Text Analytics of Student Comments in Course Evaluations

Kevyn Collins-Thompson
Adam Levick, Michael Spiegel, Florence Lee
Mika Lavaque-Manty, Phyllis Ford
Goals and Motivation

• **Goal:** Combine text analysis with traditional numeric variables for deeper understanding of:
  - Properties of classes as student learning environments
  - Student motivation and needs
  - Trends over time

• Text is (much) harder to analyze; often ignored

• Example questions:
  - What class factors as expressed in comments are associated with high and low course ratings and other variables?
  - What are students saying about their learning environment?
  - What does it mean when numeric and text analyses disagree?
The Corpus

• 11 fall/winter terms W2009-W2014
• 21 million words of comment text
• 43,000 classes
• Thanks to Phyllis Ford, UM Registrar's office
Course evaluation questions: numeric and open-ended

<table>
<thead>
<tr>
<th>Numeric response</th>
<th>Prompt text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Overall, this was an excellent course.</td>
</tr>
<tr>
<td>Q2</td>
<td>Overall, the instructor was an excellent teacher.</td>
</tr>
<tr>
<td>Q3</td>
<td>I learned a great deal from this course.</td>
</tr>
<tr>
<td>Q4</td>
<td>I had a strong desire to take this course.</td>
</tr>
<tr>
<td>Free-form comments</td>
<td></td>
</tr>
<tr>
<td>Q900</td>
<td>Comment on the quality of instruction in this course.</td>
</tr>
<tr>
<td>Q901</td>
<td>How can the instructor improve the teaching of this course?</td>
</tr>
<tr>
<td>Q902</td>
<td>Which aspects of this course did you like best?</td>
</tr>
<tr>
<td>Q903</td>
<td>Which aspects of this course did you like least?</td>
</tr>
</tbody>
</table>
Comments corpus joined with course and instructor data

• Registrar course data (Phyllis Ford) includes:
  • Subject Area
  • Class type: Lecture, Discussion, Seminar, Lab, etc.
  • Class size
  • Evaluation response counts
  • Q1-Q4 distribution (but not individual Q1-Q4)
  • Grade distribution (but not individual grades)

• Instructor data (Mika Lavaque-Manty)
  • 5,500 courses FA2008-WN2013
  • Random sample of LS&A, CoE
  • Instructor gender, rank (+contingent vs tenure-track)
Challenges of text analysis for student comments

• Comments are open-ended
  • Students can and will talk about anything
• Language structure is highly variable and noisy
  • Spelling, grammar, informal, technical, emoticons...
• Response rate for comments is highly variable
  • We only observe the language of a self-selecting group
• Many subtle linguistic issues
  • Comments often mix positive and negative aspects in the same sentence
  • Sarcasm, cultural references, idioms, ...
• Privacy
  • Students are anonymous (usually), instructors are not
• Nature of question influences nature of comments
Processing pipeline: Natural language processing (Stanford Toolkit)

• Sentence boundaries
• Part of speech tagging
  • noun, verb, adjective, ...
• Coreference resolution
  • "The GSI was great. He took care of all our questions. And although she was busy, the professor was also helpful."
• Named entity recognition
  • Location, Person, Organization, ...
• Parsing
  • Syntax and dependency structure
• Sentiment analysis
  • Positive/negative opinion levels
Processing pipeline: Topic and language models

• Statistical readability models
• Topic modeling
  • Online Latent Dirichlet Allocation
• Turbotopics
  • Finding characteristic word sequences
• Text regression
  • Input variables: n-gram frequencies
  • Output variable:
    • Q1 score
    • Instructor seniority, gender

Future:
• Aspect-based opinion mining
Comment response rates (Q900) are steady over time.

% of evaluation responders (lecture courses only) who leave a Q900 comment.
Comment response rates (Q900) are steady across class sizes

- Small: < 15
- Med-Small: < 50
- Med: < 100
- Large: over 100

- Decreases slightly for large classes
When students do write comments, who writes the most and least?

- Students in education courses write the longest comments on average.
- Economics students tend to write the shortest.
- Again, Q900
Text sentiment

• State-of-the-art model (Recursive Neural Networks) can handle more subtle mixed comments

"There are slow and repetitive parts but it has just enough spice to keep it interesting."

Contrastive conjunctions X but Y are common in student comments.

Example of most positive and negative course comments by sentiment score

<table>
<thead>
<tr>
<th>Most positive sentiment score (= 4)</th>
<th>Most negative sentiment score (&lt; 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. This was a great class that definitely helped prepare me for my career.</td>
<td>1. Prof. X went over the easy concepts too slowly, and the difficult concepts too quickly, so everyone was left both bored and confused…</td>
</tr>
<tr>
<td>2. Professor X is one of the best teachers I've ever had, and the reason I took this course.</td>
<td>2. Lecture was helpful in this class, but discussion was abysmal… The poor teaching style and lack of content ultimately persuaded me. It was an utter waste of time.</td>
</tr>
<tr>
<td>3. I loved this class and learned so much about different theories and techniques and found new and efficient ways to apply them to my own directing style.</td>
<td>3. The textbook does not cover some topics… it is difficult to do the homework with only the notes. Exam is disconnect with homework. The instructor is not clear what students should learn in this course.</td>
</tr>
<tr>
<td>4. He is an excellent language instructor and is the best one I've had in terms of being able to explain language concepts.</td>
<td>4. Way too much work for a 3 credit course. Seminar presentations were boring and repetitive…</td>
</tr>
<tr>
<td>5. A is a truly exceptional teacher who made this course very worthwhile for all of his students.</td>
<td>5. Timing and pace started off decent, but ended badly; it was too fast for the amount of work we had to do.</td>
</tr>
</tbody>
</table>
Examples of comments with an intermediate sentiment score

Leaning negative sentiment score (= 1.5)

1. I liked the structure of the course, however, it would have been nice to have more structured homework and exam problems...

2. The instruction was excellent. My only complaint is that proofs.. went by too quickly...

3. Despite the inherent nature of the material being dry, I learned a lot from this course. Too much homework.

4. Prof. X is clearly an intelligent individual. However, I do feel that he often came to class ill-prepared...

Leaning positive sentiment score (= 2.2)

1. The instruction was fine. I'm not sure about the use of time and structure...

2. Lectures were interesting. Paper assignment was worthwhile. Good instructor...

3. Decent, good. It would be nice to have more in-class problems... no complaints at all...

4. The GSI did a good job in answering questions and reviewing material..
How does text sentiment relate to numeric Q1 score?

- Subset: 8,017 lecture courses with comments from > 10 students
- Strong positive correlation with actual numeric Q1 score (r=0.69)
Sentiment vs scaled Q1 score by class size (with comment response rate)
Underappreciated faculty: when is sentiment high but Q1 low?

- "Homework grading.. Felt a bit arbitrary"
- "My one major critique is on the writing assignments.. Useful but I did not know the format required."
- "Teaching was good but grading criteria and expectations were unfair to students who are not geniuses."
- "would be even stronger if it was spread out over more time.. to allow for class discussion"
Text readability

• Traditional measures like Flesch-Kincaid rely on:
  • Mean sentence length
  • Mean word difficulty (# syllables)
  • Having lots of clean text

• Such measures are unreliable for comment text
  • Sentence structure and syntax are very noisy
  • Vocabulary may contain lots of technical or colloquial terms

• We use a statistical language modeling approach that
  learns how individual word use changes with level
  • Shown to be much more reliable for non-traditional texts
    (Collins-Thompson & Callan, 2004)
  • Produces a probability distribution over levels/grades
Comment Reading Level by Course Subject Area

- Lecture comments only, subjects with at least 100 classes (all levels)
- High: Health, Engineering, English
- Low: Sports, Music, Screen Arts, Math
Comment Reading Level by Class Type and Semester

Comment Grade Level by Class Type

Comment Grade Level by Semester
Topic modeling

• Topic models find the sets of terms that tend to occur together in the texts
• Terms that frequently occur together tend to be about the same subject or theme
• A topic modeling algorithm tries to:
  • Produce well-defined topics with high probability on a few terms
  • Attach documents to as few topics as possible

A sampling of topics derived from 1.8 million New York Times articles

What topics are evident in comment language?

<table>
<thead>
<tr>
<th>Controversial subjects</th>
<th>Language issues</th>
<th>Reading</th>
<th>Writing</th>
<th>GSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>controversial</td>
<td>grammar</td>
<td>texts</td>
<td>writing</td>
<td>GSI</td>
</tr>
<tr>
<td>encouragement</td>
<td>language</td>
<td>novels</td>
<td>english</td>
<td>discussion</td>
</tr>
<tr>
<td>equipped</td>
<td>harsh</td>
<td>forum</td>
<td>papers</td>
<td>section</td>
</tr>
<tr>
<td>violence</td>
<td>vocabulary</td>
<td>instructive</td>
<td>write</td>
<td>helpful</td>
</tr>
<tr>
<td>position</td>
<td>vocab</td>
<td>engaging</td>
<td>team</td>
<td>willing</td>
</tr>
<tr>
<td>ground</td>
<td>textbook</td>
<td>paper</td>
<td>improve</td>
<td>questions</td>
</tr>
<tr>
<td>divide</td>
<td>grammatical</td>
<td>english</td>
<td>work</td>
<td>knowledgeable</td>
</tr>
<tr>
<td>loudly</td>
<td>english</td>
<td>material</td>
<td>skills</td>
<td>prepared</td>
</tr>
<tr>
<td>impartial</td>
<td>worksheets</td>
<td>reading</td>
<td>workshop</td>
<td>knew</td>
</tr>
</tbody>
</table>

- **Tool:** Topic modeling
- Categorize language used in comments
- Latent Dirichlet Allocation
- 100 topics

## Language models: subject areas

<table>
<thead>
<tr>
<th>General law</th>
<th>Corporate law</th>
<th>Chemistry</th>
<th>General math</th>
<th>Advanced math</th>
</tr>
</thead>
<tbody>
<tr>
<td>law</td>
<td>tax</td>
<td>chemistry</td>
<td>math</td>
<td>drawing</td>
</tr>
<tr>
<td>cases</td>
<td>accelerated</td>
<td>chem</td>
<td>question</td>
<td>proof</td>
</tr>
<tr>
<td>issues</td>
<td>negotiations</td>
<td>building</td>
<td>including</td>
<td>theorems</td>
</tr>
<tr>
<td>legal</td>
<td>patent</td>
<td>organic</td>
<td>answer</td>
<td>continuous</td>
</tr>
<tr>
<td>reading</td>
<td>copyright</td>
<td>probability</td>
<td>proofs</td>
<td>measure</td>
</tr>
<tr>
<td>casebook</td>
<td>ban</td>
<td>calculate</td>
<td>pictures</td>
<td>solutions</td>
</tr>
<tr>
<td>substantive</td>
<td>bankruptcy</td>
<td>sizes</td>
<td>words</td>
<td>rigorous</td>
</tr>
<tr>
<td>discussion</td>
<td>notification</td>
<td>loads</td>
<td>concepts</td>
<td>algebraic</td>
</tr>
<tr>
<td>think</td>
<td>corporate</td>
<td>nmr</td>
<td>equations</td>
<td>manifolds</td>
</tr>
</tbody>
</table>
## Language models: class type

<table>
<thead>
<tr>
<th>Lessons</th>
<th>Labs</th>
<th>Discussions</th>
</tr>
</thead>
<tbody>
<tr>
<td>lessons</td>
<td>lab</td>
<td>class</td>
</tr>
<tr>
<td>my</td>
<td>reports</td>
<td>discussion</td>
</tr>
<tr>
<td>dates</td>
<td>work</td>
<td>voice</td>
</tr>
<tr>
<td>policy</td>
<td>data</td>
<td>leading</td>
</tr>
<tr>
<td>repertoire</td>
<td>spss</td>
<td>atmosphere</td>
</tr>
<tr>
<td>ability</td>
<td>done</td>
<td>leader</td>
</tr>
<tr>
<td>performance</td>
<td>experiments</td>
<td>discussions</td>
</tr>
<tr>
<td>vocal</td>
<td>equipment</td>
<td>enjoyed</td>
</tr>
<tr>
<td>studio</td>
<td>laboratory</td>
<td>opinions</td>
</tr>
</tbody>
</table>
### Negatives | Positives
---|---
not | very
no | great
never | good
'nt | helpful
worst | really
waste | material
nothing | you
late | always
lack | more

---

Example of courses clustered by similarity in comment language topics (topic + sentiment)
Text regression

• Problem:
  • What features in the text are predictive of an outcome variable?
  • Every n-gram is a potential feature (explanatory variable)

• Dependent variable: Property of course or instructor, e.g.
  • Median Q1-Q4 scores
  • Instructor rank, gender

• There are > 100,000 unique n-grams of interest
  • We only have ~40,000 data instances
  • Normally, this would over-fit the data

• Solution:
  • Sparse L1-regularized linear / logistic regression
  • Parallelized GraphLab linear solver
Which Q900 1- and 2-grams were associated with high/low numeric Q scores?

<table>
<thead>
<tr>
<th>Q1 (Course)</th>
<th>Q2 (Instructor)</th>
<th>Q3 (Learned a lot?)</th>
<th>Q4 (Motivated?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;always feel&quot;</td>
<td>&quot;intellectually&quot;</td>
<td>&quot;work load&quot;</td>
<td>&quot;really excited&quot;</td>
</tr>
<tr>
<td>0.29</td>
<td>0.20</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>&quot;best courses&quot;</td>
<td>&quot;best GSI&quot;</td>
<td>&quot;really well&quot;</td>
<td>&quot;best courses&quot;</td>
</tr>
<tr>
<td>0.17</td>
<td>0.17</td>
<td>0.07</td>
<td>0.17</td>
</tr>
<tr>
<td>&quot;class size&quot;</td>
<td>&quot;progress&quot;</td>
<td>&quot;practical&quot;</td>
<td>&quot;self&quot;</td>
</tr>
<tr>
<td>0.15</td>
<td>0.06</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>&quot;historical context&quot;</td>
<td>&quot;looked forward&quot;</td>
<td>&quot;difficult material&quot;</td>
<td>&quot;important concepts&quot;</td>
</tr>
<tr>
<td>0.11</td>
<td>0.08</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>&quot;group projects&quot;</td>
<td>&quot;felt uncomfortable&quot;</td>
<td>&quot;peers&quot;</td>
<td>&quot;much enjoyed&quot;</td>
</tr>
<tr>
<td>-0.14</td>
<td>-0.10</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>&quot;team homework&quot;</td>
<td>&quot;nice guy&quot;</td>
<td>&quot;controversial&quot;</td>
<td>&quot;hard time&quot;</td>
</tr>
<tr>
<td>-0.15</td>
<td>-0.11</td>
<td>-0.14</td>
<td>-0.08</td>
</tr>
<tr>
<td>&quot;sometimes confusing&quot;</td>
<td>&quot;disorganized&quot;</td>
<td>&quot;get lost&quot;</td>
<td>&quot;absolutely nothing&quot;</td>
</tr>
<tr>
<td>-0.16</td>
<td>-0.17</td>
<td>-0.07</td>
<td>-0.24</td>
</tr>
<tr>
<td>&quot;absolutely nothing&quot;</td>
<td>&quot;(couldn't) hear&quot;</td>
<td>&quot;unorganized&quot;</td>
<td>&quot;pointless&quot;</td>
</tr>
<tr>
<td>-0.28</td>
<td>-0.19</td>
<td>-0.07</td>
<td>-0.28</td>
</tr>
<tr>
<td></td>
<td>&quot;language barrier&quot;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Informative n-grams ('turbotopics')

• Distinctive multi-word phrases in topics
• Traditional methods for finding collocations test all possibilities at once
  • ...leading to dependent hypothesis tests and e.g. incorrectly rejecting potentially interesting phrases

• Turbotopics operates sequentially:
  • Finds the phrase with *maximum* log likelihood over a family of candidates
  • Then adds the candidate to the model and iterates..

Text regression with informative n-grams: Highest/lowest 6-gram weights predicting Q1 score

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>best instructors i've ever had</td>
<td>0.41</td>
</tr>
<tr>
<td>i have learned so much from</td>
<td>0.40</td>
</tr>
<tr>
<td>i am so grateful to have</td>
<td>0.35</td>
</tr>
<tr>
<td>best gsis i've ever had</td>
<td>0.34</td>
</tr>
<tr>
<td>passionate about what she was teaching</td>
<td>0.10</td>
</tr>
<tr>
<td>thorough knowledge of the subject matter</td>
<td>0.08</td>
</tr>
<tr>
<td>truly cares about his students and</td>
<td>0.07</td>
</tr>
<tr>
<td>more difficult than it needed to be</td>
<td>-0.02</td>
</tr>
<tr>
<td>good job of explaining the material</td>
<td>-0.05</td>
</tr>
<tr>
<td>willing to meet students outside of</td>
<td>-0.05</td>
</tr>
<tr>
<td>i would have learned more if</td>
<td>-0.08</td>
</tr>
<tr>
<td>i wish that there were more</td>
<td>-0.10</td>
</tr>
<tr>
<td>i had a hard time understanding</td>
<td>-0.12</td>
</tr>
<tr>
<td>i am not a fan of</td>
<td>-0.14</td>
</tr>
<tr>
<td>felt like a waste of time</td>
<td>-0.22</td>
</tr>
<tr>
<td>quality of instruction was very poor</td>
<td>-0.25</td>
</tr>
<tr>
<td>worst course i have ever taken</td>
<td>-0.55</td>
</tr>
</tbody>
</table>

Intercept: 4.15
4-grams most predictive of instructor gender  

[Sparse logistic regression]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>knows what he is</td>
<td>0.62</td>
</tr>
<tr>
<td>(intercept)</td>
<td>0.59</td>
</tr>
<tr>
<td>knew what he was</td>
<td>0.59</td>
</tr>
<tr>
<td>cares about his students</td>
<td>0.46</td>
</tr>
<tr>
<td>i wish he would</td>
<td>0.46</td>
</tr>
<tr>
<td>passionate about her subject</td>
<td>-0.71</td>
</tr>
<tr>
<td>i liked how she</td>
<td>-0.73</td>
</tr>
<tr>
<td>great deal from her</td>
<td>-0.76</td>
</tr>
<tr>
<td>knew what she was</td>
<td>-0.84</td>
</tr>
<tr>
<td>i really enjoyed her</td>
<td>-0.86</td>
</tr>
</tbody>
</table>

Male instructor: knows what he is, knew what he was, cares about his students, i wish he would, passionate about her subject, i liked how she, great deal from her, knew what she was, i really enjoyed her

Female instructor: (intercept)
4-grams most predictive of instructor position type
[Sparse logistic regression]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>0.92</td>
</tr>
<tr>
<td>i enjoyed this course</td>
<td>0.25</td>
</tr>
<tr>
<td>i really liked the</td>
<td>0.20</td>
</tr>
<tr>
<td>knowledge of the subject</td>
<td>0.20</td>
</tr>
<tr>
<td>readings were interesting and</td>
<td>0.18</td>
</tr>
<tr>
<td>learned so much</td>
<td>0.16</td>
</tr>
<tr>
<td>passion for the subject</td>
<td>0.15</td>
</tr>
<tr>
<td>just wish we had</td>
<td>-0.08</td>
</tr>
<tr>
<td>instructor s willingness to</td>
<td>-0.12</td>
</tr>
<tr>
<td>i love how he</td>
<td>-0.12</td>
</tr>
<tr>
<td>help students outside of</td>
<td>-0.16</td>
</tr>
<tr>
<td>i would hope that</td>
<td>-0.20</td>
</tr>
<tr>
<td>good at answering questions</td>
<td>-0.31</td>
</tr>
</tbody>
</table>

Tenure-track / Senior lecturers

Adjunct / post-doc / Junior lecturer
How does student comment language change over time?

Changes in technology and practice
How does student comment language change over time?

'cold / freezing' – related terms

Seasonality
Many other questions of interest

- Extend to Q2, Q3, Q4 and 90x comments
  - Summarize top suggestions, strengths
  - How these change over time
  - Instructor demographics
- Examine differences in sections of large classes
- Your suggestions?
Student Surveys: Current Limitations

• Time/quality of read-through analysis
• Change in a course over semesters is lost
• Reports not easily sharable
• Biased readers
Text Analysis: An Unbiased Note-Taker

The text analysis tool could take the role of a note-taker who can:

• Provide instant summaries
• Full knowledge of all topics
• Same way every time
  • Compare between past courses
  • Compare between different courses
• Can provide departmental/school summaries
• Allows for research
What Text Analysis Doesn’t Do

- Understand the full context of a course
- Choose interventions for your course
- Know what changes have been made in a course
SI 100: EXAMPLE OF A BIG CLASS

Q1 Which aspects of the course did you like least?

Course Summary

Students Enrolled: 239
School: Information
Term: Winter 2014
Course Number: SI100

Grade Distribution:
Mean: 73%

Student Information

Subject Majors: 15%
Freshmen: 86%
Sophomore: 10%
Junior: 3%
Senior: 1%

Prior GPA Distribution:
Mean: 3.2
SI 100: EXAMPLE OF A BIG CLASS

Q1  Which aspects of the course did you like least?

Lectures: All words used to describe lectures

Click on a bar to see more details:

- Positive sentiment
- Negative sentiment

Consistent
Like
Enjoy
Win
Illogical
Yawn
Oversimplify
Agonizing
Dislike
Confusing
Bland
Dry
Felt
Boring

Student feedback containing the word ‘boring’:

“I thought lecture was boring but liked the professor.”

“The lecture time… way too early to absorb information! I think that a lot of students did not come to lecture/fell asleep in lecture because of the time it was held. Didn’t help that the class was so boring too.”

“This could easily have been a much more interesting class but the lectures always left me bored and disengaged.”

“I was bored 90% of the time in class. Most of this stuff I already covered during my senior year of high school. Redundant.”

Course Summary

Students Enrolled: 239
School: Information
Term: Winter 2014
Course Number: SI100

Grade Distribution:
Mean: 73%

Student Information

Subject Majors: 15%
Freshmen: 86%
Sophomore: 10%
Junior: 3%
Senior: 1%

Prior GPA Distribution:
Mean: 3.2

Number of Responses: 160
Response Rate: 67%
Sentiment Ratio: 37% Positive
Feedback Types: 43% Lecture
Future steps

• More work on large-scale dataset analysis
• Toolkit for text analytics of educational discourse
• Web server for CourseBoard prototype
• Usability/User Experience Research of CourseBoard

• Thanks! Questions?

• Contact: Kevyn Collins-Thompson kevynct@umich.edu