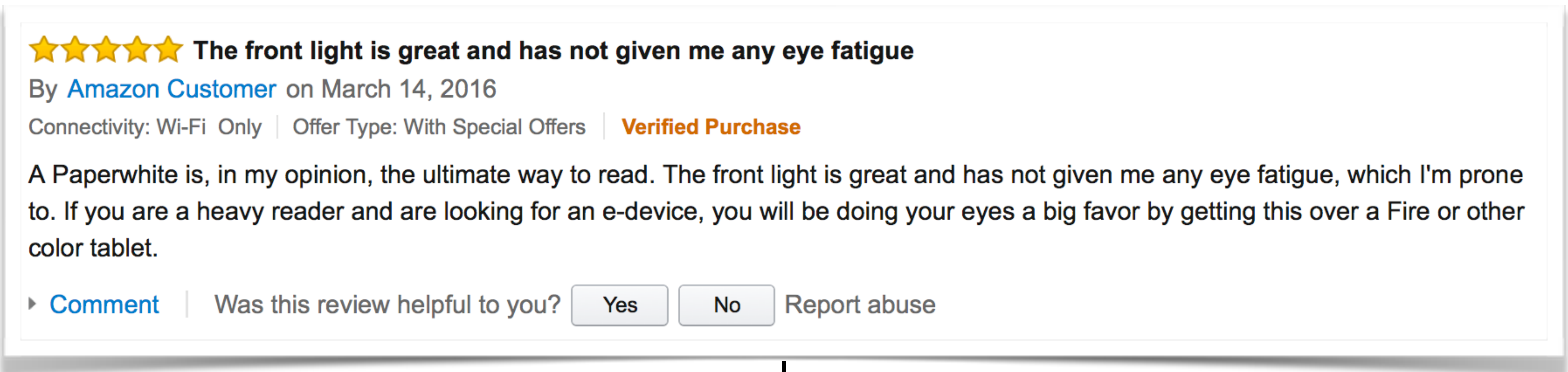
An Empirical Study of Skip-gram Features and Regularization for Learning on Sentiment Analysis

Cheng Li, Bingyu Wang, Virgil Pavlu and Javed A. Aslam
College of Computer and Information Science
Northeastern University

Sentiment Analysis

An Amazon Product Review



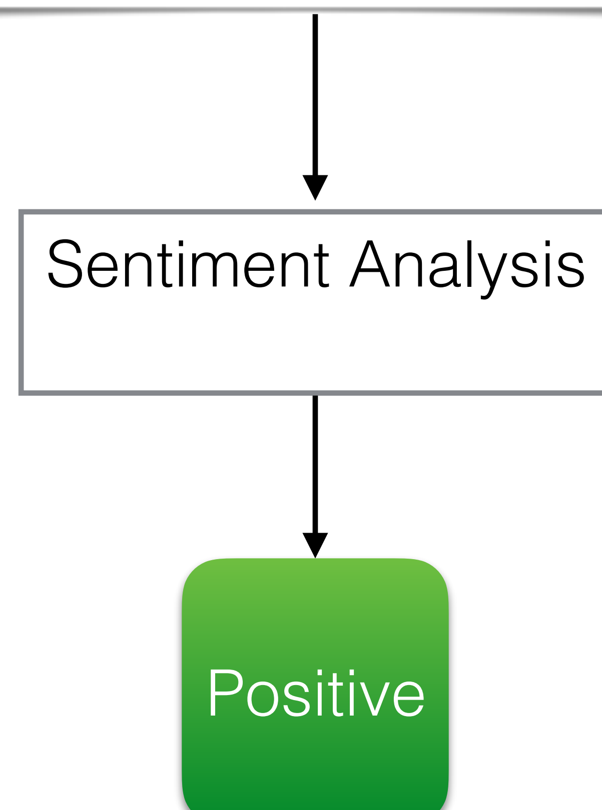
The front light is great and has not given me any eye fatigue

By Amazon Customer on March 14, 2016

Connectivity: Wi-Fi Only | Offer Type: With Special Offers | **Verified Purchase**

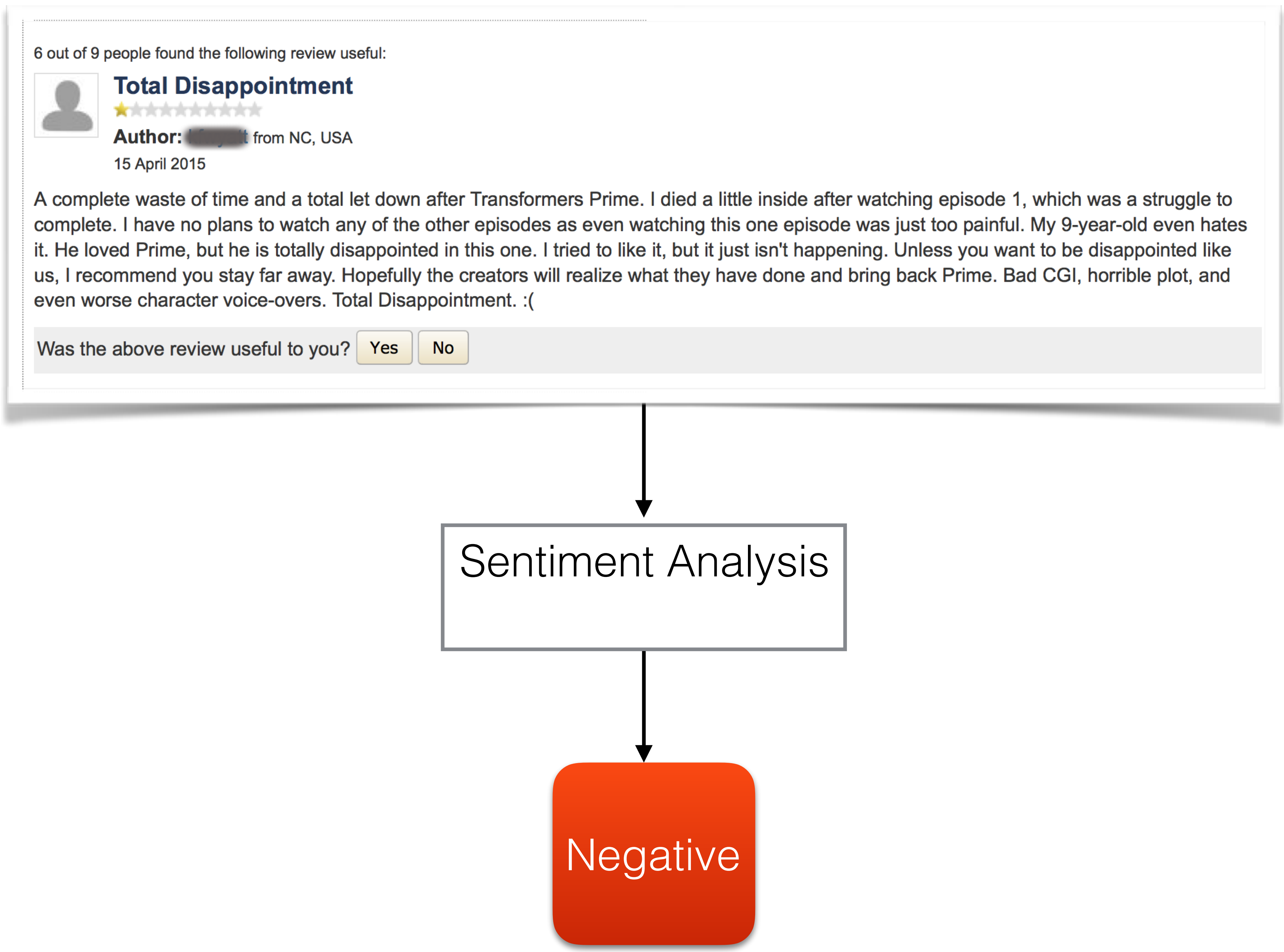
A Paperwhite is, in my opinion, the ultimate way to read. The front light is great and has not given me any eye fatigue, which I'm prone to. If you are a heavy reader and are looking for an e-device, you will be doing your eyes a big favor by getting this over a Fire or other color tablet.

Comment | Was this review helpful to you? Yes No Report abuse



Sentiment Analysis

An IMDB Movie Review



6 out of 9 people found the following review useful:

Total Disappointment

Author: [redacted] from NC, USA

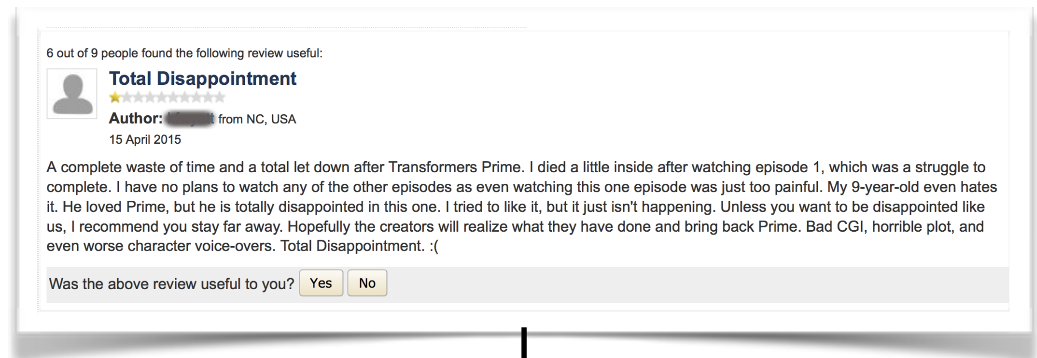
15 April 2015

A complete waste of time and a total let down after Transformers Prime. I died a little inside after watching episode 1, which was a struggle to complete. I have no plans to watch any of the other episodes as even watching this one episode was just too painful. My 9-year-old even hates it. He loved Prime, but he is totally disappointed in this one. I tried to like it, but it just isn't happening. Unless you want to be disappointed like us, I recommend you stay far away. Hopefully the creators will realize what they have done and bring back Prime. Bad CGI, horrible plot, and even worse character voice-overs. Total Disappointment. :(

Was the above review useful to you? Yes No

Sentiment Analysis with Binary Text Classification Pipeline

Review Document → Feature Vector → Classifier → Prediction


6 out of 9 people found the following review useful:

Total Disappointment

Author: [redacted] from NC, USA

15 April 2015

A complete waste of time and a total let down after Transformers Prime. I died a little inside after watching episode 1, which was a struggle to complete. I have no plans to watch any of the other episodes as even watching this one episode was just too painful. My 9-year-old even hates it. He loved Prime, but he is totally disappointed in this one. I tried to like it, but it just isn't happening. Unless you want to be disappointed like us, I recommend you stay far away. Hopefully the creators will realize what they have done and bring back Prime. Bad CGI, horrible plot, and even worse character voice-overs. Total Disappointment. :(

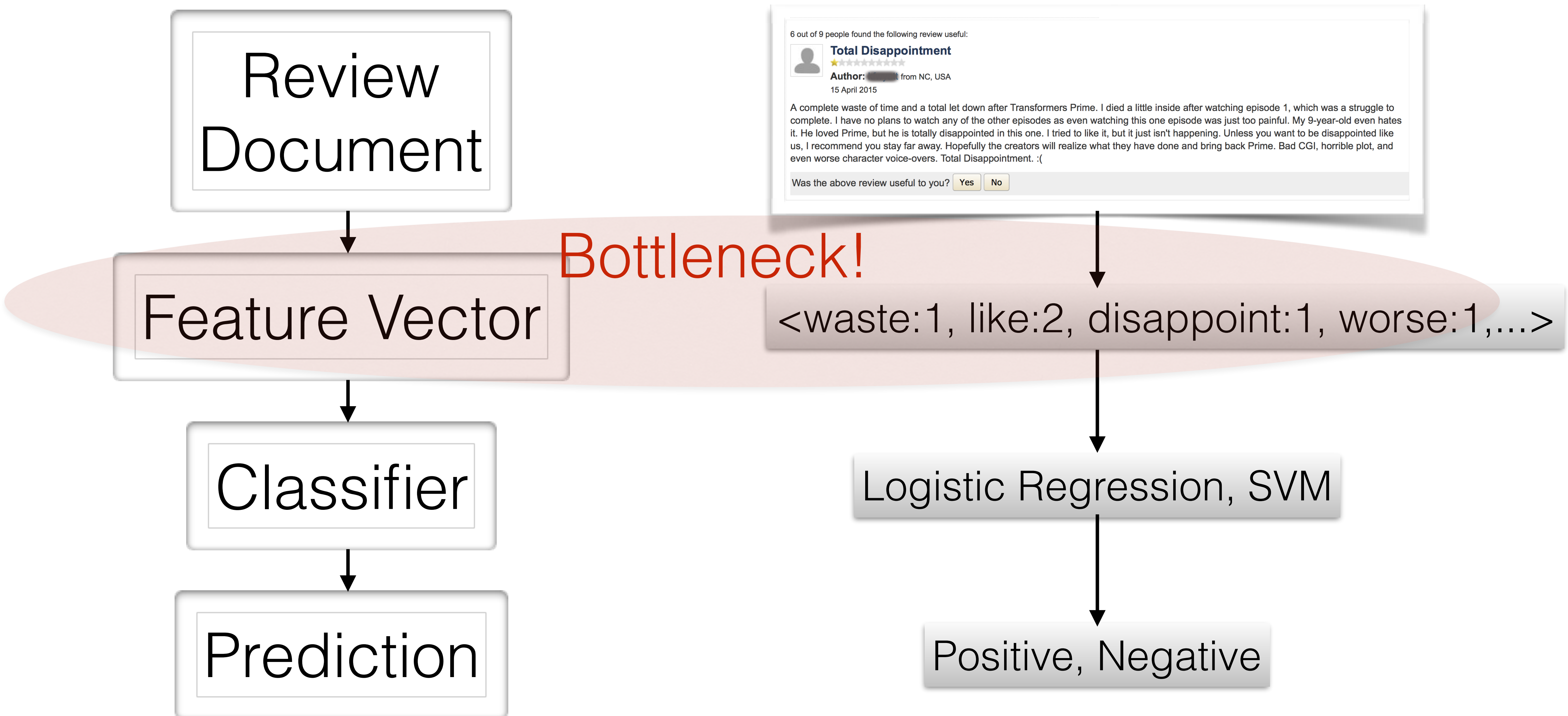
Was the above review useful to you? Yes No

<waste:1, like:2, disappoint:1, worse:1,...>

Logistic Regression, SVM

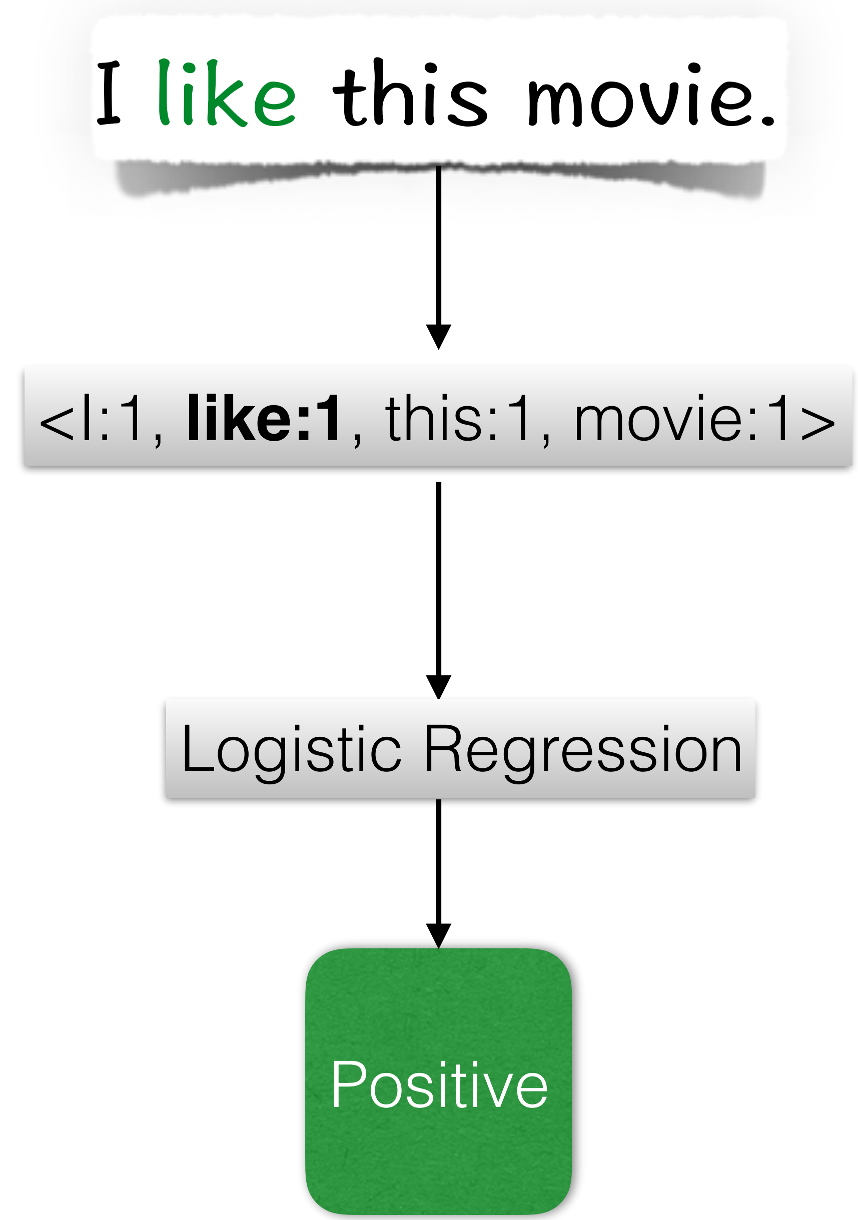
Positive, Negative

Sentiment Analysis with Binary Text Classification Pipeline



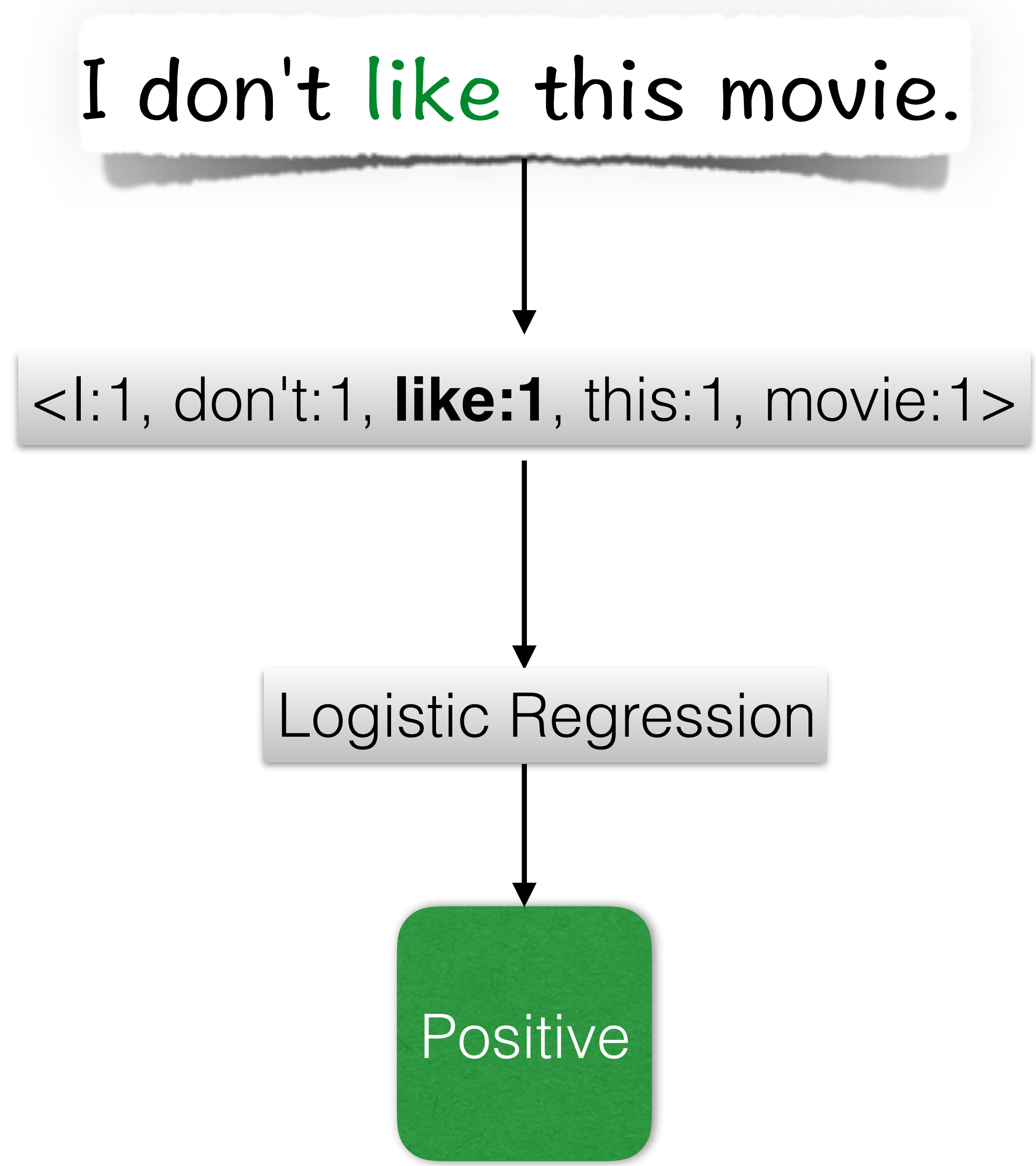
Text Representation Issues in Sentiment Analysis

- Unigram (bag of words)
  capture sentiment indicator terms



Text Representation Issues in Sentiment Analysis

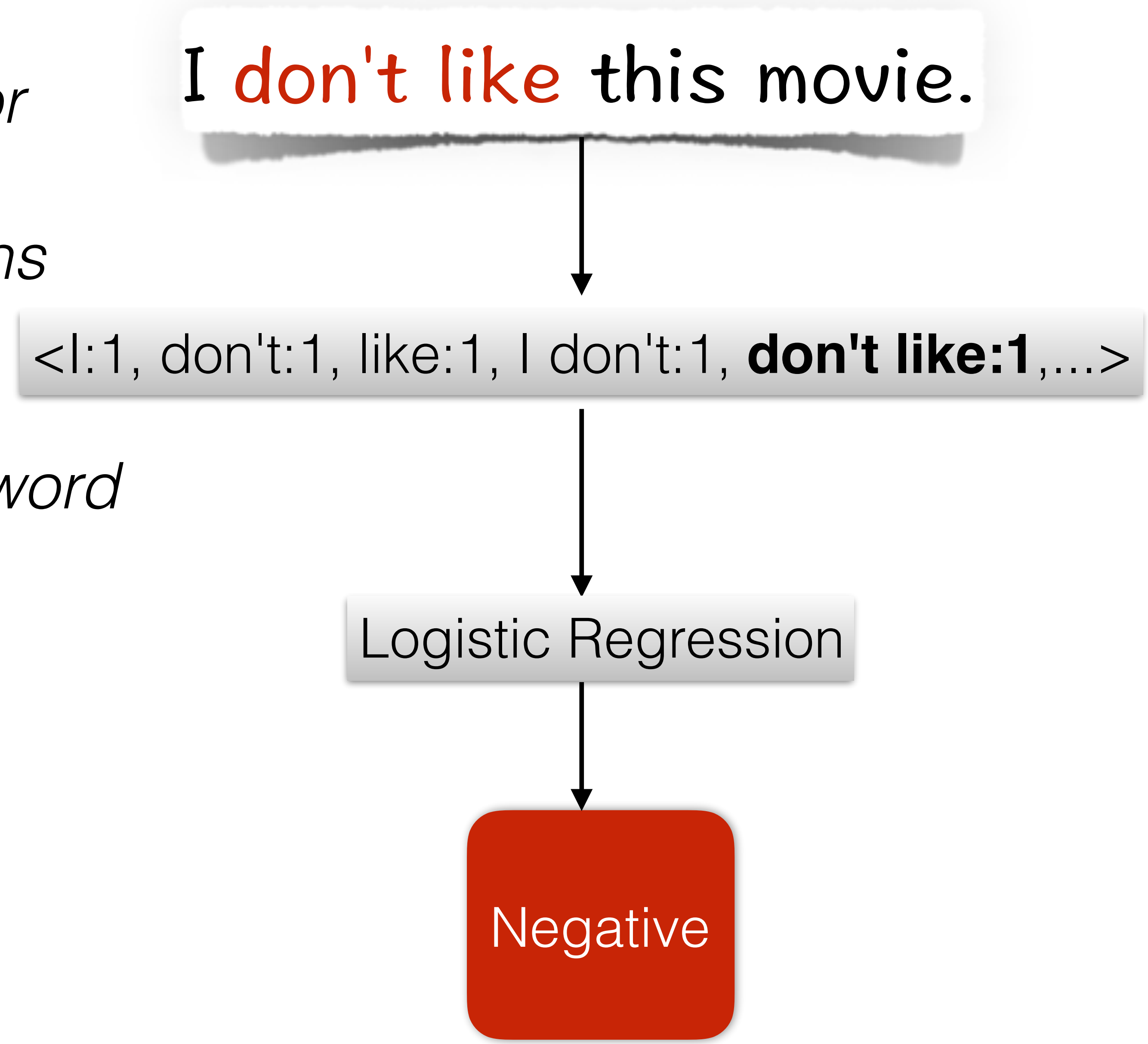
- Unigram (bag of words)
  capture sentiment indicator terms
  could not capture negations



Text Representation Issues in Sentiment Analysis

- Unigram (bag of words)
  capture sentiment indicator terms
  could not capture negations

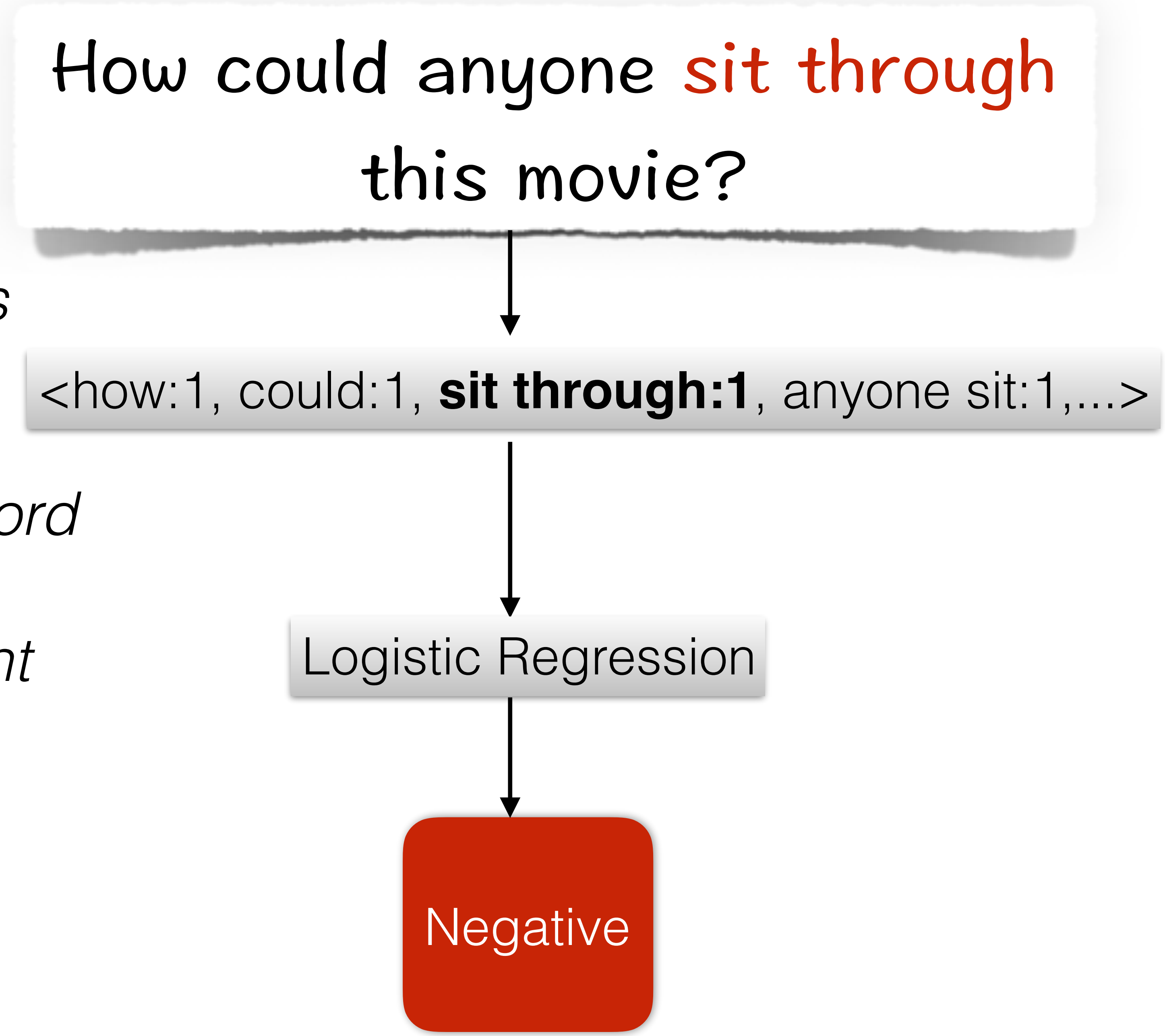
- Add Bi-grams
  capture negation-polarity word pairs



Text Representation Issues in Sentiment Analysis

- Unigram (bag of words)
  capture sentiment indicator terms
  could not capture negations

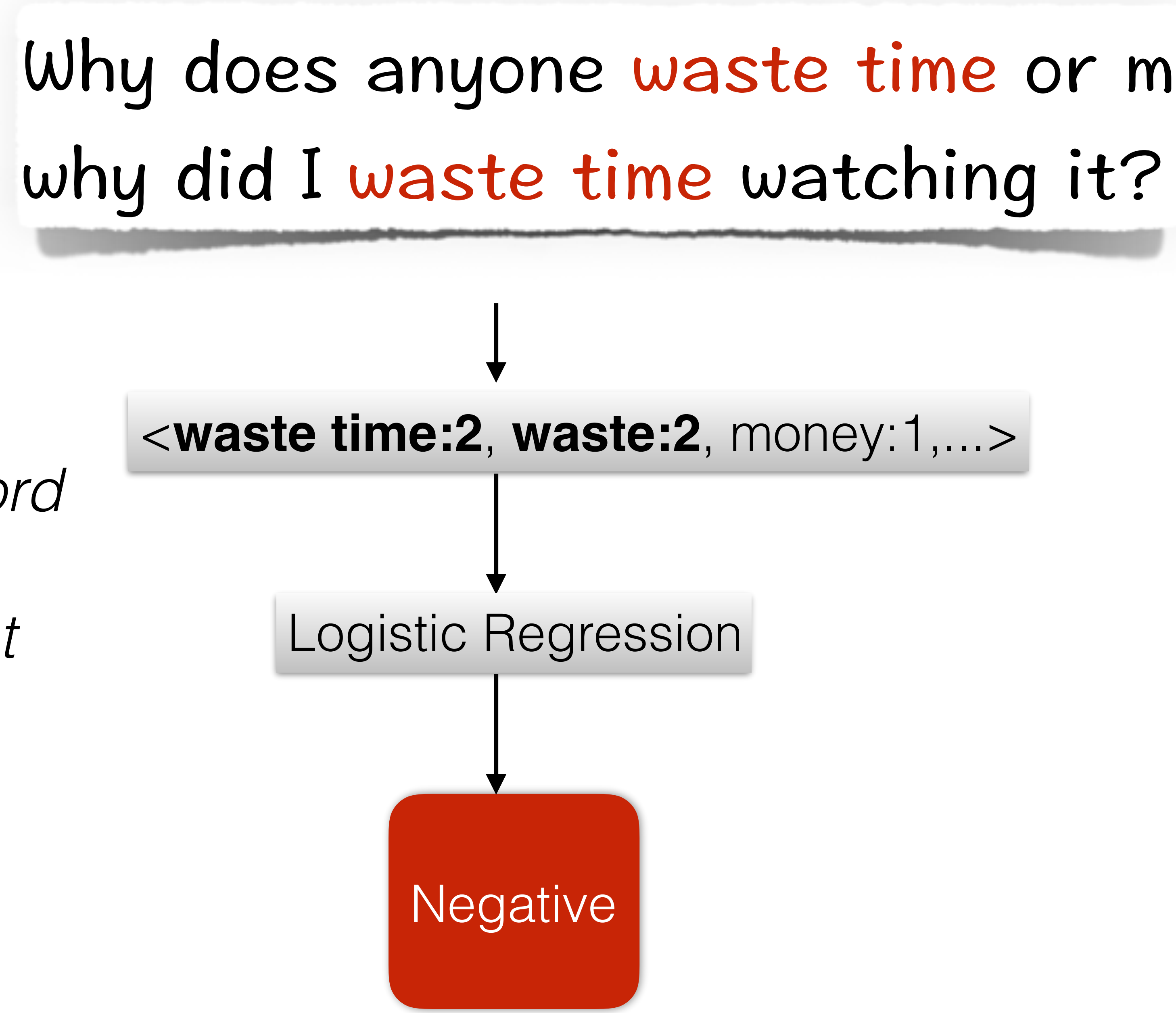
- Add Bi-grams
  capture negation-polarity word pairs
  capture two-words sentiment phrases



Text Representation Issues in Sentiment Analysis

- Unigram (bag of words)
  capture sentiment indicator terms
  could not capture negations

- Add Bi-grams
  capture negation-polarity word pairs
  capture two-words sentiment phrases



Text Representation Issues in Sentiment Analysis

- Unigram (bag of words)
  capture sentiment indicator terms
  could not capture negations

- Add Bi-grams
  capture negation-polarity word pairs
  capture two-words sentiment phrases

- Add tri-grams,quad-grams...
  capture sentiment phrases with many words

Don't waste your time on this movie.

So annoying and such a waste of my time.

A complete waste of time.

I wasted a lot of time on it.

I wasted too much time on it.

Difficulty with High Order n-grams

- Many variations

  "waste your time"

  "waste of my time"

  "waste of time"

  "wasted a lot of time"

  "wasted too much time"

  → increase the dimensionality

- rare cases

  "waste of time": 676 times in IMDB

  "waste more time": 6 times

  "waste your time": 4 times

  → insufficient data for parameter estimation

Skip-grams

- n-gram templates matched loosely
- Looseness parameterized by ***slop***, the number of additional words
- n-gram = skip-gram with ***slop*** 0

Skip-gram Examples

skipgram and count		matched ngrams and count			
skip movie (slop 2)	42	skip this movie	28	skip this pointless movie	1
		skip the movie	8	skipping all the movies (of this sort)	1
		skip watching this movie	1		
it fail (slop 1)	358	it fails	279	it completely fails	5
		it even fails	5	it simply fails	3
whole thing (slop 1)	729	whole thing	682	whole horrific thing	1
		whole damn thing	5		
waste time (slop 1)	1562	waste time	109	waste of time	676
		waste your time	4	waste more time	6
only problem (slop 1)	1481	only problem	1378	only tiny problem	4
		only minor problem	11		
never leak (slop 2)	1053	never leak	545	never a urine leak (problem)	1
		never have leak	86	never have any leak	77
no smell (slop 1)	445	no smell	340	no medicine-like smell	1
		no bad smell	13	no annoying smell	5
it easy to clean and (slop 2)	314	it is easy to wipe clean and	3	it is easy to keep clean and	3
		it is so easy to clean and	16		
I have to return (slop 2)	216	I have to return	151	I finally have to return	1
		I have never had to return	1	I do not have to return	4
good service (slop 2)	209	good service	131	good price and service	1
		good and fast service	2		

Advantages of Skip-grams

- Group infrequent n-grams into a frequent skip-gram
- Allow n-grams to borrow strength from each other
- Easier learning
- Better generalization

Difficulties with Skip-grams

- Huge number
- Many are non-informative or noisy

  skip-gram "I recommend" with *slop* 2 can match both "I highly recommend" and "I do not recommend"

Existing Use of Skip-grams in Sentiment Analysis

- Ask human assessors to pick informative skip-grams
  - ✗ limited by available domain knowledge
  - ✗ expensive
- Build dense word vectors on top of skip-grams
  - ✗ information loss
  - ✗ less interpretable

Goal of this Study

- Test whether skip-grams are helpful when used directly as features in sentiment analysis
- Test different automatic regularization/feature selection strategies
- Compare against n-grams and word vectors

Skip-gram Extraction

- Consider skip-grams with *n*<=5 and *slop*<=2 (5-grams with 2 additional words in between)

- Discard skip-grams with very low frequencies (<=5)

max *n*	max *slop*	# skip-grams on IMDB
1	0	2x10^4
2	0	1x10^5
3	0	2x10^5
5	0	4x10^5
2	1	3x10^5
3	1	9x10^5
5	1	1x10^6
2	2	6x10^5
3	2	2x10^6
5	2	3x10^6

L1 vs L2 regularization

- Skip-gram features: huge number, correlated
- L1: $\min_w \text{ loss} + \lambda ||w||_1$
  - ✓ shrink weights
  - ✓ select a subset of features → compact model
  - ✗ select one out of several correlated features
- L2: $\min_w \text{ loss} + \lambda ||w||_2^2$
  - ✓ shrink weights
  - ✗ use all features → robust model
  - ✓ spread weight among correlated features

L1+L2 regularization

- L1+L2: $\min_w \text{ loss} + \lambda\alpha||w||_1 + \lambda(1-\alpha)||w||_2^2$
  - ✓ shrink weights
  - ✓ select a subset of features → compact model
  - ✓ spread weight among correlated features → robust model

Learning and Regularization

- L2-regularized linear SVM

$$\min_w \sum_{i=1}^{N} (\max(0, 1 - y_i w^T x_i))^2 + \lambda \frac{1}{2} ||w||_2^2$$

- L1-regularized linear SVM

$$\min_w \sum_{i=1}^{N} (\max(0, 1 - y_i w^T x_i))^2 + \lambda ||w||_1$$

- L2-regularized Logistic Regression

$$\min_w -\frac{1}{N} \sum_{i=1}^{N} y_i w^T x_i + \log(1 + e^{w^T x_i}) + \lambda \frac{1}{2} ||w||_2^2$$

- L1-regularized Logistic Regression

$$\min_w -\frac{1}{N} \sum_{i=1}^{N} y_i w^T x_i + \log(1 + e^{w^T x_i}) + \lambda ||w||_1$$

- L1+L2-regularized Logistic Regression

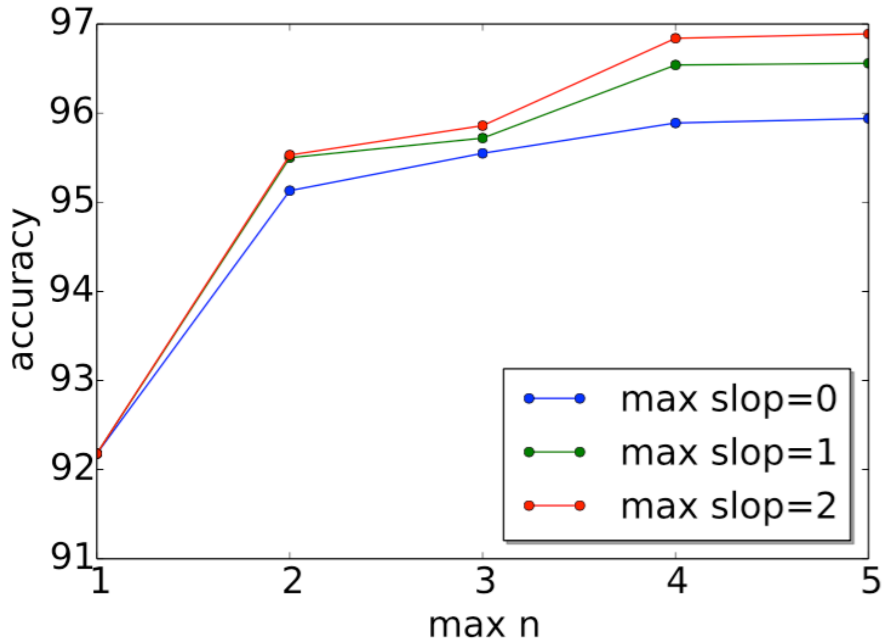
$$\min_w -\frac{1}{N} \sum_{i=1}^{N} y_i w^T x_i + \log(1 + e^{w^T x_i}) + \lambda \alpha ||w||_1 + \lambda (1 - \alpha) \frac{1}{2} ||w||_2^2$$

Datasets

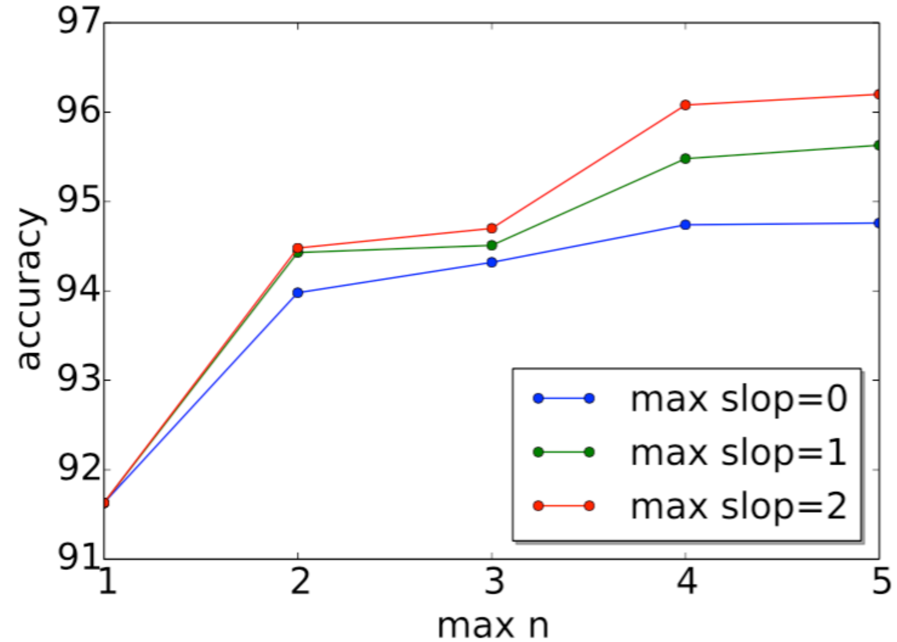
Binary classification with neutral reviews ignored

dataset	positive	negative
IMDB	25,000 reviews with ratings 7-10	25,000 reviews with ratings 1-4
Amazon Baby Product	136,461 reviews with ratings 4-5	32,950 reviews with ratings 1-2
Amazon Phone Product	47,970 reviews with ratings 4-5	22,241 reviews with ratings 1-2

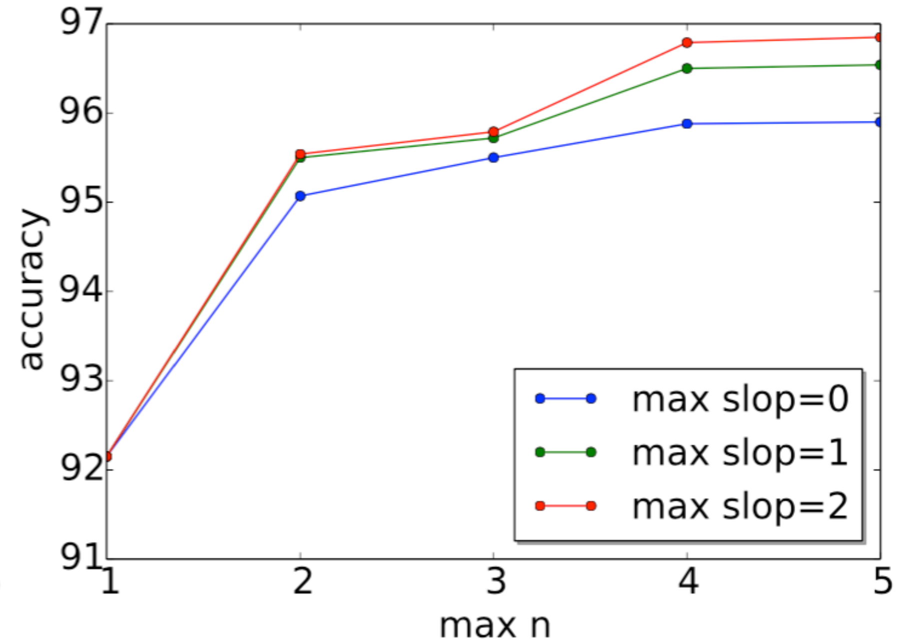
Classification Accuracy with Skip-gram Features



L2 LR



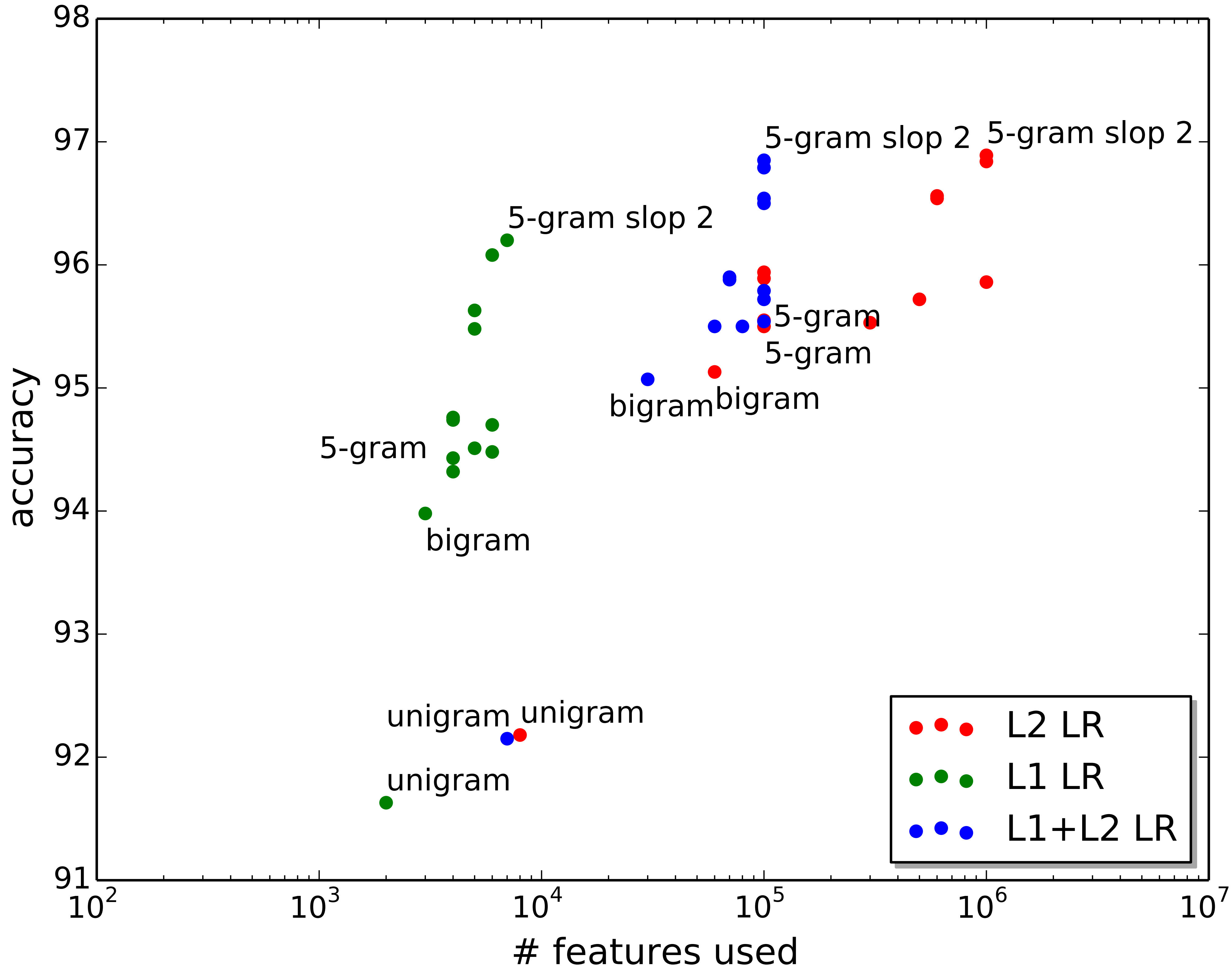
L1 LR



L1+L2 LR

- Blue line: moving from unigrams to bigrams gives substantial improvement
- Blue line: using high-order n-grams gives marginal improvement
- Green and red lines: increasing *slop* from 0 to 1 and 2 gives further improvement
- max # features selected: L2: 10^6, L1: 10^4, L1+L2: 10^5

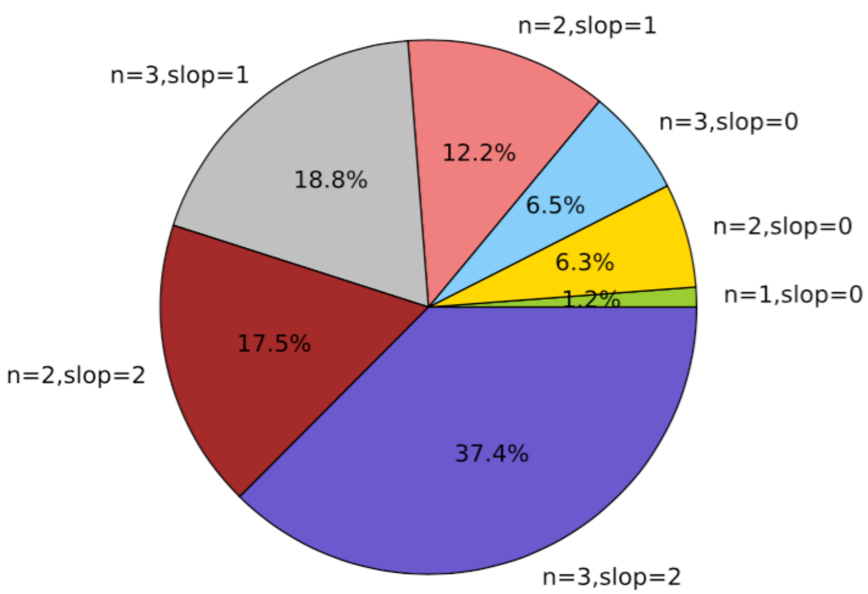
Features Used vs Accuracy



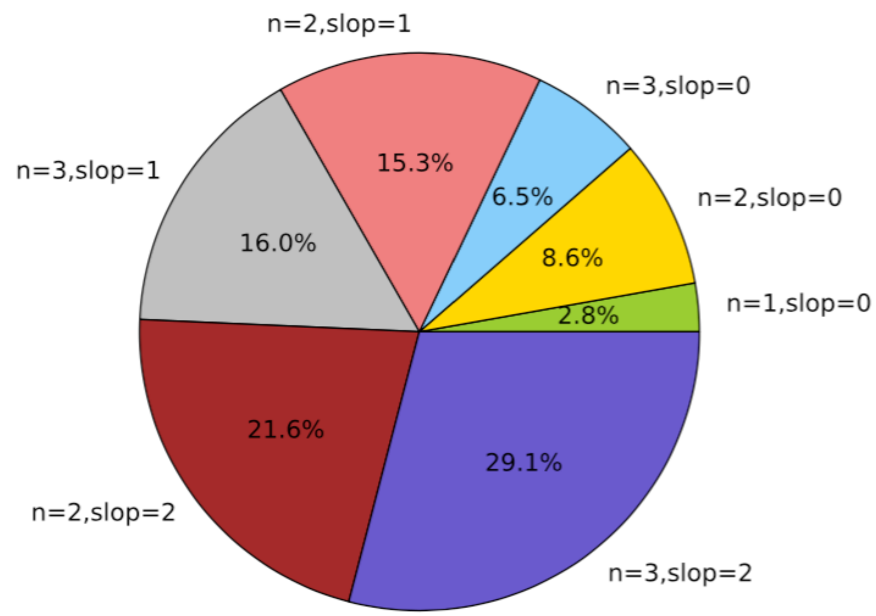
Observations on L1 vs L2

- L2: achieves better overall accuracy
  - Large training sets facilitate parameter estimation
  - Effective handling of correlated features
- L1: produces much smaller models
- L1+L2: good compromise

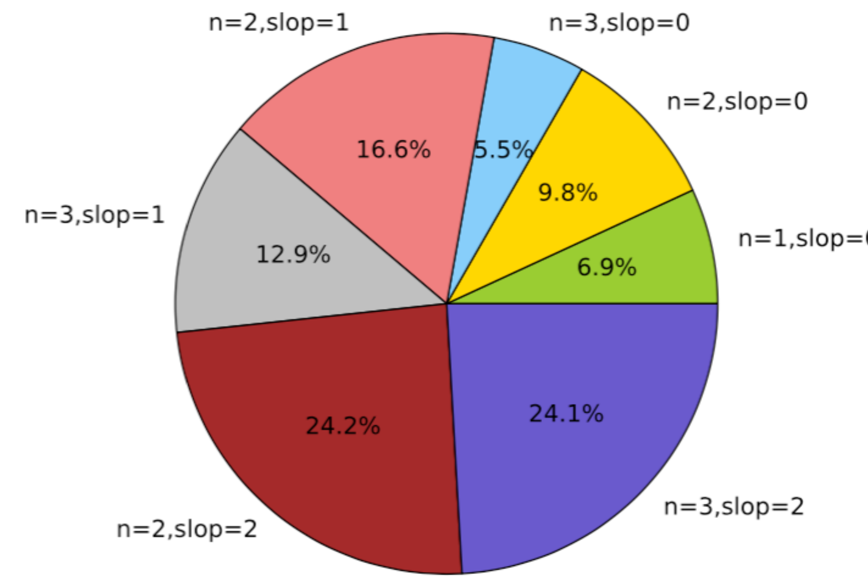
Skip-gram Feature Contribution



all features



selected features



weighted features

- Comparing left with middle: the fraction of unigrams increases;the fraction of *slop* 2 trigrams decreases. Many *slop* 2 trigrams are eliminated by L1.
- In right: The standard n-grams with *slop*=0 only contribute to 20% of the total weight, and the remaining 80% is due to skip-grams with non-zero *slops*.

Comparison with Word Vectors

	skip-gram	word vector
AMAZON BABY	96.85	88.84
AMAZON PHONE	92.58	85.38
IMDB	91.26	92.58 / 85.0

- Word vectors work extremely well on the given test set (92.58%), but poorly on random test sets (85%).

Other Results on IMDB

classifier	features	training documents	accuracy
LR with dropout regularization [21]	bigrams	25,000 labeled	91.31
NBSVM [23]	bigrams	25,000 labeled	91.22
SVM with L2 regularization	structural parse tree features + unigrams [16]	25,000 labeled	82.8
LR L1+L2 regularization	5-grams selected by compressive feature learning [20]	25,000 labeled	90.4
SVM	word vectors trained by WRRBM [6]	25,000 labeled	89.23
SVM	word vectors [15]	25,000 labeled + 50,000 unlabeled	88.89
LR with dropout regularization [21]	bigrams	25,000 labeled + 50,000 unlabeled	91.98
LR	paragraph vectors [14]	25,000 labeled + 50,000 unlabeled	92.58
LR with L2 regularization	skip-grams	25,000 labeled	91.63
SVM with L2 regularization	skip-grams	25,000 labeled	91.71
LR with L1+L2 regularization	skip-grams	25,000 labeled	91.26

- Among the methods which only use labeled data, skip-grams achieved the highest accuracy

Conclusion

- Skip-grams group similar n-grams together, facilitating learning and generalization

- Using skip-grams achieves good sentiment analysis performance

- L1+L2 regularization reduces the number of features significantly while maintaining good accuracy

- Our code will be released soon at: https://github.com/cheng-li/pyramid

Thank You