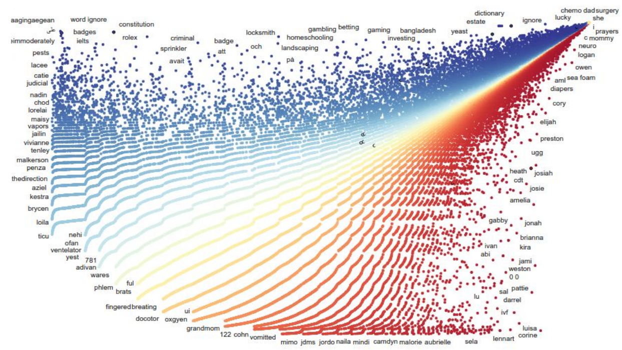
Why NLP matters to HCI researchers

June 5, 2019



Zachary Levonian z.umn.edu/zlevonian levon003@umn.edu

(Splash image was produced using the Scattertext tool, applied to a dataset of online health community posts from CaringBridge.org)

Key Links

» Ask me questions: levon003@umn.edu
» Slides: z.umn.edu/carletonNLP2019Slides
» GitHub Repository: z.umn.edu/carletonNLP2019

Agenda

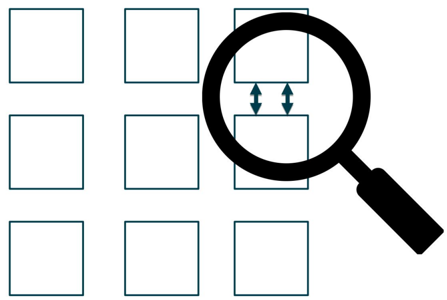
1. What is HCI & Social Computing?
2. NLP as a component of qualitative text analysis
3. Bridging qualitative themes to quantitative classification models in an online health community
4. Q&A

Content note: Discussion of cancer

Who am I?

HCI & Social Computing

- HCI = Human-Computer Interaction
- Social Computing: “technical systems mediating human-to-human communication”



Why is NLP relevant to HCI research?

- Understanding language means understanding people!
- People produce language data while using socio-technical systems
- They also produce language data when we *ask* them about socio-technical systems
- Much of our data is text!

Example: LIWC - “A person’s mental and affective state manifest in their language”
If we want to understand how people use Twitter, we have to look at the language they’re using on the platform.
Interviews, surveys, and observation all produce textual data for analysis.

Why is NLP relevant to HCI research?

- Understanding language means understanding people!
- People produce language data while using socio-technical systems
- They also produce language data when we *ask* them about socio-technical systems
- Much of our data is more text than we can read!

Why is HCI relevant to NLP research?

- Understanding people means understanding language!
- People produce language!
- Socio-technical systems are used by people!

I'm going to spend much less time discussing the impacts of HCI on NLP, but there definitely is an impact.
Understanding people = "getting a fuller view of their context"
NLP models are often used in the context of social applications

Expressive writing in OHCs

- Haiwei Ma, C. Estelle Smith, Lu He, Saumik Narayanan, Robert A. Giaquinto, Roni Evans, Linda Hanson, and Svetlana Yarosh. 2017. **Write for Life: Persisting in Online Health Communities through Expressive Writing and Social Support**. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 73 (December 2017), 24 pages. DOI: https://doi.org/10.1145/3134708
- Classification of blogs based on text data

From our lab!

Bias in sentiment analysis

- Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. **Addressing Age-Related Bias in Sentiment Analysis**. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Paper #412, 14 pages. DOI: https://doi.org/10.1145/3173574.3173986
- Correcting for bias in widely-used sentiment analysis models

Bias in word embeddings

- Hila Gonen, and Yoav Goldberg. 2019. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them**. Accepted to NAACL 2019. https://arxiv.org/abs/1903.03862
- Is correcting for bias even possible?
- Existing bias removal techniques are insufficient

Feminist textual analysis using topic models

- Shauna Julia Concannon, Madeline Balaam, Emma Simpson, and Rob Comber. 2018. **Applying Computational Analysis to Textual Data from the Wild: A Feminist Perspective**. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Paper 226, 13 pages. DOI: https://doi.org/10.1145/3173574.3173800
- Linking prevalence of topics with metadata (SES of region in England)

Quantitative vs Qualitative Research Methods

Quantitative	Qualitative
Focus on generalizable results Numerous data points Less context associated with each observation	Focus on in-depth analysis Specific, local phenomena Intention of generalizing to other sites and other people
+ Good for demonstrating differences + Can be extended/combined − Need to know relevant metrics ahead of time	+ Good for gathering rich description and understanding + Inspires “next steps” − Subjective, time-consuming, and non-replicable

» Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *GROUP ‘16*.

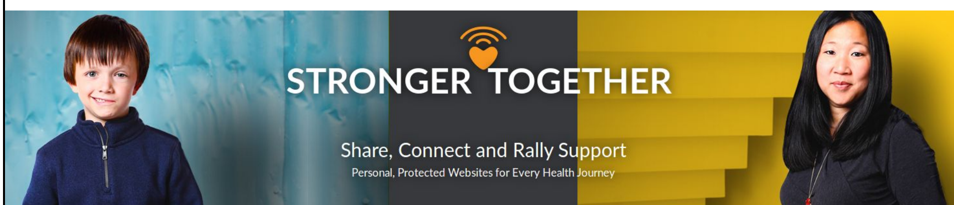
In this talk, I’m focusing on the use of NLP methods to address HCI research questions
Specifically, we’ll explore the use of NLP methods to bridge from qualitative methods to quantitative methods.
Quant: generalize from limited info. Anwers: How much?
Qual: Examine specifics to understand. Answers: How, Why
(Some language borrowed from slides by Lana Yarosh)

My research

- Social support in OHCs
  - OHC = “Online health community”
- Specifically: patient use of OHCs for communicating labor (over time)

CaringBridge

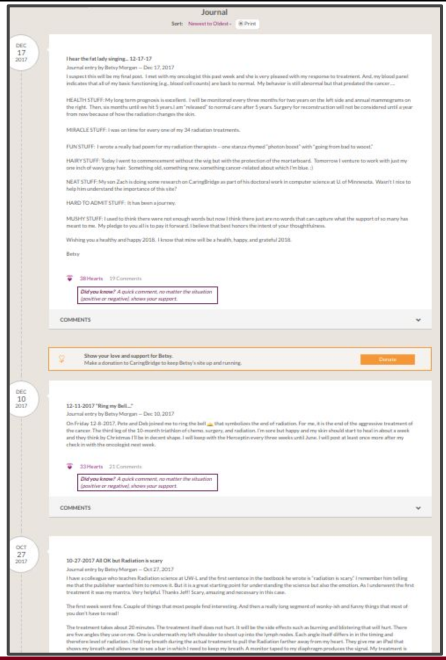
- Personal, protected place for health journeys
- Authors include patients and non-professional caregivers



CaringBridge

CaringBridge

» Site journals have text updates



CaringBridge

CaringBridge

Each update:

- Title text
- Body text
- Creation date/time

Dataset & Ethics of Use

- Data provided directly by CaringBridge
- 500,000+ individual sites
- Most data public… but a lot is private!
- Terms of Service covers this use

- Fiesler, C., & Proferes, N. (2018). “Participant” Perceptions of Twitter Research Ethics. *Social Media + Society*.

CaringBridge

4,946 sites containing 158,597 journal updates

Journal Updates	Median: 22 updates M=32.1; SD=43.7	
Site Visits	Median: 1017 visits M=2099.2; SD=4136.9	
Survival Time	Median: 8.2 months M=12.9; SD=13.3	

Breast	2752 (55.6%)	Leukemia	209 (4.2%)
Lymphoma	597 (12.1%)	Ovarian	169 (3.4%)
Other	380 (7.7%)	Lung	168 (3.4%)
Not Specified	257 (5.2%)	Myeloma	120 (2.4%)
Colorectal	225 (4.5%)	Brain	69 (1.4%)

Classification of patient updates

» To do classification, need:
  - Taxonomy of classes
  - Automated classification method

Class: labels that can be applied to each update
Taxonomy: Necessarily rigid boundaries between classes
Let’s think first about the problem of identifying a taxonomy

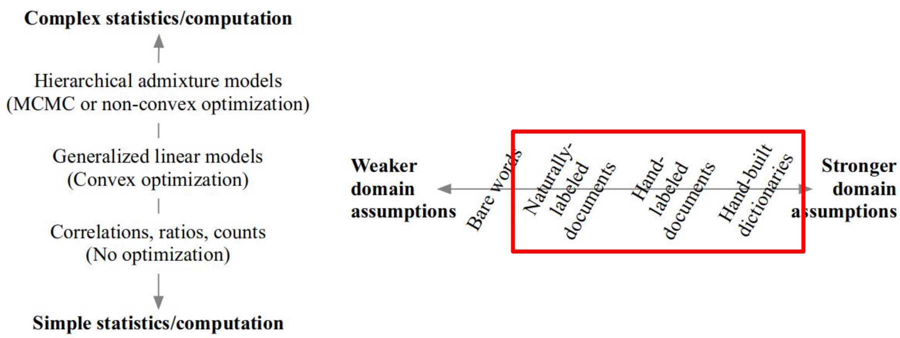
Identifying a taxonomy

- From unsupervised machine learning
- From experts
- From qualitative research

- Singer, P.; Lemmerich, F.; West, R.; Zia, L.; Wulczyn, E.; Strohmaier, M.; and Leskovec, J. 2017. Why We Read Wikipedia. In Proc. of *WWW '17*, 1591–1600. ACM.

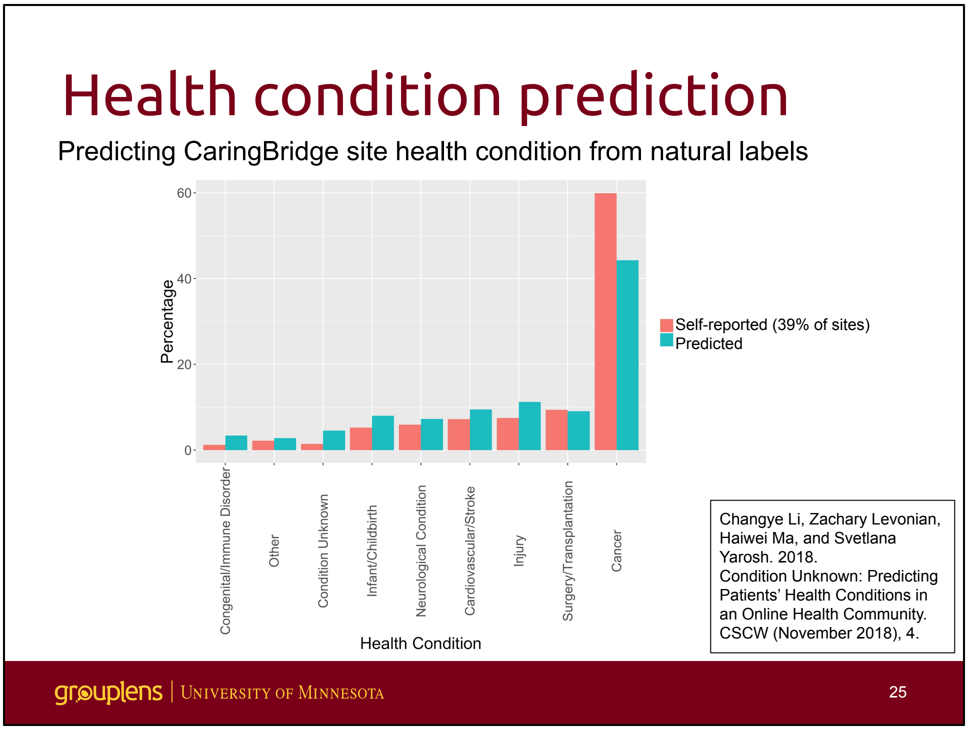
Unsupervised: no domain assumptions, hard to validate relevance
Experts (includes crowd): ignores novel categories, domain expertise may not exist
Qualitative: What if we don't have the money/expertise to conduct qual work, what if there isn't existing qual work on the target population
This is the problem we're interested in.

From taxonomy to classification



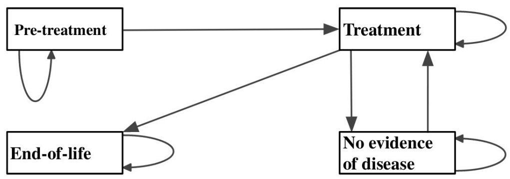
» Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. Computational text analysis for social science: Model assumptions and complexity. In *NeurIPS'11*.

Even once we have our taxonomy, need to figure how we're doing classification.
On the modeling side, we have a number of options ranging from more to less complex.
But the more interesting question is on the domain assumptions side:
how much are we assuming about the domain, as per our qualitatively-informed priors
Domain assumptions: "how much knowledge of the substantive issue in question is used in the analysis." i.e. how much prior knowledge is used in the analysis



Assume sites that don't self-report a health condition use similar language to sites that do report a health condition.
Error analysis: Lots of errors related to comorbidity.
In other words, hand labels disagreed with natural labels.
We probably needed stronger domain assumptions!
We're going to explore hand-labeled documents as a potential middle-ground that enables us to make use of domain assumptions in the qualitative work while being responsive to the specific context.
Next, let's talk about the actual qualitative work.

Cancer phases (Hayes et al.)



Methods

-Exploratory, qualitative
-Contextual inquiry
-Websites (n=42)
-Listservs (n=12)
-Artifact analysis
-Interviews (n=21, 7 patients)

- “Personal journey with cancer” as a significant metaphor
- Journey “allows for divergent, convergent, and even circular paths”
- Gillian R. Hayes, Gregory D. Abowd, John S. Davis, Marion L. Blount, Maria Ebling, and Elizabeth D. Mynatt. 2008. Opportunities for Pervasive Computing in Chronic Cancer Care. In *Pervasive Computing*. Springer, Berlin, Heidelberg, 262–279.

Next: operationalization.

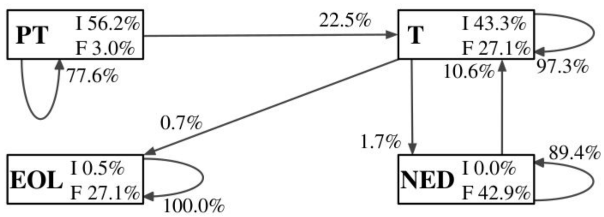
Cancer phase operationalization

Phase	Occurrence	Disagreement	κ
PT	7.4%	5.5%	0.91
T	69.7%	7.4%	0.94
EOL	1.9%	0.2%	—
NED	6.4%	3.6%	0.95
Overall	99.62%	10.2%	0.93

Taxonomy & Annotation

-2 rounds of codebook iteration
-IRR: 31 sites (619 updates)

-Single pre-treatment phase
-Transitions: allow multi-phase
-Uncertainty label
-EOL site sampling
-200 sites annotated



Multilabel classification

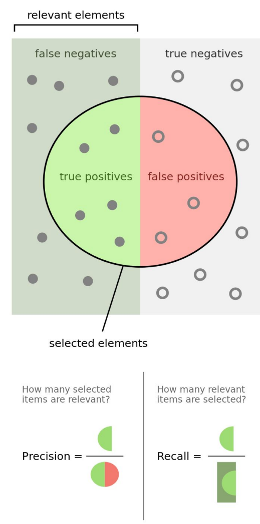
- Prediction target: 4x1 vector of labels
- Input: Title/body text of updates
- Features: hashed unigrams and bigrams
- Vowpal Wabbit online learner

- Beygelzimer, A.; Langford, J.; and Zadrozny, B. 2005. Weighted one-against-all. In *Proc. of AAAI '05*, 720–725.

See also: “Baselines and Bigrams”

Evaluating predictive performance

» Precision and recall are core to evaluating NLP classifiers
» F1 score is the harmonic mean of precision and recall and is widely used and reported



Aside: Transfer Learning

- We tried ULMFiT…
- Worse than the linear models!
- Possibly due to labeled data size?

- Howard, J., and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. arXiv:1801.06146 [cs].

Cancer phase classification

Phase	P	R	F1
PT	0.91	0.95	0.93
T	0.96	0.99	0.97
EOL	0.55	0.96	0.70
NED	0.86	0.86	0.86
Mean	0.94	0.97	0.95
B_{SA}	0.74	0.86	0.79
B_{FM}	0.74	0.99	0.81

Weighted macro average

Subset accuracy (Treatment only)

F-Measure baseline (All phases)

ML Classifier

Next: If we were to use keywords, how much predictive performance would we be giving up?

Keyword classification

- Identify list of words for each class
- Assign class to document if document contains any word in class list
- Two approaches

Max-precision keyword lists

- Constraint: Use only words uniquely associated with each class
- Goal: Achieve best possible recall
- Max *k*-Cover: Select *k* sets to maximize number of elements covered
- In our case: Select *k* words to maximize number of updates labeled with this class
- NP-Hard! Use a greedy approximation
  - Recall at worst 63% of optimal.
- Feige, U. 1998. A Threshold of Ln N for Approximating Set Cover. J. ACM 45(4):634–652.

Representative keyword lists

- Use *most representative* words of each class
- Frequency-based odds ratio:

$$OR(w, c) = \frac{f_c(w) \times f_{\bar{c}}(\bar{w})}{f_c(\bar{w}) \times f_{\bar{c}}(w)}$$

$f_c(w)$ = # of updates assigned class *c* that contain word *w*

- MacLean, D.; Gupta, S.; Lembke, A.; Manning, C.; and Heer, J. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proc. of CSCW '15*, CSCW '15, 1511–1526.

Cancer phase classification

Phase	P	R	F1
PT	0.91	0.95	0.93
T	0.96	0.99	0.97
EOL	0.55	0.96	0.70
NED	0.86	0.86	0.86
Mean	0.94	0.97	0.95
B_{SA}	0.74	0.86	0.79
B_{FM}	0.74	0.99	0.81

ML Classifier

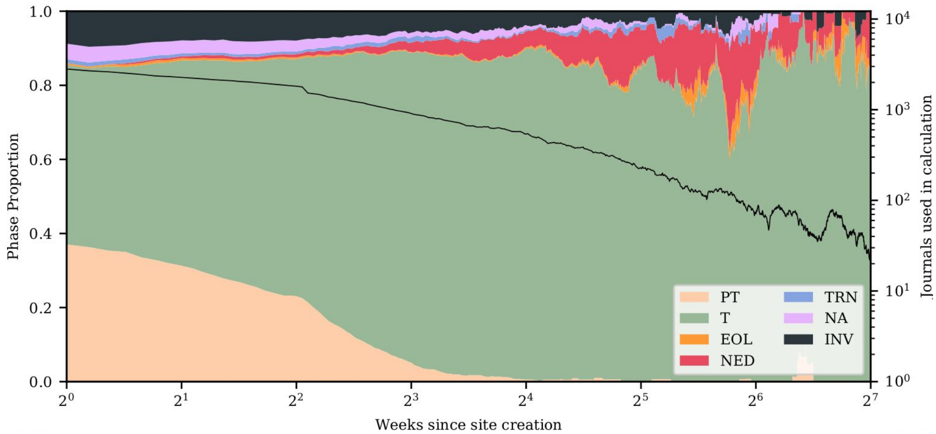
Class	k=10				k=100			
	Train		Test		Train		Test	
Label	R	F1	R	F1	R	F1	R	F1
PT	.08	.15	.01	.02	.45	.62	.03	.04
T	.13	.23	.05	.09	.49	.66	.31	.46
EOL	.39	.56	.21	.31	.99	.99	.26	.31
NED	.11	.20	.00	.01	.52	.69	.03	.04

Max-precision keywords

Class	k=10						k=100	
	Train			Test			Train	Test
Label	P	R	F1	P	R	F1	F1	F1
PT	.12	.72	.21	.12	.71	.20	.13	.14
T	.88	.92	.89	.88	.90	.88	.92	.92
EOL	.10	.73	.18	.11	.72	.18	.03	.03
NED	.06	.97	.12	.07	.97	.13	.11	.12

Representative keywords

Cancer phase classification



Next: responsibilities

Cancer journey framework (Jacobs et al.)

	Responsibilities *Patient work; health tasks placed on patients*	**Challenges** *Barriers to care*	**Personal Journey** *The effects of cancer on one's personal, daily life*
Screening and Diagnosis	• Communicating the disease to others	• Information gaps • Emotional impacts • Dealing with others' reactions	• Attitude changes • Major life events
Information Seeking	• Information filtering and organization • Clinical decisions • Preparation	• Overwhelming amount of information • Understanding treatment options	• Coping strategies
Acute Care and Treatment	• Symptom management • Support management • Compliance • Managing clinical transitions • Financial management	• Inability to work • Transportation • Lack of support • Reluctance to ask for help • Unexpected complications	• Relationship changes • Responsibilities of daily life • Social behavior changes • Loss of independence • Asserting control • Health milestones • Personal goals
No Evidence of Disease	• Continued monitoring • Giving back to the community • Health behavior changes	• Worry about recurrence	• Survivor identity • Return to normal

- Patient-centered cancer experience, captured in three categories
- Maia Jacobs, James Clawson, and Elizabeth D. Mynatt. 2016. A Cancer Journey Framework: Guiding the Design of Holistic Health Technology. In *PervasiveHealth '16*.

Cancer journey framework (Jacobs et al.)

Code	Responsibility	Phase
CO	Communicating the disease to others	PT
IF	Information filtering and organization	PT
CD	Clinical decisions	PT
PR	Preparation	PT
ST	Symptom tracking	T
CS	Coordinating support	T
SM	Sharing medical information	T
CP	Compliance	T
MT	Managing clinical transition	T
FM	Financial management	T
CM	Continued monitoring	NED
GB	Giving back to the community	NED
BC	Health behavior changes	NED

Methods

-Single cancer clinic in Georgia
-Breast cancer survivors
-Majority still receiving treatment
-Interviews (n=17)
-Focus groups (n=14)

- Responsibilities: “multiple tasks that are placed on patients”
- Responsibilities mapped to cancer phases
- Maia Jacobs, James Clawson, and Elizabeth D. Mynatt. 2016. A Cancer Journey Framework: Guiding the Design of Holistic Health Technology. In *PervasiveHealth ’16*.

Patient responsibility operationalization

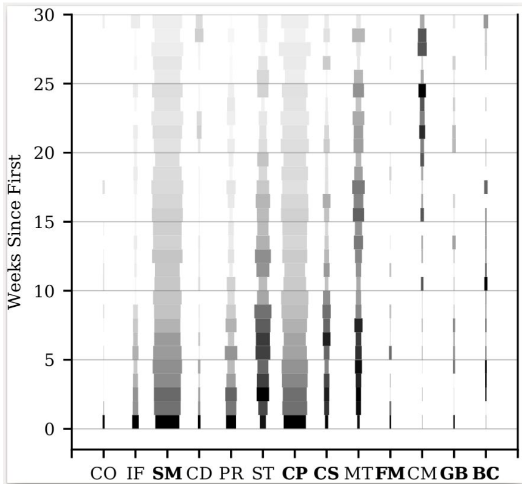
Responsibility	Occurrence	Disagreement	κ
CO	1.3%	2.3%	0.00
IF	7.5%	17.0%	0.06
CD	3.4%	6.1%	0.21
PR	14.4%	26.2%	0.22
ST	20.4%	32.9%	0.15
CS	9.2%	12.9%	**0.43**
SM	52.4%	16.7%	**0.57**
CP	46.6%	26.8%	**0.45**
MT	12.3%	22.9%	0.13
FM	1.8%	2.6%	**0.42**
CM	5.0%	7.4%	0.32
GB	2.6%	4.8%	**0.42**
BC	2.6%	4.4%	**0.44**
Overall	96.19%	85.2%	0.10

Taxonomy & Annotation
-4 rounds of codebook iteration
-IRR: 20 sites (471 updates)
-Support management split into CS and SM
-Disagreement discussion process
-25% of discussed disagreements were irresolvable
-105 sites annotated

Kappa Statistic	*Strength of Agreement*
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

(Landis & Koch, 1977)

Patient responsibility operationalization



Patient responsibility classification

Resp.	P	R	F1
CS	0.75	0.83	0.80
SM	0.93	0.98	0.95
CP	0.90	0.97	0.93
FM	0.47	0.92	0.58
GB	0.19	0.87	0.68
BC	0.32	0.41	0.34
Mean	0.89	0.96	0.92
B_{SA}	0.70	0.86	0.77
B_{FM}	0.72	0.99	0.80

ML Classifier

Class	k=10				k=100			
	Train		Test		Train		Test	
Label	R	F1	R	F1	R	F1	R	F1
CS	.19	.32	.04	.08	.87	.93	.14	.16
SM	.34	.50	.30	.46	.90	.95	.73	.81
CP	.22	.36	.20	.32	.79	.88	.58	.68
FM	.47	.64	.07	.11	.95	.97	.09	.09
GB	.39	.56	.00	.00	.99	.99	.05	.08
BC	.30	.46	.02	.02	.99	.99	.03	.03

Max-precision keywords

Class	k=10						k=100	
	Train			Test			Train	Test
Label	P	R	F1	P	R	F1	F1	F1
CS	.24	.88	.37	.23	.86	.36	.26	.26
SM	.86	.98	.92	.86	.98	.92	.93	.93
CP	.77	.99	.87	.77	.99	.87	.87	.87
FM	.22	.87	.35	.20	.77	.30	.06	.07
GB	.16	.65	.25	.12	.50	.19	.08	.08
BC	.14	.69	.23	.08	.42	.13	.08	.08

Representative keywords

Representative keyword precision is actually better than the phases, and drops in test performance are relatively marginal.

Takeaways

- Qualitative themes can be adapted for classification in similar contexts
- Choosing a taxonomy is important and hard
- Complex phenomena are hard to capture with keywords
- Linear models with many features are really effective
- Lots of boundaries makes designing unambiguous annotation codebooks challenging

Ask me about this work, HCI, grad school, etc.

Q & A

Backup

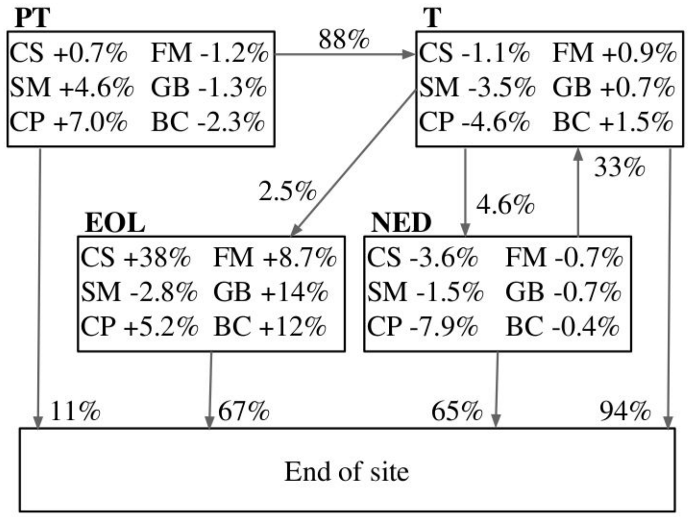
- Other details that might be useful
- (Feel free to ask me about these)

Responsibility model validation

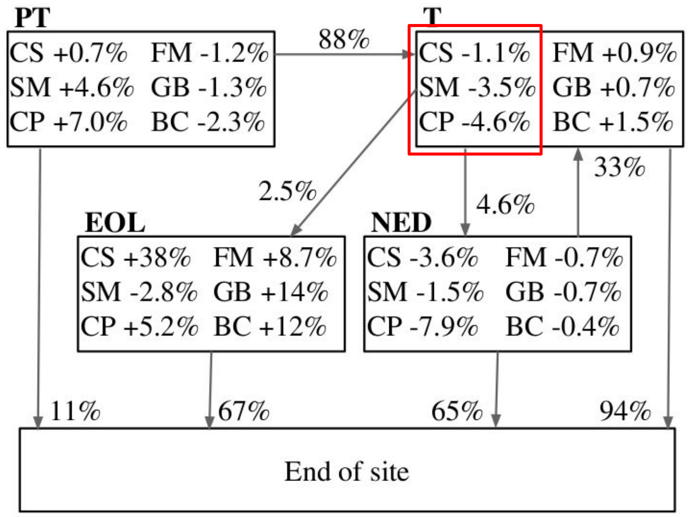
	Contains r?	Baseline rate of r	G^2 (df=30287)
CS	1.48 ± 0.06	1.031 ± 0.001	22122.01
SM	1.21 ± 0.03	1.011 ± 0.001	4460.91
CP	1.26 ± 0.03	1.011 ± 0.001	7171.64
FM	2.16 ± 0.50	1.053 ± 0.006	11078.42
GB	1.85 ± 0.22	1.043 ± 0.003	15279.27
BC	1.88 ± 0.24	1.047 ± 0.003	14357.79
Mean	1.64	1.033	—

Poisson regression
When an update is predicted to contain a responsibility, other updates in that week are predicted to contain that responsibility at a rate 1.64 times greater than if the update is predicted not to contain that responsibility.

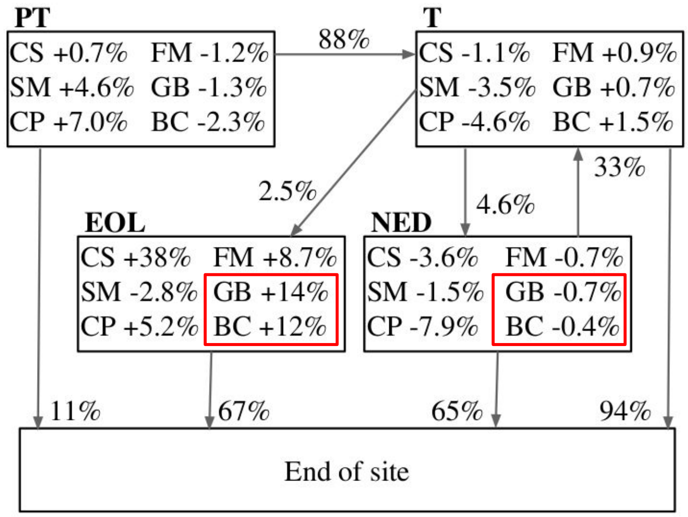
Integrating model predictions



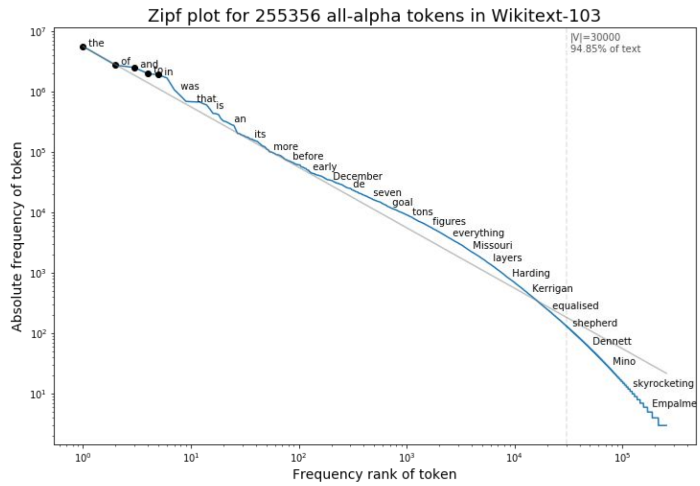
Integrating model predictions



Integrating model predictions

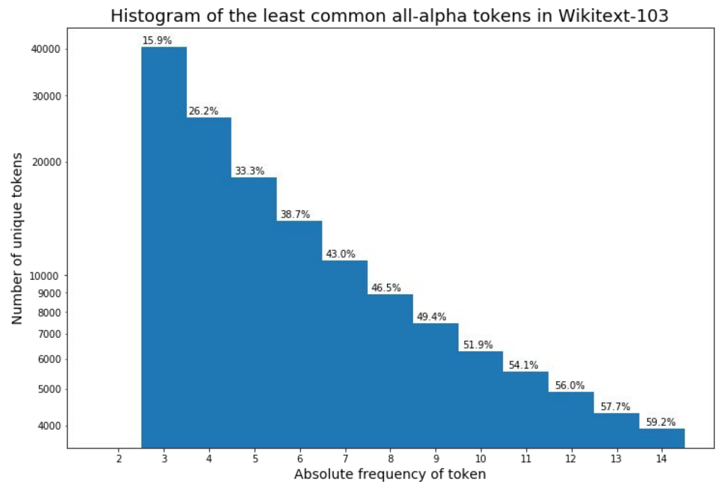


Zipf's Law



Only words that occur 3+ times!

Zipf’s Law



“cumulative percentage of unique words with this number or fewer occurrences”

Hashing Trick

» Map each word to a single column index

Hashing Trick

» Map each word to a single column index ✗
» Hash each word, use the hash as the column index
  • (Actually, hash multiple times and add the hashes to decrease collision odds)
  • (Same theory as Bloom Filters)
» Zipf's Law: When words collide, very unlikely to be two frequent words!
» No such thing as an out-of-vocab word
» Vocab can be set as small as you want
  • But collisions will start to become very frequent
» Read more: link1 link2

Software: My recommendations

- Preprocessing
  - SpaCy (Python)
  - NLTK (Python)
- Classification
  - scikit-learn (Python)
  - Vowpal Wabbit (C++)
  - SpaCy (Python)
- Topic modeling
  - Gensim (Python)
  - MALLET (Java)
  - LDAvis (R / Python)
- Visualization
  - Scattertext (Python)
  - t-SNE
- Word embeddings
  - Many existing options
  - For training: FastText, Gensim
  - For use: Gensim/SpaCy
- Lexical content
  - Empath (Python)
- Deep learning
  - PyTorch (Python)
  - Keras + Tensorflow (Python)
  - Specific options:
    - fast.ai ULMFiT
    - OpenAI Transformer

Be wary and clever!

David Mimno @dmimno · 2/10/19

The space between problems where counting words is good enough and problems that require full linguistic and cultural knowledge is much smaller than anyone expected