Coding And Tagging of Qualitative Interview Transcripts Using a Natural Language Processing-Based Machine Learning Algorithm

Kevin A. Wilson, MS, MPH

Introduction and Background

- It is estimated that the amount of stored data in the world is doubling every two years and it is widely cited that about 80% of all data is stored as text (Gantz & Reinsel, 2012).
- Traditional manual approaches to organizing and analyzing text can be costly and time intensive for a large and growing number of datasets.
- Qualitative data analysis relies on manual approaches that, to date, have been difficult to automate.
  - Crowston et al. (2012) estimated that manually coding 700 messages from an online discussion forum took approximately 100 hours of effort for one coder.
- This study aimed to develop an automated machine learning algorithm to code text-based qualitative research data.

Research Questions

1. Is it possible to develop an automated, machine learning-based algorithm to code qualitative research data?
2. How does the performance of the automated algorithm compare to that of trained qualitative analysts?
3. How much training data is needed to achieve strong performance?
4. What factors should be considered prior to using an automated algorithm?

The Qualitative Coding Process

Context: qualitative interviews for a cross-site evaluation of a public health program

"We have been asked to do the GPRA. The whole process is quite long. Just completing the screen would be much quicker rather than the other research activities."

- GPRA/Data Collection[1]
- Barriers
- Service
- Screening

"After completing the screening, I calculate the score and provide a brief intervention if the patient scores between a 4 and 26 for drugs or 11 and 26 for alcohol."

- Service
- Screening
- Brief Intervention

[1]Government Performance and Results Act

Training Analysts to Code

Code	Brief description
Barriers	Factors that challenge or hinder successfully implementing, integrating, or sustaining the program.
Brief Intervention	A specific service component involving a time-limited effort (usually 5 to 20 minutes) to provide information or advice, increase motivation to avoid substance use, or teach behavior change skills that will reduce substance use and the chances of negative consequences.
GPRA/Data Collection	Federally-mandated data collection and reporting requirements associated with the Government Performance and Results Act.
Screening	A specific service component involving a preliminary procedure to evaluate the likelihood that an individual has a substance use disorder or is at risk for negative consequences from substance use.
Service	Activities associated with delivering or supporting program services.

Detailed Coding Guidelines

The *GPRA/Data Collection* code should be assigned to all passages that refer to federally-mandated data collection and reporting requirements (and associated activities) that result from the Government Performance and Results Act (GPRA). These include references to the following:

- Screening quotas and other service delivery benchmarks;
- Administration of the GPRA baseline interview modules (which differ as a function of screening score);
- The collection of discharge data;
- Patient selection and recruitment (including the consent process) for participation in the 6-month follow-up interview, and patient tracking for follow-up;
- GPRA data cleaning and entry (baseline [including screen negatives], discharge, and follow-up data);
- The use of GPRA data as a program monitoring tool and the generation of reports based on GPRA data;
- GPRA training and quality assurance;
- Descriptions of how GPRA interfaces with service delivery;
- Barriers and facilitators with respect to GPRA requirements;
- Passages that refer to screening score thresholds are also relevant because they relate to the definition of screen positives for GPRA purposes and dictate the number and content of GPRA interview modules to be administered.

Passages that refer to local evaluation and other local data collection efforts (including data processing and analysis) should also be assigned this code.

- Descriptions of local evaluation plans, including how the local evaluations are integrated into the program.
- Information regarding comparisons with untreated control groups, patient tracking, follow-up, cost or financing evaluations, treatment or service utilization tracking.
- References to data storage systems that have been put in place to accommodate data collection on-site (for local evaluation), as well as for the funder's information needs.

About the Dataset

- We used interview data coded by trained qualitative analysts from a large cross-site evaluation.
- The evaluation studied a public health approach to delivering early intervention and treatment services for persons with or at risk for substance use disorders.
- A total of 171 hour-long interviews were conducted with program administrators, practitioners, local evaluators, and other key stakeholders.
- Transcripts were coded in ATLAS.ti, a popular computer-assisted qualitative data analysis software program.
- Prior to coding in ATLAS.ti, a qualitative codebook was developed based on the evaluation questions.
- The full dataset consists of 9,255 passages, where a “passage” is defined as a section of text to which at least one code was applied.

Overview of Methods

Text and code extraction from Atlas.ti

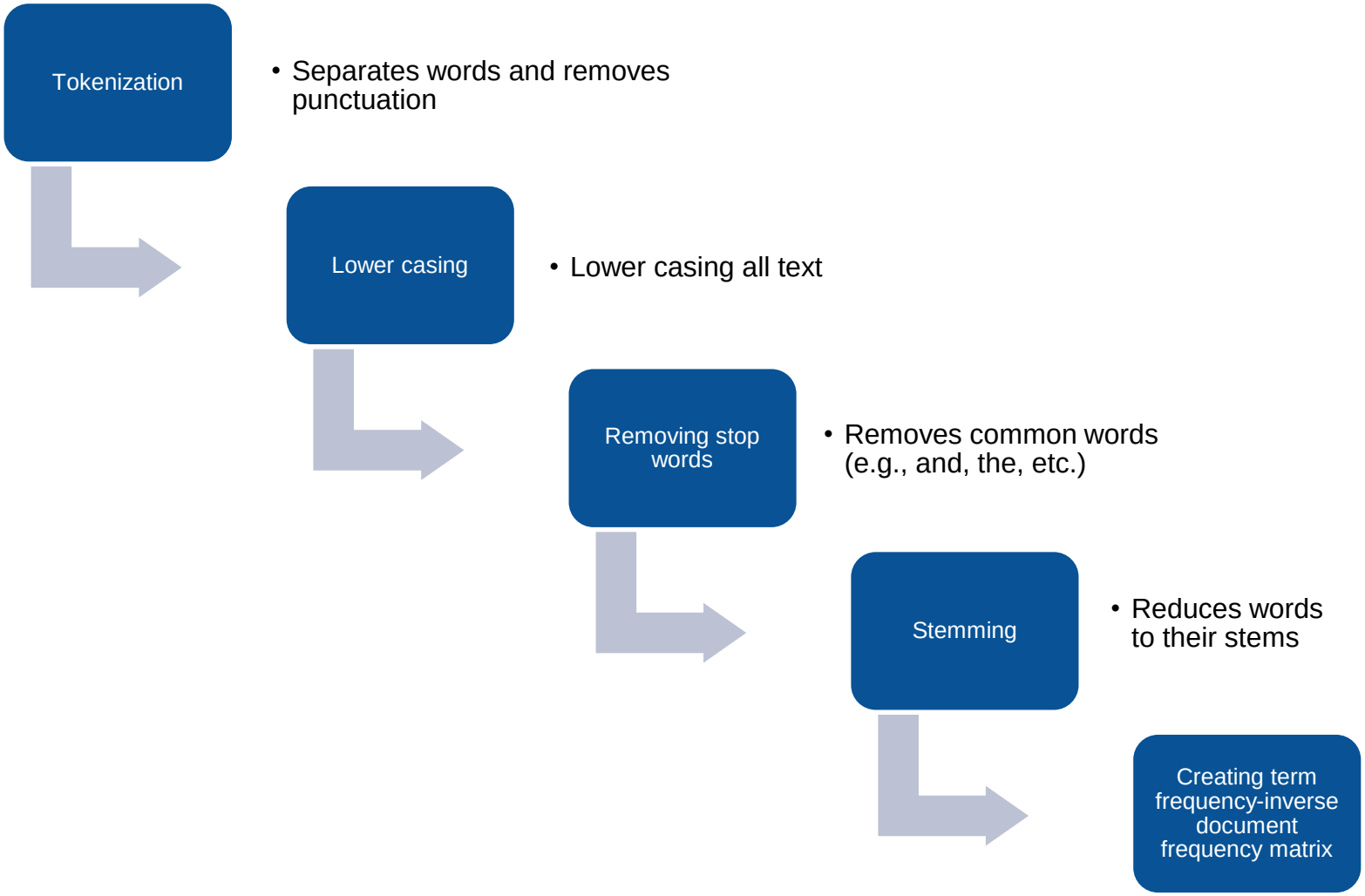
Text preprocessing

Split data into training (80%) and test (20%) datasets

Train a series of binary classifiers using Support Vector Machines (SVM)

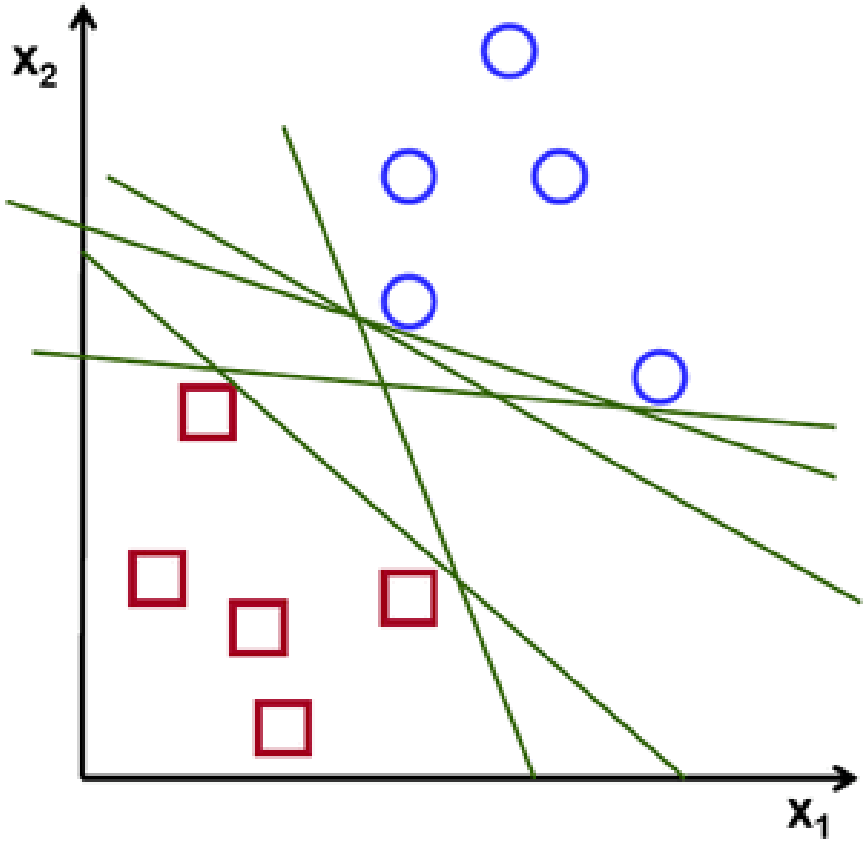
Evaluate classifiers using precision, recall, f-score and accuracy

Data Preprocessing



Support Vector Machine Coding Algorithm

- SVM is a binary classification algorithm.
- SVM is ideally suited for text classification because of the sparse high-dimensional nature of text (Joachims, 1998).
- SVM algorithm effectiveness in text classification has been demonstrated experimentally in several studies.
- We used R with the 'RTextTools' package



Evaluation Metrics Overview

- Definitions
  - True Positive (TP) – the algorithm predicts that the code applies to a passage and this is correct
  - True Negative (TN) – the algorithm predicts that the code does not apply to a passage and this is correct
  - False Positive (FP) – the algorithm predicts that the code applies to a passage but this is not correct
  - False Negative (FN) – the algorithm predicts that the code does not apply to a passage but it, in fact, does apply
- Accuracy doesn't tell the whole story
  - e.g., if a code only applies to 10% of all passages, the algorithm can easily guess the negative cases and will be correct 90% of the time
  - Precision and Recall metrics address this issue
  - F-Score provides a measure of overall accuracy, based on precision and recall

Evaluation Metrics Details

- Accuracy
  - Proportion of predictions that are correct
  - Generally, not a useful measure
- Precision
  - Proportion of positive cases that are correct
  - Precision = TP / (TP + FP)
  - Also known as positive predictive value
- Recall
  - Proportion of actual positive cases that are identified correctly
  - Recall = TP / (TP + FN)
  - Also known as sensitivity
- F-Score
  - Harmonic mean of precision and recall

Results – Algorithm Performance (N=9,255)

Code	Number of positive examples in training set (N = 7,404)	Number of positive examples in the test set (N = 1,851)	Precision	Recall	F-score	Accuracy
Time/Duration	494	124	100%	54%	70%	97%
Screening	717	322	88%	54%	67%	94%
Service	2,241	542	77%	58%	67%	83%
Goals	453	121	87%	50%	63%	96%
GPRA/Data Collection	653	164	81%	50%	62%	95%
Brief Treatment	276	71	79%	48%	60%	98%
Prescreening	683	174	79%	47%	59%	94%
Brief Intervention	996	249	81%	46%	59%	96%
Sustainability	524	140	80%	46%	58%	95%
Involvement	1,285	308	82%	44%	57%	89%
Referral to Treatment	378	105	80%	45%	57%	96%
Implementation	1,960	512	80%	44%	57%	81%
Change	655	164	93%	40%	56%	94%
Costs/Funding	583	125	71%	46%	56%	95%
Performance Site	1,284	315	83%	42%	56%	89%
Patients	923	260	83%	38%	52%	90%
Barriers	1,778	456	76%	40%	52%	82%
QA	626	162	76%	38%	50%	94%
Facilitators	1,416	389	84%	34%	49%	85%
Model	951	224	80%	35%	48%	91%
Integration	1,072	274	76%	35%	48%	89%
Staffing	1,065	282	84%	33%	48%	89%
Milestones	392	104	81%	33%	47%	96%
Suggestion	529	147	88%	24%	37%	94%
Risk Factors	412	87	64%	24%	35%	96%
Impact	488	134	95%	16%	27%	94%
Z_Interesting	41	4	n/a	0%	n/a	100%

Results – Comparison to Trained Analysts

		Average across seven trained analysts				SVM algorithm			
Code	Number of positive examples in the test set (N = 130 examples)[1,2]	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score	Accuracy
Integration	55	68%	41%	50%	68%	100%	67%	80%	91%
Barriers	53	67%	61%	63%	72%	100%	58%	74%	89%
Service	52	78%	88%	81%	85%	100%	67%	80%	91%
Facilitators	44	61%	39%	45%	71%	100%	46%	63%	84%
Prescreening	37	83%	95%	88%	94%	100%	29%	44%	89%
Performance Site	29	67%	53%	58%	82%	100%	67%	80%	95%
Implementation	25	36%	75%	48%	67%	80%	50%	62%	89%
Costs/Funding	24	92%	38%	51%	89%	75%	50%	60%	91%
Patients	24	58%	30%	37%	83%	100%	60%	75%	95%
Involvement	23	59%	55%	56%	84%	80%	50%	62%	89%
GPRA/Data Collection	21	96%	65%	76%	94%	67%	33%	44%	89%
Model	20	32%	16%	19%	78%	100%	17%	29%	89%

Results – Number of Training Examples

Passages with Word Count Between…	Average Accuracy Across All 27 Codes
14–30 words	94.5% (N = 37)
31–65 words	93.9% (N = 243)
66–100 words	92.7% (N = 283)
101–150 words	91.7% (N = 409)
151–200 words	92.0% (N = 308)
201–300 words	91.5% (N = 359)
301–550 words	91.9% (N = 192)
551–828 words	89.1% (N = 20)

Discussion

- Overall, an automated coding approach is promising, as the machine learning algorithm showed good performance on a number of codes
- For most codes the algorithm equaled or outperformed the trained qualitative analyst (caveat – small evaluation set)
  - An analyst may change their interpretation of the same definition over time (“drift”)
  - Analysts may interpret passages differently and also interpret code definitions differently
  - The algorithm was able to “average” the interpretations and perspectives of the analysts over thousands of passage
- While the results are encouraging, further research needs to be performed to determine if the algorithm will generalize to other datasets

Key Factors to Consider / Open Research Questions

- Choosing codes wisely
  - More narrowly defined codes exhibit stronger performance
- Choosing the lesser of two errors
  - Which is more important, precision or recall?
- Determining the level of effort required to create an automatic coding algorithm
  - Is the effort needed to process the text and train the models justified by the performance of the algorithm?
- Determining the appropriate training size
  - How many training examples are needed?
- Achieving desired performance with fewer examples
  - Can the algorithm be optimized further?
- Coding smaller passages
  - Does coding smaller passages yield better results?

References

Crowston, K., Liu, X., & Allen, E. E. (2010). Machine learning and rule-based automated coding of qualitative data. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1-2.

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East. *IDC iView: IDC Analyze the Future*.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features* (pp. 137-142). Springer Berlin Heidelberg.

Collaborators

Kevin A. Wilson
(919) 485-5521
kwilson@rti.org

Justin Landwehr
(919) 990-8345
jlandwehr@rti.org

Mark Pope
(919) 485-5701
mpope@rti.org