Performing Sentiment Analysis on Princeton Course Evaluations: Investigating How Students in Humanities, Social Sciences and STEM Classes Express Satisfaction through Post-Course Surveys

Isabelle Ingato
Class of 2017
Adviser: Professor Christiane Fellbaum

Motivation and Goal

“[professor] has prodigious ability as a lecturer and really helped me to think critically”

“one of the defining experiences of your Princeton career”

“super well-designed”

- Domain Sensitivity in Sentiment Analysis
- Student Methodologies of Evaluation
- Feature Engineering and Classifier Tuning
- Study of a New Data Set

“a lot of work but worth it”

“tells a story and keeps your attention”

“[professor is] great… accessible, funny, and makes history come alive”

Problem Background and Related Work

Natural Language Processing with Python (Bird et al. 2015)

Sentiment Analysis & Opinion Mining (Liu 2012)

“Genre and Domain Dependencies in Sentiment Analysis” (Remus 2015)

“Mining Opinions in User-Generated Contents to Improve Course Evaluation” (El-Halees 2011)

“Sentiment Analysis in MOOC Discussion Forums: What does it tell us?” (Wen et al. 2014)

Approach

- Collected data set from Easy Princeton Course Evaluations (easypce.com)
- Used 25 highest ranked courses (with some exclusions) in each domain (larger data set compared to Wen's analysis)
- Interested particularly in satisfied reviews
- No neutral class
- Amazon Mechanical Turk for gold standard
- Chi-Squared for feature selection
- Naive Bayes for classification

Data

Size of Data Set

	Reviews	Words	Characters
Humanities	778	10543	62386
Social Sciences	986	15525	91818
STEM	772	12583	74980
All	**2536**	38651	229184

Average Number of Reviews per Course

Humanities	31.12
Social Sciences	**39.44**
STEM	30.88
All	33.8

Average Length of a Review

Humanities	**13.55 words**
Social Sciences	15.75 words
STEM	**16.3 words**
All	15.24 words

Humanities
Total Adjectives: 736
Types: 208
Most Frequent:

great	84
interesting	67
interested	43
good	41
easy	**22**
worth	**21**
wonderful	13
enjoyable	**12**

STEM
Total Adjectives: 938
Types: 256
Most Frequent:

interesting	81
good	65
interested	50
great	49
sure	29
useful	**27**
difficult	**27**
worth	**22**

Social Sciences
Total Adjectives: 1200
Types: 298
Most Frequent:

interesting	148
great	111
good	81
interested	73
worth	**32**
easy	**23**
other	17
political	**17**

Implementation

- Preprocessing
  - Removing bad ASCII characters
  - Removing personal identifiers
  - Reviews of length within one standard deviation from the mean
  - Tokenization, POS tagging
- Tools
  - NLTK
  - NumPy

- Bag of (unstemmed, case insensitive) adjectives representation of reviews
- Chi-Squared
- Training data set (75%) and testing data set (25%)
- Naive Bayes
- Precision, recall, f-measure

Bag of Words (Adjectives) Representation

Results

Humanities

Precision .86

Recall .92

F-measure .89

STEM

Precision .84

Recall .9

F-measure .87

Precision

STEM on HUM .85

HUM on STEM .79

SOC on HUM .85

HUM on SOC .85

Social Sciences

Precision .85

Recall .89

F-measure .87

All (Tested on mixed data)

Precision .7

Recall .85

F-measure .77

Preliminary Results using Chi-Squared

STEM

Precision .77
Recall .97
F-measure .86

Incorrectly Classified Examples

Incorrectly Classified by STEM Classifier:

“Take this class! Start programs early and study **hard** for tests. Algorithm design questions are very **difficult** so make sure you look at old exams.” (Satisfied misclassified as Not Satisfied)

Incorrectly Classified by Social Sciences Classifier:

“If you like ___ 's teaching style and have an interest in public economics, then definitely take this class!” (Satisfied misclassified as Not Satisfied)

Incorrectly Classified by Humanities Classifier:

“The readings just aren't that **interesting**.” (Not Satisfied misclassified as Satisfied)

Conclusions… so far!

- Adjectives are a good feature for identifying satisfied reviews in course evaluations corpora (but we need more)
- Domain sensitivity is apparent at this level
- Exist significant domain differences among STEM, SOC, and HUM regarding satisfaction expression and evaluation methodology
- Domain sensitivity seems to have greatest effect on number of false positives (precision) here
- Naive Bayes is an effective model

Next Steps and Future Work

- Exploring additional features, selection techniques and classifier tuning
  - Identify and use evaluative adjectives only (using SentiWordNet)
  - “Extracting Resource Terms for Sentiment Analysis” (Zhang & Liu 2011)
  - “Using Word Lengthening to Detect Sentiment in Microblogs” (2011)
  - “Scope of Negation Detection in Sentiment Analysis” (Dadvar, Hauff & Jong 2011)
- Performing 2-fold cross validation
- Investigating correspondence between class size, response and satisfaction
- Other potential quantitative and qualitative evaluations
  - Rank courses instead by percent satisfied reviews it receives from my classifier and compare this ranking to that of Easy Princeton Course Evaluations

Acknowledgements

Thank you to my adviser, Professor Fellbaum, for her support and guidance throughout my project.

Thank you to Dean Bogucki and SEAS for the funding to complete this project.

Additional References

“An Introduction to Concept Level Sentiment Analysis” (Cambria 2013)

“Building Affective Lexicons from Specific Corpora for Automatic Sentiment Analysis” (Bestgen 2008)

“Qualitative Evaluation of Multimodal e-Learning Content using Sentiment Analysis” (Sadanandan et al.)

“Identifying E-learner’s Opinion using Automated Analysis in E-learning” (Bharathisindhu & Brunda 2014)