Exploring automated coding of open-ended survey responses using machine learning

Rebecca Weiss\textsuperscript{1}, Justin Grimmer\textsuperscript{2}, Matt Berent\textsuperscript{3}, Jon A. Krosnick\textsuperscript{4}, and Arthur Lupia\textsuperscript{5}

May 14th, 2012
Overview

Project motivation

- Big picture: how does machine learning compare to human coders with respect to classifying open-ended responses?

Take-aways from this talk
Project motivation

- Big picture: how does machine learning compare to human coders with respect to classifying open-ended responses?
- **Question 1:** How good are machines on the basis of transcriptions alone?

Take-aways from this talk
Overview

Project motivation

▶ Big picture: how does machine learning compare to human coders with respect to classifying open-ended responses?

▶ **Question 1**: How good are machines on the basis of transcriptions alone?

▶ **Question 2**: How much better do they get with additional human-generated data?

Take-aways from this talk
Overview

Project motivation

- Big picture: how does machine learning compare to human coders with respect to classifying open-ended responses?
- Question 1: How good are machines on the basis of transcriptions alone?
- Question 2: How much better do they get with additional human-generated data?

Take-aways from this talk

- Learn about machine learning: unsupervised methods vs supervised methods
Overview

Project motivation

- Big picture: how does machine learning compare to human coders with respect to classifying open-ended responses?
- **Question 1**: How good are machines on the basis of transcriptions alone?
- **Question 2**: How much better do they get with additional human-generated data?

Take-aways from this talk

- Learn about machine learning: unsupervised methods vs supervised methods
- Why response length matters
Overview

Project motivation

- Big picture: how does machine learning compare to human coders with respect to classifying open-ended responses?
- **Question 1**: How good are machines on the basis of transcriptions alone?
- **Question 2**: How much better do they get with additional human-generated data?

Take-aways from this talk

- Learn about machine learning: unsupervised methods vs supervised methods
- Why response length matters
- The practical reality of computational methods
Motivation

Using open-ended survey questions

- A brief history of open-ended questions
Motivation

Using open-ended survey questions

- A brief history of open-endeds
- The Good
Using open-ended survey questions

▸ A brief history of open-endeds
▸ The Good
   ▸ Get a deeper understanding of respondent’s opinions
Using open-ended survey questions

- A brief history of open-ended questions
- The Good
  - Get a deeper understanding of respondent’s opinions
  - Respondents use their own words: avoid “forcing the issue”
Motivation

Using open-ended survey questions

- A brief history of open-ended surveys
- The Good
  - Get a deeper understanding of respondent’s opinions
  - Respondents use their own words: avoid “forcing the issue”
- The Bad
Using open-ended survey questions

- A brief history of open-ended surveys
- The Good
  - Get a deeper understanding of respondent’s opinions
  - Respondents use their own words: avoid “forcing the issue”
- The Bad
  - Extremely human-intensive analysis, from speech \(\Rightarrow\) transcription \(\Rightarrow\) codes
Motivation

Using open-ended survey questions

▶ A brief history of open-endedds
▶ The Good
  ▶ Get a deeper understanding of respondent’s opinions
  ▶ Respondents use their own words: avoid “forcing the issue”
▶ The Bad
  ▶ Extremely human-intensive analysis, from speech ⇒ transcription ⇒ codes
  ▶ Very expensive process, in terms of time and money
Open-ended coding as a **classification problem**

- Transcription and coding are systematic and repetitive actions
Open-ended coding as a **classification problem**

- Transcription and coding are systematic and repetitive actions
- Generally, train humans to be as machine-like as possible; “take input, apply rules, produce output”
Open-ended coding as a **classification problem**

- Transcription and coding are systematic and repetitive actions
- Generally, train humans to be as machine-like as possible; “take input, apply rules, produce output”
- Very similar to other repetitive tasks, like genetic sequencing and spam detection
Open-ended coding as a **classification problem**

- Transcription and coding are systematic and repetitive actions
- Generally, train humans to be as machine-like as possible; “take input, apply rules, produce output”
- Very similar to other repetitive tasks, like genetic sequencing and spam detection
- **Solution:** *machine learning*
Machine Learning

Given enough data, use statistical and/or algorithmic methods to “evolve” intelligent behaviors

- Two flavors: unsupervised and supervised
Machine Learning

Given enough data, use statistical and/or algorithmic methods to “evolve” intelligent behaviors

- Two flavors: unsupervised and supervised
- How to tell if your classification task is unsupervised or supervised: is your data unlabeled or labeled?
Machine Learning

Given enough data, use statistical and/or algorithmic methods to “evolve” intelligent behaviors

- Two flavors: unsupervised and supervised
- How to tell if your classification task is unsupervised or supervised: is your data unlabeled or labeled?
- If unlabeled, unsupervised: attempt to infer classification from the data’s structure alone
Machine Learning

Given enough data, use statistical and/or algorithmic methods to “evolve” intelligent behaviors

- Two flavors: unsupervised and supervised
- How to tell if your classification task is unsupervised or supervised: is your data unlabeled or labeled?
- If unlabeled, unsupervised: attempt to infer classification from the data’s structure alone
- If labeled, supervised: use the train-test paradigm to infer a classification function
Given enough data, use statistical and/or algorithmic methods to “evolve” intelligent behaviors

- Two flavors: *unsupervised* and *supervised*
- How to tell if your classification task is unsupervised or supervised: is your data *unlabeled* or *labeled*?
- If unlabeled, **unsupervised**: attempt to infer classification from the data’s structure alone
- If labeled, **supervised**: use the train-test paradigm to infer a classification function
- How does this work with open-ended survey data?
The Data

- The ANES 2008 Time Series Study
- $n$ of 2100, face-to-face
- Several open-ended questions on many different topics
- For each response:
  - Professional transcription of interview
  - Interviewer notes
  - ANES provided the coding scheme
Question 1: How good are machines at coding human responses on the basis of transcriptions alone?

In other words, can we infer codes from the response language itself?
On Terrorism

A sample question

“As you know, on September 11th 2001, a group of Terrorists took control of several U.S. commercial airplanes and crashed them into the World Trade Center in New York and the Pentagon in Washington. What do you think the Terrorists were trying to accomplish by their actions?”

Some sample transcriptions

▶ “I think they were trying to cause dissension in America... and to prove that they could do it.”
▶ “I THINK THEIR INSANE, THEY ANGRY BUT MISDIRECTED, I AM NOT SURE THEY EVEN KNEW WHAT THEY WERE DOING”
▶ “obviously to disrupt our way of living our security and all that good stuff”
▶ “strking at the white Satan and givin us a black eye”
Natural language is messy, but we can make it cleaner
Natural language is messy, but we can make it cleaner

The basic idea: consider each response in terms of *word counts*
Natural language is messy, but we can make it cleaner

The basic idea: consider each response in terms of *word counts*

How to do this; first, preprocessing steps:

- Spell-check: ‘strking’ ⇒ ‘striking’; ‘givin’ ⇒ ‘giving’
- Remove whitespace, remove punctuation, convert to lower-case
- Remove stopwords (analysts’ choice)
  - More on this shortly
Natural language is messy, but we can make it cleaner

The basic idea: consider each response in terms of *word counts*

How to do this; first, preprocessing steps:

- Spell-check: 'strking' ⇒ 'striking'; 'givin' ⇒ 'giving'
- Remove whitespace, remove punctuation, convert to lower-case
  - Example: ('striking', 'at', 'the', 'white', 'satan', 'and', 'giving', 'us', 'a', 'black', 'eye')
- Remove stopwords (analysts’ choice)
  - Example: ('striking', 'white', 'satan', 'giving', 'us', 'black', 'eye')
  - More on this shortly

Now what? Using responses-as-word counts, create a **document-term matrix**
Document-Term Matrix

A 2D vector-space representation of all respondents and their responses
Analogue: a giant spreadsheet, respondents on rows, all possible words as columns

- R1: (‘I’, ‘like’, ‘cats’)
- R2: (‘I’, ‘like’, ‘dogs’)

<table>
<thead>
<tr>
<th>Respondents</th>
<th>“like”</th>
<th>“cats”</th>
<th>“dogs”</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>R2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>R3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
From the document-term matrix, you can determine how similar respondents are to each other by the similarities between their word counts (e.g., vector distance between R1 and R2).
From the document-term matrix, you can determine how similar respondents are to each other by the similarities between their word counts (e.g., vector distance between R1 and R2).

We can create a scatterplot from these similarities.
From the document-term matrix, you can determine how similar respondents are to each other by the similarities between their word counts (e.g., vector distance between R1 and R2).

We can create a scatterplot from these similarities.

Coding becomes a scatterplot coloring problem, where colors = codes.
Are you convinced?

- Choice of clustering algorithm is arbitrary
Respondents in 2D: MultiMixEM

Clusters
1
2
3
4
5
6
7
8
9

Automated Classification, AAPOR 2012
May 14th, 2012 16 / 26
Clusters Defined by Word Frequencies

Words in order of log(Scores)

1. destroy
   - attention
   - unifying
   - people
   - score
2. government
   - attack
   - security
   - power
   - military
3. afraid
   - intimidated
   - health
   - whole
   - body
4. destruction
   - strike
   - start
   - start
   - destruction
5. fear
   - war
   - fear
   - war
6. noticed
   - murder
   - wounded
7. prove
   - warning
   - citizens
   - lives
   - man
8. president
   - bush
   - invincible
9. revenge
   - terror
   - terror

log(Scores)

3.0 3.5 4.0
Are you convinced?

- Choice of clustering algorithm is arbitrary
- Number of clusters is arbitrary
Respondents in 2D: k−means

Clusters

1
2
3
4
5
6

Automated Classification, AAPOR 2012
May 14th, 2012 19 / 26
Clusters Defined by Word Frequencies

Words in order of log(Scores):

1. attention
   - feel
   - call
   - serious
   - governments
   - drawing
   - unsafe
   - attract
   - none
   - cause
   - country
   - organization
   - draw
   - noticed
   - bushes
   - repulse
   - eyes
democracy
stop
country
no
cause
none
attract
unsafe
drawing
governments
serious
call
feel
attention

2. destroy
   - trying
   - country
   - economy
   - spirit
   - country
   - economy
   - proved
   - weaknesses
   - world
   - world
   - nations
   - hurt
   - freedom
   - terror
   - us
   - none
   - government
   - cause
   - america
   - that
towers

3. kill
   - americans
   - trying
   - scare
   - destroy
   - americans
   - trying
   - none
   - power
   - us
   - nation
   - hurt
   -翅
   - freedom
   - terror
   - government
   - cause
   - america
   - that

4. fear
   - americans
   - strike
   - hearts
   - american
   - instill
   - provoke
   - americans
   - none
   - establish
   - terror
   - destruction
   - inspire
   - hate
   - harts
disarmament
americans
that

5. vulnerable
   - don't
   - bush
   - attack
   - home
   - government
   - strong
   - pride
   - powerful
   - send
   - message
   - war
   - hate
   - harts
disarmament
american
that

6. 
   - attention
   - americans
   - feel
   - suppose
   - week
   - don't
   - job
   - touching
   - reminding
   - broken
   - getting
   - painfully
   - anything
   - whatever
   - protected
   - unsafe
   - sending
   - watching
   - background

log(Scores)
Are you convinced?

- Choice of clustering algorithm is arbitrary
- Number of clusters is arbitrary
- Distances are based upon co-occurrence of words, and thus collinearity in matrix
Are you convinced?

- Choice of clustering algorithm is arbitrary
- Number of clusters is arbitrary
- Distances are based upon co-occurrence of words, and thus collinearity in matrix
- If words don’t co-occur amongst respondents, their distances are **bigger** and increase the likelihood of being **coded differently**
Are you convinced?

- Choice of clustering algorithm is arbitrary
- Number of clusters is arbitrary
- Distances are based upon co-occurrence of words, and thus collinearity in matrix
- If words don’t co-occur amongst respondents, their distances are **bigger** and increase the likelihood of being **coded differently**
- **Response length matters**: the more words, the better
Takeaways

- Totally automated coding using unsupervised learning: still a dream

Not without value: useful for exploratory data analysis

Still require humans for interpretation of results
Takeaways

- Totally automated coding using unsupervised learning: still a dream
- Not without value: useful for exploratory data analysis
Takeaways

- Totally automated coding using unsupervised learning: still a dream
- Not without value: useful for **exploratory data analysis**
- Still require humans for interpretation of results
Question 2: How much better do machines get with additional human-generated data?

More importantly, what kind of human-generated data do you need?
<table>
<thead>
<tr>
<th>Statement</th>
<th>n responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>*No/None/Nothing</td>
<td>15</td>
</tr>
<tr>
<td>*Personal feelings mentions</td>
<td>7</td>
</tr>
<tr>
<td>*Spanish responses</td>
<td>6</td>
</tr>
<tr>
<td>*The terrorist wanted to be taken serious/force to be reckoned with</td>
<td>14</td>
</tr>
<tr>
<td>*The terrorist wanted to defeat the president/leadership</td>
<td>5</td>
</tr>
<tr>
<td>*The terrorist wanted to get power/take over United States/the world</td>
<td>66</td>
</tr>
<tr>
<td>*The terrorists wanted to destroy/destruct/damage (unspecified)</td>
<td>13</td>
</tr>
<tr>
<td>*The terrorists were angry with the United States</td>
<td>10</td>
</tr>
<tr>
<td>A respondent’s comment about himself or herself</td>
<td>4</td>
</tr>
<tr>
<td>I don’t know</td>
<td>217</td>
</tr>
<tr>
<td>I’m not going to answer</td>
<td>6</td>
</tr>
<tr>
<td>Other comment</td>
<td>113</td>
</tr>
<tr>
<td>Other comments about the terrorists or attacks</td>
<td>5</td>
</tr>
<tr>
<td>The Muslim religion made the terrorists do it</td>
<td>12</td>
</tr>
<tr>
<td>The terrorists are crazy</td>
<td>7</td>
</tr>
<tr>
<td>The terrorists are jealous of the United States</td>
<td>2</td>
</tr>
<tr>
<td>The terrorists hate the United States</td>
<td>32</td>
</tr>
<tr>
<td>The terrorists wanted to cause fear among people in the United States</td>
<td>340</td>
</tr>
<tr>
<td>The terrorists wanted to cause fear among people outside the United States</td>
<td>1</td>
</tr>
<tr>
<td>The terrorists wanted to cause pain among people in the United States</td>
<td>32</td>
</tr>
<tr>
<td>The terrorists wanted to create dissension between people in the United States.</td>
<td>49</td>
</tr>
<tr>
<td>The terrorists wanted to damage the United States’ economy</td>
<td>83</td>
</tr>
<tr>
<td>The terrorists wanted to destroy a symbol of the United States</td>
<td>4</td>
</tr>
<tr>
<td>The terrorists wanted to destroy democracy</td>
<td>11</td>
</tr>
<tr>
<td>The terrorists wanted to destroy the twin towers</td>
<td>10</td>
</tr>
<tr>
<td>The terrorists wanted to destroy the United States</td>
<td>85</td>
</tr>
<tr>
<td>The terrorists wanted to destroy the United States government</td>
<td>54</td>
</tr>
<tr>
<td>The terrorists wanted to destroy Western civilization</td>
<td>22</td>
</tr>
<tr>
<td>The terrorists wanted to get a reward</td>
<td>6</td>
</tr>
<tr>
<td>The terrorists wanted to get even with the United States</td>
<td>52</td>
</tr>
<tr>
<td>The terrorists wanted to get noticed</td>
<td>305</td>
</tr>
<tr>
<td>The terrorists wanted to influence the United States’ involvement in the Middle East</td>
<td>35</td>
</tr>
<tr>
<td>The terrorists wanted to kill George Bush</td>
<td>10</td>
</tr>
<tr>
<td>The terrorists wanted to kill people</td>
<td>170</td>
</tr>
<tr>
<td>The terrorists wanted to prove that the United States is vulnerable</td>
<td>130</td>
</tr>
<tr>
<td>The terrorists wanted to prove that the United States is weak</td>
<td>33</td>
</tr>
<tr>
<td>The terrorists wanted to prove that they are strong</td>
<td>59</td>
</tr>
<tr>
<td>The terrorists wanted to prove they could hurt the United States</td>
<td>20</td>
</tr>
<tr>
<td>The terrorists wanted to start a war with the United States</td>
<td>51</td>
</tr>
<tr>
<td>The U.S. government made the terrorists do it</td>
<td>4</td>
</tr>
</tbody>
</table>
Supervised classification

How to think about it

- Codes as features of the respondent, not of the response

**PROBLEM**: Too many labels, leaving not enough data to train on!

- Option 1 (Change the task): Collapse labels
Supervised classification

How to think about it

- Codes as features of the **respondent**, not of the **response**

**PROBLEM**: Too many labels, leaving not enough data to train on!

- Option 1 (Change the task): Collapse labels
- Option 2 (Going backwards): Use previous years’ data to train and test
Supervised classification

How to think about it

▶ Codes as features of the **respondent**, not of the **response**

**PROBLEM**: Too many labels, leaving not enough data to train on!

▶ Option 1 (Change the task): Collapse labels
▶ Option 2 (Going backwards): Use previous years’ data to train and test
▶ Option 3 (Going wide): Use covariate features of the respondents (e.g., demographic features, responses to other questions)
How to think about it

- Codes as features of the **respondent**, not of the **response**

**PROBLEM**: Too many labels, leaving not enough data to train on!

- Option 1 (Change the task): Collapse labels
- Option 2 (Going backwards): Use previous years’ data to train and test
- Option 3 (Going wide): Use covariate features of the respondents (e.g., demographic features, responses to other questions)
- Option 4 (Going long): Get more responses!
rjweiss@stanford.edu

Questions?