


Analyzing radio data for epidemic surveillance

Benjamin Q. Huynh
Biomedical Informatics
Stanford University
November 13, 2019

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Introduction

- Joint work between Stanford University, UN Global Pulse, and WHO
 - Goal: How to leverage radio data for epidemic intelligence.
- About me:
 - Biostatistician by training.
 - Focus on public health + artificial intelligence.
- This is early and ongoing work that is subject to change.
 - Questions/comments/suggestions very much welcome.

Background: Radio as a source of information

- Radio is a principal source of communication in remote and rural areas across the world.
 - Mobile phones + community radio stations = a form of social media.
 - Growing community radio stations + low internet connectivity = radio will still be here in the future
- Radio data is thus a valuable source of information for epidemic intelligence.

Background: Radio in Uganda

- More than half of Ugandans rely on radio as their primary source of information.
 - ~25,000 daily callers into radio talk stations
- Previous work by United Nations Global Pulse:
 - Speech recognition algorithms applied to radio data
 - Result: daily transcriptions of radio data across Uganda

Goal

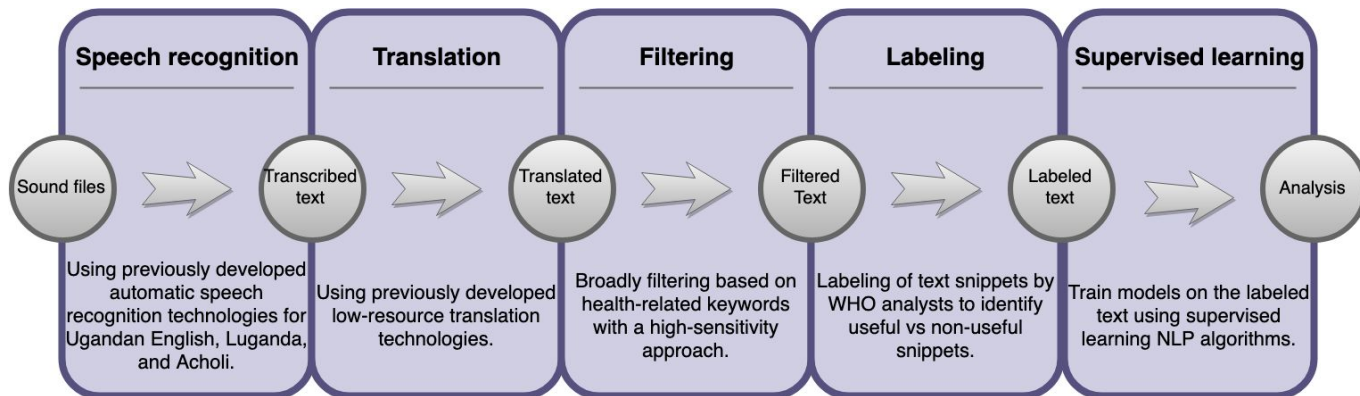
- Given transcribed radio data in Uganda, can we use it to improve epidemic surveillance?
 - How do we incorporate it into EIOS?
 - How do we find the useful information amongst the noise?
- This is an iterative process:
 - Project aims and proposed methods will continually shift based on feedback between Stanford, UNGP, and WHO.

Data

- Our dataset:
 - 685,358 transcriptions of 5-minute radio snippets
 - Dates range from 2015 to 2019
 - Languages: Luganda (71%), Acholi (11%), English (18%)
- About 10-15% of snippets include disease related terms.
 - Cholera, malaria, hemorrhagic fever, etc.

Methods: Overview

- Two-step approach:
 - Step 1: Filtering with EIOS system of keyword patterns
 - Step 2: Labeling of new training set based on prior filtering
 - Step 3: Classification via machine learning



Methods: Keyword patterns

- Initial approach: treat radio snippets the way EIOS treats news articles.
 - Categorization of snippets using keyword patterns
- Problem: radio snippets are messier than news articles!
 - Possible errors in transcription or translation
 - Many false positives: some conversations about disease might not be relevant for epidemic intelligence.

Methods: Labeling

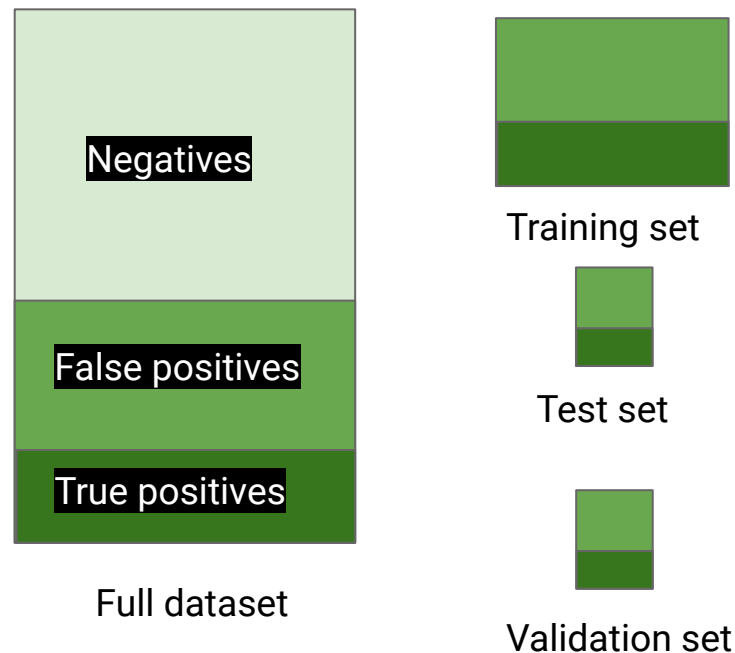
- Machine learning can be used to identify relevant vs irrelevant snippets.
 - Machine learning algorithms require **labeled** training data.
 - Our raw dataset does not come with useful labels for our task.
- Subject matter experts/WHO analysts are great at determining which articles are relevant and which aren't.
 - Keyword patterns can't do this on their own very well.
- Task: have subject matter experts label the filtered snippets as relevant/irrelevant.

Methods: Machine learning

- With labeled data, machine learning algorithms can mimic subject matter experts.
- Use supervised learning to identify relevant snippets.
- We are predicting $y = f(x)$, where y is the label, and x is the set of words in a snippet.
 - Each observation is a transcript of a 5-minute snippet.
- Will try a variety of models:
 - Iterate from simpler models to more complex ones.
 - Naive bayes, support vector machines, deep neural networks, etc.

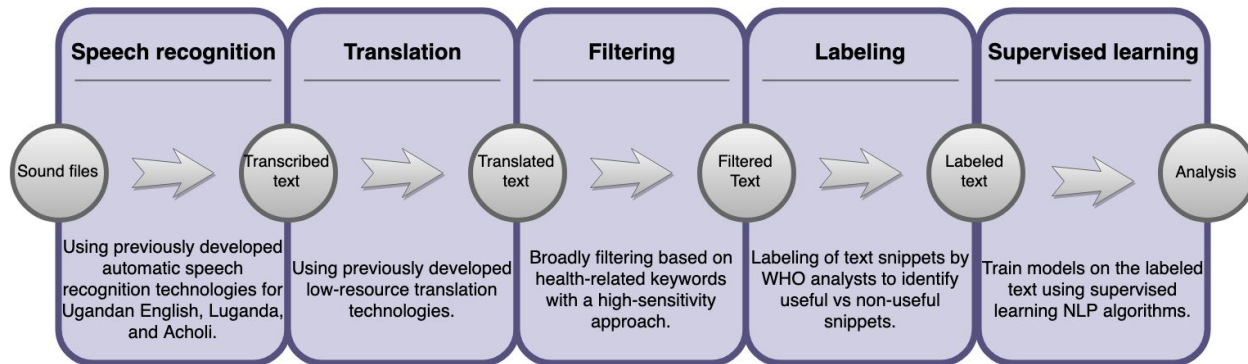
Evaluation

- Evaluate models as binary classification: relevant vs irrelevant.
- Standard metrics of evaluation: area under ROC curve, accuracy, sensitivity, specificity, etc.
- Split the data into training, test, and validation sets for model tuning.



Conclusion

- Radio data provide unique opportunity for epidemic intelligence.
- Messy, unstructured nature of the data requires machine learning to filter it out.
- Ongoing, iterative process with WHO/UN Global Pulse to figure out most useful way to improve + integrate into EIOS.



Thank you!

- Questions, comments, and/or suggestions very much welcome.
- Contact info
 - E: benhuynh@stanford.edu
 - T: @benqhuynh

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE – 1656518.



Stanford University