Words Can Weight

Aaron F. McKenny

August 9, 2019

KELLEY SCHOOL OF BUSINESS
INDIANA UNIVERSITY

What would the response be…?

I love you…

I like you…

Words have both meaning and weight

- "I love you"
- "Dammit Janet"
- "We are an innovative company"

- "I like you"
- "Gosh Janet"
- "We are a software company"

Review of management CATA research

- Journals
  – 12 usual suspects (AMJ, ASQ, JAP…)
- Years
  – 2000-2018
- Search criteria (18)
  – Technique (“CATA”, “computerized text”, “computer-aided text”…)
  – Tools (“LIWC”, “Diction”, “CAT Scanner”…)
  – Process (“dictionary”, “word list”, “word count”…)

Review of management CATA research

- Initial sample: 167
- Use dictionary-based coding: 124 (74%)

- Report that weights were used… 4 (2%)
- Produce their own weights… 2 (1%)
- Document how weights were determined… 1 (0.6%)

Note: Just for dictionary-based CATA research, but I suspect a broader search of management content analysis research would yield similar numbers.

Current state of the literature

- Uniform term weighting
  - All words count equally
  - …but should they?

- Why?
  - Institutionalized
  - Easy/convenient
  - How to weight?
  - Theory should drive methods

So how do we weight? (Manual)

- Can end up a lot like a survey

- Semantic differential: **How socially oriented is the author of this text?**

  Prosocial |______|______|______|_____|______|______| Antisocial

- Likert scale: **The author of this text is socially oriented.**

  Strongly agree | Agree | Don't Know | Disagree | Strongly Disagree

So how do we weight? (Dictionary-based CATA: Individual words)

- Unclear… so let’s look at options

- Term Frequency-Inverse Document Frequency (TF-IDF)
  - Commonly used in Information Retrieval (e.g., Google search)
  - Words discriminate best when they:
    - Are used frequently in some texts (term frequency)
    - Are not used in all texts (inverse document frequency)

- The challenge:
  - Penalizes common-but-relevant words (“optimistic” vs “panglossian”)
  - Isn’t concerned with *polarity* (“like” vs “love”)

So how do we weight? (Dictionary-based CATA: Individual words)

Uses a Bayesian algorithm to assign each word a value from 0-100

Kovács et al (2013) – AllOurIdeas.org

So how do we weight?
(Dictionary-based CATA: Individual words)

Table 1 Authenticity Scores Assigned to Keywords

Keyword	Score	Keyword	Score	Keyword	Score	Keyword	Score
Authentic	95	Truthful[a]	68	Usual	53	Bogus	13
Genuine	92	Unmistakable[a]	68	Decent[a]	51	Forgery	13
Real	88	Artisan[a]	67	Unusual	51	Fake	12
Skilled[a]	83	Unpretentious[a]	67	Caring[a]	49	Hoax	11
Faithful	81	Heartful[a]	66	Ambitious[a]	48	Cheat	10
Legitimate[a]	81	Delicious	65	Replica[a]	46	Dishonest	10
Original[a]	80	Virtuous	64	Offbeat	43	Feigned	10
Traditional	79	Normal[a]	63	Atypical	41	Ersatz	9
Pure	78	Creative[a]	62	Unassuming[a]	37	Faked	9
Historical[a]	77	Interesting[a]	62	Invented	36	Imitation	9
Sincere	77	Orthodox[a]	62	New[a]	36	Quack	9
Master chef	75	Artful[a]	60	Unconventional	36	Unreal	8
Craftsmanship	74	Special[a]	60	Peculiar	35	Humbug	7
Honest[a]	74	Righteous	58	Outlandish	32	Impostor	7
Integrity[a]	74	Substantial[a]	57	Assumed	30	Sham	7
Quintessential	74	Authoritative	56	Idiosyncratic	30	Unauthentic	7
Expert	73	Typical	56	Quirky	29	Deceptive	6
Iconic[a]	73	Awesome[a]	55	Extroverted[a]	28	Inauthentic	6
Inspiring[a]	73	Moral	55	Modern	27	False	6
Unique[a]	72	Eccentric	54	Unorthodox[a]	27	Phony	5
Wholesome	72	Ethical[a]	54	Pretentious	19	Scam	4
Professional	70	Fresh[a]	53	Untraditional[a]	17		
Skillful	70	Old-fashioned[a]	53	Artificial	14		

[a]Keywords added by participants.

So how do we weight? (Dictionary-based CATA: Entire dictionaries)

Table 1. Ten Most Important Linguistic Features in Measuring Extraversion

Linguistic feature	Source	Description	Relation with extraversion	Weights
Unique	LIWC	Measure of repetition of words in a given text.	−	.6457
MEANP	MRC	Paivio meaningfulness, defined as the mean value of written associations people list with a word in 30 seconds. (Paivio, 1968)	+	.3553
We	LIWC	The relative number of times the first-person plural is used, e.g., "we," "us," "our" (11 words).	+	.2845
T-L-FREQ	MRC	Measure of how frequently words are used in the English language. (Thorndike and Lorge, 1944)	−	.2544
Number	LIWC	The relative frequency of numbers in the text, e.g., "one," "thirty," "million" (29 words).	−	.2468
Motion	LIWC	The relative frequency of words related to motion in the text, e.g., "walk," "move," "go" (73 words).	+	.2464
Insight	LIWC	The relative frequency of words related to insight, e.g., "think," "know," "consider" (116 words).	−	.2355
Up	LIWC	The relative frequency of words like "up," "above," "over" (12 words).	−	.2296
NLET	MRC	Average number of letters in a word.	−	.2282
WPS	LIWC	Average number of words per sentence.	+	.2219

Uses Machine Learning to assign weights to Dictionary results

Malhotra et al (2018) citing Mairesse et al (2007)

Project In-Progress (Alphabetical)

Jason Kiley

Tim Michaelis

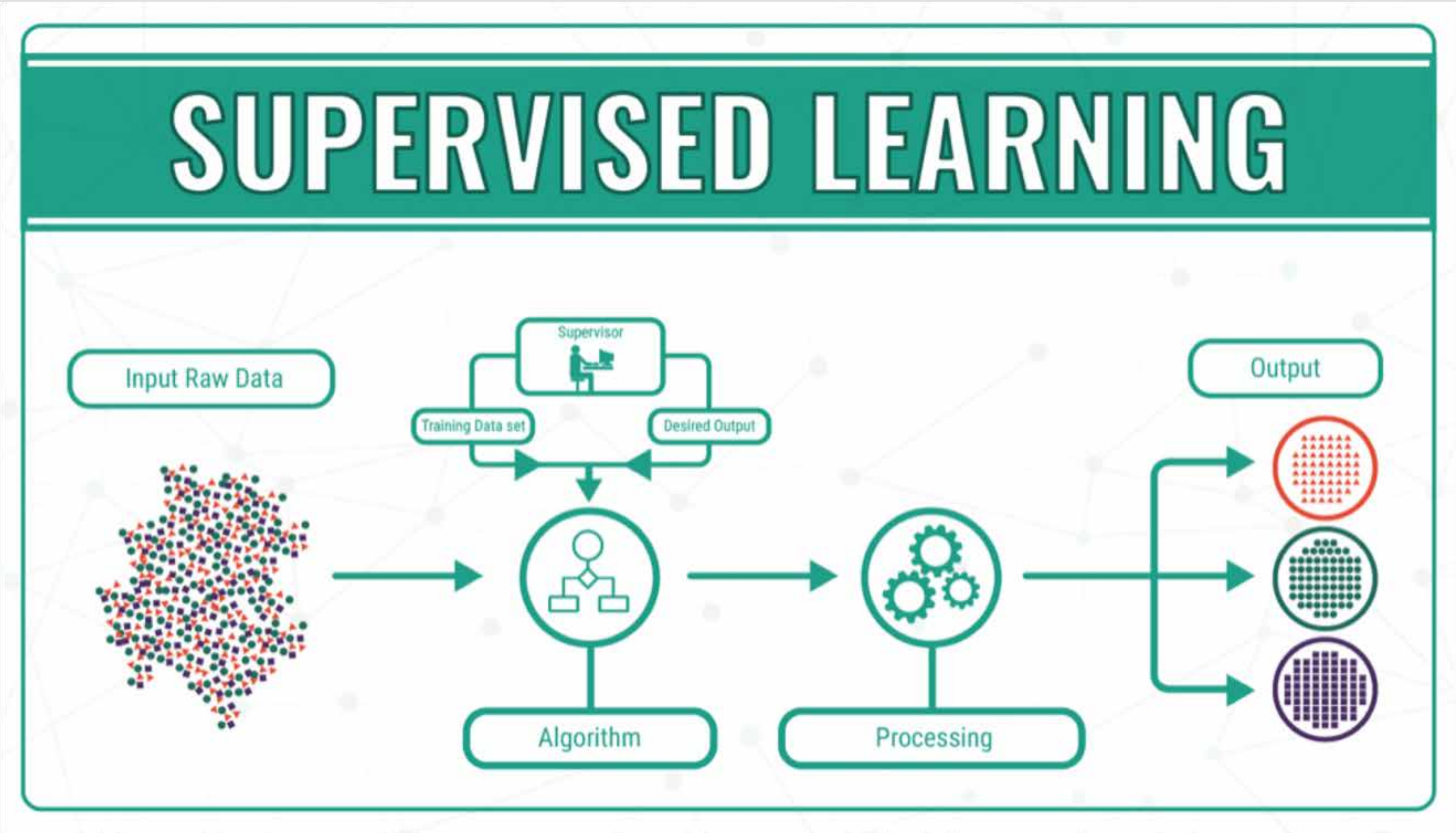
Clay Posey

Project In-Progress

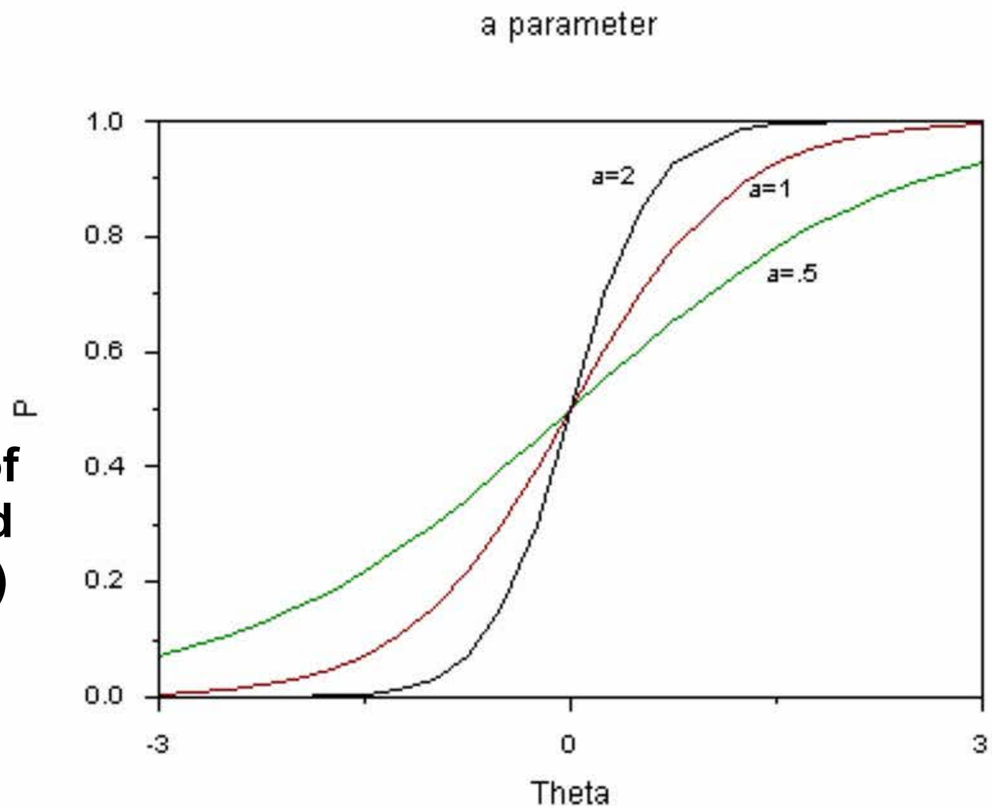
- Comparison of term weighting approaches

- Existing approaches
  - Unweighted
  - TF-IDF
  - AllOurIdeas.org
- New approaches
  - Machine Learning
  - Item Response Theory

Machine Learning



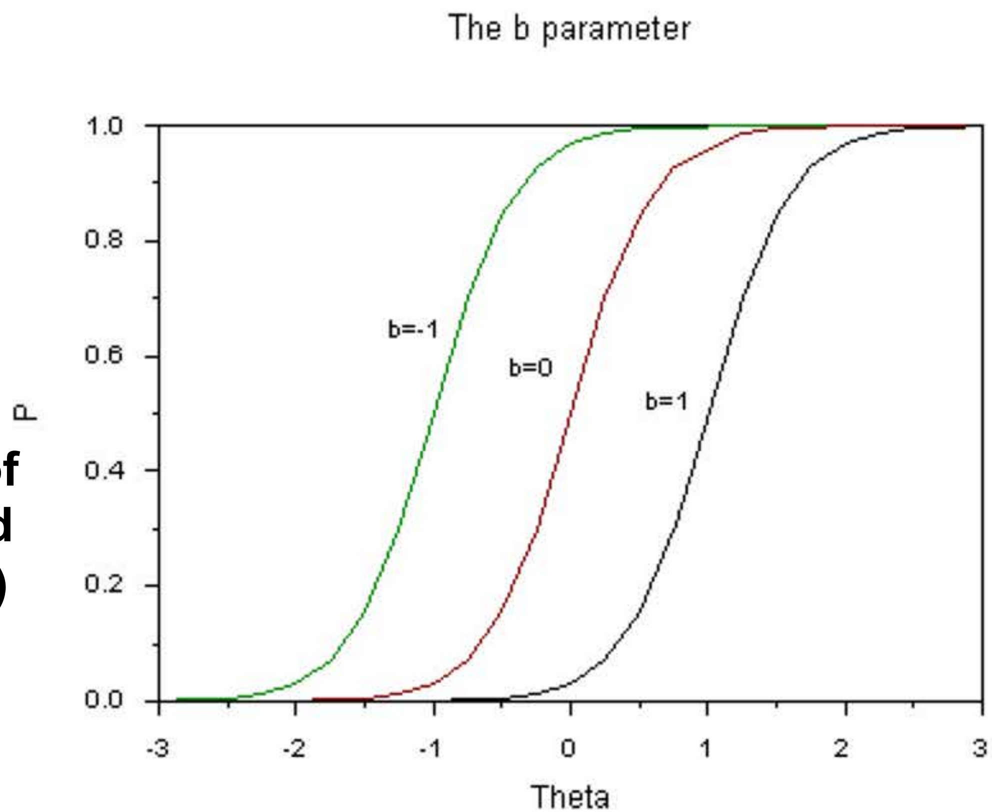
Item Response Theory: Discrimination



Probability of using a word (e.g., “love”)

Your “true” score (e.g., positive affect)

Item Response Theory: Difficulty



Probability of using a word (e.g., "love")

Your "true" score (e.g., positive affect)

Questions?