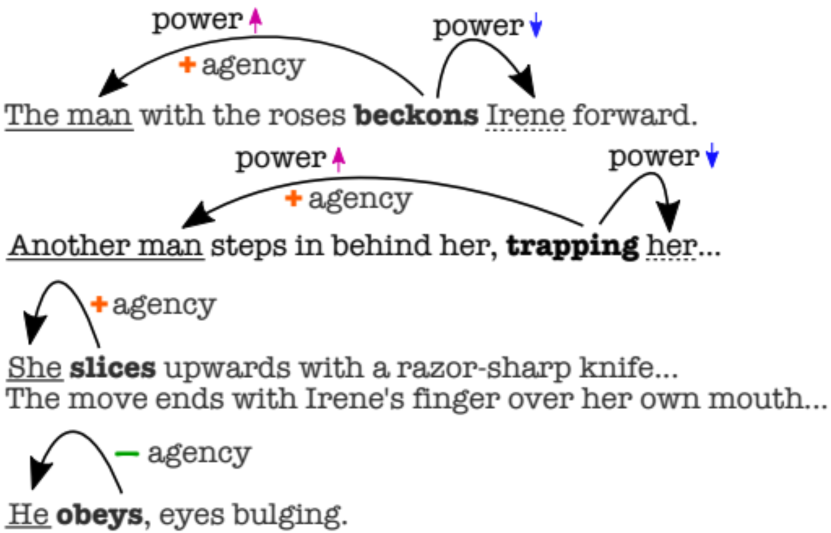
JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Alternate view: Continuous representation of affect

- Osgood et al. (1957) asked human participants to rate words along dimensions of opposites such as heavy– light, good–bad, strong–weak
- Factor analysis of these judgments revealed that the three most prominent dimensions of meaning:
  - **Valence**/Evaluation/Sentiment (good–bad)
  - **Dominance**/Power/Potency (strong–weak)
  - **Activity**/Arousal/Agency (active–passive)

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11(3):273–294.

Connotation Frames of Power, Agency, and Sentiment



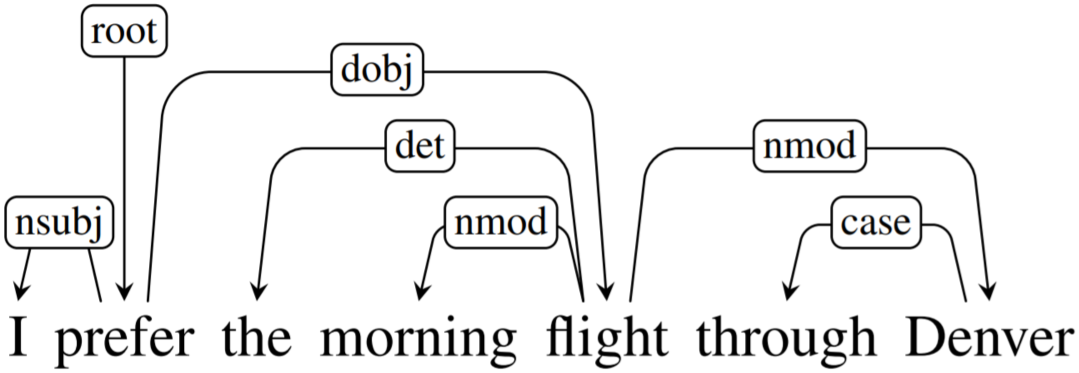
power ↑
+ agency
power ↓
The man with the roses beckons Irene forward.
power ↑
+ agency
power ↓
Another man steps in behind her, trapping her...
+ agency
She slices upwards with a razor-sharp knife...
The move ends with Irene's finger over her own mouth...
– agency
He obeys, eyes bulging.

Unlike agency, power is considered to be relative: one entity has power over the other

Connotation Frames of Power, Agency, and Sentiment

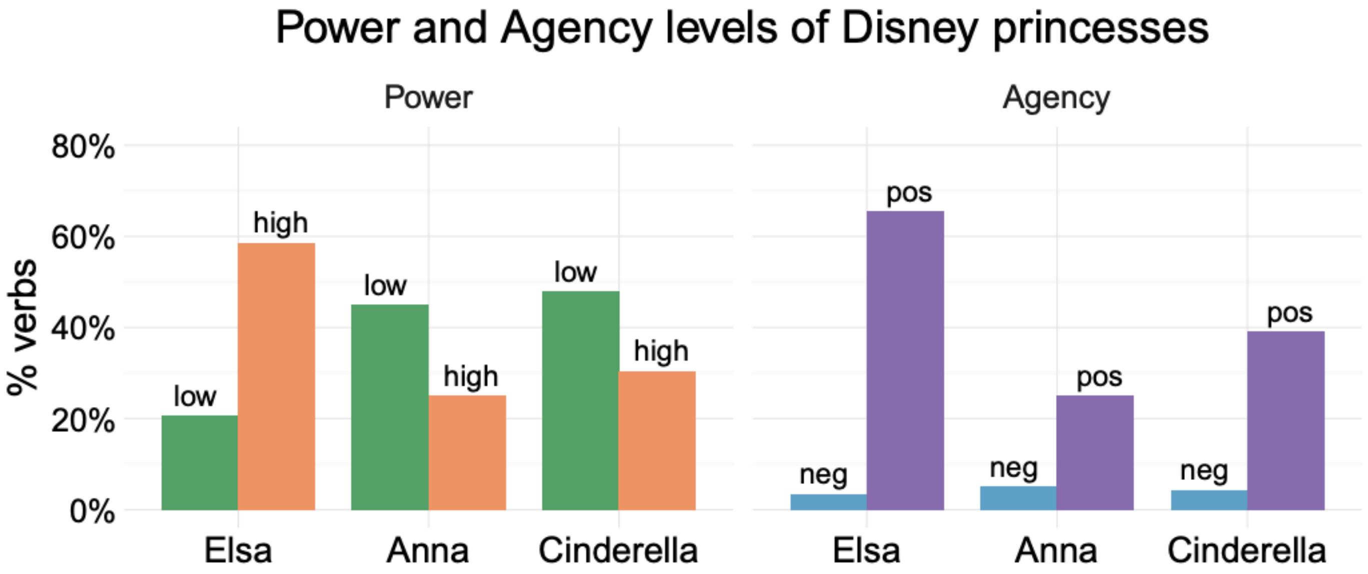
- We have verbs annotated for sentiment, power and agency - how do we use them?
  - We can't just count verbs – we need to resolve agent/theme or subject/object

- Dependency parsing (alternative: semantic role labeling):

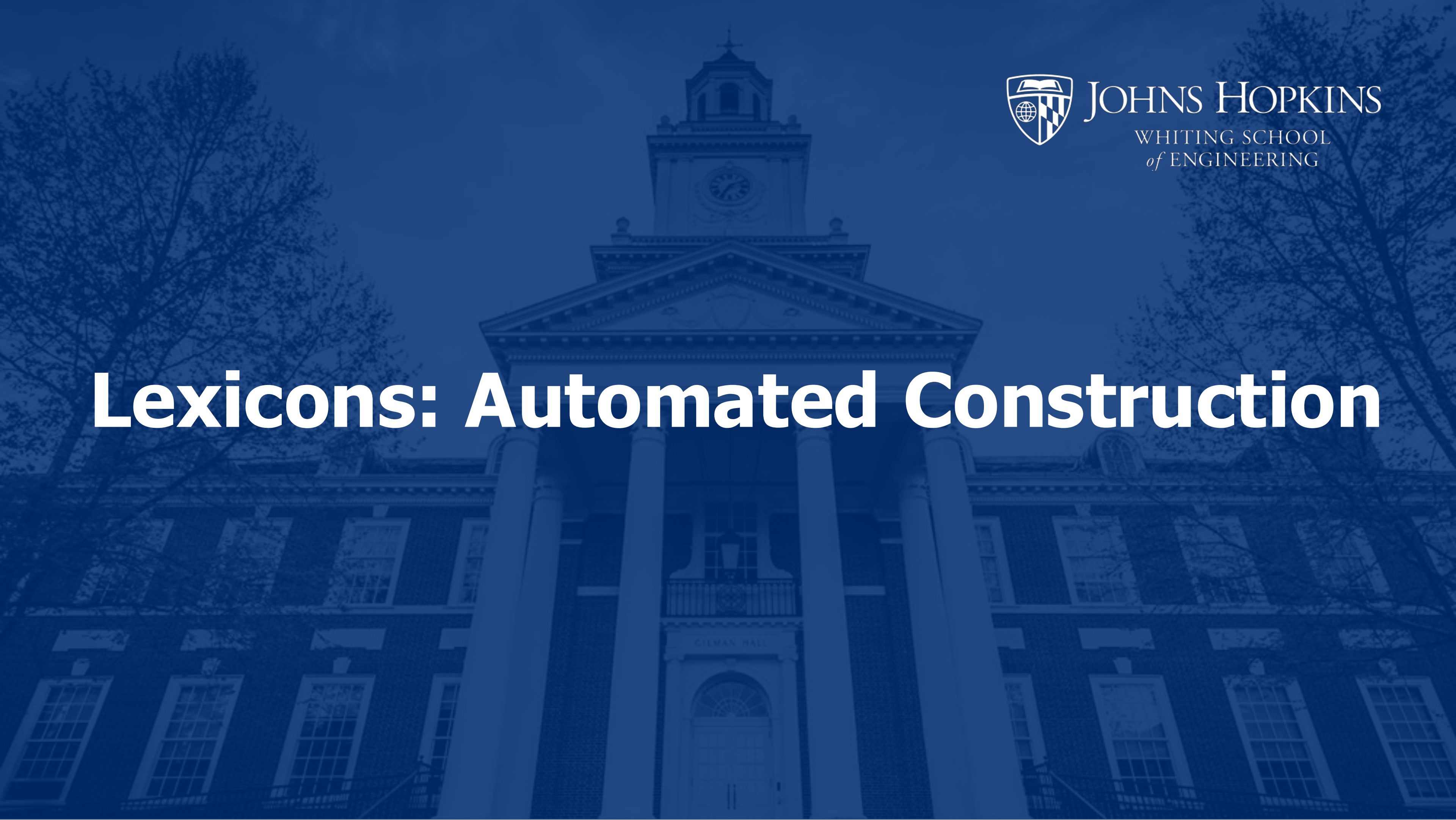


https://www.analyticsvidhya.com/blog/2021/12/dependency-parsing-in-natural-language-processing-with-examples/

Connotation frames: Movie Analysis



Available in python package: https://github.com/maartensap/riveter-nlp



Johns Hopkins
Whiting School
of Engineering

Lexicons: Automated Construction

Inducing Domain-specific lexicons

- A word’s sentiment (or connotation or emotion) depends on the domain in which it is used
  - Words can change meaning over time
  - Connotations can be domain-specific: NRC lexicons associate “police” with “trust”
    - Not the association you would expect in a social movement about police brutality

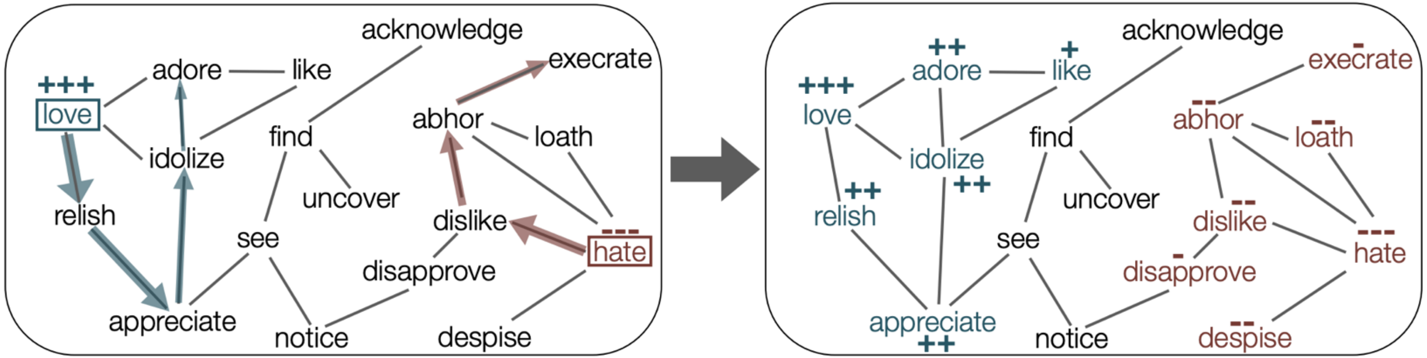
- What can we do about this?
  - Annotate a new lexicon for every domain of interest? → Time consuming and expensive

Hamilton, William L., et al. "Inducing domain-specific sentiment lexicons from unlabeled corpora." EMNLP. Vol. 2016. NIH Public Access, 2016.

SentProp: Algorithm for Domain-specific sentiment lexicons

- Starting point: small seed set of negative and positive words (e.g. ~10 each)
- Construct word embeddings (they use matrix-decomposition approach)
- Construct a graph representation
    - Words are nodes
    - Edges are between each node’s k-nearest neighbors (based on embedding similarity)
    - Run a random walk (with transition matrix defined by edges)
    - Polarity scores are based on random walk visits

SentProp: Algorithm for Domain-specific sentiment lexicons



a. Run random walks from seed words.

b. Assign polarity scores based on frequency of random walk visits.

Figure 3: Visual summary of the SENTPROP algorithm.

- Evaluation: recreating existing lexicons

Example differing domain-induced lexicons: two subreddits

"big men are very soft"

"freakin raging animal"

"went from the ladies tees"

"two dogs fighting"

"being able to hit"

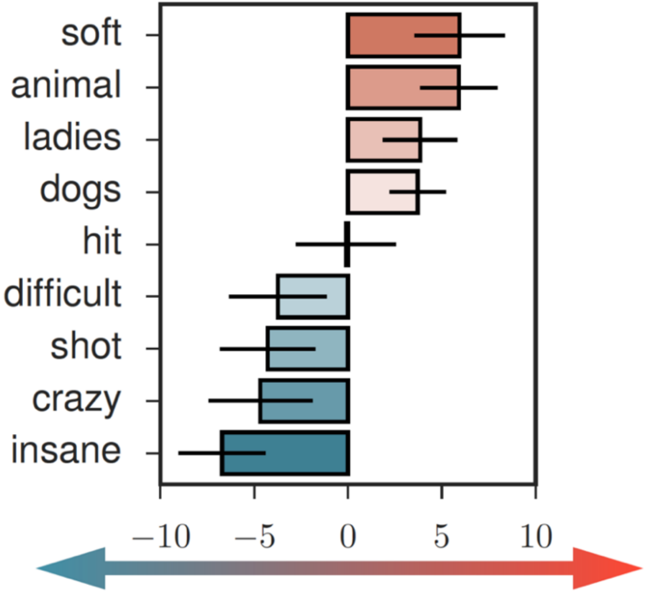
"insanely difficult saves"

"amazing shot"

"he is still crazy good"

"his stats are insane"

Ex. contexts in r/sports



more positive in r/sports,
more negative in r/TwoX

more positive in r/TwoX,
more negative in r/sports

"some soft pajamas"

"stuffed animal"

"lovely ladies"

"hiking with the dogs"

"it didn't really hit me"

"a difficult time"

"totally shot me down"

"overreacting crazy woman"

"people are just insane"

Ex. contexts in r/TwoX

Alternative approaches to lexicon induction

- Word co-occurrence PMI scores (Turney and Littman, 2003)
- Variants of the propagation approach or embedding construction (Velikovich et al. 2010)
- DENSIFIER (Rothe et al. 2016): condenses word embeddings into a single dimension

Recap

- Emotions:
  - Different models of emotions in psychology
- Lexicons:
  - Commonly used lexicons
    - LIWC, NRC lexicons, connotation frames
  - When lexicons are useful and when they are not
  - Different was of constructing them
    - Manual vs. automated, categorical vs. continuous, directed (connotation frames) vs. not
- Data annotating:
  - Likert scale, Best-worst scaling

References

- Giovanna Colombetti (2009) From affect programs to dynamical discrete emotions, Philosophical Psychology, 22:4, 407, DOI: 10.1080/09515080903153600
- Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Saif M. Mohammad. ACL 2018.
- Tausczik, Yla R., and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of language and social psychology* 29.1 (2010): 24-54.
  - https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf