Decoding Morality in Social Media

Ruijie Xi

August 1, 2023

North Carolina State University
Department of Computer Science

NC STATE UNIVERSITY

About Me

Education

- Ph.D. in Computer Science, North Carolina State University, September 2019 to present
  - Cognitive Linguistic Influences on Blame Assignment, qualified January 2022
- M.Sc. in Computer Science, University of Delaware, Newark, DE, May 2017
- B.Sc. in Computer Engineering, Harbin Institute of Technology, Heilongjiang, China, May 2015

Publications

- The Blame Game: Understanding blame assignment in social media (IEEE Transactions on Computational Social Systems, March, 2023)
- Morality in the Mundane: Categorizing moral reasoning in real-life social situations (International AAAI Conference on Web and Social Media, 2024)

Table of Contents

Introduction

Morality in Social Psychology

Morality evaluates behavior as right or wrong and involves measuring the conformity of a person's actions to a code of conduct

Morality in Social Media

Prior research on morality has primarily been theoretical and lacks systematic exploration of online platforms

Morality in social media

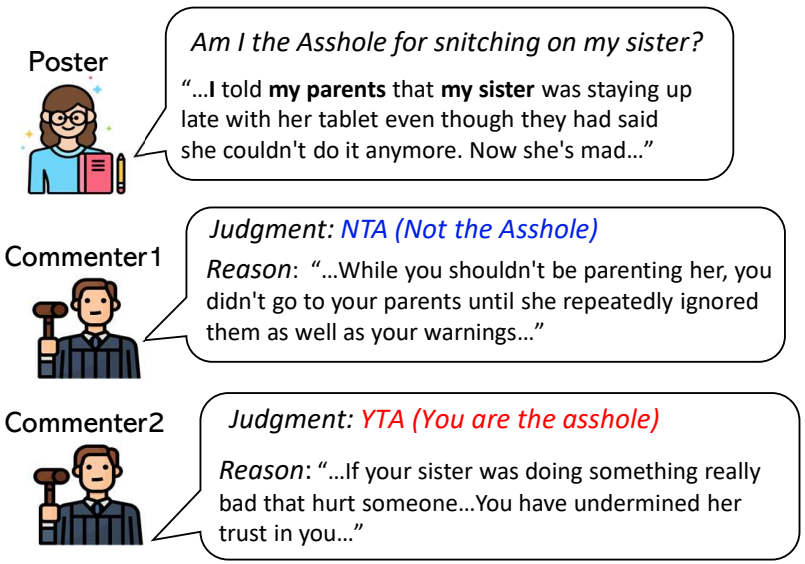
- Uncover real-life ethical behaviors
- Understand morality from language

Our contributions

- Investigate morality in online platforms
  - Linguistic features in posted moral situations and the posts' comments
  - Users' social factors
- Use computational methods to understand morality
  - Statistical analysis
  - Machine Learning and Natural Language Processing

Blame Assignment and Moral Reasoning

Real-life moral situations from r/AmItheAsshole (AITA)



Blame assignment

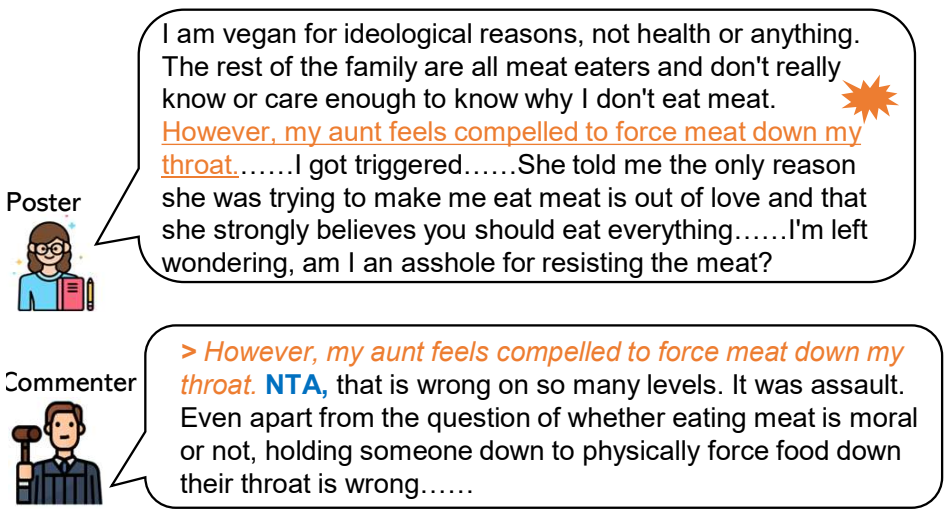
- Investigate the influence of linguistic features and social factors on the attribution of blame

Moral reasoning

- Examine the prevalent components in moral reasoning when assessing the appropriateness of behaviors

Elements that Captivate Moral Attention

Quoted sentences (moral sparks) captivate the commenter's attention



Proposed: Moral sparks

- Uncover the inherent commonsense present in moral situations that captures users' attention

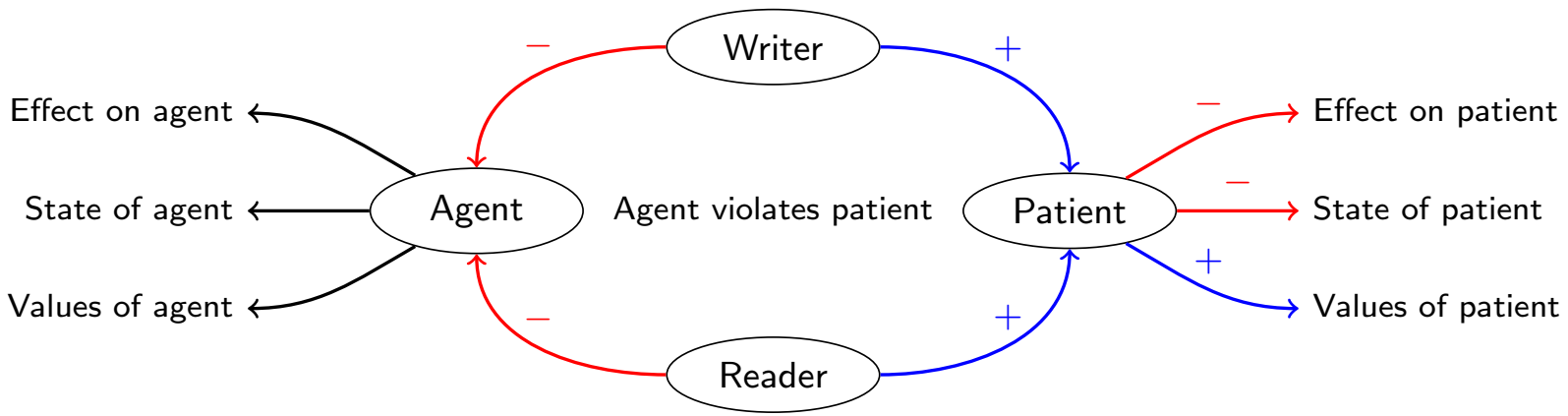
The Blame Game: Understanding Blame Assignment in Social Media

Moral Blame

Norm-violating behaviors in the Blame Theory [4]

- *Agent*: who caused the harm
- *Patient*: who perceived the harm

Assigning blame is a cognitive process that requires individuals to foresee the negative outcomes of agentive behaviors



Agent is powerful

Patient suffers

Portrayals of Entities

Our hypothesis

- The use of attributive and predicative words when the authors describe themselves and others are different

Entity-centric stories in AITA

- Author: 24 years old male
- The author's girlfriend: 23 years old female
- Subjective-verb-objective tuples such as *(i, upset, her) and (she, told, me)*
- Adjective-noun pairs such as *a terrible aunt*

Research Questions

- $RQ_{features}$: What are the cognitive-affective linguistic features in blame assignment?
  - Content (post–level): facts extracted from the posts
  - Tone (post–level): author's attitude towards the post
  - Connotation (entity–level): linguistic indicators, including attributive and predicative words describing agents and patients
- RQ_{social}: What social factors besides linguistic influence blame assignment?
  - Social factors associated with moral development: education, religion, gender, and age
  - Gender stereotypes impair moral evaluations

Content: Sample Topics of Posts

Topic Label	Top Weighted Words
Relationship with family (20.8%)	life, relationship, mother, ex, child, father, life, wife, partner, son
Intimate relationship (17.3%)	girlfriend, boyfriend, relationship, dating, upset, feel, pretty, lot, love, guy
Living in shared accommodation (16.5%)	apartment, rent, live, room, living, house, lease, stay, bedroom
Money (7.3%)	pay, rent, saving, buy, job, account, car, loan, afford, cost
Pregnancy concerns in pets (5.5%)	dog, child, husband, child, pregnant, puppy, cat, law, animal, birth
Work (4.4%)	hour, work, boss, company, manager, job, employee, office, shift, week

Tone: Post–Level Features

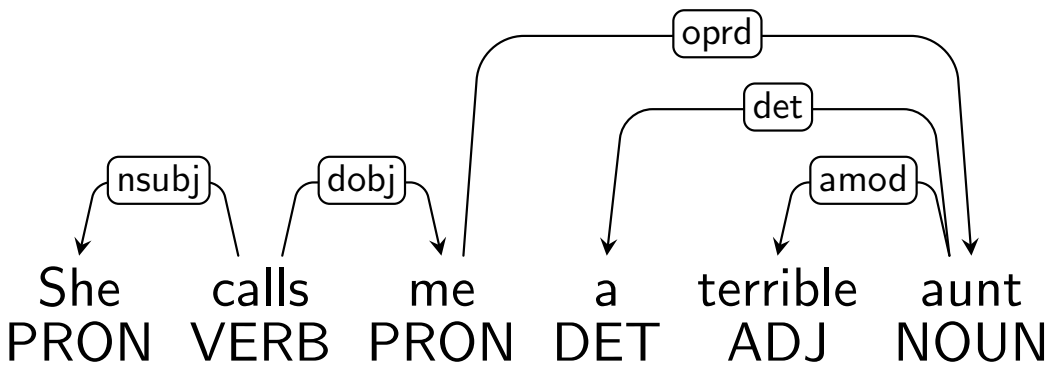
- Subjectivity: count of a subjectivity lexicon and first, second, third-person pronouns
- Modal: percentage of modal words
- Hedge: percentage of hedge words
- Sentiment: standardized VADER compound score of each post and nominal sentiment categories (positive, neutral, and negative)
    - VADER is a lexicon and rule-based sentiment analysis tool that is specifically for analyzing sentiments in social media data

Connotation: Agent and Patient

Semantic Role Labeling (SRL)
Identify the semantic roles of words or phrases within a sentence such as "*who* did *what* to *whom* and *when* and *how* and *why*"

- **They** (`ARG0`) claimed **me** (`ARG1`) a dependent even though **I** (`ARG0`) have been financially independent for about a year

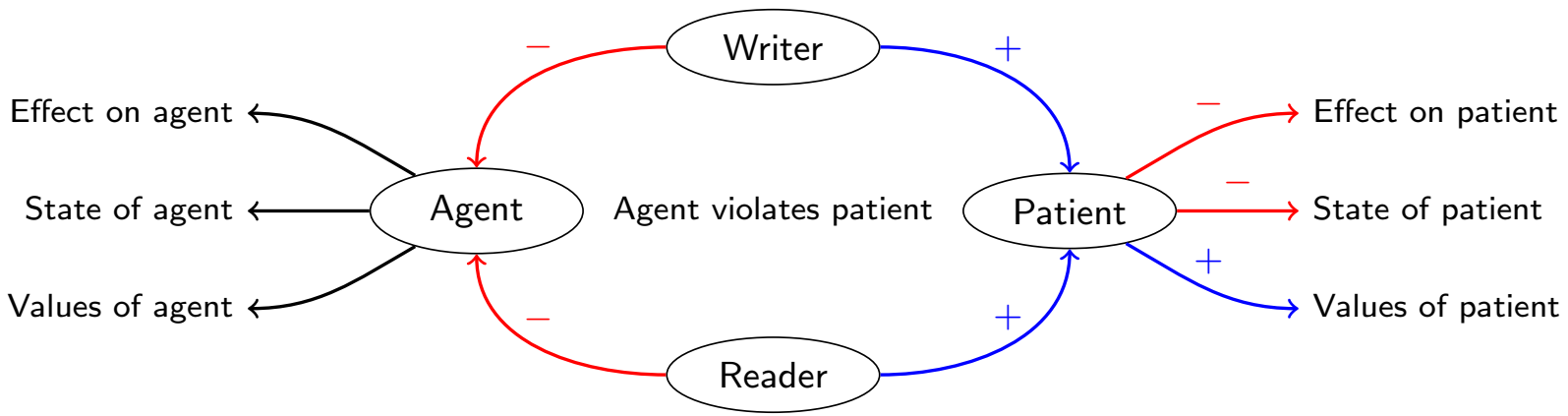
Dependency Parsers
Extract subjective-verb-objective (SVO) tuples and adjective-noun (AN) pairs describing entities using



SVO: *(she, calls, me)*, AN: *(terrible, aunt)*

Connotation: Portrayals

A pragmatic formalism organized to model how different levels of power and agency are implicitly projected on people through their actions



This lexicon contains binary labels of each verb, which are positive (1), equal (0), and negative (–1)

Connotation: Moral Content

Moral Foundation Theory (MFT), a psychological theory proposed by social psychologists, includes five foundational dimensions ranging from virtues (-1) to vices (1): care to harm, fairness to cheating, loyalty to betrayal, (respect for) authority to subversion, and sanctity to degradation

Examples of MFT lexicon (2,041) words

Care virtue	empathy, kindness, caring
Care vice	murder, attacker, destroyer
Fairness virtue	integrity, honesty, objectiveness
Fairness vice	deceived, distrust, betrayers

Connotation: Valence, Arousal, and Dominance

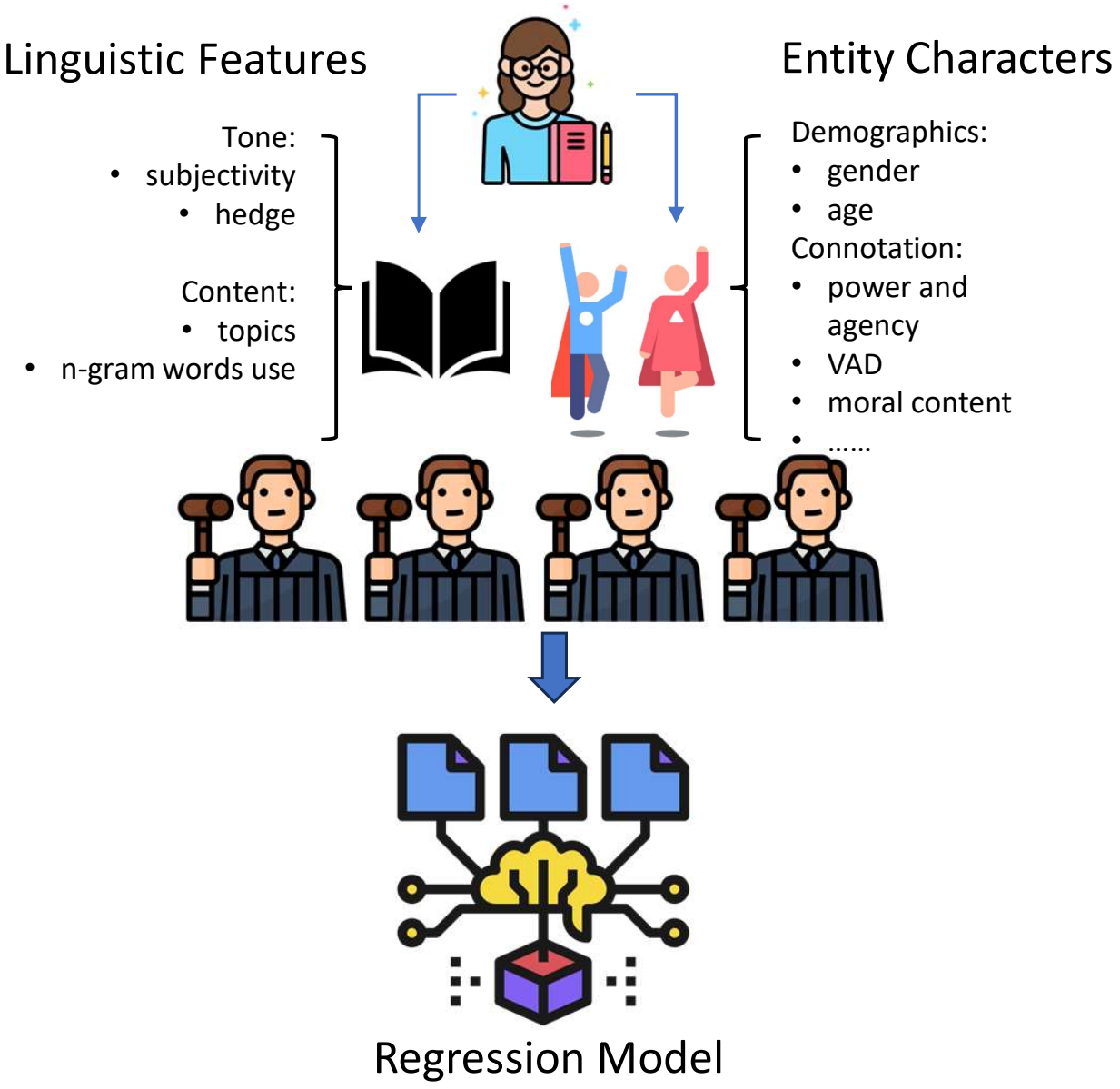
Valence, Arousal, and Dominance (VAD) scores ranging from 0 to 1 for each category

- Valence: pleasantness of word
- Arousal: intensity of emotion provoked by a word
- Dominance: degree of control exerted by a word

Example for *abandon*:

- Valence: 0.052
- Arousal: 0.519
- Dominance: 0.245

Prediction of Blame Assignment



$RQ_{Feature}$: Content, Connotation and Tone are Cognitive-Affective Linguistic Features

Method	F1		Recall		Precision	
	DEV	TEST	DEV	TEST	DEV	TEST
Random	0.49	0.50	0.50	0.50	0.50	0.49
Length	0.39	0.38	0.50	0.50	0.49	0.52
LR	0.66	0.65	0.65	0.64	**0.66**	**0.65**
(X) Tone	0.63	0.62	0.60	0.61	0.63	0.62
(X) Content	0.53	0.53	0.54	0.54	0.60	0.60
(X) Connotation	0.48	0.49	0.52	0.53	0.59	0.59
BERT-LR	**0.71**	**0.72**	**0.68**	**0.69**	**0.66**	**0.65**

Prediction accuracy (macro-average scores), all scores have standard deviations between 0.01 and 0.03

$RQ_{Feature}$: Content and Tone Associate with Blame

Topic	OR	P-value
Relationship with family	1.11	0.02
Intimate relationship	1.07	0.07
Living in shared accommodation	0.79	0.02
Money	0.82	0.20
Pregnancy concerns in pets	1.46	0.03
Work	0.98	0.03
Appearance judgment	1.16	0.07
Neighborhood	0.71	0.14

Feature	OR	P-value
Subjectivity	1.09	0.006
Hedge	0.66	0.04
First pronoun	1.45	0.10
Second pronoun	1.01	0.0009
Third pronoun	1.96	0.003
Sentiment score	1.78	0.001
Sentiment: positive	1.18	0.005
Sentiment: neutral	0.99	0.08
Sentiment: negative	3.18	0.01

$RQ_{Feature}$: Connotation Associates with Blame

Feature	Authors		Others	
	Blame OR	p-value	Blame OR	p-value
Agent	1.93	0.05	0.93	0.002
Patient	0.53	0.03	1.01	0.001
WP	1.01	0.006	0.81	0.031
Value	0.99	0.13	1.02	0.13
Power	1.04	0.006	0.97	0.003
Agency	2.00	0.003	0.96	0.002
Care	0.99	0.03	1.03	0.03
Harm	0.97	0.08	1.00	0.07
Betrayal	0.95	0.06	1.11	0.16
Loyalty	0.97	0.08	1.03	0.17
Valence	0.99	0.14	1.22	0.13
Arousal	1.04	0.11	1.21	0.15
Dominance	1.23	0.14	1.09	0.15
Joy	0.98	0.09	0.98	0.13
Sadness	0.31	0.05	1.28	0.005
Anger	1.05	0.01	0.11	0.03
Fear	2.33	0.01	1.06	0.03
Trust	1.10	0.08	0.20	0.09
Disgust	1.06	0.07	2.16	0.02
Anticipation	1.34	0.05	1.74	0.04

RQ_{Social}: Gender and Age Associate with Blame Assignment

- Females submitted 55.5% (10,284) of the 18,530 posts
- Of the posts blaming authors, 44.5% were submitted by females compared to 55.5% submitted by males
- Of the posts blaming others, 63% were submitted by females

	Age Ranges			
Metrics	15–25	26–35	36–45	46–55
Number of posts	3,554	1,951	410	136
χ^2	76.56	50.89	13.46	2.96
Cramer's ϕ	0.15	0.16	0.18	0.15

RQ_{Social}: Moral Judgments are Biased in Some Situations

- Aim: examining whether the audience has opposite attitudes towards different genders in similar situations
- Clusters: ten semantically similar clusters in where gender has strong ($p \leq 0.001$ and Cramer's $\phi \geq 0.3$ [2]) association

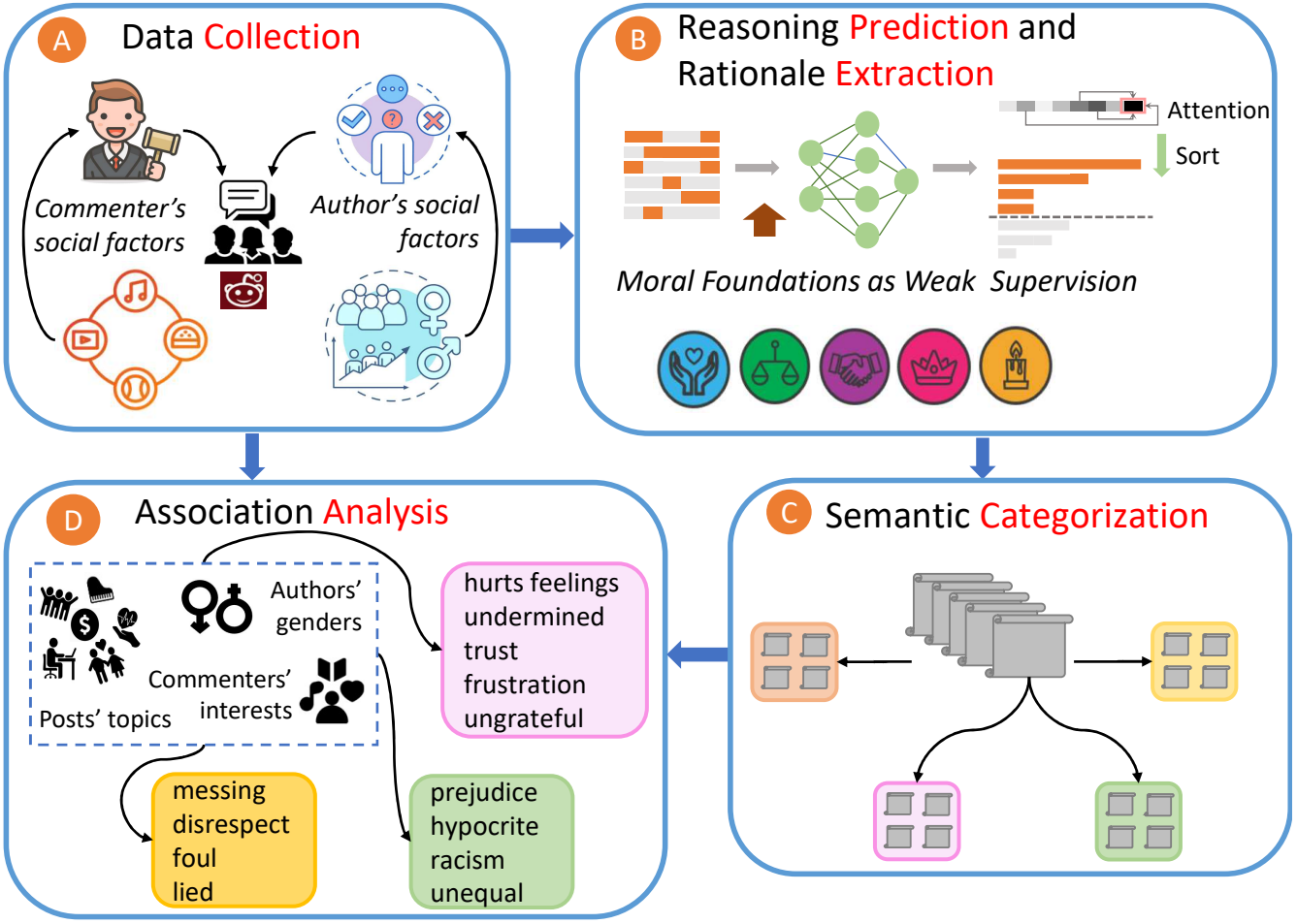
Example of biased situations' (ordered by ϕ) semantic tags[1]

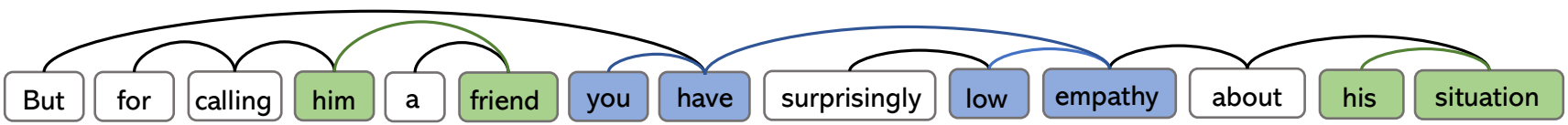
Semantic Tag	Example Words
kin	cousin, son, daughter
relationship: intimate and sexual	wife, husband, girlfriend
groups and affiliation	personal, society, company
anatomy and physiology	sweat, foot, tooth
work and employment	job, teacher, profession
sports	goal, baseball, football
games	hockey, golf, gym
money	cost, dollar, cashier
judgment of appearance	grace, beauty, pretty

[1] https://ucrel.lancs.ac.uk/usas/

Morality in the Mundane: Categorizing Moral Reasoning in Real-Life Social Situations

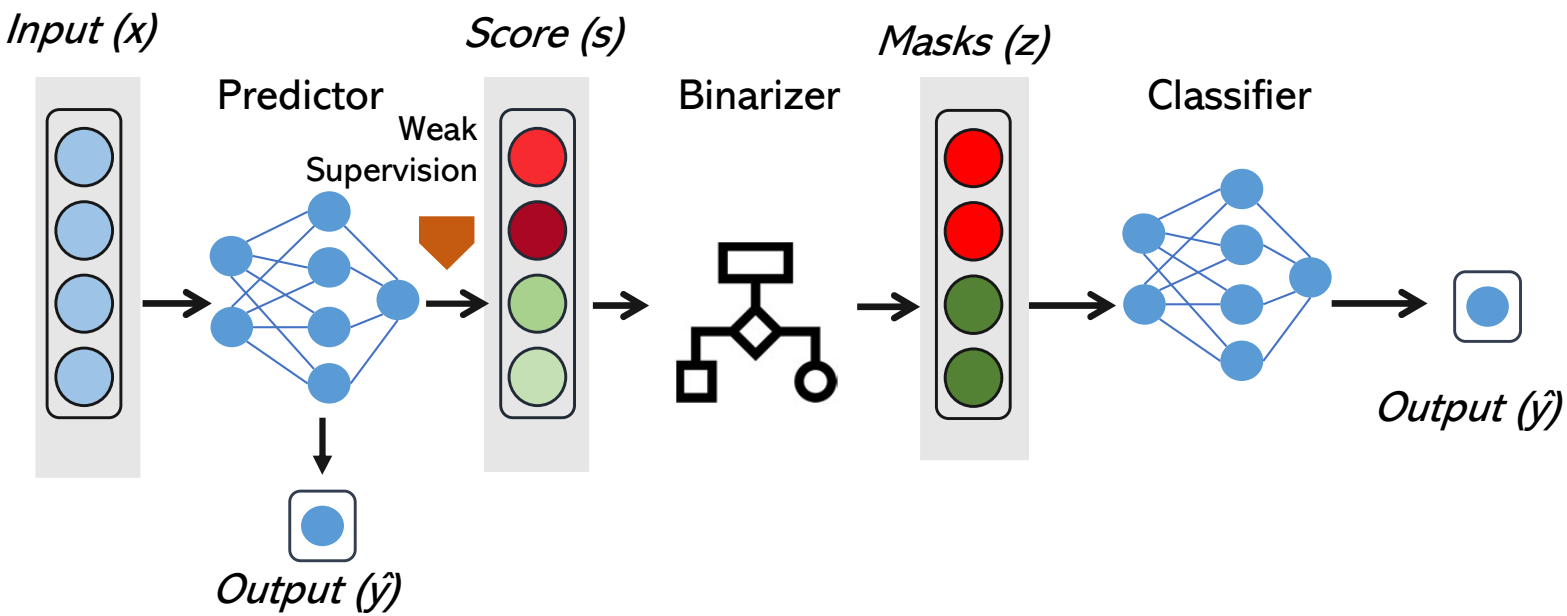
Moral Reasoning from Comments





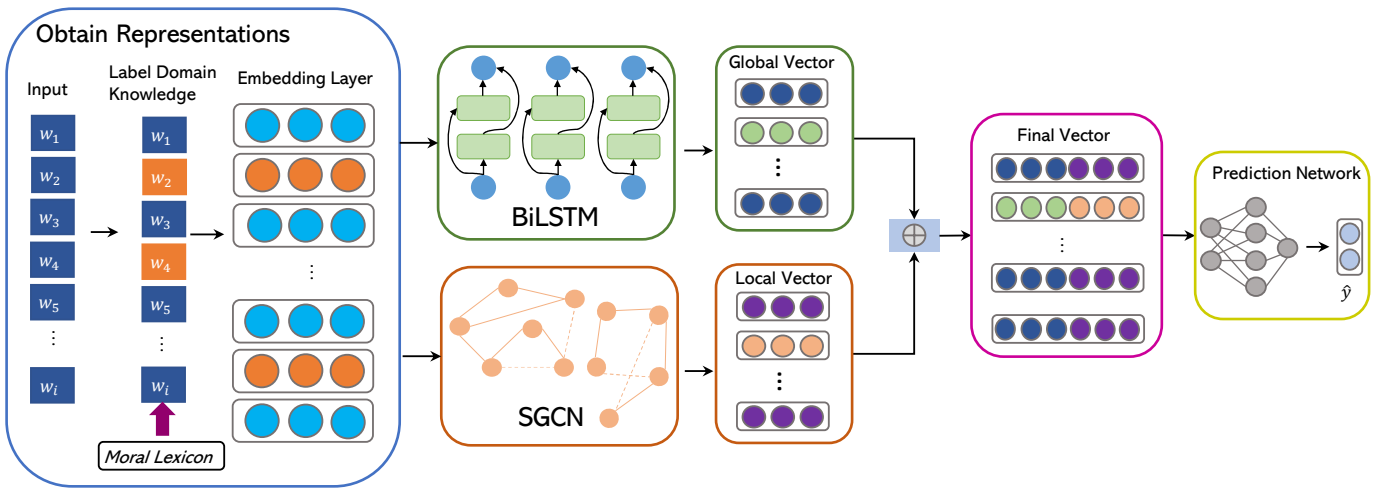
Rationalization in Deep Learning Models

- Predictor outputs $\hat{y}$ and importance scores s (**rationale**)
- Binarizer assigns masks to tokens z with low importance scores
- Classifier takes unmasked tokens to predict y again to evaluate a rationale's accuracy

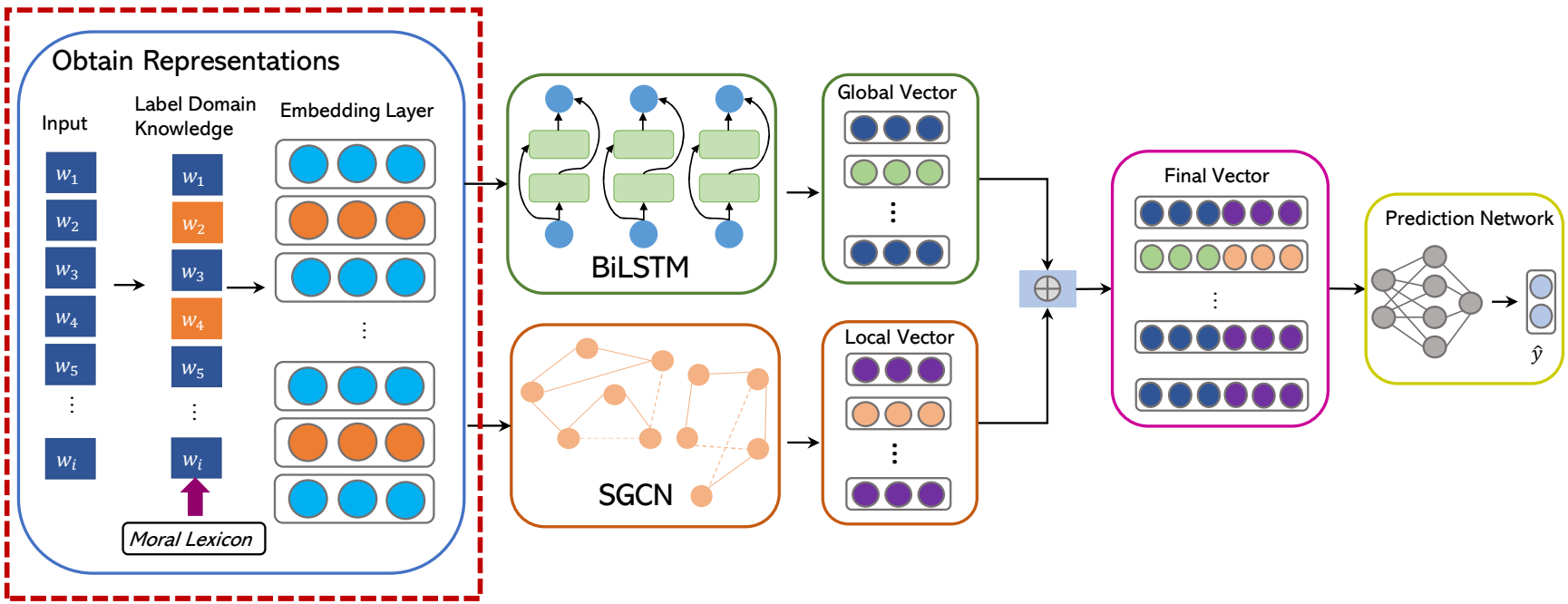


Predictor: Encoding Comments

- Concatenate global and local vectors
  - Global context features are multidimensional embeddings encoded using BERT
  - Local context features capture the neighboring syntactic context of entities, containing words or phrases modifying those entities. For example, *(has, low empathy)* represents the local syntactic context for the subject entity *he*
- Predict a verdict on a comment via a dense network that uses softmax



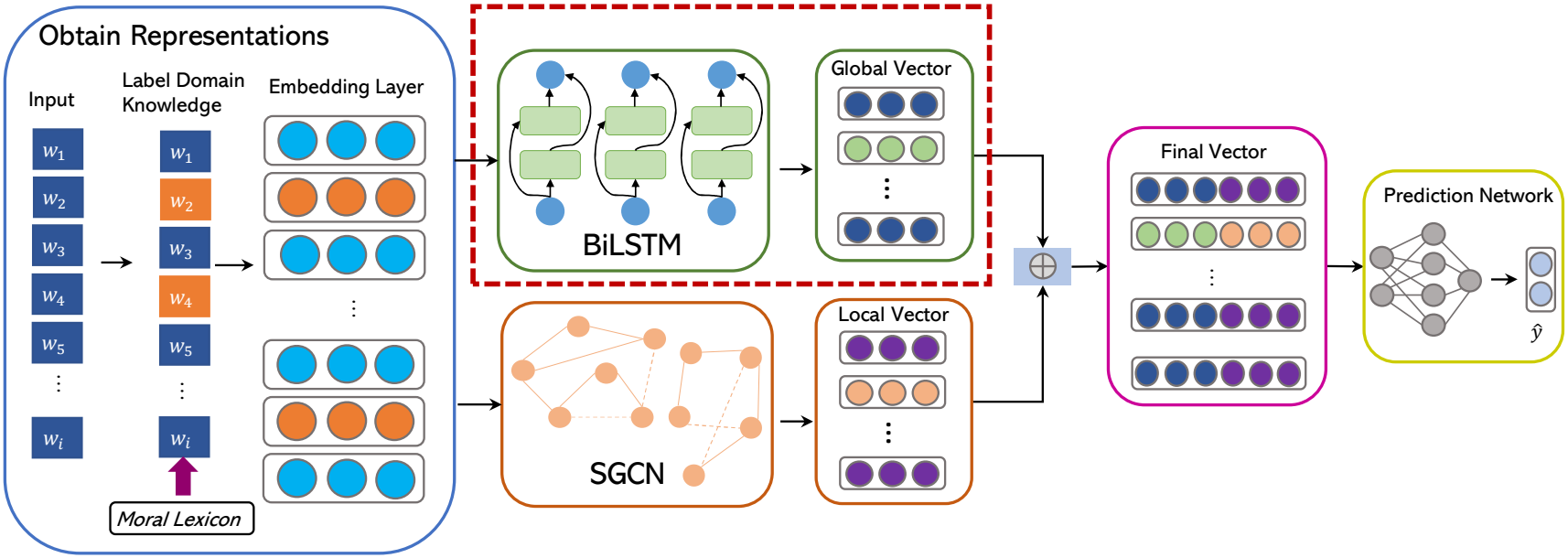
Additional Lexicon: Moral Foundation Theory (MFT)



Weak labeling of rationales via MFT lexicon of 2,041 words:

$z_d = [z_d^i] \in \{0, 1\}$, where $z_d^i = 1$ if w^i is in the lexicon (tokens in yellow)

Obtain Representations: Global Vectors

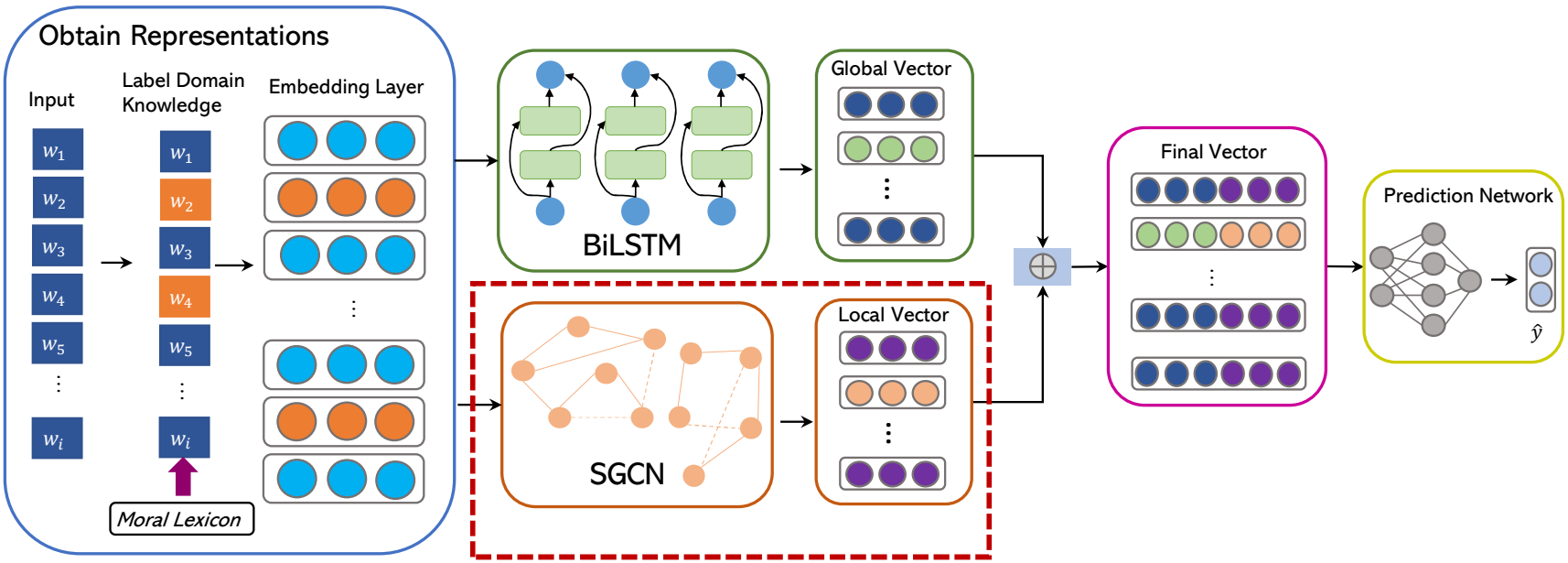


Hidden states $h_{g,1}$ and $h_{g,2}$: BERT-encoded output to a stacked BiLSTM:

$$\overleftarrow{h_{g,i}}; \overrightarrow{h_{g,i}} = \text{BiLSTM}(S), i = 1, 2, \quad (1)$$

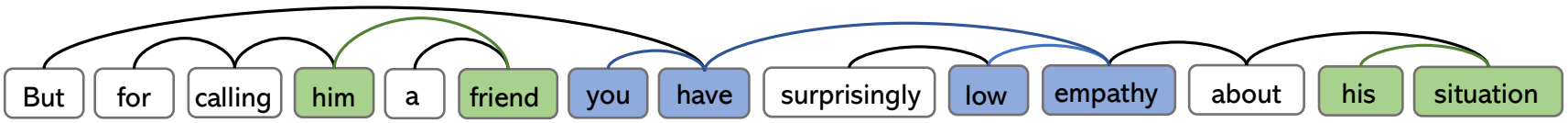
S represents the encoding output of the last layer of BERT and i denotes the direction, the vectors capture contextual representations of input texts

Obtain Representations: Local Vectors



Based on dependency graphs, local vectors represent syntactic context of words and phrases modifying the entities

Obtain Representations: Local Vectors



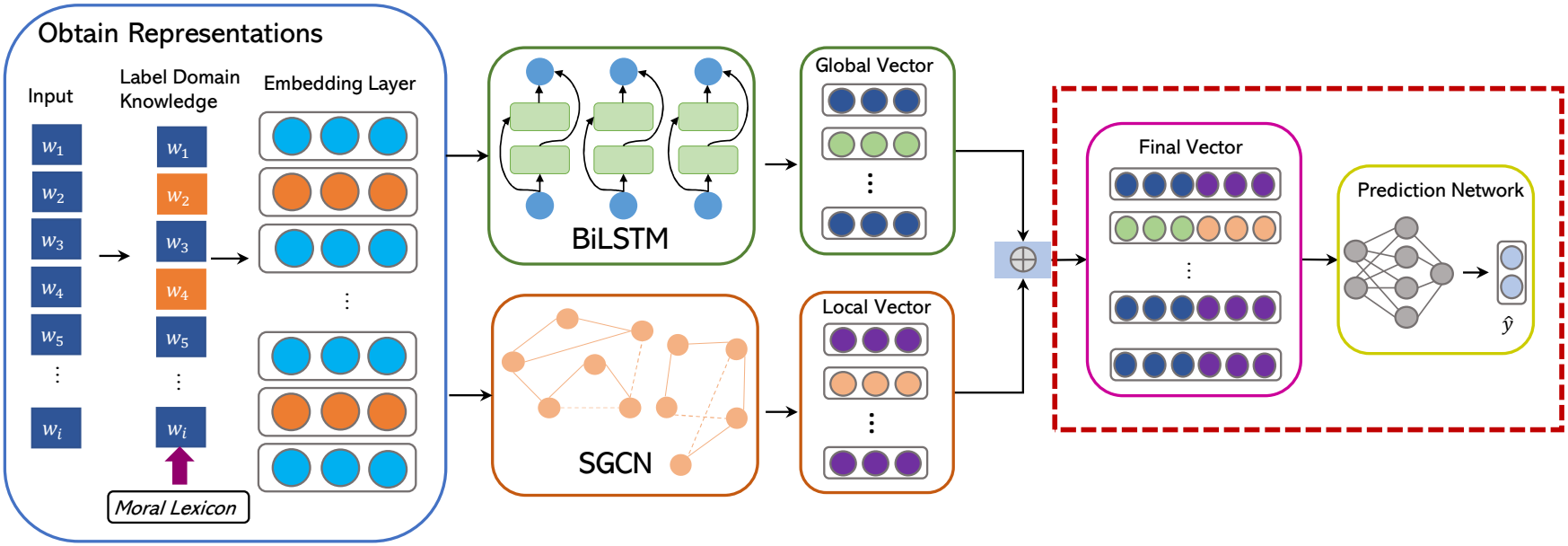
Syntactic Graph Convolutional Network

Given a vertex v in G and its neighbors $\mathcal{N}(v)$, the representation of v on the $(j+1)$st layer is:

$$h_v^{j+1} = \sum_{u \in \mathcal{N}(v)} W^j h_u^j + b^j, \tag{2}$$

where $W^j \in \mathbb{R}^{d^{j+1} \times d^j}$ and $b^j \in \mathbb{R}^{d^{j+1}}$ are trainable parameters, and d^{j+1} and d^j denote latent feature dimensions of the $(j+1)$st and the jth layers

Prediction Optimization Network



Cross-entropy objective: $\arg\min L(y, \hat{y}) + \lambda L_d(z, z_d)$

- $L(y, \hat{y})$ captures the accuracy of prediction $\hat{y}$
- $L_d(z, z_d) = -\sum_i |a^i| z_d^i$ captures the importance (the *attention* a^i) given to the words that appear in the moral lexicon
- λ controls the relative importance of the above

Results: Evaluation of Prediction

Methods	Precision (%)	Recall (%)	F1 (%)
LR-Length	53.9	53.8	53.8
LR-GloVe	57.7	56.2	57.0
Random Forest	60.8	62.4	61.6
SVM	63.2	65.3	64.2
BERT	83.7	82.6	83.1
BERT-Domain	82.8	82.5	82.6
Global	83.0	82.8	83.0
Global-Domain	83.7	81.5	82.6
Local	82.9	84.2	83.5
Local-Domain	83.6	83.6	83.6
Global-Local	85.6	**86.9**	86.2
Global-Local-Domain	**86.8**	86.1	**86.4**

Feature Scoring Methods for Rationale Extraction

- Random (RAND): Randomly allocate importance scores
- Attention (α): Normalized attention weights
- Scaled Attention ($\alpha\nabla\alpha$): Attention weights multiplied by the corresponding gradients
- Integrated Gradients (IG): The integral of the gradients from the baseline (zero embedding vector) to the original input
- Flexible (FLX): A flexible instance-level rationale selection method, under which each instance selects one of the above scoring methods [1]

Results: Adding Moral Knowledge Improves Rationales

Metrics for evaluating Rationales

- reverse-Macro F1 (revF1): Performance differences when using full input and rationale-reduced input; lower is better
- Normalized Sufficiency (NS): Differences between predicting full input text and rationales; higher is better
- Normalized Comprehensiveness (NC): Differences between predicting full input text and rationale-reduced text; higher is better

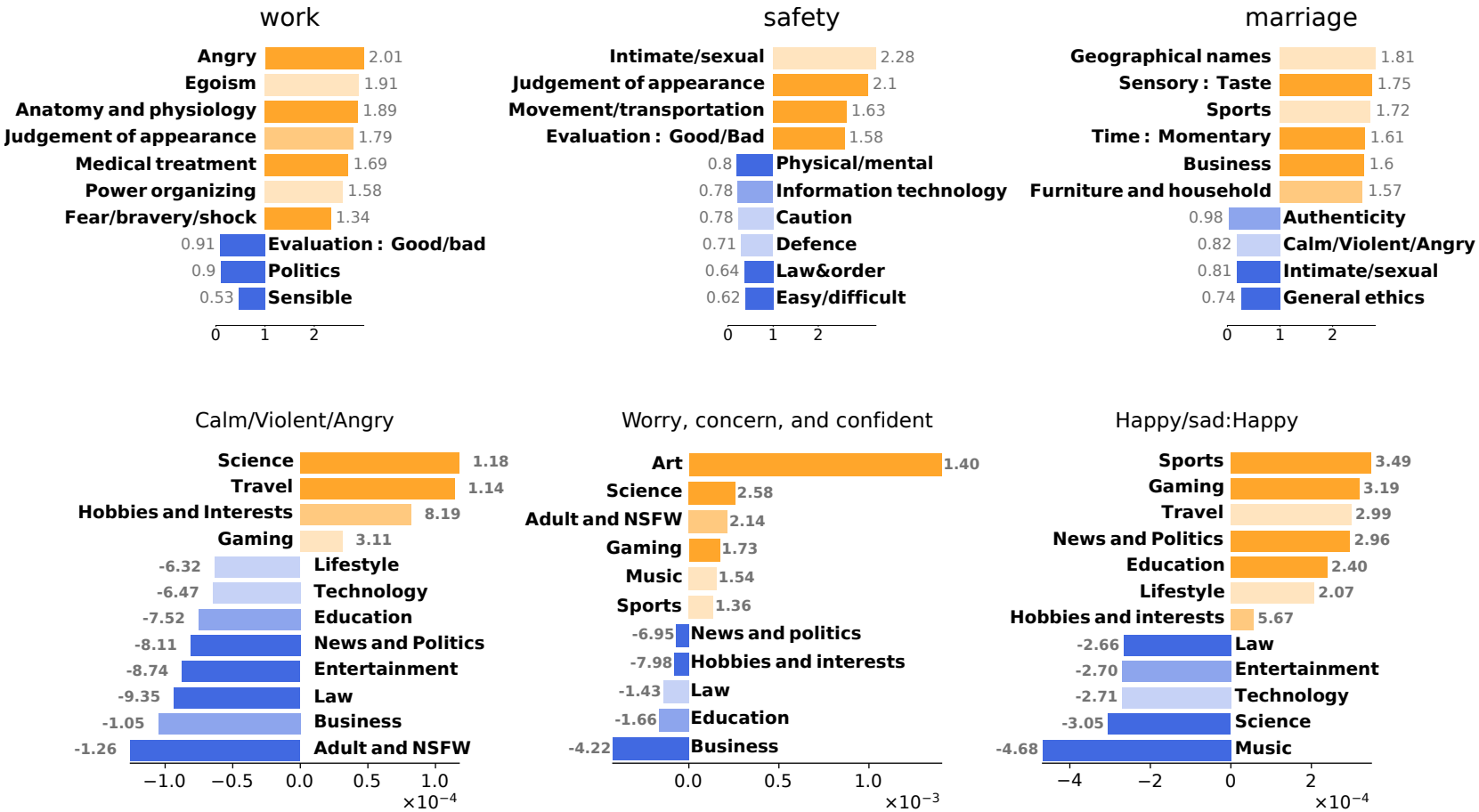
		Global			Local			Global-Local		
	Methods	revF1	NS	NC	revF1	NS	NC	revF1	NS	NC
Domain	RAND	85.9	0.26	0.27	79.3	0.25	0.27	84.0	0.20	0.34
	α	59.2	0.31	0.42	56.0	0.33	0.53	52.5	0.45	0.64
	$\alpha\nabla\alpha$	58.1	**0.47**	0.61	45.8	0.50	0.77	42.9	0.47	**0.81**
	IG	66.8	0.31	0.54	65.2	0.35	0.54	65.9	0.37	0.50
	FLX	**42.3**	0.52	**0.72**	**41.3**	**0.59**	0.77	**38.9**	**0.50**	0.80
No Domain	RAND	79.0	0.25	0.30	86.6	0.24	0.29	88.0	0.21	0.33
	α	62.1	0.29	0.39	62.5	0.37	0.61	63.6	0.38	0.61
	$\alpha\nabla\alpha$	57.2	0.37	0.65	56.9	0.45	0.64	54.1	0.45	0.70
	IG	68.2	0.32	0.53	63.9	0.30	0.53	62.2	0.28	0.45
	FLX	44.6	0.44	0.69	42.6	0.46	**0.78**	41.3	0.49	0.77

$RQ_{reasoning}$: Moral Reasoning Embeds Distinct Meaning Clusters

Commenters use different adjectives, verbs, and nouns to emphasize their concerns based on a given situation, while employing similar adverbs to express their emotions

Clusters	Topics	Examples
Judgment of appearance	Work	skinny, curly, chubby, lean, eat, meat, bodied
	Safety	underwear, panties, bikini, clingy, thong, boudoir
Evaluation: Good or Bad	Work	derogatory, *extremely terrible*, derisive, *awful*
	Safety	*awful*, *extremely terrible*, incredible
Angry and Violent	Marriage	kick, spitting, stomped, slapping, wasted
	Education	picky, lived, mortified, resentful, nauseous, baffled
Law and order	Education	punish, punishment, jails, prison, inability
	Safety	abuse, harassment, bullying, sexual, neglect

$RQ_{reasoning}$: An Author's Gender and Commenters' interests Associate with Moral Reasoning



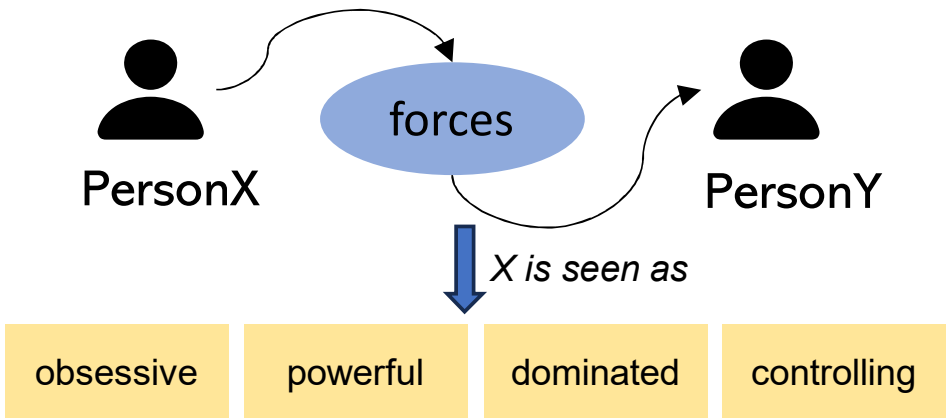
The odds ratio (OR) values: darkest colors indicate p-values ≤ 0.005, middle colors indicate p-values ≤ 0.001, lightest colors indicate p-values ≤ 0.01

Proposed: Unveiling Moral Sparks: Exploring Real-life Moral Narratives through Social Commonsense

Moral Sparks: What Captivates a Commenter's Attention?

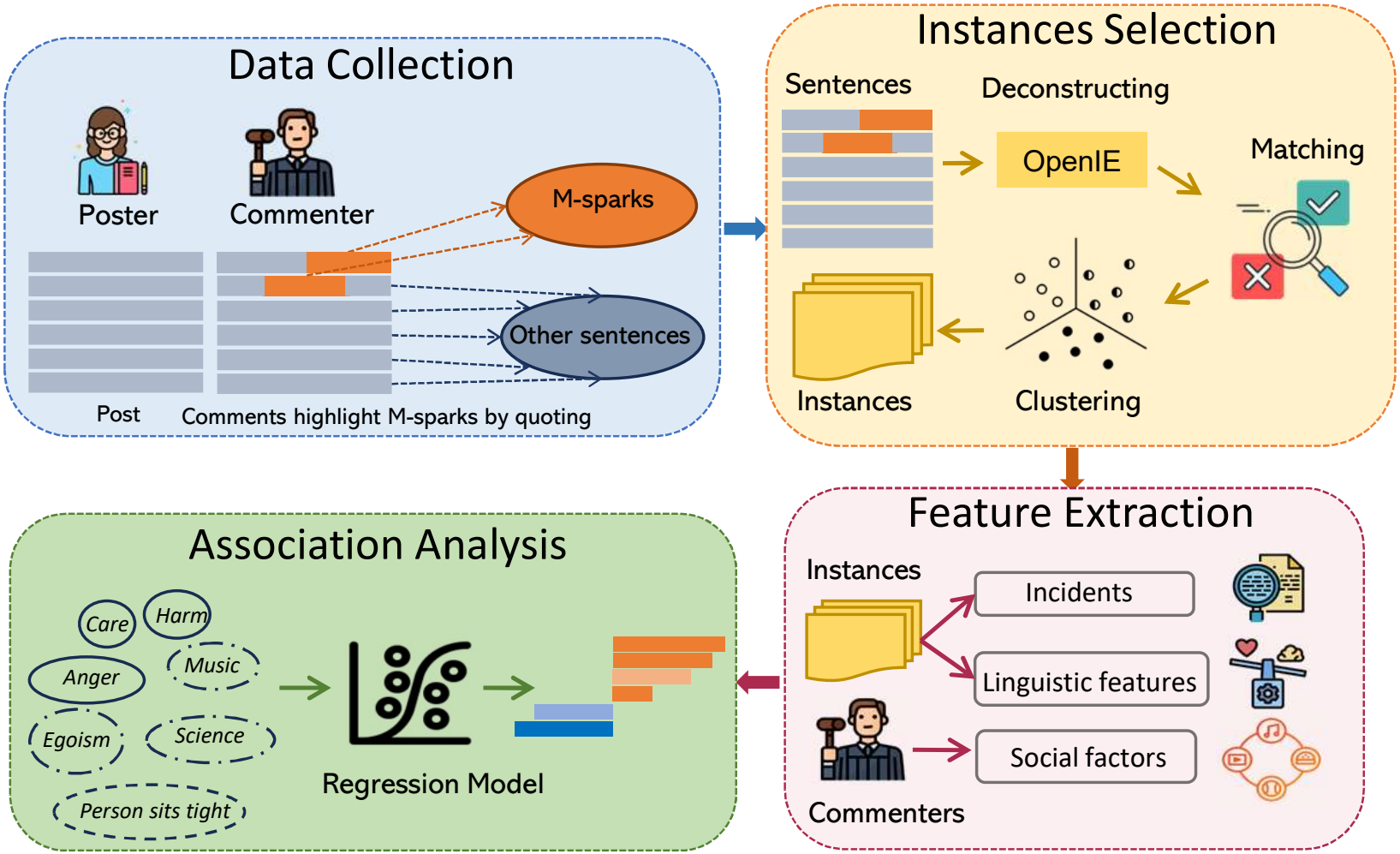
- An instance refers to a sentence in a post
- A moral spark (m-spark) refers to a quoted instance, which are focal points that offer insights into the moral aspects of the situations.

Inferred attributes of an entity from *commonsense* knowledge when a quoted sentence indicates a person's behavior as *forces*



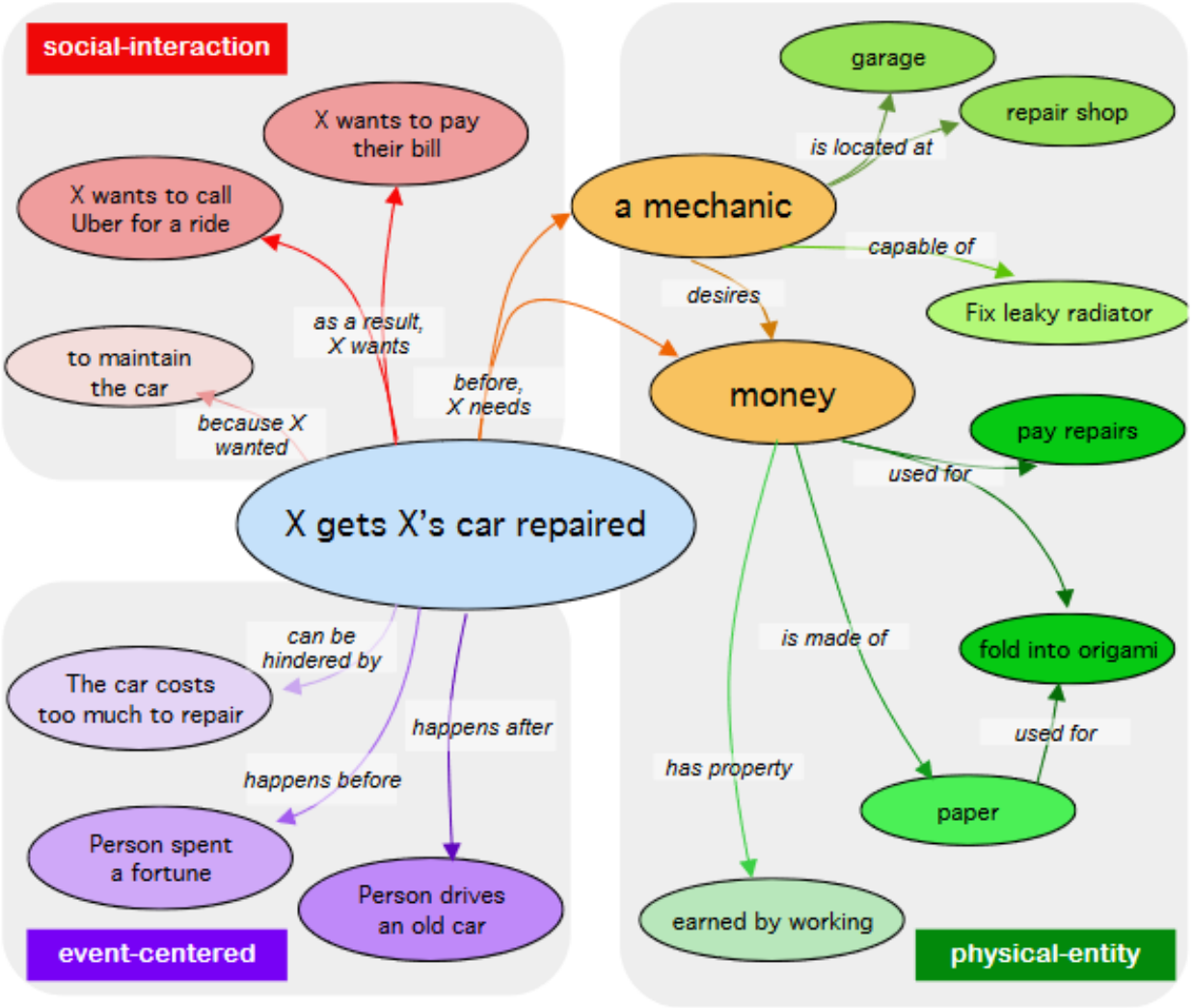
However, my aunt feels compelled to force meat down my throat

Framework for Moral Sparks



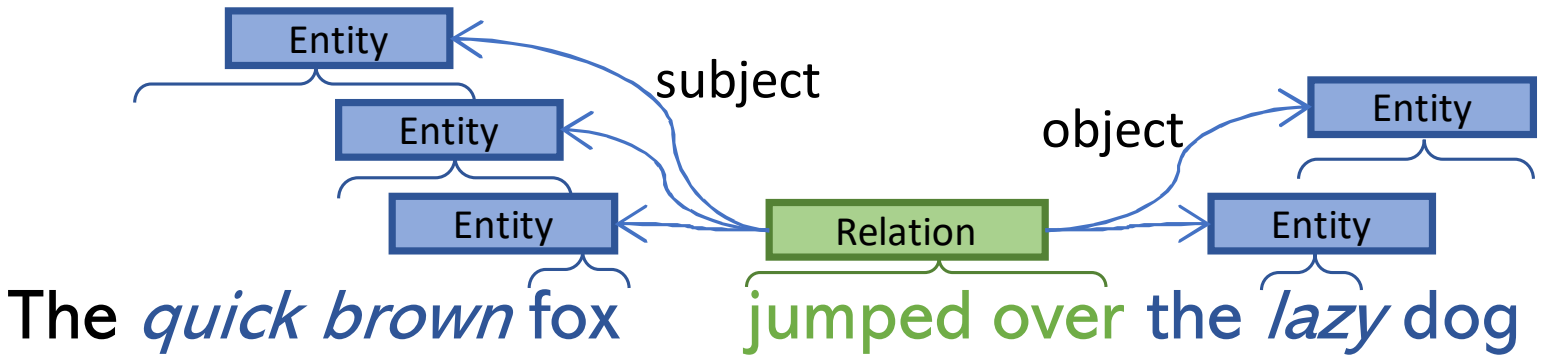
Commonsense Graph

Commonsense if–then reasoning graph from Atomic [3]

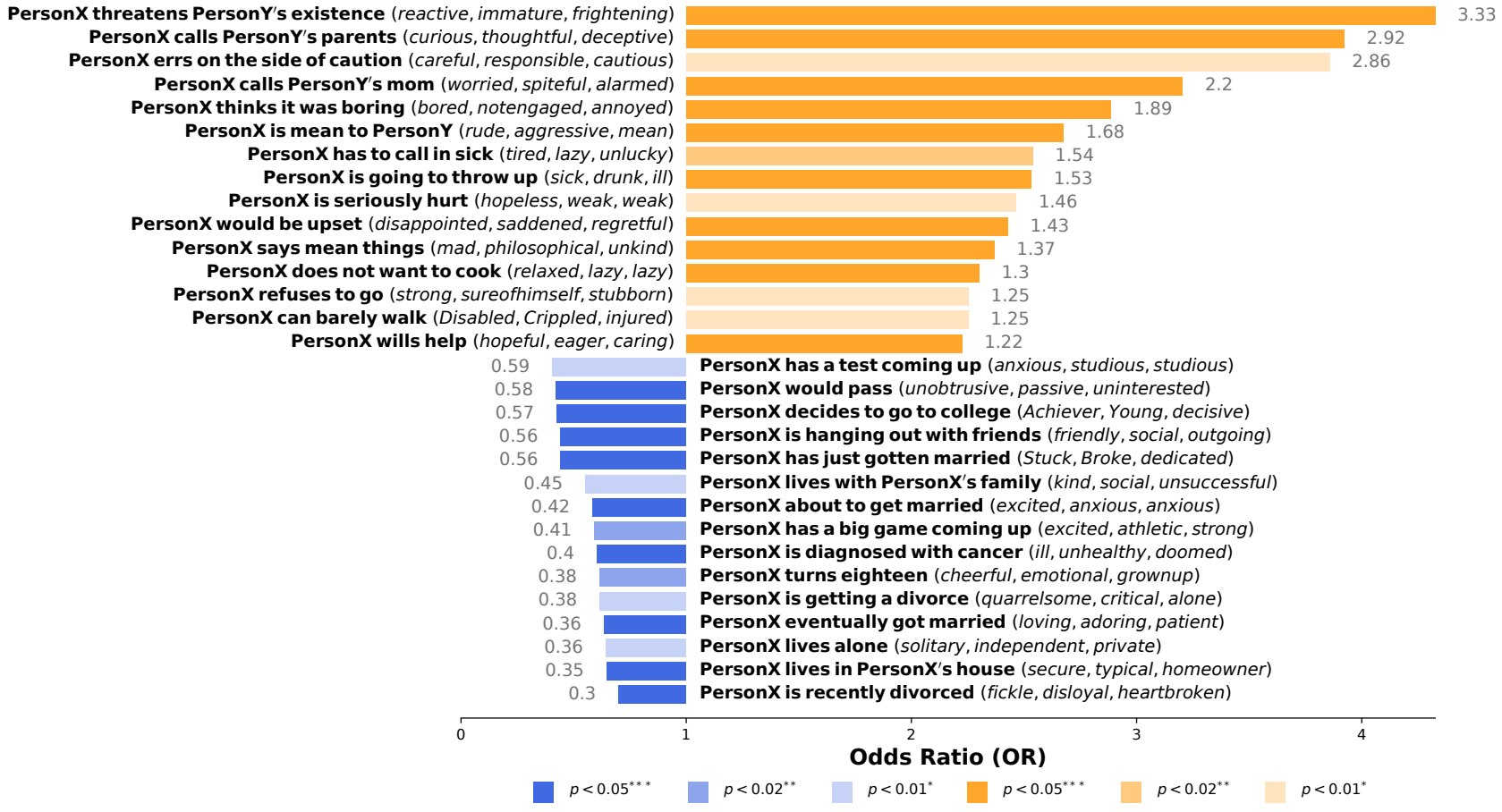


Extracting Commonsense from Instances

Compact Open Information Extraction system generates events triples from instances



RQmsparks: Social Commonsense Differentiate M-Sparks



Conclusion

Timeline and Plan of Work

Timeline

	Task	Status	Estimated Completion
1	Blame Game	Complete	March 2023
2	Morality in the Mundane	Complete	July 2023
3	Unveiling Moral Sparks	Ideation	January 2024

Plan of Work

Moral sparks: Investigate how commenters' social factors affect M-sparks

Moral sparks: Investigate linguistic features' effects on M-sparks

Thank you!

References

[1] George Chrysostomou and Nikolaos Aletras. "Flexible Instance-Specific Rationalization of NLP Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Online: AAAI Press, Feb. 2022, pp. 10545–10553.

[2] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates, 1988.

[3] Jena D Hwang et al. "(Comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2021, pp. 6384–6392.

[4] Bertram Malle, Steve Guglielmo, and Andrew Monroe. "A Theory of Blame". In: *Psychological Inquiry* 25 (Apr. 2014), pp. 147–186. DOI: 10.1080/1047840X.2014.877340.