Natural Language Processing in the Context of Qualitative Research

Eungang (Peter) Choi

Aug. 07, 2022

Table of Contents

Introduction

Expectations

- What I am NOT trying to do:
  - Define qualitative research
  - Promote NLP for all qualitative research
  - Argue that using NLP is qualitative research
  - Argue that humans are useless, and that AI will take over the world

- What I AM trying to do:
  - Introduce NLP tools
  - Show how it can help/assist *some* qualitative research
  - Demonstrate that human intervention is key in using NLP for research

Motivation

- Rise in big data and computational social science (CSS), has opened up opportunities for social scientists to conduct research in innovative ways (Lazer et al., 2009)
- Among the many developments in CSS that has impacted researchers in the social science domain, text-as-data is gaining popularity (Grimmer et al., 2022)
- Using text as data:
  - Causal inference (Egami et al., 2018)
  - Examine culture (Bail, 2014)
  - Study the public discourse (Bail, 2016)

What is NLP?

- Natural Language Processing (NLP) focuses on making computers analyze (understand) and use (speak and write) text and languages as humans would.
  - AltaVista → Google
  - Siri, Alexa
- Qualitative research?
  - Understand and analyze text data to infer findings
  - How can NLP benefit qualitative researchers?

Human Intervention is Key

- NLP to assist researchers with Qualitative research
- Not do everything for you automatically (Unsupervised learning)
- But do the tasks that you have set under the terms you have set (Supervised learning)

Experimenting with NLP in Qualitative Research

Data & Methods

- Study on Eviction (Kepes and Kempler, Forthcoming)
- 14 Interview transcripts (each interview ranging from 20 ~ 90 mins)
- Two coders (1Month)

- How can NLP be used in Qualitative Research?
  - Case 1: Did not collect the data, not aware of what to expect from the text.
  - Case 2: Conducted the interviews, collected all the data, know exactly what to expect.

Case 1: Finding Topics with NLP

- Received the data with minimal knowledge on what the interviews talked about.
- Can I detect the same topics (codes) as the qualitative researchers do?
- How fast and accurate (close to the codebook the researchers established) is it?

NLP

- 'Corex algorithm' (Gallagher et. al., 2017)
  - looks at the correlation of all the content to produce the optimum number of topics
  - Human Intervention 1: deciding if the proposed number of topics is good. If not make changes.
  - Human Intervention 2: Label each topic using the keywords for each topic the model provides
  - Human Intervention 3: Add 'anchor words' (what should belong in the topic) to each topic.

Case 1: Results

Intervention 2

Eviction Court
Strive Efforts
Understanding the System
Paying Rent Landlord
Community

Intervention 1

```
Total correlation: 14.014663715253612
Topic #1: court, legal aid, legal, work, aid, need, case, right, everybody, talk
Topic #2: try, call, time, even, stuff, happen, somebody, ive, lot people, day
Topic #3: lot, system, really, kind, gon na, gon, theres, na, explain, look
Topic #4: rent, pay, money, pay rent, landlord, theyre, youre, impact, tenant, fi
Topic #5: thing, person, eviction, one, sort, say, different, whatever, communit
```

Intervention 3

```
anchors = [
    ['evicition', 'court', 'legal aid'],
    ['try', 'call'],
    ['system', 'explain'],
    ['rent', 'pay', 'landlord', 'money', 'pay rent'],
    ['person', 'community', 'belong', 'person', 'people']
]
```

Case 1: Compare with Qual Researchers

- Overall
  - Not as granular as the researchers but identified similar topics/codes
    - Researchers: 12 Main Codes vs. NLP: 5 Main Codes
  - Researchers had sub-codes for further distinctions, NLP model did not
  - Time: Two researchers: 1month vs. NLP: 3hours

- Shortcomings
  - Cannot have multiple codes for a single sentence
  - Limitations in going into further details
  - Potentially could be overcome by running multiple iterations (re-running the model for each topic)

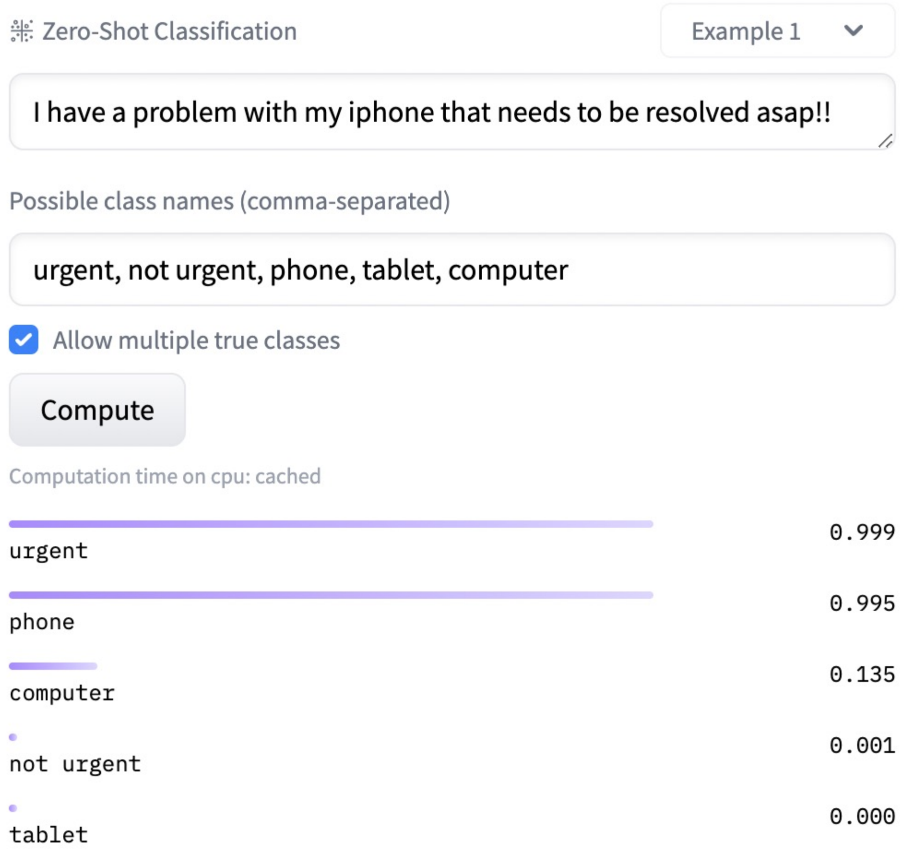
Case 2: Analyze Data With Codebook

- With access to the codebook, how efficiently can NLP analyze large amounts of data efficiently and accurately?

NLP

- Zero-shot classification (Bujel, Yannakoudakis, and Marek, 2021)
  - Sentence-level labeler using a pretrained model.
  - Input: list of expected labels (Human Intervention)
  - Output: labeled sentences

Case 2: Zero-shot Classification (Example)



Input: List of potential labels (codes)

- Urgent
- Not urgent
- Phone
- Tablet
- Computer

Intervention

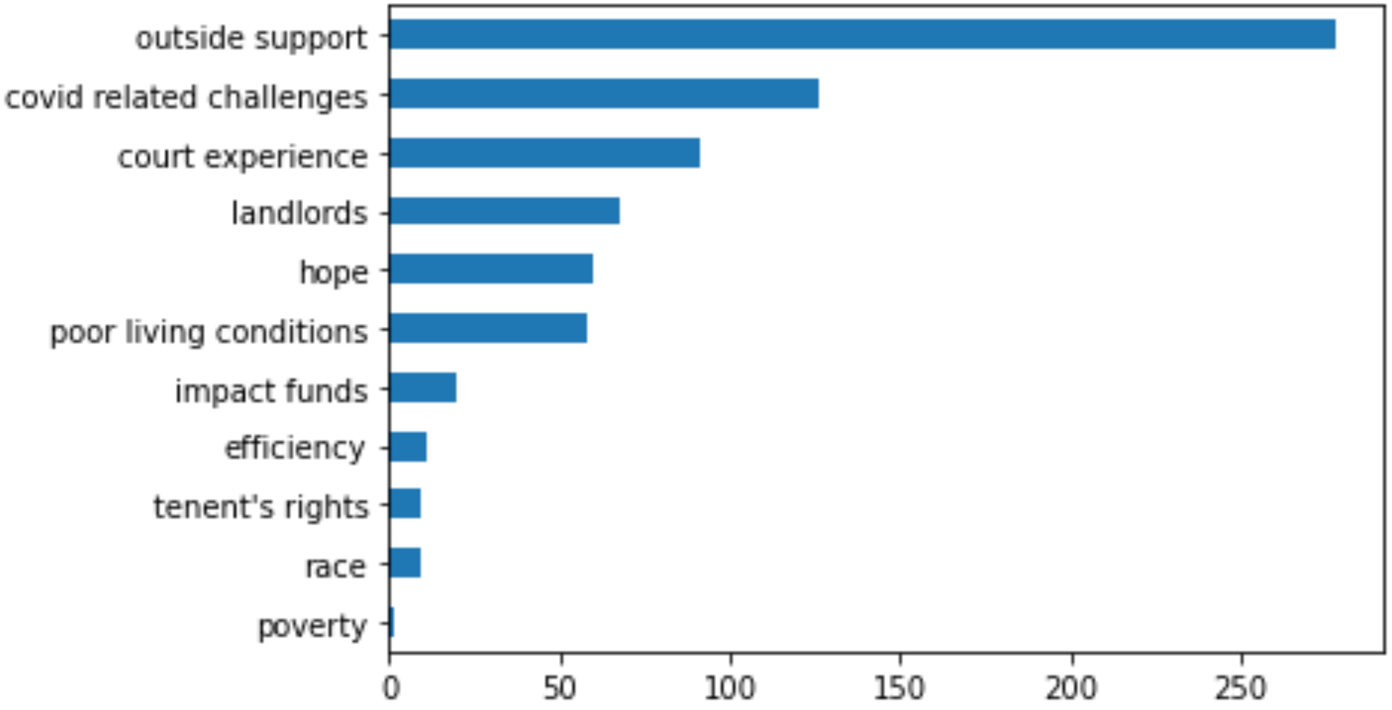
Data

"I have a problem with my iphone that needs to be resovled asap!!"

Results: Probability of the label (code) that the data corresponds to.

Case 2: Results

- While direct comparison of the results from researchers is not possible, it was able to detect and label each sentence in correspondence to the codebook the researchers set
- Total runtime: 1hour



Discussion

Overview of Results

- Is it the ultimate solution to qualitative research?
  - No. Not perfect and human intervention is very important
- Can it help analyze text data fast?
  - Very fast. It will make it possible for qualitative researchers to work with vast amounts of data.
- Barriers to Entry?
  - Coding knowledge needed (Python or R)
  - Lots of resources out there.
  - Its Free! (open-source)

More Use, Better Fitting for Researchers

- Whether you like it or not, these are methods that are used widely in everyday lives (led by tech companies)
- Relatively less use is focused on research – especially qualitative research.
- With more use, the better the models will be fitting for research uses
- Better control over how these models work and impact society

Bibliography

Christopher A Bail. The cultural environment: Measuring culture with big data. Theory and Society, 43 (3-4):465–482, 2014.

Christopher Andrew Bail. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. Proceedings of the National Academy of Sciences, 113(42):11823–11828, October 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1607151113. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1607151113.

Bujel, Kamil, Helen Yannakoudakis, and Marek Rei. "Zero-shot Sequence Labeling for Transformer-based Sentence Classifiers." arXiv preprint arXiv:2103.14465 (2021).

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. arXiv preprint arXiv:1802.02163, 2018.

Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. Transactions of the Association for Computational Linguistics, 5:529–542, 2017.

Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. Text as data: A new framework for machine learning and the social sciences. Princeton University Press, 2022.

Kepes, Jacob and Kempler Alex. COVID-19 and Evictions in Columbus, Ohio: A Mixed Methods Approach, 2022. *Forthcoming*

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-L´aszl´o Barab´asi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. Science, 323(5915):721–723, 2009.

Amir Marvasti. Qualitative research in sociology. Sage, 2004.

Thank you.