# Baselines and Bigrams: Simple, Good Sentiment and Topic Classification

Using bag of features + linear classifiers

Sida Wang and Chris Manning

Stanford University

# Sentiment and Topical Classification

- Positive vs. negative:
  - "a sentimental mess that never rings true"

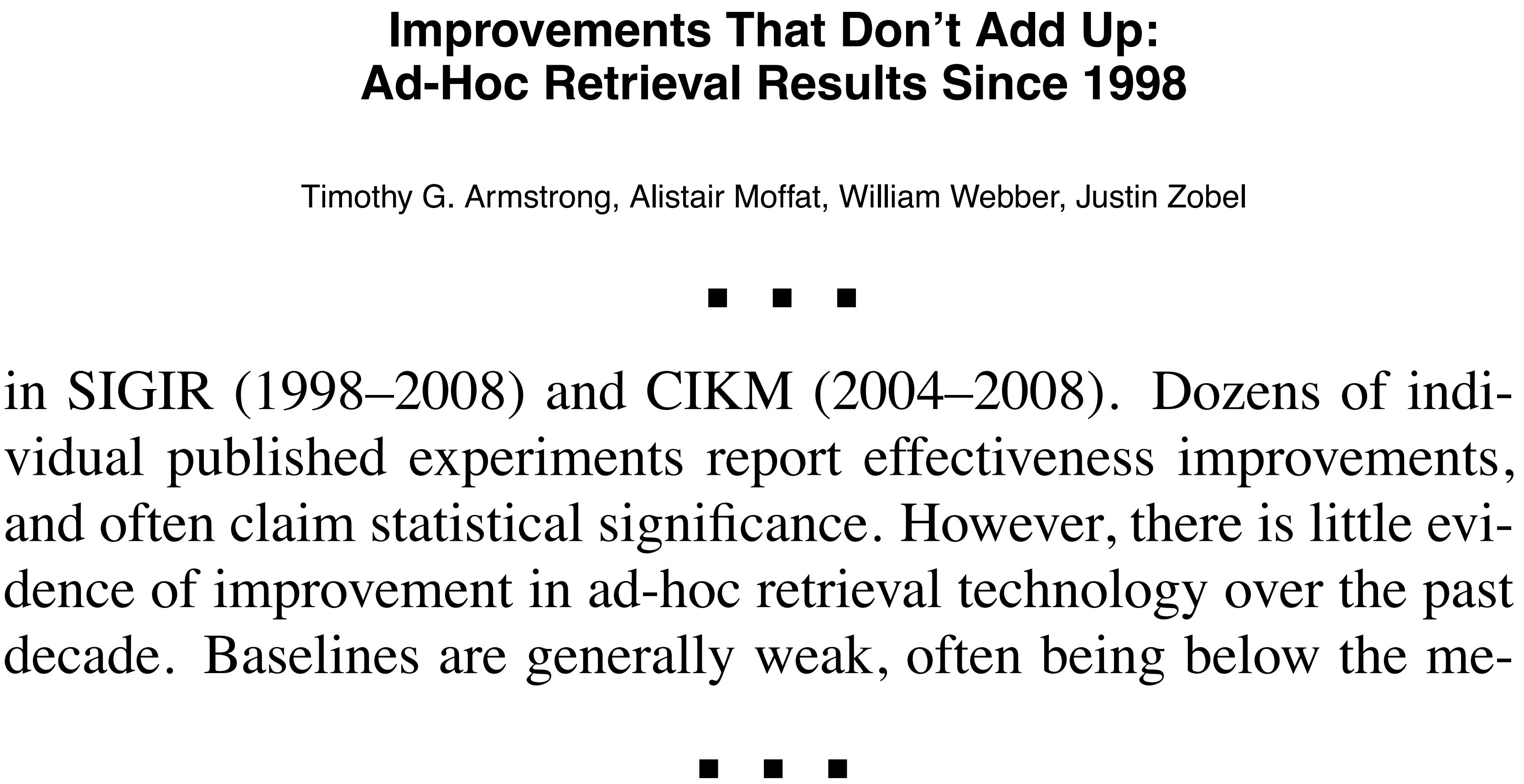
- Objective or subjective?
  - "the movie takes place in mexico, 2002“

- Classify atheism vs. christianity, baseball vs. cryptography, full IMDB reviews, and etc.

# Standard datasets

- Snippets
  - Rotten tomato movie reviews, Subjective vs. Objective
  - Customer reviews, MPQA polarity
- Longer reviews
  - Long IMDB reviews, large IMDB reviews
- Topical classification
  - Subtopics within 20-newsgroups, such as religion vs. atheism, cryptography vs. baseball

# How are we doing?



**Improvements That Don't Add Up:**
**Ad-Hoc Retrieval Results Since 1998**

Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel

. . .

in SIGIR (1998–2008) and CIKM (2004–2008). Dozens of individual published experiments report effectiveness improvements, and often claim statistical significance. However, there is little evidence of improvement in ad-hoc retrieval technology over the past decade. Baselines are generally weak, often being below the me-

. . .

# Sentiment Snippets

- Recent works on sentiment analysis tries to model compositionality and sentence structure

- Strong case that linear, bag of words classifiers are insufficient for classifying sentiments [Moilanen and Pulman, *RANLP07]*

- Examples:
  - "*not an inhumane monster*"
  - "*killing cancer*"
  - *"Interesting **BUT** not compelling"*

# On snippet datasets

- The most thorough comparisons in the literature are provided by [Nakagawa et al, ACL2010]
- Recursive autoencoder [Socher et al, EMNLP11]

| Methods | CR | MPQA | RTs |
|---|---|---|---|
| Voting * | 71.4 | 80.4 | 62.9 |
| Rule-Based * | 74.3 | 81.8 | 62.9 |
| BoF-w/Rev. * | 81.4 | 84.1 | 76.4 |
| Tree-CRF | 81.4 | 86.1 | 77.3 |
| RAE | N/A | 85.7 | 76.8 |
| RAE-pretrain* | N/A | 86.4 | 77.7 |

*: uses extra data or resource;  Blue is the best

# Linear classifiers perform well

- Linear BoW models did not do that badly

| Methods | CR | MPQA | RTs |
|---|---|---|---|
| Best previous result | 81.4 | *86.4 | *77.7 |
| **Multinomial Naïve Bayes (MNB)** | 79.8 | 85.3 | 77.9 |
| **Linear SVM** | 79.0 | 86.1 | 76.2 |

*: uses extra data or resource

Blue is the best

# Bigrams are very helpful

- (Uni): unigram, (Bi): bigram
- MNB: multinomial Naïve Bayes

| Methods | CR | MPQA | RTs |
|---|---|---|---|
| Previous SotA | 81.4 | *86.1 | *77.3 |
| **MNB (Uni)** | 79.8 | 85.3 | 77.9 |
| **MNB(Bi)** | 80.0 | 86.3 | 79.0 |
| **SVM (Uni)** | 79.0 | 86.1 | 76.2 |
| **SVM (Bi)** | 80.8 | 86.7 | 77.7 |

*: uses extra data or resource

Blue is the best

# In this work

- **Propose a simple but novel method that often gives state of the art performance**
  - **Can we do better?**
- Establish strong baselines for several standard sentiment analysis datasets of various types
- Analyze methods vs. type of datasets

# A simple method

Take Multinomial Naïve Bayes, and fit w discriminatively with regularization:

$$p(y|x;w) \propto p(y) \prod_{i;x_i=1} p(x_i|y)^{w_i}$$

Maximize the margins (SVM), or maximize the log-likelihood (Logistic regression) of this discriminant function

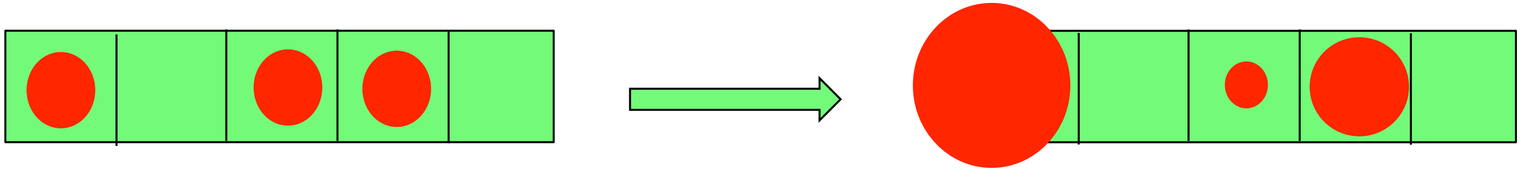
There are 2 equivalent interpretations.

# Interpretation 1: features

- Regular SVM uses the indicator vector $x_i = I\{v_i\}$
- We use:

$$x_i = I\{v_i\} \log \frac{p(x_i|y=1)}{p(x_i|y=0)}$$

- Just train a discriminative classifier with these feature vectors



[1 0 1 1 0] [3 0 .2 .9 0]

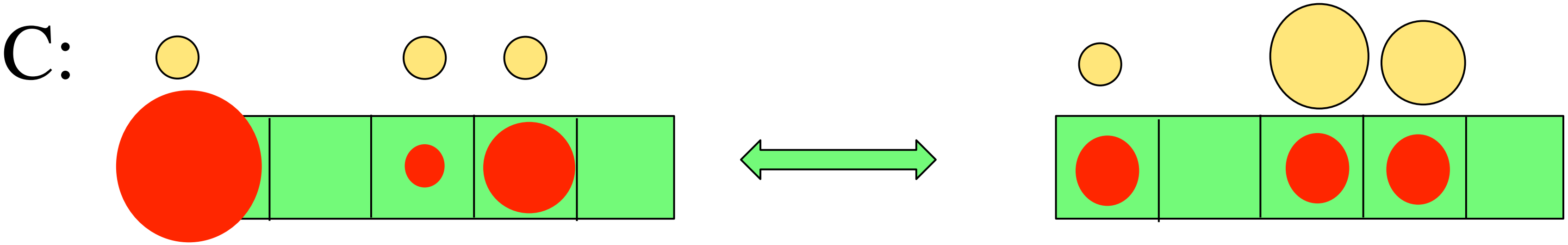
# Interpretation: regularization

- Use different regularization strengths

$$L'(y, x; w) = L(y, x; w) + \sum_i C_i w_i^2$$

$$\frac{1}{C_i} = |\log \frac{p(x_i|y=1)}{p(x_i|y=0)}| + \epsilon$$

Stronger regularization for features that are uninformative by themselves



# In this work

- Propose a simple, new method that often gives state of the art performances
  - Let’s call it NBSVM
- **Establish strong baselines for several standard sentiment analysis datasets**
  - **Use the same hyperparameter for all dataset. Experiment with many different datasets**
- Analyze methods vs. type of datasets

# NBSVM is good for snippets

- Compare with NBSVM with previous SotA
- No tuning for specific dataset

| Methods | CR | MPQA | RTs |
|---|---|---|---|
| Previous SotA | 81.4 | *86.1 | *77.3 |
| **NBSVM (Uni)** | 80.5 | 85.3 | 78.1 |
| **NBSVM (Bi)** | 81.8 | 86.3 | 79.4 |

*: uses extra data or resource

Blue is the best

# On longer documents

| Methods | PL04 (long) | IMDB (long) |
|---|---:|---:|
| Best baseline[1] | 85.4 | 87.80 |
| LDA[1] | 66.7 | 67.42 |
| Full+Unlab'd+BoW[1]* | 88.9 | 88.89 |
| WRRBM+BoW[2]* | N/A | 89.23 |
| MNB (Bi) | 85.8 | 86.59 |
| SVM (Bi) | 87.4 | 89.16 |
| NBSVM (Bi) | 89.4 (90.4) | 91.22 |

1 [Maas et al, ACL 2011] BoW vector space model
2 [Dahl et al, ICML 2012] 5-gram vector space model

*: uses extra data or resource

Blue is the best, red also tunes parameter

# NBSVM is a good baseline

| Dataset | Cases | l | Best baseline | SotA |
|---|---|---|---|---|
| RT-s | 10662 | 21 | Y | Y |
| CR | 3772 | 20 | Y | Y? |
| MPQA | 10624 | 3 | N? | N? |
| Subj. | 10000 | 24 | N? | N? |
| PL04 | 2000 | 787 | Y | N?/Y? |
| IMDB | 50000 | 231 | Y | Y |
| AthR | 1427 | 345 | Y | Y |
| XGraph | 1953 | 261 | Y | Y |
| BbCrypt | 1987 | 269 | Y | Y |

?: not statistically significant at the p=5% level

With hyperparameter tuning

# In this work

- Propose a simple new method that often gives state of the art performance
- Establish strong baselines for several standard sentiment analysis datasets
- **Analyze methods vs. type of datasets**

# Bigrams are more useful for sentiment analysis

| Dataset | Cases | l | Bi>Uni |
|---|---|---|---|
| RT-s | 10662 | 21 | Y |
| CR | 3772 | 20 | Y |
| MPQA | 10624 | 3 | Y |
| Subj. | 10000 | 24 | Y |
| PL04 | 2000 | 787 | Y |
| IMDB | 50000 | 231 | Y |
| AthR | 1427 | 345 | ? |
| XGraph | 1953 | 261 | ? |
| BbCrypt | 1987 | 269 | ? |

?: not statistically significant at the p=5% level

# Naïve Bayes is better for snippets

| Dataset | Cases | l | NB>SVM |
|---|---|---|---|
| RT-s | 10662 | 21 | Y |
| CR | 3772 | 20 | N? |
| MPQA | 10624 | 3 | N? |
| Subj. | 10000 | 24 | Y |
| PL04 | 2000 | 787 | N |
| IMDB | 50000 | 231 | N |
| AthR | 1427 | 345 | Y? |
| XGraph | 1953 | 261 | Y |
| BbCrypt | 1987 | 269 | Y |

?: not statistically significant at the p=5% level

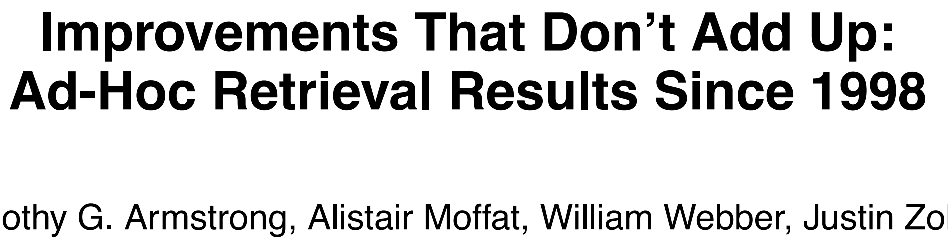
# Methods vs. datasets

- What is better? NB or LR/SVM?
  - NB is better for snippets
  - SVM/LR is better for full reviews
- Bigram vs. Unigram
  - Bigrams are more helpful for sentiment analysis than for topical classification
- Which NB? Which SVM?
  - Multinomial NB
  - L2-loss L2-regularized SVM
  - Or combine them

# Related works

- Jason Rennie et al. *Tackling the poor assumptions of naive bayes text classifiers*. ICML03
    - Use discounting (tf.idf), normalization
- Andrew McCallum et al. *A comparison of event models for naive bayes text classification*. AAAI98
    - MNB vs. Bernoulli NB: use MNB, and indicators
- Andrew Y Ng and Michael I Jordan. *On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes.* NIPS02

# So how are we doing?



**Improvements That Don't Add Up:**
**Ad-Hoc Retrieval Results Since 1998**

Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel

> "continuing to provide the same quantum improvement over the same modest baselines inspires neither confidence in our experimental methods nor conviction of our contribution."

- Very useful to explore models that are interesting, or linguistically plausible

- How important is performance?

# Better methods should exist

- Still no existing method outperforms linear classifier with bag of bigrams features…
  - We can afford a 1000 times slow down

- Opinion by many previous authors that linear BoW is very limited is still correct:

  - “There is an excellent 90 min film here, but it runs for 180” - real data in rt-polarty.negative

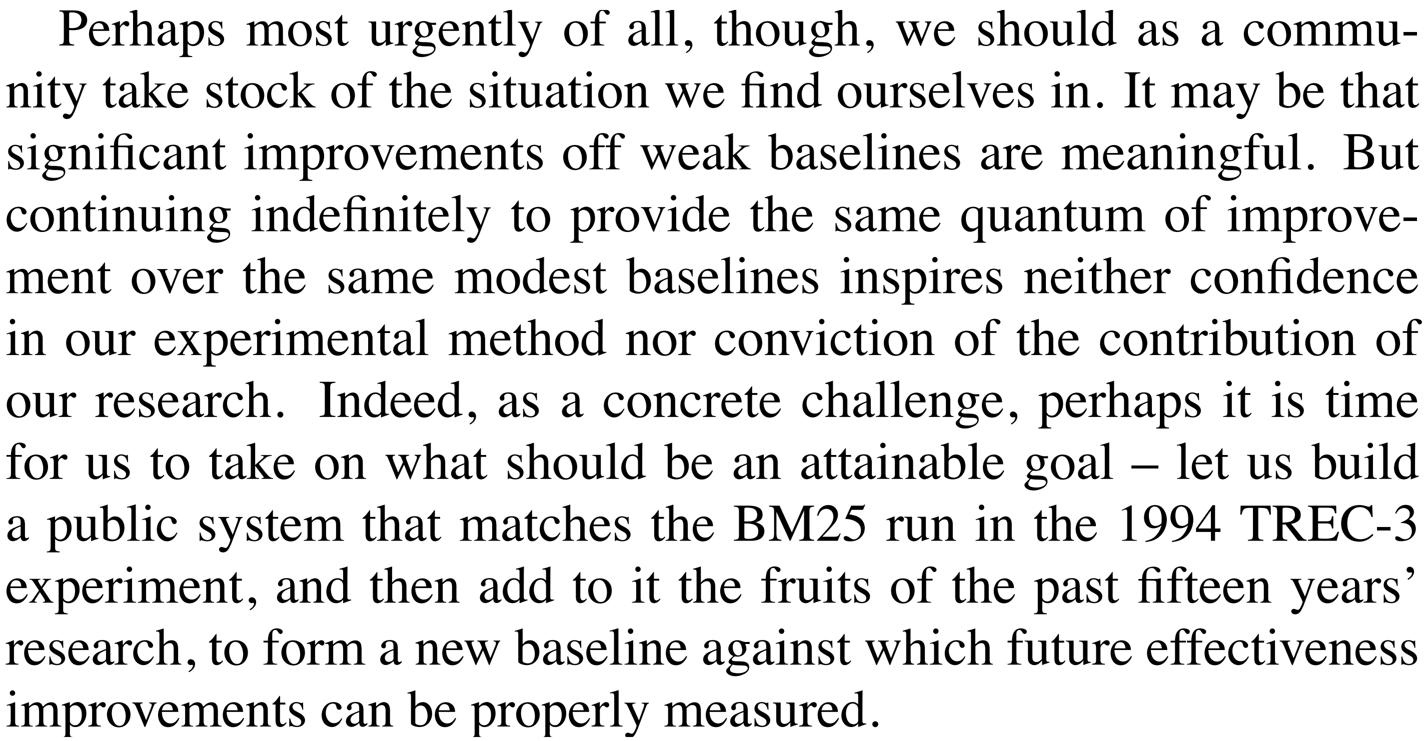
# Questions?

- Thank you!
- Code to replicate all our results available on my website www.stanford.edu/~sidaw

# Extra-slides

- Beginning of extra slides

# Armstrong’s conclusion



Perhaps most urgently of all, though, we should as a community take stock of the situation we find ourselves in. It may be that significant improvements off weak baselines are meaningful. But continuing indefinitely to provide the same quantum of improvement over the same modest baselines inspires neither confidence in our experimental method nor conviction of the contribution of our research. Indeed, as a concrete challenge, perhaps it is time for us to take on what should be an attainable goal – let us build a public system that matches the BM25 run in the 1994 TREC-3 experiment, and then add to it the fruits of the past fifteen years’ research, to form a new baseline against which future effectiveness improvements can be properly measured.

# Snippet full results

| **Method** | RT-s | MPQA | CR | Subj. |
|---|---|---|---|---|
| MNB-uni | 77.9 | 85.3 | 79.8 | **92.6** |
| MNB-bi | **79.0** | **86.3** | 80.0 | **93.6** |
| SVM-uni | 76.2 | 86.1 | 79.0 | 90.8 |
| SVM-bi | 77.7 | **86.7** | 80.8 | 91.7 |
| NBSVM-uni | **78.1** | 85.3 | 80.5 | 92.4 |
| NBSVM-bi | **79.4** | **86.3** | **81.8** | **93.2** |
| RAE | 76.8 | 85.7 | – | – |
| RAE-pretrain | 77.7 | **86.4** | – | – |
| Voting-w/Rev. | 63.1 | 81.7 | 74.2 | – |
| Rule | 62.9 | 81.8 | 74.3 | – |
| BoF-noDic. | 75.7 | 81.8 | 79.3 | – |
| BoF-w/Rev. | 76.4 | 84.1 | **81.4** | – |
| Tree-CRF | 77.3 | 86.1 | **81.4** | – |
| BoWSVM | – | – | – | 90.0 |

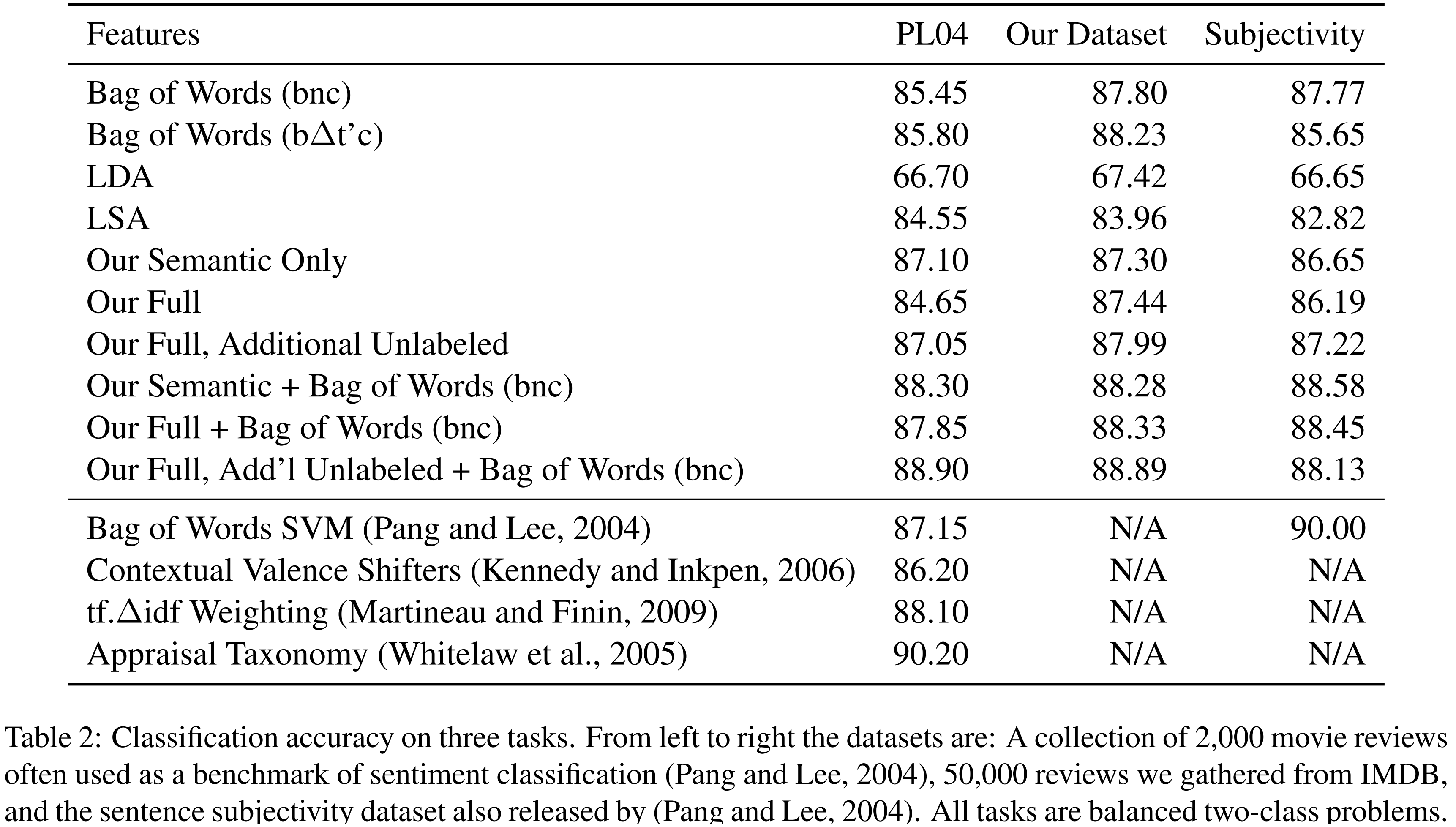
Table 2: Results for snippets datasets. Tree-CRF: (Nakagawa et al., 2010) RAE: Recursive Autoencoders (Socher et al., 2011). RAE-pretrain: train on Wikipedia (Collobert and Weston, 2008). "Voting" and "Rule": use a sentiment lexicon and hard-coded reversal rules. "w/Rev": "the polarities of phrases which have odd numbers of reversal phrases in their ancestors". The top 3 methods are in **bold** and the best is also **underlined**.

# Long doc results

| **Our results** | RT-2k | IMDB | Subj. |
|---|---|---|---|
| MNB-uni | 83.45 | 83.55 | **92.58** |
| MNB-bi | 85.85 | 86.59 | **93.56** |
| SVM-uni | 86.25 | 86.95 | 90.84 |
| SVM-bi | 87.40 | **89.16** | 91.74 |
| NBSVM-uni | 87.80 | 88.29 | 92.40 |
| NBSVM-bi | **89.45** | **91.22** | **93.18** |
| BoW (bnc) | 85.45 | 87.8 | 87.77 |
| BoW (b$\Delta t'$c) | 85.8 | 88.23 | 85.65 |
| LDA | 66.7 | 67.42 | 66.65 |
| Full+BoW | 87.85 | 88.33 | 88.45 |
| Full+Unlab'd+BoW | **88.9** | 88.89 | 88.13 |
| BoWSVM | 87.15 | – | 90.00 |
| Valence Shifter | 86.2 | – | – |
| tf.$\Delta$idf | 88.1 | – | – |
| Appr. Taxonomy | **90.20** | – | – |
| WRRBM | – | 87.42 | – |
| WRRBM + BoW(bnc) | – | **89.23** | – |

(Maas et al., 2011). **BoW**: linear SVM on bag of words features. **bnc**: **b**inary, **n**o idf, **c**osine normalization. **$\Delta t'$**: smoothed delta idf. **Full**: the full model. **Unlab'd**: additional unlabeled data. **BoWSVM**: bag of words SVM used in (Pang and Lee, 2004). **Valence Shifter**: (Kennedy and Inkpen, 2006). **tf.$\Delta$idf**: (Martineau and Finin, 2009). **Appraisal Taxonomy**: (Whitelaw et al., 2005). **WRRBM**: Word Representation Restricted Boltzmann Machine (Dahl et al., 2012).

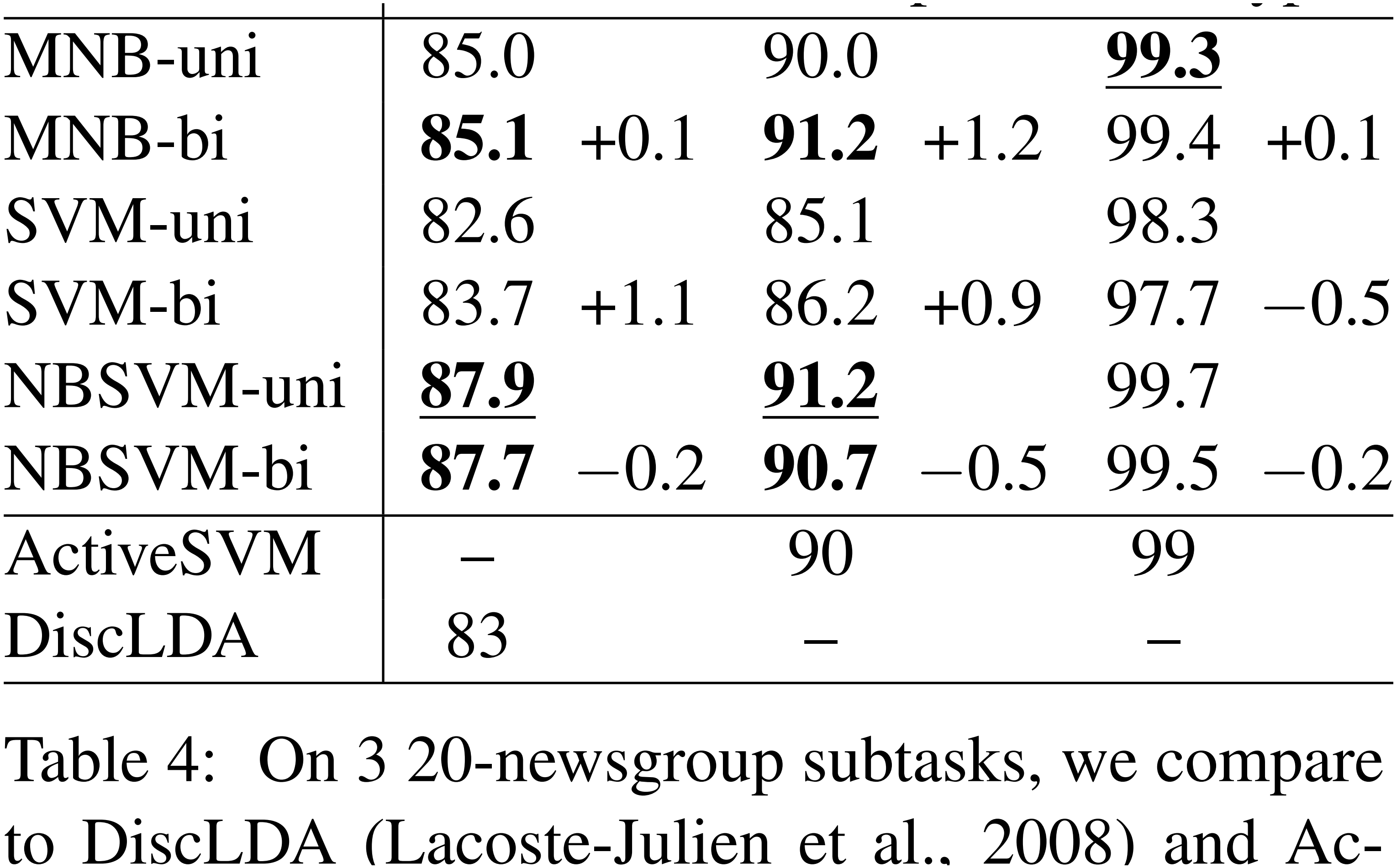
# Maas et al.



| Features | PL04 | Our Dataset | Subjectivity |
|---|---|---|---|
| Bag of Words (bnc) | 85.45 | 87.80 | 87.77 |
| Bag of Words (bΔt'c) | 85.80 | 88.23 | 85.65 |
| LDA | 66.70 | 67.42 | 66.65 |
| LSA | 84.55 | 83.96 | 82.82 |
| Our Semantic Only | 87.10 | 87.30 | 86.65 |
| Our Full | 84.65 | 87.44 | 86.19 |
| Our Full, Additional Unlabeled | 87.05 | 87.99 | 87.22 |
| Our Semantic + Bag of Words (bnc) | 88.30 | 88.28 | 88.58 |
| Our Full + Bag of Words (bnc) | 87.85 | 88.33 | 88.45 |
| Our Full, Add'l Unlabeled + Bag of Words (bnc) | 88.90 | 88.89 | 88.13 |
| Bag of Words SVM (Pang and Lee, 2004) | 87.15 | N/A | 90.00 |
| Contextual Valence Shifters (Kennedy and Inkpen, 2006) | 86.20 | N/A | N/A |
| tf.Δidf Weighting (Martineau and Finin, 2009) | 88.10 | N/A | N/A |
| Appraisal Taxonomy (Whitelaw et al., 2005) | 90.20 | N/A | N/A |

Table 2: Classification accuracy on three tasks. From left to right the datasets are: A collection of 2,000 movie reviews often used as a benchmark of sentiment classification (Pang and Lee, 2004), 50,000 reviews we gathered from IMDB, and the sentence subjectivity dataset also released by (Pang and Lee, 2004). All tasks are balanced two-class problems.

# Topical



| | | | | | | |
|---|---|---|---|---|---|---|
| MNB-uni | 85.0 | | 90.0 | | **99.3** | |
| MNB-bi | **85.1** | +0.1 | **91.2** | +1.2 | 99.4 | +0.1 |
| SVM-uni | 82.6 | | 85.1 | | 98.3 | |
| SVM-bi | 83.7 | +1.1 | 86.2 | +0.9 | 97.7 | −0.5 |
| NBSVM-uni | **87.9** | | **91.2** | | 99.7 | |
| NBSVM-bi | **87.7** | −0.2 | **90.7** | −0.5 | 99.5 | −0.2 |
| ActiveSVM | – | | 90 | | 99 | |
| DiscLDA | 83 | | – | | – | |

Table 4: On 3 20-newsgroup subtasks, we compare to DiscLDA (Lacoste-Julien et al., 2008) and Ac-

# Datasets

| **Dataset** | $(N_+, N_-)$ | $l$ | CV | $\|V\|$ | $\Delta$ |
|---|---|---|---|---|---|
| RT-s | (5331,5331) | 21 | 10 | 21K | 0.8 |
| CR | (2406,1366) | 20 | 10 | 5713 | 1.3 |
| MPQA | (3316,7308) | 3 | 10 | 6299 | 0.8 |
| Subj. | (5000,5000) | 24 | 10 | 24K | 0.8 |
| RT-2k | (1000,1000) | 787 | 10 | 51K | 1.5 |
| IMDB | (25k,25k) | 231 | N | 392K | 0.4 |
| AthR | (799,628) | 345 | N | 22K | 2.9 |
| XGraph | (980,973) | 261 | N | 32K | 1.8 |
| BbCrypt | (992,995) | 269 | N | 25K | 0.5 |

# Findings

- Establish strong baselines for several standard sentiment analysis datasets
  - A well chosen linear classifier performs very well
- Propose a simple new method that often gives state of the art performance
  - By combining NB and SVM/LR, this method is very robust
- Systematically analyze methods vs. datasets
  - MNB is better than Bernoulli NB, bigram is often better
  - The answer often depends on the class of dataset