Bag-of-Words Models and Beyond

Sentiment, Subjectivity, and Stance
Ling 575
April 8, 2014

Roadmap

- Polarity classification baselines
  - Common features, processing, models
  - ‘Sentiment aware’ modifications
- Baseline vs state-of-the-art
- Improving the baseline
  - Incorporating linguistic features
  - Incorporating context features
- Topics and resources

Baseline Approaches

- Early approaches: Intuitive
  - Use lexicon of positive/negative words
  - Heuristic:
    - Count: |P| = # positive terms, |N| = # negative terms
    - If |P| > |N|, assign positive, else negative
  - Simple!
  - Can work surprisingly well!

Sentiment Lexicon Analysis

- Many issues still unresolved
- Possible solution for domain sensitivity:
  - Learn a lexicon for the relevant data
  - Range of approaches:
    - Unsupervised techniques
    - Domain adaptation
    - Semi-supervised methods
- However, still fundamentally limited

Machine Learning Baselines

- Similar to much of contemporary NLP
- Sentiment analysis explosion happened when
  - Large datasets of opinionated content met
  - Large-scale machine learning techniques
- Polarity classification as machine learning problem
  - Features?
  - Models?

Baseline Feature Extraction

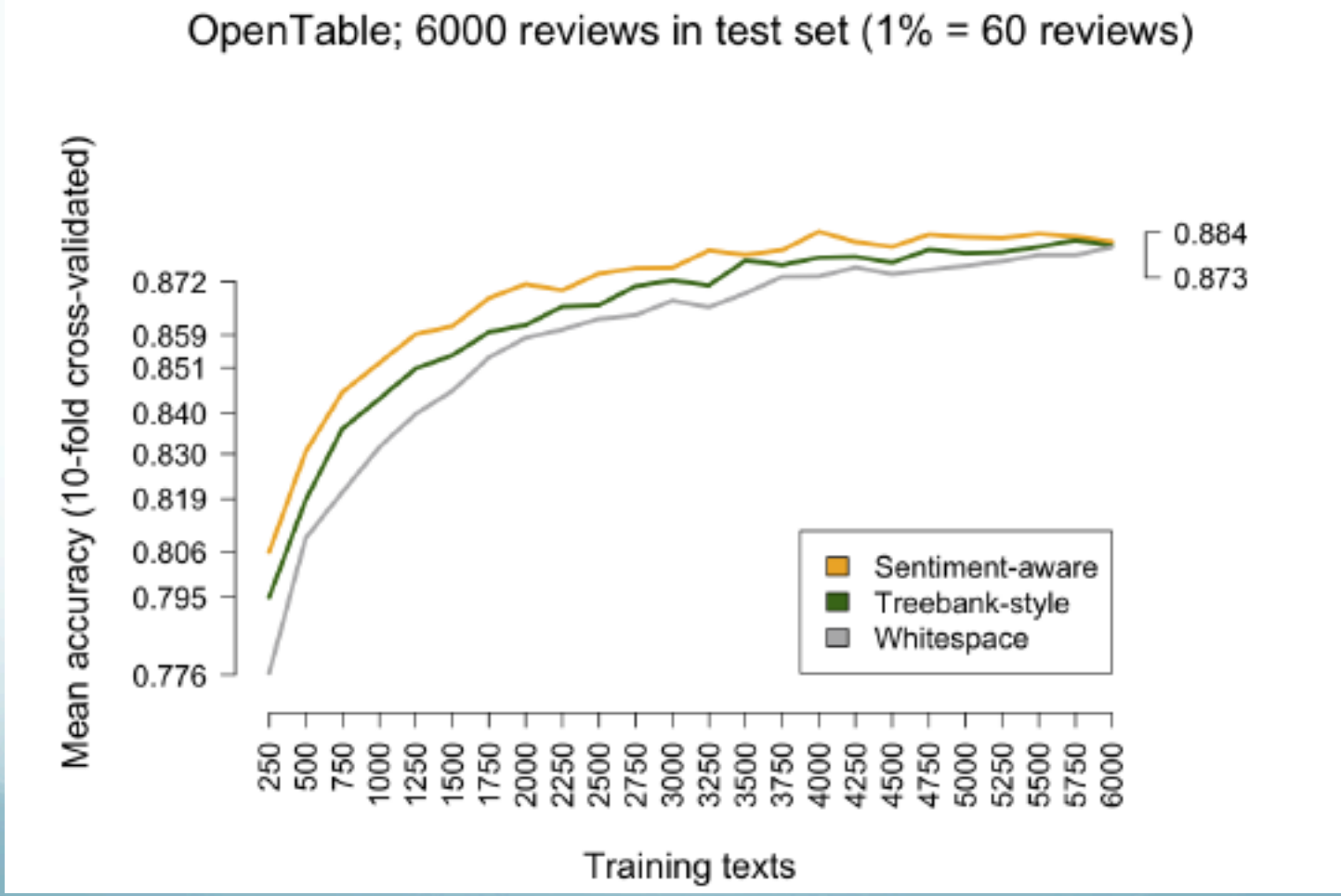
- Basic text features?
  - Bag-of-words, of course
    - N-grams
- Basic extraction:
  - Tokenization?
  - Stemming?
  - Negation?

Tokenizing

- Relatively simple for well-formed news
- Sentiment analysis needs to work on:
  - Sloppy blogs, tweets, informal material
  - What's necessary?
    - Platform markup handling/extraction
    - Emoticons ☺
    - Normalize lengthening
    - Maintain significant capitalization
    - Handle swear masks (e.g. %$^$ing)
- Comparisons on 12K OpenTable reviews: 6K: 4,5; 6K: 1,2
    - Results from C. Potts

Sentiment-Aware Tokenization

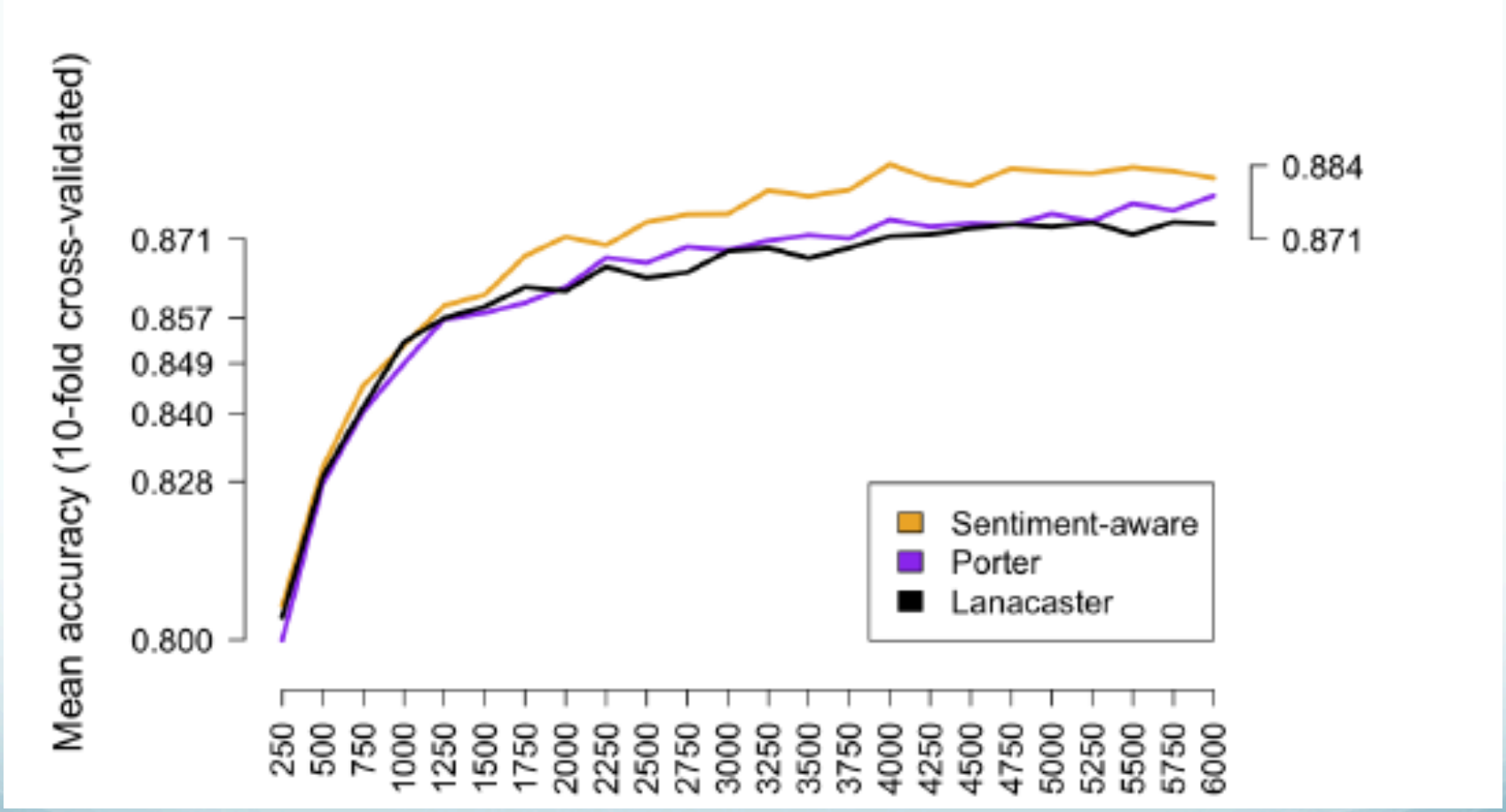
- From C. Potts



Stemming

- Should we stem?
  - Pros:
    - Reduces vocabulary, shrinks feature space
    - Removes irrelevant distinctions
  - Cons:
    - Can collapse relevant distinctions!

Stemming Impact on Sentiment Classification



Take home: Don't just grab a stemmer for sentiment analysis

Sentiment meets the Porter Stemmer

- Porter stemmer:
  - Classic heuristic rule cascade
    - Repeatedly strips off suffixes based on patterns
    - Highly aggressive
  - Applied to the General Inquirer
    - Destroys key contrasts

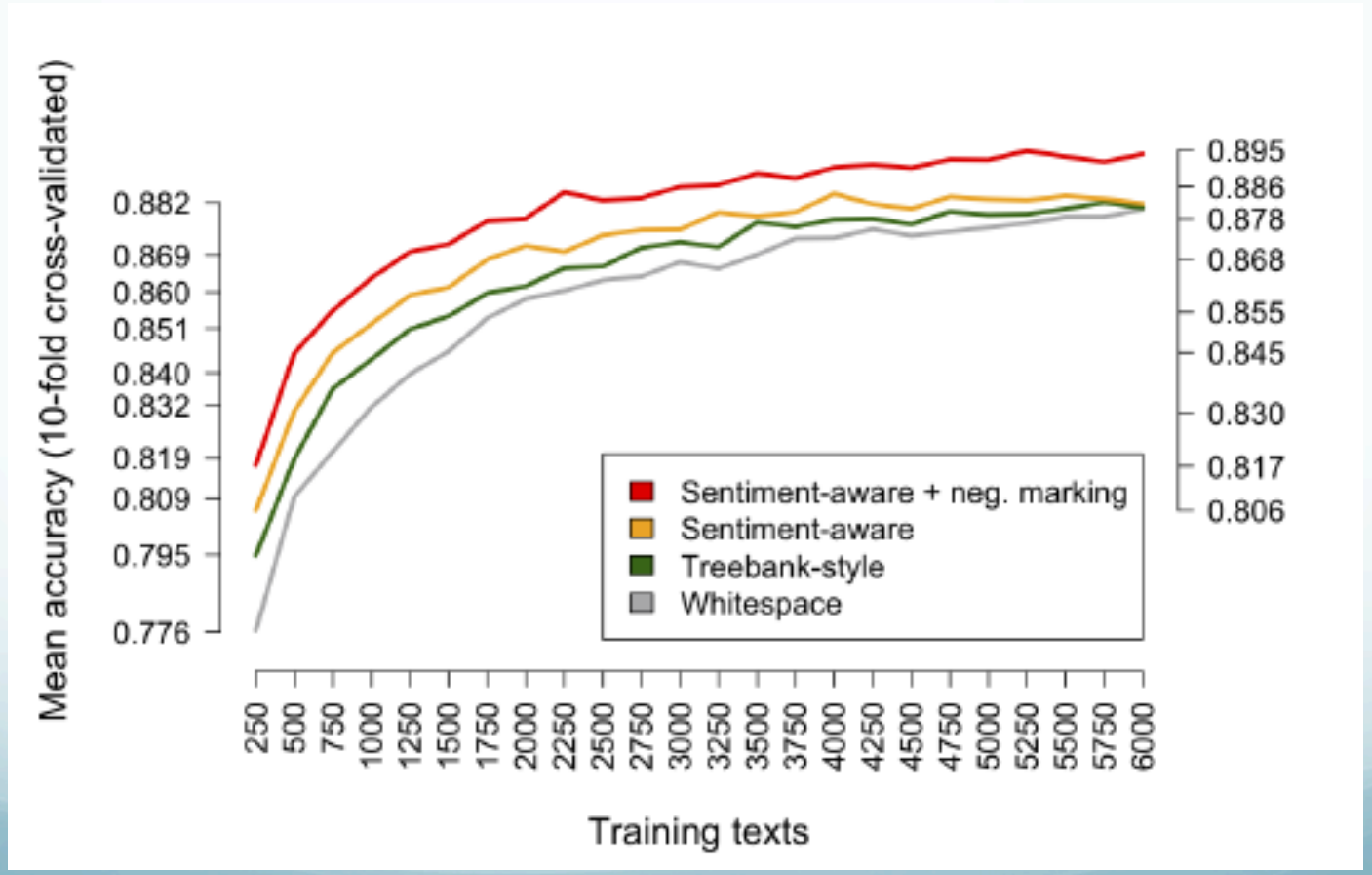
Positiv	Negativ	Porter stemmed
defense	defensive	defens
extravagance	extravagant	extravag
affection	affectation	affect
competence	compete	compet
impetus	impetuous	impetu
objective	objection	object

Naïve Negation Handling

- Negation:
  - The book was not good.
  - I did not enjoy the show.
  - No one enjoyed the movie.
- Approach due to Chen & Das, 2001
  - Add _NEG to each token between negation and end of clause punctuation
    - I did not enjoy the show. →
      - I did not enjoy_NEG the_NEG show_NEG

Impact of Negation Marking on Sentiment Analysis

- Even simple handling provides a boost



Bag-of-Words Representation

- Do polarity classification on:

Jane	so	want	from	over	that
can't		beat	madden	shinbone	up
read	my	Austen	Prejudice	reader	her
frenzy	Pride	conceal		I	
and	books		Everytime	with	dig
the	own	skull		to	me

Full text: Jane Austen's book madden me so that I can't conceal my frenzy from the reader. Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shinbone. - Mark Twain

Bag-of-Words Representation

- Choices:
  - Binary (0/1) vs Frequency?
- For text classification?
  - Prefer frequency
    - Associated with ‘aboutness’ relative to topic
- For sentiment?
  - Prefer binary
    - Multiple words with same polarity, not same words
- For subjectivity detection?
  - Prefer *hapax legomena* : singletons
    - Unusual, out-of-dictionary words: e.g. “bugfested”

Baseline Classifiers

- MaxEnt:

$$P(class \mid text, \lambda) = \frac{\exp\left(\sum_i \lambda_i f_i(class, text)\right)}{\sum_{class'} \exp\left(\sum_i \lambda_i f_i(class', text)\right)}$$

  - Discriminative classifier
  - Can handle large sets of features with internal dependencies
  - Select highest probability class
    - Typically with little regard to score

Other Classifiers

- Support Vector Machines (SVMs)
  - Performance typically similar to or slightly better
    - Relative to MaxEnt (see Pang et al, 2002)

- Boosting
  - Combination of weak learners
  - Applied in some cases

Classification vs Regression

- What about the non-binary case?
  - I.e. positive, negative, neutral, or
  - 1-5 stars
- It depends:
  - For 3-way positive/negative/neutral
    - Classification performs better
  - More fine-grained labels
    - Regression is better
- Why?
  - Hypothesis: More distinct vocab. in 3-way

Naïve Bayes vs MaxEnt

- OpenTable data; in-domain train/test

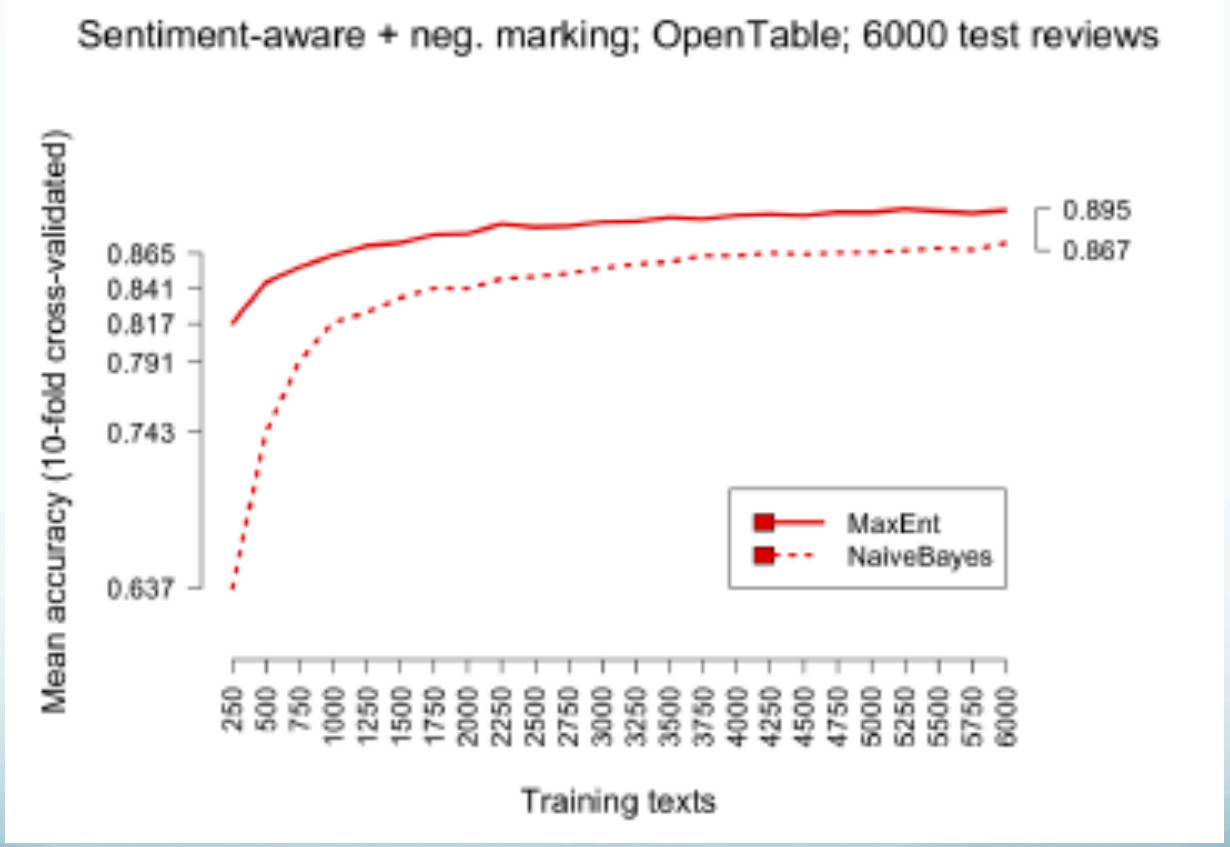
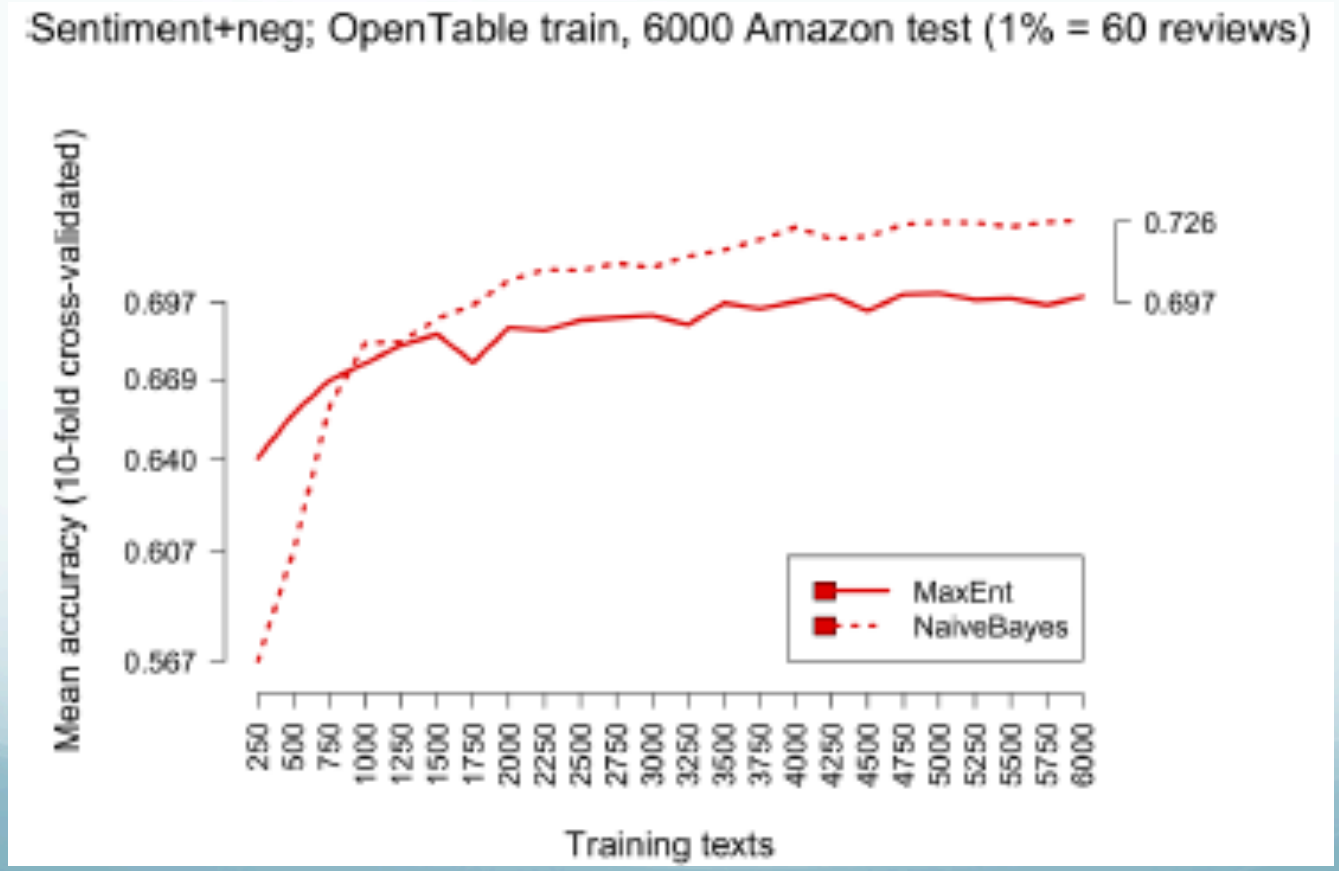


Figure from C. Potts

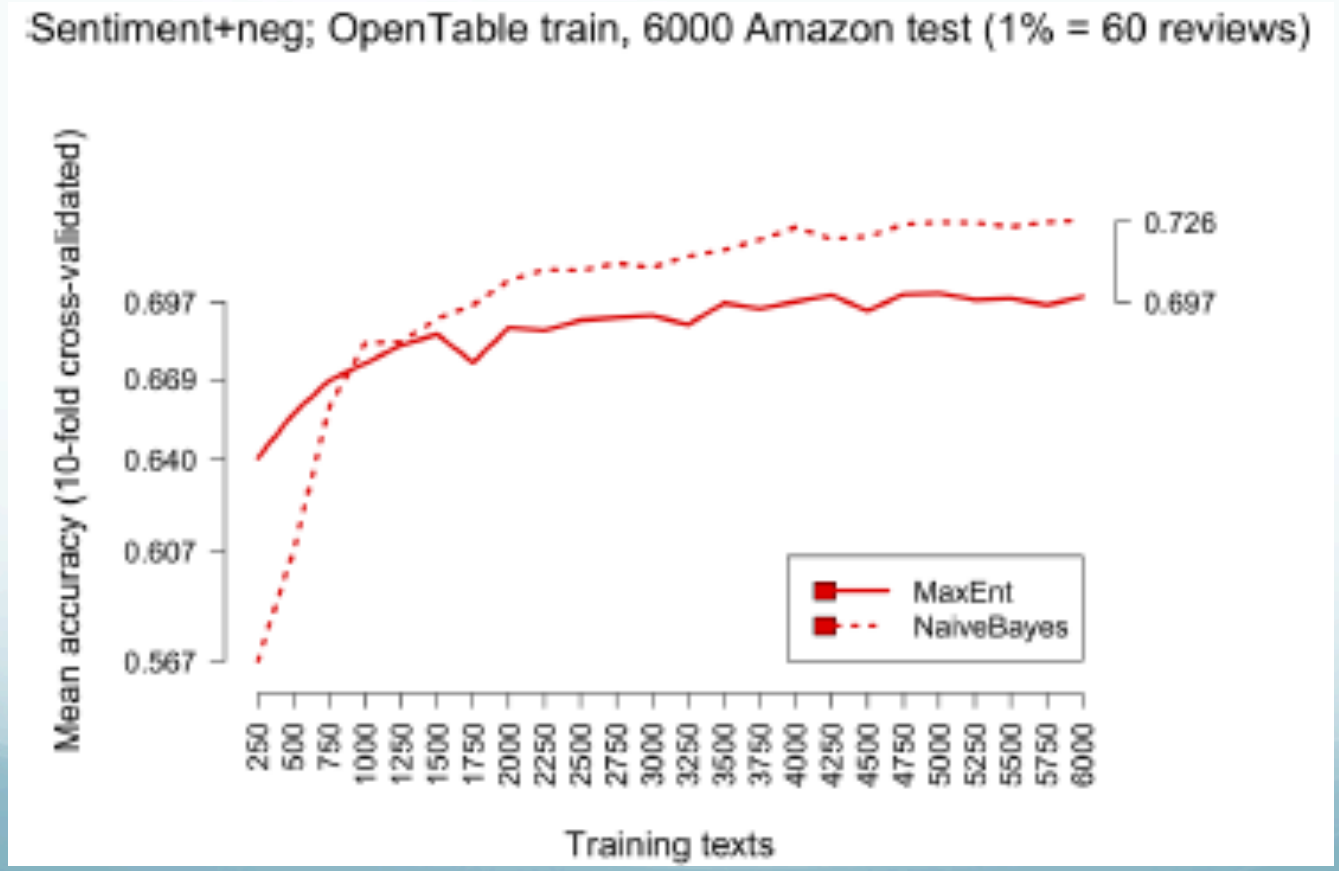
Naïve Bayes vs MaxEnt

- Cross-domain data:
  - OpenTable → Amazon



Naïve Bayes vs MaxEnt

- Cross-domain data:
  - OpenTable → Amazon ➔ MaxEnt overfits



Avoiding Overfitting

- Employ some feature selection
  - Threshold:
    - Most frequent features
    - Minimum number of occurrences
      - Sensitive to setting
  - Alternative criteria:
    - Mutual information, χ^2, etc
      - Some measures too sensitive to rare cases
  - Sentiment lexicons

Bag-of-Words

- Clearly, bag-of-words can not capture all nuances
  - Polarity classification hard for humans on that basis
- However, forms the baseline for many systems
- Can actually be hard to beat
  - MaxEnt classifiers with unigrams: >= 80%
    - On many polarity classification tasks
  - Current best results on polarity classification in dialog:
    - Combination of word, character, phoneme n-grams ~90% F-measure

Current Approaches

- Aim to improve over these baselines by
  - Better feature engineering
    - Modeling syntax, context, discourse, pragmatics
  - More sophisticated machine learning techniques
    - Beyond basic Naïve Bayes or MaxEnt models
- Recent state-of-the-art results (Socher et al)
  - Large-scale, fine-grained, crowdsourced annotation
  - Full parsing, syntactic analysis
  - Deep tensor network models

State-of-the-Art

- Rotten Tomatoes movie review data
  - 'Root'= sentence level classification

Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.7**	**87.6**	**85.4**

Table 1: Accuracy for fine grained (5-class) and binary predictions at the sentence level (root) and for all nodes.

Integrating Linguistic Evidence

- Sources of evidence:
  - Part-of-speech
  - Negation
  - Syntax
  - Topic
  - Dialog
  - Discourse

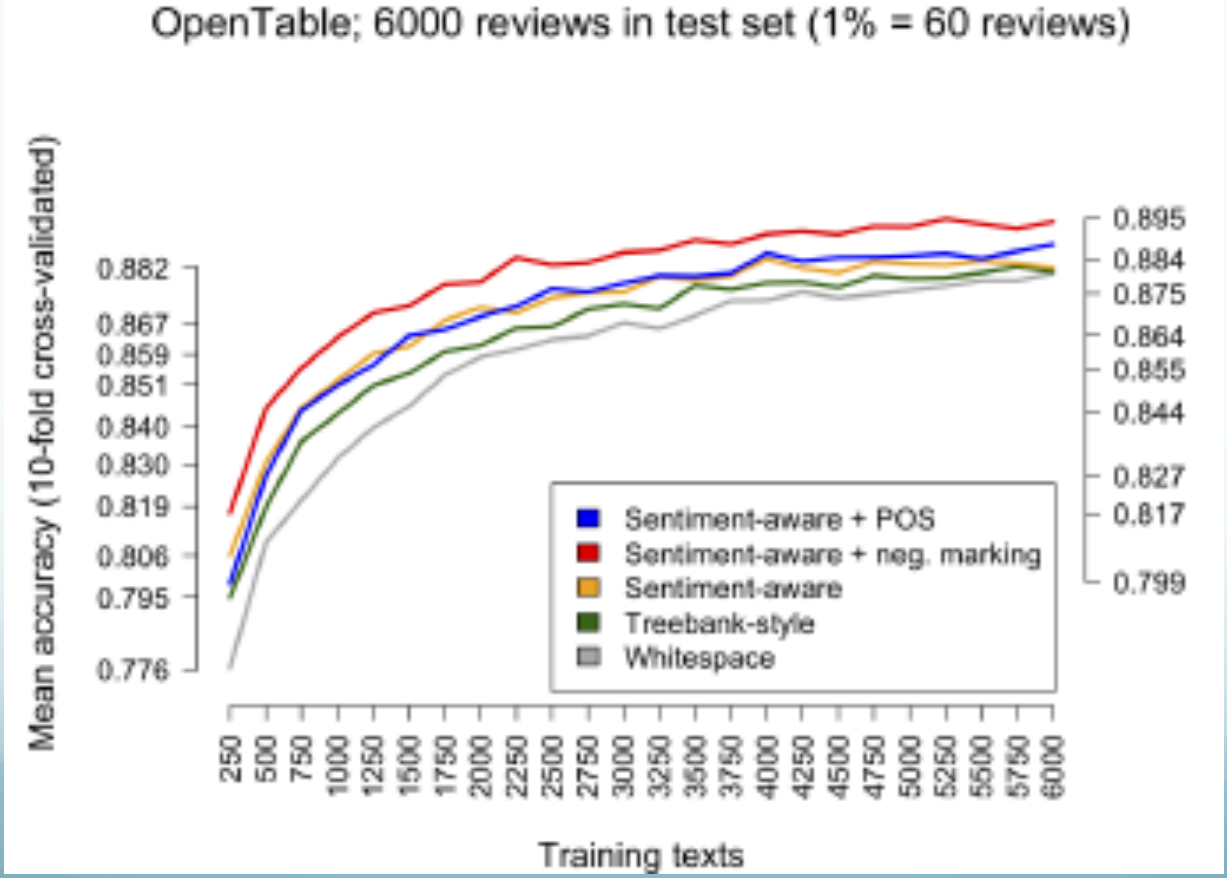
Part-of-Speech

- Why use POS?
  - Sentiment varies by word POS
    - Many sentiment-bearing words are adjectives
      - Just adjectives?
  - Simple, accurate form of WSD

Word	Tag1	Val1	Tag2	Val2
arrest	jj	Positiv	vb	Negativ
even	jj	Positiv	vb	Negativ
order	nn	Positiv	vb	Negativ
pass	nn	Negativ	vb	Positiv

Impact of POS Features

- Append POS tags to each word
  - It's a wash...



POS Ngram Features

- Bridge to syntax
  - Are some POS sequences good sentiment cues?
    - (Gentile, 2013)
  - Strongly positive:
    - PRP VBP PRP: (156/11) : I love it.
    - PRP RB VB DT NN: (83/1): I highly recommend this product
    - PRP RB VB PRP: (70/0) : I highly recommend it.
  - Strongly negative:
    - VBP RB VB PRP NN: (82/0): Don't waste your money.
    - VBP RB VB DT NN: (59/3): Don't buy this product.
    - VBP PRP NN: (59/13): Save your money.

Syntax

- Two main roles:
  - Directly as features: dependency structures
    - E.g. modifier relations in sentiment
      - Amod(book, good), advmod(wonderful, absolutely)
    - Structure in subjectivity
      - Xcomp(think, VERB)
    - Results somewhat variable

Syntax & Negation

- Another key role
  - Determining scope of *valence shifters*
    - E.g. scope of negation, intensifiers, diminishers
      - I really like this book vs
      - I don't really like this book vs
      - I really don't like this book
  - Simple POS phrase patterns improve by > 3% (Na et al)
  - Significant contributor to Socher's results
    - Phrase-level tagging/analysis
    - Compositional combination based on constituent parse
      - Handles double-negation, 'but' conjunction, etc

Negation & Valence Shifters

- Degree modification:
  - Very, really: enhance sentiment
- Intensifiers:
  - Incredibly: apply to lower sentiment terms
    - Confuse models
- Attenuators:
  - Pretty: weaken sentiment of modified terms
- Negation:
  - Reverses polarity of mid-level terms: good vs not good
  - Attenuates polarity of high-level terms: great vs not great

Incorporating Topic

- Why does topic matter?
  - Influences polarity interpretation
    - Walmart's profit rose:
      - Article is about Walmart → Positive
    - Target's profict rose:
      - Article is about Walmart → Negative
  - Within an opinionated document:
    - May not be all about a single topic
      - Blogs wander, may compare multiple items/products
      - To what does the sentiment apply

Incorporating Topic

- Common approach:
  - Multipass strategy
    - Search or classify topic
    - Then perform sentiment analysis
  - Document level:
    - Common approach to TREC blog task
  - Sentence-level:
    - Classify all sentences in document:
      - On/off-topic or label multiple topics
    - Perform polarity classification of sentences
      - Target of sentiment? Topic

Datasets

- Diverse data sets:
  - Web sites: Lillian Lee's and Bing Liu's
- Movie review corpora
- Amazon product review corpus
- Online and Congressional floor debate corpora
- Multi-lingual corpora: esp. NTCIR
- MPQA subjectivity annotation news corpus