Yelp Elitism!

Predicting Annual Yelp Elite User Selections

Pratibha Rathore, Rachel Lee, Michael Peng

Project Goal

Given a user's reviews for a year, will he/she be awarded elite status in the following year?

"The Yelp Elite Squad is our way of recognizing and rewarding people who are active in the Yelp community and role models on and off the site" [1]

Hypothesis & Data Exploration

- How does text in elite reviews differ from text in normal reviews?
- How does average number of votes per review for users change over time?
- Are elite users first to review a new business?
- Does a user's metadata indicate his/her status?
- Does the social network structure suggest whether a user is elite user?

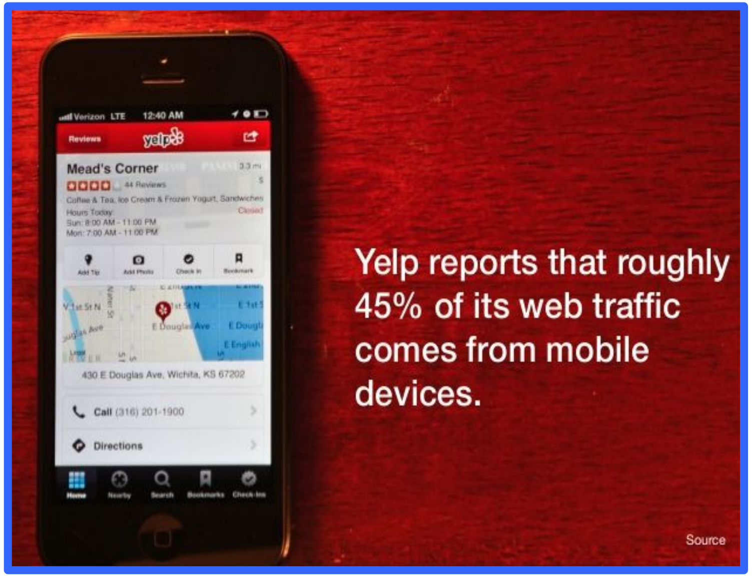
Yelp Academic Dataset[2,3]

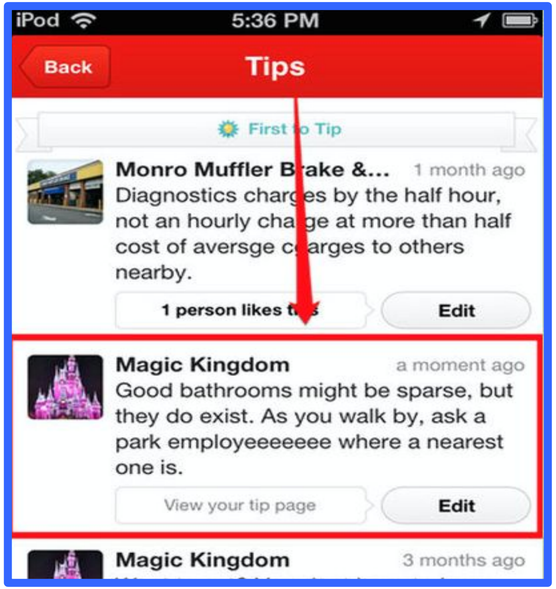
JSON objects, 1.6M reviews, 366K users, 61K businesses[1]

Statistics	Elite Users	Non-Elite Users
Review Count (per user)	245	16
Review Length (# words)	98.8	64.9
Vocabulary (size across all reviewers of type)	125,137	95,428
Votes (# on user's reviews)	1336	32
Friends	55	3
Compliments	8	0.0000
Fans	16	0.0000

User Tips

- Tip: short chunk of text user can submit to restaurant via mobile[1]
- Relatively newer, not taken into particular consideration[4]





User Metadata

- Review count: total number of reviews
- Number of friends: total number of friends of user
- Number of fans: total fans of user
- Average rating: avg star rating [1-5] user gives to businesses
- Number of compliments: separate from reviews, given by other users
- "Network metadata" e.g. friends, fans, compliments found to be insufficiently explanatory for status[5]

Review Metadata

```
{   'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},   }
```

- Stars: no correlation between rating given and elite status
- Text: worth further analysis -- NLP!
- Date: relevant for annual user selection parsing
- Votes: relevant in user status determination

Review Metadata (continued)

- Average number of votes greater for elite users vs normal users
- Trend observed in all three vote types[2]

Elite vs Normal users Statistics

	useful votes	funny votes	cool votes
elite users	616	361	415
normal users	20	7	7

Temporal Analysis

- User and review object metadata
  - Reviews’ timestamps
  - User activity over time
  - Average number of votes received per review
  - Review grouping and ordering by date posted
  - Users’ social graphs
- Temporal analysis ultimately inconclusive[3]

Language Model

Background	Normal	Elite
the	gorsek	uuu
and	forks)	aloha!!!
a	yu-go	**recommendations**
i	sabroso	meter:
to	(***	**summary**
was	eloff	carin
of	-/+	no1dp
is	jeph	(lyrics
for	deirdra	friends!!!!!
it	ruffin'	**ordered**
in	josefa	8/20/2011
that	ubox	rickie
my	waite	kuge
with	again!!	;]]]
but	optionz	#365
this	ecig	g
you	nulook	*price
we	gtr	visits):
they	shiba	r_
on	kenta	ik

Lee and Massung, 2014[3]*:*

- Unigram Language Model: freq. dist. of top 20 unigram tokens
- Individual (non-stopword) token significance low, random, and user-biased
- Elite users likely to segment reviews into different sections, discussing different aspects of the business

Review Textual Features

- Average review length: # of tokens, chars across all user's reviews
- Average review sentiment: sentiment valence scores, opinion mining[3]
  - Computational cost-benefit infeasible
- Paragraph rate: paragraph segmentation, rate of multiple newline characters per review per user
- All caps: high rate might indicate spam or useless reviews
- Bad punctuation: to detect less serious reviews, new sentence not starting with capital letter
- Capitalization/punctuation: minimal impact, computationally intensive
  - Preprocessing filters out most "low-effort" posts (see preprocessing)

Review Textual Features (continued)

- Readability scores: based off character, syllable, word, complex word, and sentence counts in a text
  - Flesch-Kincaid Grade Level
  - Automated Readability Index
  - Coleman-Liau Index
  - Flesch Reading Ease
  - Gunning Fog Index
  - LIX
  - SMOG Index
  - RIX

Readability scores were ultimately not informative features for Yelp reviews. Review texts appear to be generally too short to form informative scores differentiating elite from non-elite user reviews.

Review Textual Features (continued)

- Parts of speech counts: for sentence beginnings, and for general word usage
    - pronoun
    - conjunction
    - interrogative
    - preposition
    - to be verb
    - auxiliary verb

Parts of speech, especially at sentence beginnings seemed intuitively useful, but our results were inconclusive. It is possible, with further testing, that a different combination of features may prove informative.

Data Preprocessing

- Only consider users w/ 20+ reviews
  - Only 0.083% of elite users have <20 reviews
  - Confounds impact of review text
- Only consider reviews in US cities
  - Other cities often have foreign-language reviews[2]
  - More review text standardization

Feature Selection

```
feature_dict = {
        1: "total reviews",
        2: "total characters",
        3: "total paragraphs",
        4: "total cool votes",
        5: "total funny votes",
        6: "total useful votes",
        7: "total sentences",
        8: "total words",
        9: "total size of vocabulary (unique words)",
        10: "chars per review",
        11: "paragraphs per review",
        12: "cool votes per review",
        13: "funny votes per review",
        14: "useful votes per review",
        15: "sentences per review",
        16: "words per review",
        17: "size of vocabulary per review"
}
```

17 features selected

Main feature engineering limitations:

- Processing power: many NLP features computationally infeasible in context of machine learning featurization
- Feature utility: many NLP-related features, e.g. unigram tokens, judged ineffective as seen in papers or through data examination

Feature Selection (continued)

Top Features

Rank	Feature	Importance Score (0-1)
1	Total Paragraphs	.2332
2	Total Characters	.1368
3	Paragraphs Per Review	.1271
4	Total Cool Votes	.1204
5	Characters Per Review	.0936
6	Total Useful Votes	.0702

Remainder of features: <.0400 importance score, omitted

Training, Development, & Test Sets

- Training set: 95,575 reviews
- Development set: 23,894 reviews
- Test set: 34,770 reviews
    - Only and all reviews in 2014
    - Train & dev sets contain only reviews posted prior
- How does model perform on most recent year with labeled data?
- Feature normalization

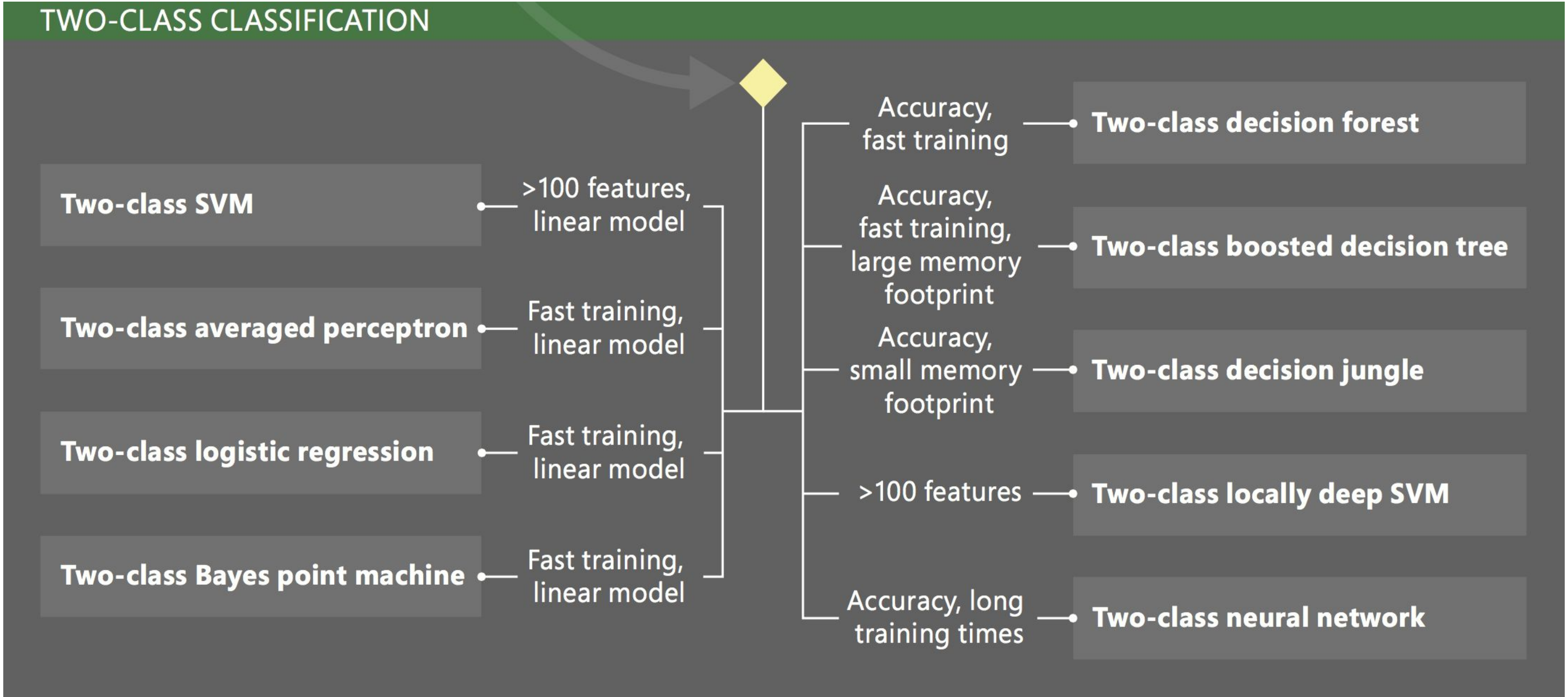
It’s a Classification Problem

- Binary classification: elite versus non-elite years of reviews
- scikit-learn learning machines considered:
  - Naive Bayes
  - SVM
  - Logistic Regression
  - Random Forests

Learning Model Selection

- Binary classifier selection[6,7]
- Naive Bayes: classic bag of words model w/ NB insufficient[2]
- SVM: training time infeasible (1 hour+ per run)
- Logistic Regression: promising, but insufficient for this purpose
  - Running on *training* data yields same accuracy as guessing!
- Random Forests: highest overall performance metrics
  - On dev set: ~80% accuracy, 70% precision, 15% recall

Learning Model Selection[6]



Learning Model Selection (continued)

- Random Forests: selected for accuracy, quick training times
- Important: resistant to overfitting
- Hyperparameters used:
    - n_estimators=40
    - max_depth=5

Feature Selection (continued)

Top Features

Rank	Feature	Importance Score (0-1)
1	Total Paragraphs	.2332
2	Total Characters	.1368
3	Paragraphs Per Review	.1271
4	Total Cool Votes	.1204
5	Characters Per Review	.0936
6	Total Useful Votes	.0702

Remainder of features: <.0400 importance score, omitted

Results

Confusion Matrix

	Elite User	Non-Elite User
Classified Elite	T Positive: 1,143	F Positive: 574
Classified Non-Elite	F Negative: 4,680	T Negative: 28,373

Results (continued)

- Selected model: Random Forest on non-normalized data
- Model accuracy: 0.8489
- Model precision: 0.6657
- Model recall: 0.1963
- Model F_1 Score: 0.3032
- Baseline: ~76% users non-elite, so 76% accuracy achievable by always guessing "non-elite"
- Accuracy seems OK, precision mediocre, recall low[8]
    - Bias for elite underestimation

Discussion

- Paragraphing (indicated by double newline) most significant (total, avg)
  - Inference: Elite users use “narratives”, sectioning
- Characters counts also significant (total, avg)
  - Inference: Elite users simply have high raw output
- Cool votes, useful votes (totals only)
  - Relation to “cliquey-ness” and utility?
- Noteworthy feature impact (lack thereof): total reviews
  - Importance score: 0.0034
  - Inference: past threshold of 20 reviews total for user (ever), review count low impact in decision tree

Discussion (continued)

- For the Q: "Has user X *ever* been Elite?", 96.4% accuracy achievable
  - With single feature: reviewCount[2]
- Object metadata, quantitative data apparently more predictive
  - At least for data considered as whole
  - Most NLP-related features low contribution, high cost
- Lack of computational power w.r.t. data size
- Continuation: utilize spaCy et al for quicker, more accurate NLP analysis
  - 45 mins+ algorithm text processing time

Questions?

Appendix

1. "Yelp." Dataset Challenge. Yelp, n.d. Web. 5 Dec. 2015.
2. Costa, Gian, Arturo Aguilar, and Eric Jiang. "Evaluating The Yelp Elite Squad." (2015): n. pag. University of California, San Diego. Web.
3. Lee, Cheng Han, and Sean Massung. "Multidimensional Characterization of Expert Users in the Yelp Review Network ∗." (n.d.): n. pag. Web.
4. "What Is Yelp's Elite Squad?" What Is Yelp's Elite Squad? Yelp, n.d. Web. 11 Dec. 2015.
5. Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval (n.d.): n. pag. 2008. Web.
6. Rohrer, Brandon. "Machine Learning Algorithm Cheat Sheet for Microsoft Azure Machine Learning Studio." Microsoft Azure. Microsoft, 13 Oct. 2015. Web.
7. Baharudin, Baharum, Lam Hong Lee, and Khairullah Khan. "A Review of Machine Learning Algorithms for Text-Documents Classification." Journal of Advances in Information Technology JAIT 1.1 (2010): n. pag. Web.
8. Brownlee, Jason. "Classification Accuracy Is Not Enough: More Performance Measures You Can Use - Machine Learning Mastery." Machine Learning Mastery. N.p., 21 Mar. 2014. Web. 10 Dec. 2015.

Review Activity Window

- Distribution of user's activity over time
- Window examined: user's first review to last review posted in the data-set
- Based on interval in days for each review, a score is calculated
- Hypothesis: score low for elite users compared to normal users

$$score = \frac{var(intervals) + avg(intervals)}{days\_on\_yelp}$$

Source : multidimensional characterisation of expert review in yelp by Chan Han lee