More About Me

Minhwa Lee

Updated Jan 2023

Research Interests

| I am interested in computational approaches to improve health equity, including but not limited to the following topics:

- Data Science for Health Equity
  - (1) Data mining for public health monitoring (e.g., social and environmental determinants of health)
  - (2) Data science for social systems and policymaking

- Equitable AI for Health
  - (1) ML/NLP for Public Health
  - (2) Social media for public health
  - (3) Debiasing current ML/NLP models
  - (4) Examining biases in social media texts/EHR, etc.

Most of my recent work

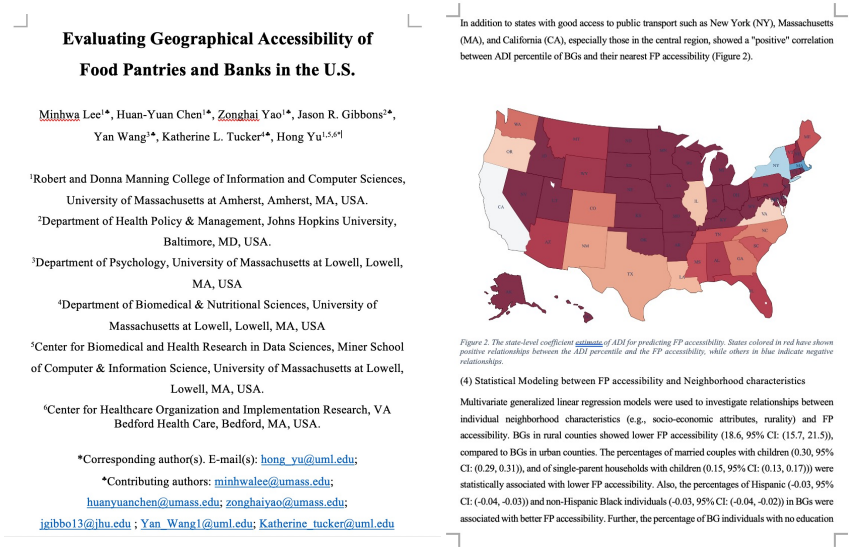
: *finding computational approaches to identify social and behavioral determinants of health*

*Social & Behavioral Determinants of Health (SBDH)
: the environmental and behavioral conditions that impact health outcomes
(e.g., socio-economic status, education, neighborhood environment, access to healthcare, substance use, alcohol, physical activity, etc.)

Research Experience

| BioNLP group, UMass Amherst (Supervisor: Prof. Hong Yu) [Aug. 2021 – Present]

- **(1)** Data Science approaches to social & behavioral determinants of health (SBDH)
  - Title: “Evaluating Geographical Accessibility of Food Pantries and Banks in the U.S.”
  - Food insecurity (FI) is a major SBDH -> examining whether current social systems that resolve FI are easily accessible.
  - Major Contribution to work
    - I led the research by defining several experimental protocols and implementing ideas with statistical modeling.
    - Conducted a large amount of literature review and currently writing the manuscript as the first author.
    - Presented findings with possible guidelines for national policymaking on improving accessibility to emergency food resources in rural areas.
    - Created a website to host interactive visuals (using Tableau) that present the findings.



Evaluating Geographical Accessibility of Food Pantries and Banks in the U.S.

Minhwa Lee[1♣], Huan-Yuan Chen[1♣], Zonghai Yao[1♣], Jason R. Gibbons[2♣], Yan Wang[3♣], Katherine L. Tucker[4♣], Hong Yu[1,5,6*]

[1]Robert and Donna Manning College of Information and Computer Sciences, University of Massachusetts at Amherst, Amherst, MA, USA.
[2]Department of Health Policy & Management, Johns Hopkins University, Baltimore, MD, USA.
[3]Department of Psychology, University of Massachusetts at Lowell, Lowell, MA, USA
[4]Department of Biomedical & Nutritional Sciences, University of Massachusetts at Lowell, Lowell, MA, USA
[5]Center for Biomedical and Health Research in Data Sciences, Miner School of Computer & Information Science, University of Massachusetts at Lowell, Lowell, MA, USA.
[6]Center for Healthcare Organization and Implementation Research, VA Bedford Health Care, Bedford, MA, USA.

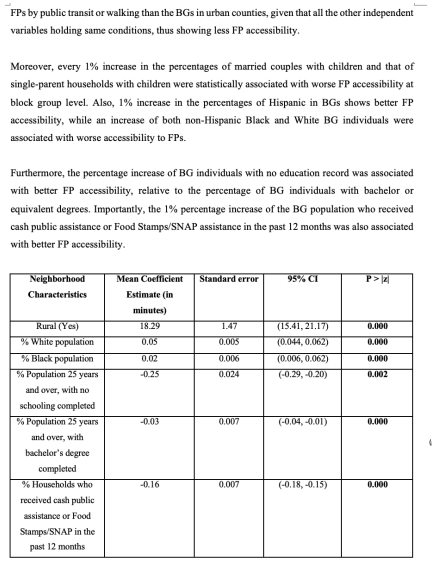
*Corresponding author(s). E-mail(s): hong_yu@uml.edu;
♣Contributing authors: minhwalee@umass.edu; huanyuanchen@umass.edu; zonghaiyao@umass.edu; jgibbo13@jhu.edu ; Yan_Wang1@uml.edu; Katherine_tucker@uml.edu

In addition to states with good access to public transport such as New York (NY), Massachusetts (MA), and California (CA), especially those in the central region, showed a "positive" correlation between ADI percentile of BGs and their nearest FP accessibility (Figure 2).

Figure 2. The state-level coefficient estimate of ADI for predicting FP accessibility. States colored in red have shown positive relationships between the ADI percentile and the FP accessibility, while others in blue indicate negative relationships.

(4) Statistical Modeling between FP accessibility and Neighborhood characteristics

Multivariate generalized linear regression models were used to investigate relationships between individual neighborhood characteristics (e.g., socio-economic attributes, rurality) and FP accessibility. BGs in rural counties showed lower FP accessibility (18.6, 95% CI: (15.7, 21.5)), compared to BGs in urban counties. The percentages of married couples with children (0.30, 95% CI: (0.29, 0.31)), and of single-parent households with children (0.15, 95% CI: (0.13, 0.17))) were statistically associated with lower FP accessibility. Also, the percentages of Hispanic (-0.03, 95% CI: (-0.04, -0.03)) and non-Hispanic Black individuals (-0.03, 95% CI: (-0.04, -0.02)) in BGs were associated with better FP accessibility. Further, the percentage of BG individuals with no education

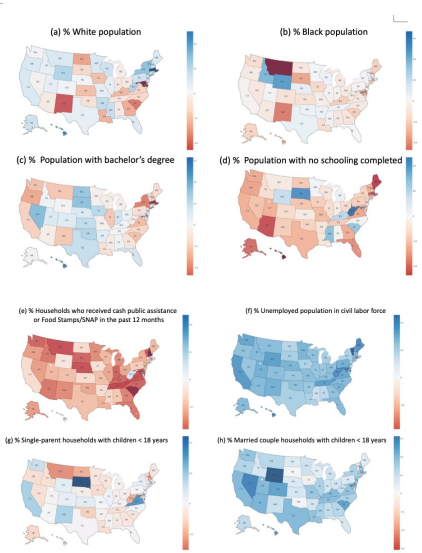


FPs by public transit or walking than the BGs in urban counties, given that all the other independent variables holding same conditions, thus showing less FP accessibility.

Moreover, every 1% increase in the percentages of married couples with children and that of single-parent households with children were statistically associated with worse FP accessibility at block group level. Also, 1% increase in the percentages of Hispanic in BGs shows better FP accessibility, while an increase of both non-Hispanic Black and White BG individuals were associated with worse accessibility to FPs.

Furthermore, the percentage increase of BG individuals with no education record was associated with better FP accessibility, relative to the percentage of BG individuals with bachelor or equivalent degrees. Importantly, the 1% percentage increase of the BG population who received cash public assistance or Food Stamps/SNAP assistance in the past 12 months was also associated with better FP accessibility.

Neighborhood Characteristics	Mean Coefficient Estimate (in minutes)	Standard error	95% CI	P > |z|
Rural (Yes)	18.29	1.47	(15.41, 21.17)	**0.000**
% White population	0.05	0.005	(0.044, 0.062)	**0.000**
% Black population	0.02	0.006	(0.006, 0.062)	**0.000**
% Population 25 years and over, with no schooling completed	-0.25	0.024	(-0.29, -0.20)	**0.002**
% Population 25 years and over, with bachelor's degree completed	-0.03	0.007	(-0.04, -0.01)	**0.000**
% Households who received cash public assistance or Food Stamps/SNAP in the past 12 months	-0.16	0.007	(-0.18, -0.15)	**0.000**



(Left)
Screenshots of the draft that I am currently working on.

Research Experience

| BioNLP group, UMass Amherst [Aug. 2021 – Present]

- **(1)** Data Science approaches to social & behavioral determinants of health (SBDH)

  - Procedures
    - Analyzed the spatial distributions of food pantries (FPs) (per state, county, and block groups).
    - Computed the travel time from any block group to its nearest FP by public transit, using Google Map API.
    - Conducted generalized linear regression analysis between the deprivation indices of block groups and their travel time to FP.
    - Developed other regression models that examine the relationship between the neighborhood characteristics of block groups and FP accessibility, using the Statsmodels package.
      - (1) rurality of block groups.
      - (2) rurality + census characteristics (e.g., % black population per block group, % households of receiving SNAP per block group, etc.).

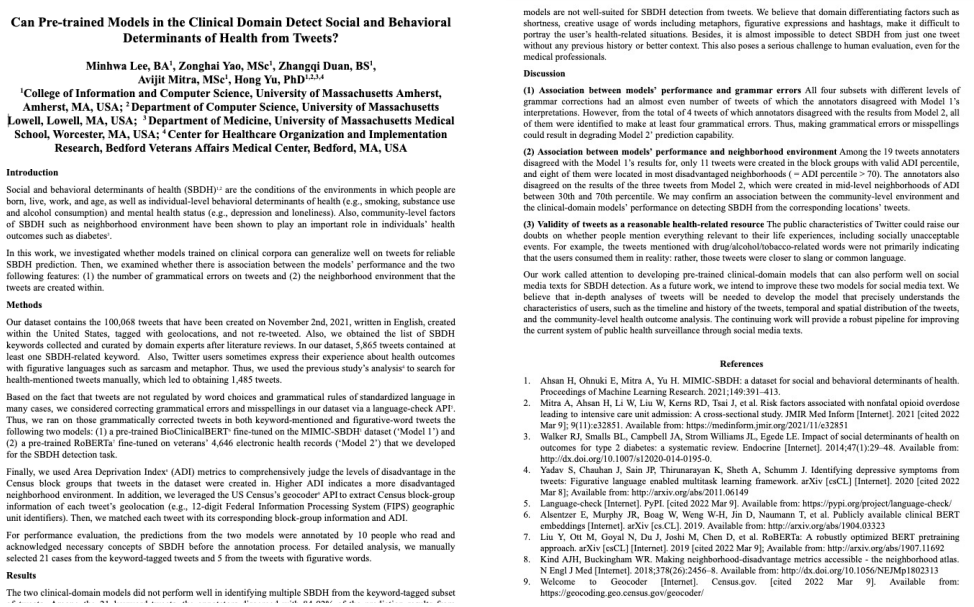
    - Urged policy reforms on re-distributing FPs equitably based on the availability of public transit systems, particularly in rural regions.

Research Experience

| BioNLP group, UMass Amherst [Aug. 2021 – Present]

- **(2)** Natural Language Processing (NLP) approaches to social & behavioral determinants of health (SBDH)
  - Title: “Can Pre-trained clinical language models Detect SBDH from Tweets?”
  - Major Contribution to work

    - Led the project, under the supervision by a Ph.D. student in the BioNLP group.
    - Tested the existing language models in the clinical domain on our collection of tweets.
    - Discovered a need for developing a robust NLP model for the SBDH identification task.
    - Submitted the work to AMIA 2022 Annual Symposium as the first author.



Can Pre-trained Models in the Clinical Domain Detect Social and Behavioral Determinants of Health from Tweets?

Minhwa Lee, BA[1], Zonghai Yao, MSc[1], Zhangqi Duan, BS[1], Avijit Mitra, MSc[1], Hong Yu, PhD[1,2,3,4]
[1]College of Information and Computer Science, University of Massachusetts Amherst, Amherst, MA, USA; [2]Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, USA; [3]Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA; [4]Center for Healthcare Organization and Implementation Research, Bedford Veterans Affairs Medical Center, Bedford, MA, USA

Introduction

Social and behavioral determinants of health (SBDH)[1,2] are the conditions of the environments in which people are born, live, work, and age, as well as individual-level behavioral determinants of health (e.g., smoking, substance use and alcohol consumption) and mental health status (e.g., depression and loneliness). Also, community-level factors of SBDH such as neighborhood environment have been shown to play an important role in individuals' health outcomes such as diabetes[3].

In this work, we investigated whether models trained on clinical corpora can generalize well on tweets for reliable SBDH prediction. Then, we examined whether there is association between the models' performance and the two following features: (1) the number of grammatical errors on tweets and (2) the neighborhood environment that the tweets are created within.

Methods

Our dataset contains the 100,068 tweets that have been created on November 2nd, 2021, written in English, created within the United States, tagged with geolocations, and not re-tweeted. Also, we obtained the list of SBDH keywords collected and curated by domain experts after literature reviews. In our dataset, 5,865 tweets contained at least one SBDH-related keyword. Also, Twitter users sometimes express their experience about health outcomes with figurative languages such as sarcasm and metaphor. Thus, we used the previous study's analysis[4] to search for health-mentioned tweets manually, which led to obtaining 1,485 tweets.

Based on the fact that tweets are not regulated by word choices and grammatical rules of standardized language in many cases, we considered correcting grammatical errors and misspellings in our dataset via a language-check API[5]. Thus, we ran on those grammatically corrected tweets in both keyword-mentioned and figurative-word tweets the following two models: (1) a pre-trained BioClinicalBERT[6] fine-tuned on the MIMIC-SBDH[7] dataset ('Model 1') and (2) a pre-trained RoBERTa[8] fine-tuned on veterans' 4,646 electronic health records ('Model 2') that we developed for the SBDH detection task.

Finally, we used Area Deprivation Index[9] (ADI) metrics to comprehensively judge the levels of disadvantage in the Census block groups that tweets in the dataset were created in. Higher ADI indicates a more disadvantaged neighborhood environment. In addition, we leveraged the US Census's geocoder[9] API to extract Census block-group information of each tweet's geolocation (e.g., 12-digit Federal Information Processing System (FIPS) geographic unit identifiers). Then, we matched each tweet with its corresponding block-group information and ADI.

For performance evaluation, the predictions from the two models were annotated by 10 people who read and acknowledged necessary concepts of SBDH before the annotation process. For detailed analysis, we manually selected 21 cases from the keyword-tagged tweets and 5 from the tweets with figurative words.

Results

The two clinical-domain models did not perform well in identifying multiple SBDH from the keyword-tagged subset

models are not well-suited for SBDH detection from tweets. We believe that domain differentiating factors such as shortness, creative usage of words including metaphors, figurative expressions and hashtags, make it difficult to portray the user's health-related situations. Besides, it is almost impossible to detect SBDH from just one tweet without any previous history or better context. This also poses a serious challenge to human evaluation, even for the medical professionals.

Discussion

(1) Association between models' performance and grammar errors All four subsets with different levels of grammar corrections had an almost even number of tweets of which the annotators disagreed with Model 1's interpretations. However, from the total of 4 tweets of which annotators disagreed with the results from Model 2, all of them were identified to make at least four grammatical errors. Thus, making grammatical errors or misspellings could result in degrading Model 2' prediction capability.

(2) Association between models' performance and neighborhood environment Among the 19 tweets annotaters disagreed with the Model 1's results for, only 11 tweets were created in the block groups with valid ADI percentile, and eight of them were located in most disadvantaged neighborhoods (= ADI percentile > 70). The annotators also disagreed on the results of the three tweets from Model 2, which were created in mid-level neighborhoods of ADI between 30th and 70th percentile. We may confirm an association between the community-level environment and the clinical-domain models' performance on detecting SBDH from the corresponding locations' tweets.

(3) Validity of tweets as a reasonable health-related resource The public characteristics of Twitter could raise our doubts on whether people mention everything relevant to their life experiences, including socially unacceptable events. For example, the tweets mentioned with drug/alcohol/tobacco-related words were not primarily indicating that the users consumed them in reality: rather, those tweets were closer to slang or common language.

Our work called attention to developing pre-trained clinical-domain models that can also perform well on social media texts for SBDH detection. As a future work, we intend to improve these two models for social media text. We believe that in-depth analyses of tweets will be needed to develop the model that precisely understands the characteristics of users, such as the timeline and history of the tweets, temporal and spatial distribution of the tweets, and the community-level health outcome analysis. The continuing work will provide a robust pipeline for improving the current system of public health surveillance through social media texts.

References

1. Ahsan H, Ohnuki E, Mitra A, Yu H. MIMIC-SBDH: a dataset for social and behavioral determinants of health. Proceedings of Machine Learning Research. 2021;149:391–413.
2. Mitra A, Ahsan H, Li W, Liu W, Kerns RD, Tsai J, et al. Risk factors associated with nonfatal opioid overdose leading to intensive care unit admission: A cross-sectional study. JMIR Med Inform [Internet]. 2021 [cited 2022 Mar 9]; 9(11):e32851. Available from: https://medinform.jmir.org/2021/11/e32851
3. Walker RJ, Smalls BL, Campbell JA, Strom Williams JL, Egede LE. Impact of social determinants of health on outcomes for type 2 diabetes: a systematic review. Endocrine [Internet]. 2014;47(1):29–48. Available from: http://dx.doi.org/10.1007/s12020-014-0195-0.
4. Yadav S, Chauhan J, Sain JP, Thirunarayan K, Sheth A, Schumm J. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. arXiv [csCL] [Internet]. 2020 [cited 2022 Mar 8]; Available from: http://arxiv.org/abs/2011.06149
5. Language-check [Internet]. PyPI. [cited 2022 Mar 9]. Available from: https://pypi.org/project/language-check/
6. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings [Internet]. arXiv [cs.CL]. 2019. Available from: http://arxiv.org/abs/1904.03323
7. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv [csCL] [Internet]. 2019 [cited 2022 Mar 9]; Available from: http://arxiv.org/abs/1907.11692
8. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the neighborhood atlas. N Engl J Med [Internet]. 2018;378(26):2456–8. Available from: http://dx.doi.org/10.1056/NEJMp1802313
9. Welcome to Geocoder [Internet]. Census.gov. [cited 2022 Mar 9]. Available from: https://geocoding.geo.census.gov/geocoder/

(Left)
Screenshots of the submitted abstract

Research Experience

| BioNLP group, UMass Amherst [Aug. 2021 – Present]

- **(2)** Natural Language Processing (NLP) approaches to social & behavioral determinants of health (SBDH)

    - But… Why SBDH on Twitter?
        - Openly available data, thus a good resource for public health surveillance.
        - Several previous literature on applying NLP to health-related tweets.
            - However, there is little prior work on identifying SBDH from general tweets.

Research Experience

| BioNLP group, UMass Amherst [Aug. 2021 – Present]

- **(2)** NLP approaches to social & behavioral determinants of health (SBDH)

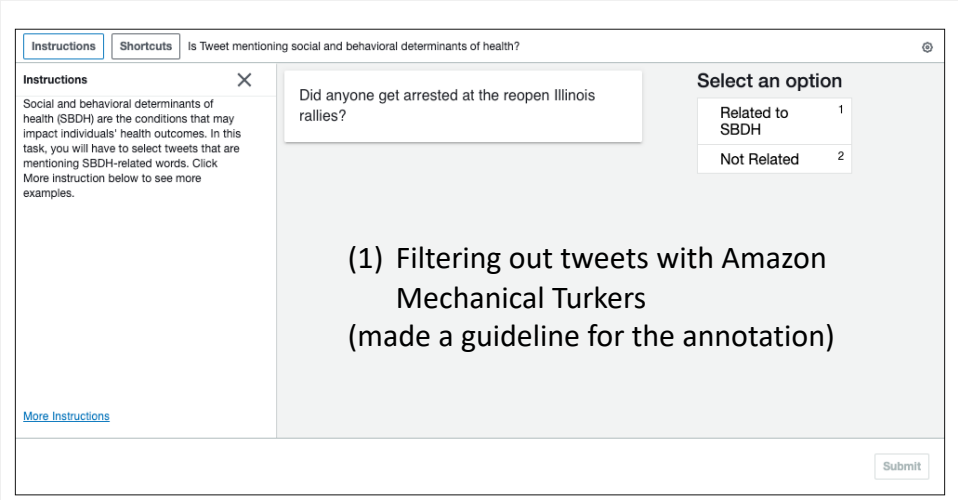
  - Procedures
    - Collected 100K English-written tweets that were created within the U.S. (geo-tagged tweets), using Twarc2 API.
    - Selected SBDH-related keywords (curated by experts) to be used in the search query of Twarc2 API.
    - Manually selected SBDH-related tweets that used figurative language (e.g., metaphor and sarcasm) and tweets that exactly mentioned the keywords.
    - Corrected grammatical errors and misspellings in our tweets by using language-check API.
    - Analyzed the spatial distributions of the collected tweets (e.g., # of tweets that were created in higher/lower socio-economic block groups in the U.S., etc.) and language characteristics (e.g., average # of corrections, length of tweets).
    - Tested zero-shot transferability of the existing clinical language models on these collected tweets.
    - Evaluated the models' performance by manually annotating the models' results with humans.

  - Lessons
    - Though negative results (as expected, because we did not finetune the models on the tweets), we found the significance of using a longitudinal set of tweets to develop an NLP model that can effectively find SBDH from tweets.
    - Motivated to the next steps, with the lessons from this initial work.

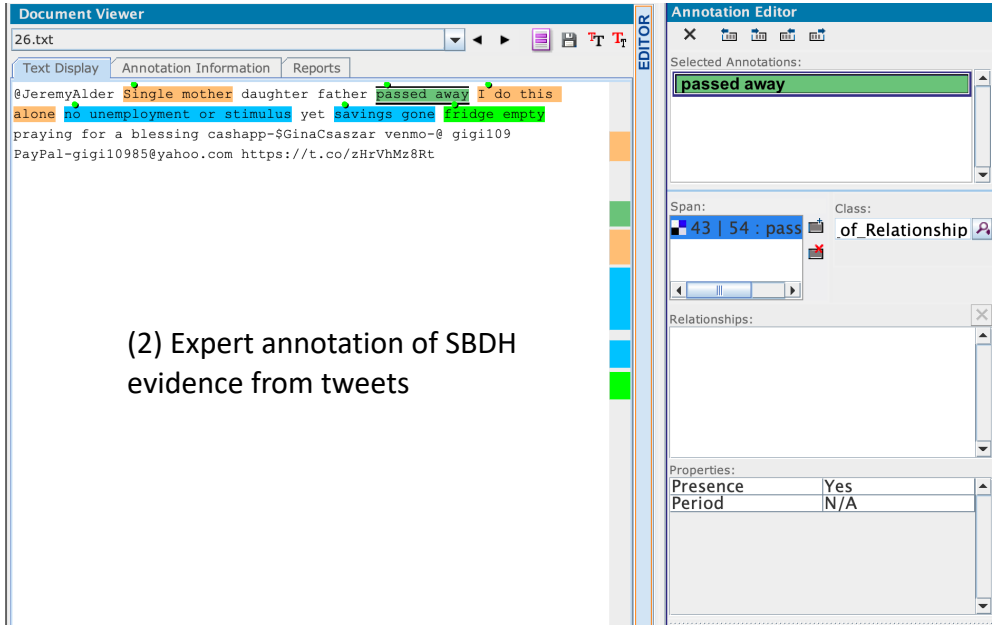
Research Experience

| BioNLP group, UMass Amherst [Aug. 2021 – Present]

- **(2)** Natural Language Processing (NLP) approaches to social & behavioral determinants of health (SBDH)
  - Current work:
    - Set up Amazon MTurk for filtering SBDH-related tweets, provided with the detailed annotation guideline.
    - Currently annotating the filtered tweets with SBDH evidence (e.g., word span, presence, and period).
    - Hydrating a longitudinal set of tweets of users who were identified from our expert annotation, using Twarc2 API.
    - Will be developing language models for the SBDH NER task.



(1) Filtering out tweets with Amazon Mechanical Turkers
(made a guideline for the annotation)



(2) Expert annotation of SBDH evidence from tweets

… and more works

Research Experience

| Honor Thesis Researcher, The College of Wooster [Aug. 2020 – April. 2021]

- **Theis Title: Statistical and Machine Learning Approaches to Depressive Disorders among adults in the U.S.: From Factor Discovery to Prediction Evaluation** (Department of Computer Science, Department of Mathematics) [Paper]

  - Major Contribution
    - Identified SBDH of depressive disorders (including major depression) among different age groups of US adults in 2018. (younger adults [18-39], middle-aged adults [40-60], and older adults [61-85]).
    - Used a tabular survey (BRFSS) from the CDC data published in 2018.
    - Leveraged supervised machine learning methods (e.g., CART, logistic regression, support vector machines) to identify factors that impact the diagnosis of depressive disorders and build predictive models using these identified SBDH for different age groups.
    - Provided the statistical foundations of the models.
    - Validated the results with statistical tests (e.g., Wald-test, drop-in-deviance test, t-test, classification metrics, etc.).
    - Wrote a 107-page thesis and passed the oral defense.

Research Experience

| Honor Thesis Researcher, The College of Wooster [Aug. 2020 – April. 2021]

- **Theis Title: Statistical and Machine Learning Approaches to Depressive Disorders among adults in the U.S.: From Factor Discovery to Prediction Evaluation** (Department of Computer Science, Department of Mathematics)

  - Honors & Awards
    - Departmental Honors (Computer Science, Mathematics)
    - Exemplar thesis
    - Honorable Mention at MAA Undergraduate Student Poster Session during JMM 2021
    - Nominated as a finalist at ACM Student Research Competition at Grace Hopper Celebration 2021

Statistical and Machine Learning Approaches to Depressive Disorders Among Adults in the United States: From Factor Discovery to Prediction Evaluation

Independent Study Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Bachelors of Arts in the Computer Science and Mathematics in the Department of Mathematical & Computational Sciences at The College of Wooster

by

Minhwa Lee

The College of Wooster

2021

Advised by:

Dr. Sofia Visa (Computer Science)

Dr. Marian Frazier (Mathematics)

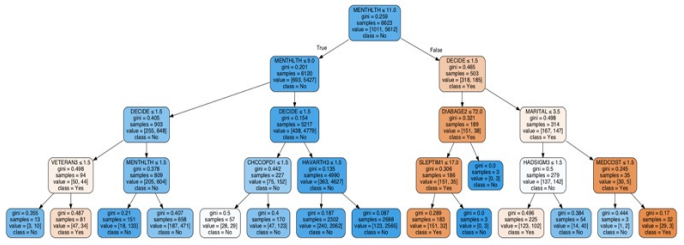
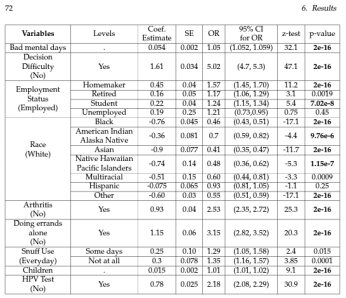


Figure 6.3: A decision tree for the old-aged adult group

6.1. Decision Trees



72 6. Results

Variables	Levels	Coef. Estimate	SE	OR	95% CI for OR	z-test	p-value
Bad mental days	.	0.054	0.002	1.05	(1.052, 1.059)	32.1	**2e-16**
Decision Difficulty (No)	Yes	1.61	0.034	5.02	(4.7, 5.3)	47.1	**2e-16**
Employment Status (Employed)	Homemaker	0.45	0.04	1.57	(1.45, 1.70)	11.2	**2e-16**
	Retired	0.16	0.05	1.17	(1.06, 1.29)	3.1	0.0019
	Student	0.22	0.04	1.24	(1.15, 1.34)	5.4	**7.02e-8**
	Unemployed	0.19	0.25	1.21	(0.73,0.95)	0.75	0.45
Race (White)	Black	-0.76	0.045	0.46	(0.43, 0.51)	-17.1	**2e-16**
	American Indian Alaska Native	-0.36	0.081	0.7	(0.59, 0.82)	-4.4	**9.76e-6**
	Asian	-0.9	0.077	0.41	(0.35, 0.47)	-11.7	**2e-16**
	Native Hawaiian Pacific Islanders	-0.74	0.14	0.48	(0.36, 0.62)	-5.3	**1.15e-7**
	Multiracial	-0.51	0.15	0.60	(0.44, 0.81)	-3.3	0.0009
	Hispanic	-0.075	0.065	0.93	(0.81, 1.05)	-1.1	0.25
	Other	-0.60	0.03	0.55	(0.51, 0.59)	-17.1	**2e-16**
Arthritis (No)	Yes	0.93	0.04	2.53	(2.35, 2.72)	25.3	**2e-16**
Doing errands alone (No)	Yes	1.15	0.06	3.15	(2.82, 3.52)	20.3	**2e-16**
Snuff Use (Everyday)	Some days	0.25	0.10	1.29	(1.05, 1.58)	2.4	0.015
	Not at all	0.3	0.078	1.35	(1.16, 1.57)	3.85	0.0001
Children	.	0.015	0.002	1.01	(1.01, 1.02)	9.1	**2e-16**
HPV Test (No)	Yes	0.78	0.025	2.18	(2.08, 2.29)	30.9	**2e-16**

Table 6.4: Coefficient estimate, standard error (SE), the odds ratio and its 95% CI, and results from Wald-Test statistics in M_{young} (Most statistically significant variables are marked as bold, as presented in Table 5.2)

- The odds of the depressive disorders for people who have difficulties in doing errands alone are 215% higher than the odds of the disorders for those who do not.
- One increase in the number of children is associated with 1% higher odds of the depressive disorders.
- The odds of the depressive disorders for people who have HPV test records are 118% higher than the odds for those who do not.

For simplicity, we interpret 95% confidence intervals for odds ratios of only two statistically significant predictors. Holding all other variables constant, we are 95% confident that young adult participants with decision-making difficulties will have between 370% and 430% higher odds of the depressive disorders diagnosis than those without. Also, we are 95% confident that the young

CHAPTER 7

Summary & Discussion

In this chapter, we summarize and discuss the important findings from the previous chapter. First, we compare and discuss our interpretations of the characteristics of U.S. adults diagnosed with depressive disorders, based on both decision trees and logistic regression models for each adult group. Then, we compare the performance of the three methods as predictors of depressive disorders diagnosis for each of the three adult groups.

7.1 Factor Discovery

7.1.1 Young adults

Decision tree	Logistic regression
- # of bad mental days per month (↑) - Decision-making issues (Yes) - Arthritis or related illnesses (Yes) - Difficulty of doing errands alone (Yes) - # of children in household (↑)	
- Employment status (student, homemaker, unemployed) - Tobacco & Snuff history (Yes)	- Employment status (student, homemaker) - Race (White) - HPV test records (Yes)

Table 7.1: The most important variables for the **young adult group**, selected by decision trees and logistic regression: when holding all other conditions constant, categories in parentheses for each variable in the table are relevant factors that increase the likelihood of a depressive disorders diagnosis for the young adults.

Both the decision tree model and the logistic regression model ('M_{young}') present that the following five variables are relevant factors of the depressive disorders diagnosis among the U.S. young adult group in the 2018 BRFSS sample: the number of bad mental days, decision-making difficulty,

90

(Left)
Screenshots of the honor thesis

Research Experience

| Sophomore Research Assistant, The College of Wooster [Fall 2018, Fall 2019]

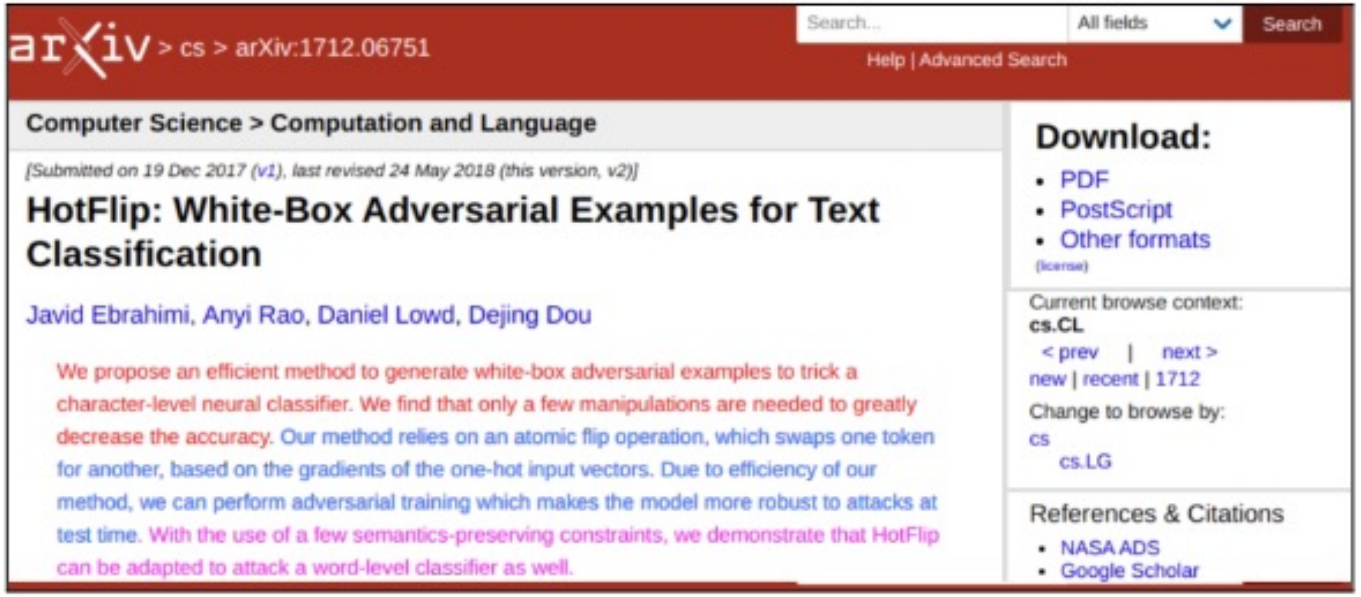
- **A short-term research program in which students work as paid research assistants to Wooster faculty members.**
- Only available to students in the second semester of first year through the first semester of their junior year.
- Participated in Fall 2018 (with Prof. Hyokyeong Lee) and Fall 2019 (with Prof. Robert Kelvey).

    - Responsible Tasks
        - Pre-processed genomic dataset for Prof. Lee's research, using R.
        - Designed and implemented a computational algorithm for Prof. Kelvey's research on combinatorics, using Python.

Other Experience

| Graduate Student Researcher, UMass Amherst [Jan. 2022 – May. 2022]

- Collaborated with Bloomberg as an industry-affiliated research course at UMass Amherst
  - Course Name: COMPSCI 696DS: IS – Data Science
- Topic: Sequential Sentence Classification for Longer Documents



(Excerpted from the project report)

Figure 1: An example of how sequential sentence classification would be performed on an abstract from the CS Abstruct dataset. The lines in red, blue and pink belong to background, method and result respectively.

Other Experience

| Graduate Student Researcher, UMass Amherst [Jan. 2022 – May. 2022]

- Major Contribution
  - Collected five datasets (e.g., PubMed, MIMIC-III discharge notes, full scientific papers, short CS paper abstracts, etc.).
  - Analyzed the patterns of label transitions of sequential sentences in the datasets.
  - Extended the previous work of Cohan et al.*** to longer documents.
  - Developed a long-range language model using Longformer* and SciBERT** to label sequential sentences.
  - Tested the applications of transfer learning on a few-shot setup.
  - Had a weekly meeting with Dr. Yuval Marton (industry mentor) and a Ph.D. mentor at UMass Amherst.
  - Presented the findings to Prof. Andrew McCallum (course instructor) and Dr. Marton (Bloomberg) at the end of the semester.
  - Received a grade of A on the course.

*Iz Beltagy, Kyle Lo, and Arman Cohan. 2019b. Scibert: A pretrained language model for scientific text.

**Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150.

***Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Other Experience

| Organization Committee, UMass Amherst [Jan. 2022 – May. 2022]

- Interested in promoting diversity & inclusion in Computer Science fields.
- Hosted an annual event: Voices of Data Science 2022 at UMass Amherst.
  - A student-run tech conference that provides a platform to broadcast the voices of data scientists, especially from underrepresented communities.

- Responsibilities
  - Presented talks from current women and non-binary data scientists working in application areas including healthcare, social sciences, etc.
  - Held a networking session for current UMass students.
  - Led to a successful event with 157 participants.
  - Managed logistics for technical troubleshooting on the remote platform as a logistics head.

Me
(with other committee members)

Featured in the UMass Newsletter😊

Voices of Data Science

Voices of Data Science (VODS), to be held online March 25, 2022, is a student-run tech conference that aims to amplify the voices of data scientists, especially those from traditionally underrepresented communities. With the goal of building a strong and inclusive network of data scientists with a shared vision to use data science for the common good, the conference works to highlight the work of women (cis and trans) and non-binary data scientists. Conference organizers welcome and value attendees of all identities.

The conference will feature talks from data scientists working in application areas including healthcare, the social sciences, and business, as well as a panel discussion, "Navigating the World of Data Science." The panel is especially for aspiring data scientists who want to learn academic and career pathways from data scientists in the field. Participants can also take advantage of a networking hour with fellow attendees, speakers, sponsors, and panelists.

VODS is a free and inclusive event that invites attendence from data science enthusiasts from all over the world. The leadership team (pictured below) would like to thank Erika Dawson-Head, Jennifer Shiao, Thomas Bernardin, Elena Hayes, and Pracheta Amarnath for their guidance and support.

I appreciate your time and consideration.
Thank you!

Best,
Minhwa Lee