RECURSIVE DEEP MODELS FOR SEMANTIC COMPOSITIONALITY [1]

Zhicong Lu

DGP Lab

luzhc@dgp.toronto.edu

[1]Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts. **Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.** *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*

OVERVIEW

- Background
- Stanford Sentiment Treebank
- Recursive Neural Models
- Experiments

SENTIMENT ANALYSIS

- Identify and extract subjective information
- Crucial to business intelligence, stock trading, ...

[1]Adapted from: http://www.rottentomatoes.com/

Zootopia excels on so many levels that it stands with the finest of the Disney classics.
Full Review... | March 10, 2016
Lou Lumenick
New York Post
★ Top Critic

There's a lot here for kids to like and nearly as much to keep parents from fidgeting.
Full Review... | March 6, 2016
James Berardinelli
ReelViews
★ Top Critic

What saves this big-budget cartoon behemoth is its modest, old-fashioned storytelling.
Full Review... | March 4, 2016
David Edelstein
New York Magazine/Vulture
★ Top Critic

"Zootopia," like its heroine, is zesty, bright, and breakneck, with chase scenes and well-tuned gags where you half expect songs to be.
Full Review... | March 7, 2016
Anthony Lane
New Yorker
★ Top Critic

In many ways a conventional movie, but it unfolds with so much wit, panache, and visual ingenuity that it outstrips many a more high-concept film.
Full Review... | March 6, 2016
Christopher Orr
The Atlantic
★ Top Critic

Gorgeous to look at, clever, funny and with a solid and atmospheric mystery at its core. But there's more here in the film's timely and relevant thematic content.
Full Review... | March 13, 2016

RELATED WORK

- Semantic Vector Spaces
  - Distributional similarity of single words (e.g., tf-idf)
    - Do not capture the differences in antonyms
- Neural word vectors (Bengio et al.,2003)
  - Unsupervised
  - Capture distributional similarity
  - Need fine-tuning for sentiment detection

RELATED WORK

- Compositionally in Vector Spaces
  - Capture two word compositions
  - Have not been validated on larger corpora
- Logical Form
  - Mapping sentences to logic form
  - Could only capture sentiment distributions using separate mechanisms beyond the currently used logic forms

RELATED WORK

- Deep Learning
  - Recursive Auto-associative memories
  - Restricted Boltzmann machines etc.

SENTIMENT ANALYSIS AND BAG-OF-WORD MODELS[1]

- Most methods use bag of words + linguistic features/ processing/lexica
- Problem: such methods can't distinguish different sentiment caused by word order:
- + white blood cells destroying an infection
- - an infection destroying white blood cells

[1]Adapted from Richard Socher's slides: https://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

SENTIMENT DETECTION AND BAG-OF-WORD MODELS[1]

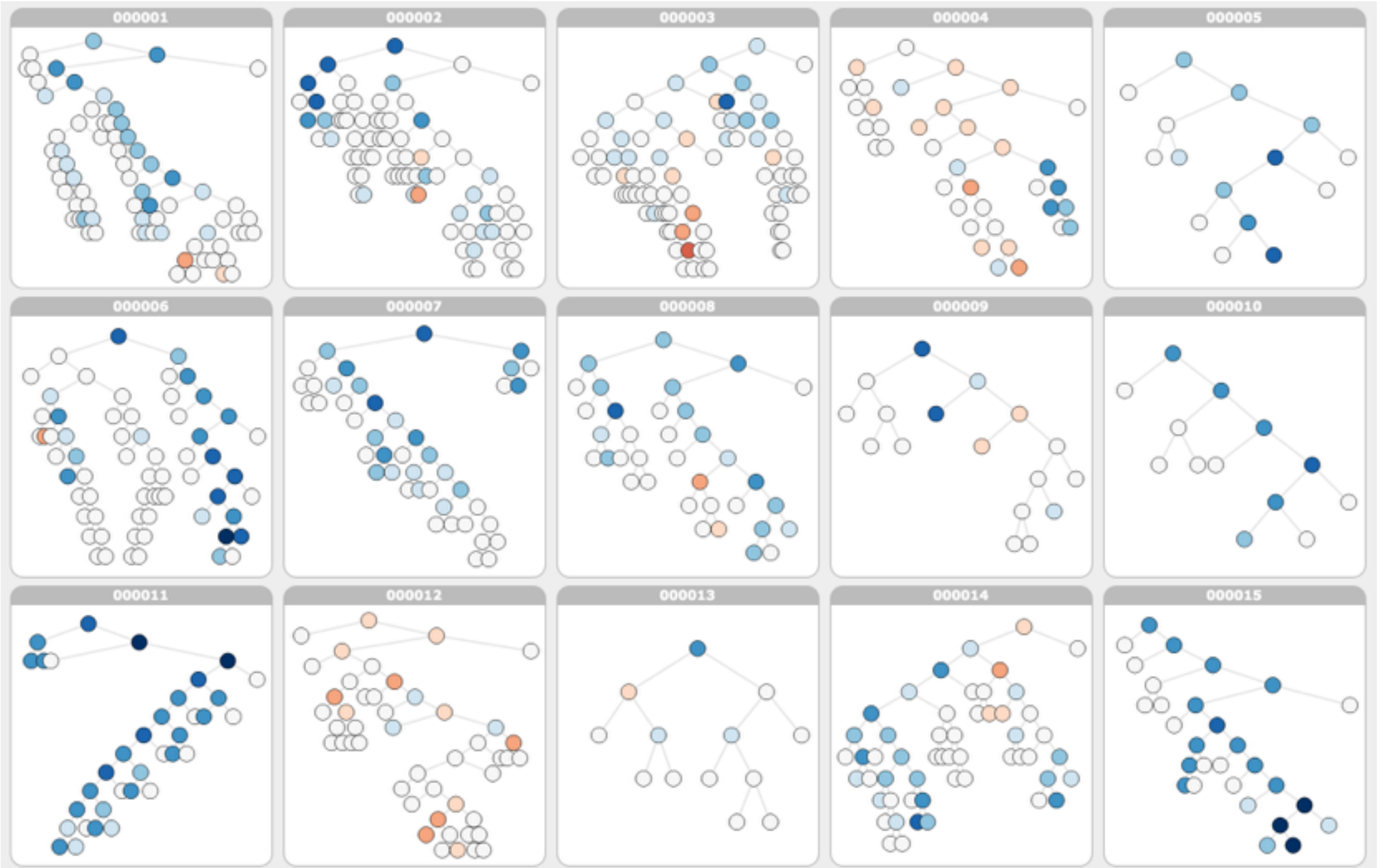
- Sentiment detection seems easy for some cases
- Detection Accuracy for *longer documents* reaches 90%
- Many easy cases, such as horrible or awesome
- For dataset of single sentence movie reviews (Pang and Lee, 2005), accuracy never reached >80% for >**7 years**
- Hard cases require actual understanding of **negation and its scope** + other semantic effects

[1]Adapted from Richard Socher's slides: https://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

TWO MISSING PIECES FOR IMPROVING SENTIMENT DETECTION

- Large and labeled compositional **data**
  - Sentiment Treebank
- Better **models** for semantic compositionality
  - Recursive Neural Networks

STANFORD SENTIMENT TREEBANK



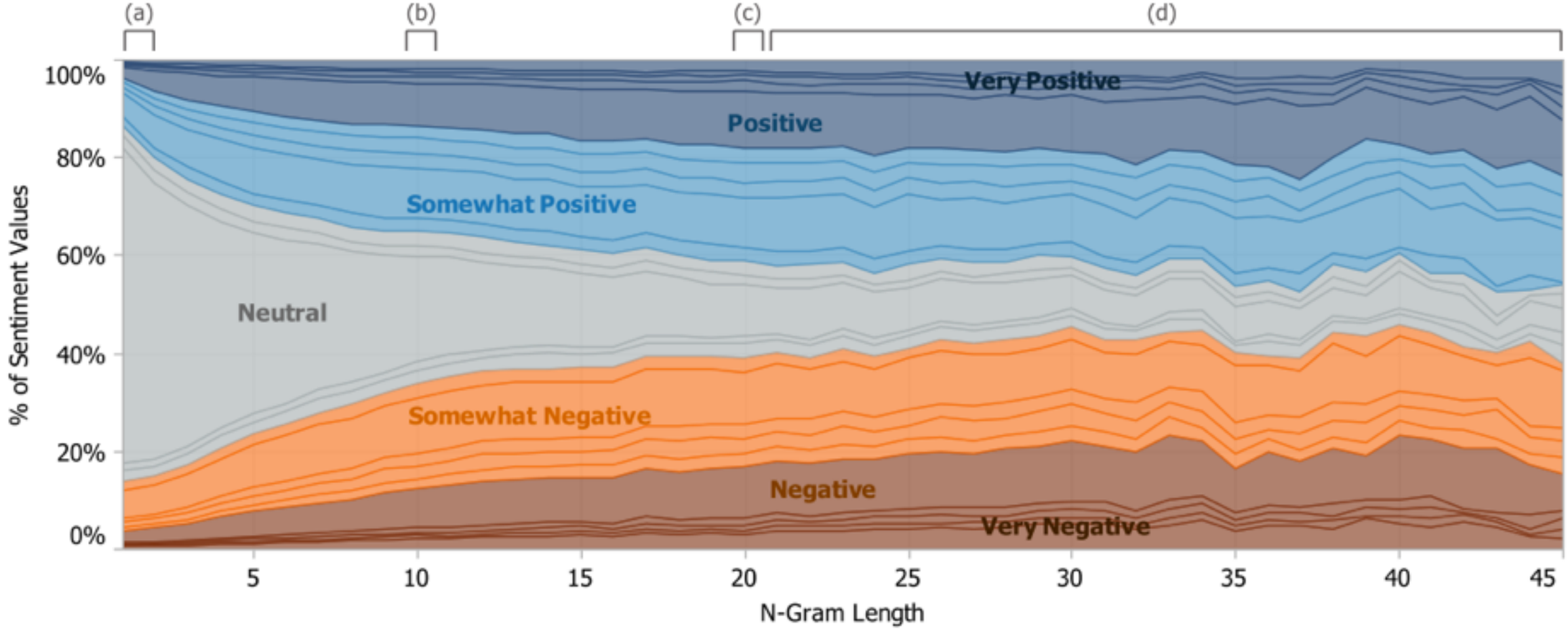
[1]Adapted from http://nlp.stanford.edu/sentiment/treebank.html

DATASET

- 215,154 phrases with labels by Amazon Mechanical Turk
- Parse trees of 11,855 sentences from movie reviews
- Allows for a complete analysis of the compositional effects of sentiment in language.

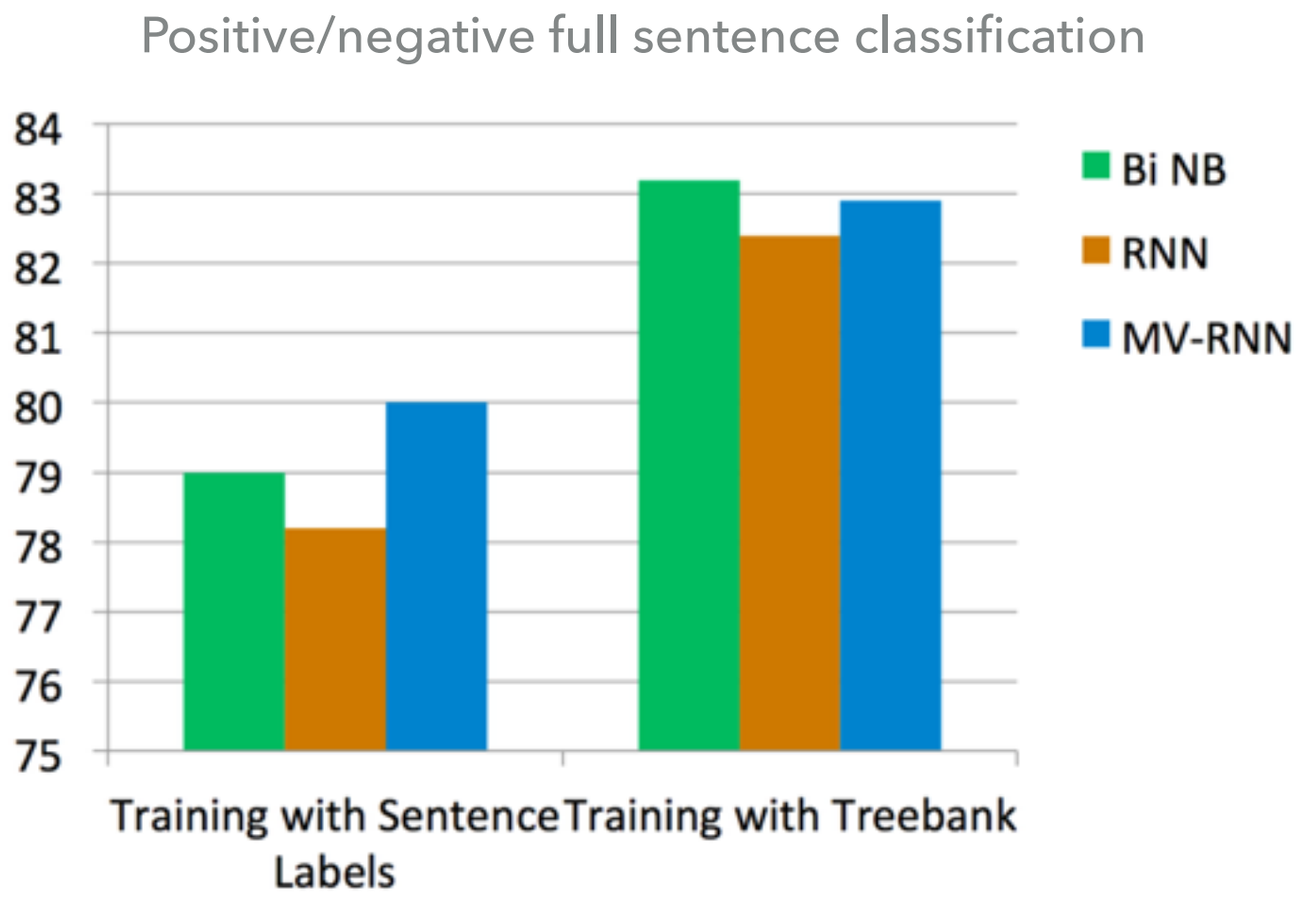
FINDINGS

- Stronger sentiment often builds up in longer phrases and the majority of the shorter phrases are neutral
- The extreme values were rarely used and the slider was not often left in between the ticks



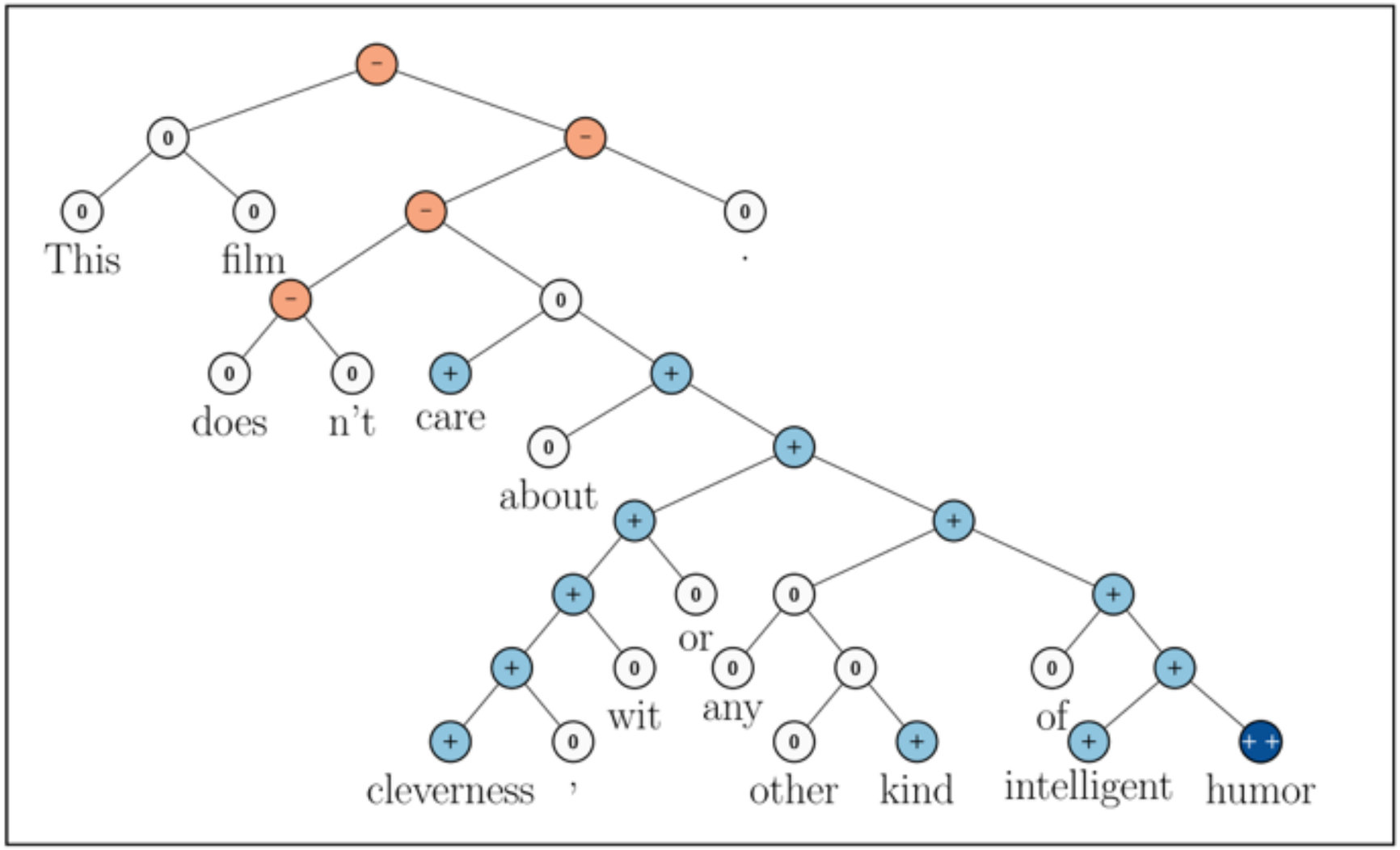
BETTER DATASET HELPED[1]

- Performance improved by 2-3%
- Hard negation cases are still mostly incorrect
- Need a more powerful model



[1]Adapted from Richard Socher's slides: https://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

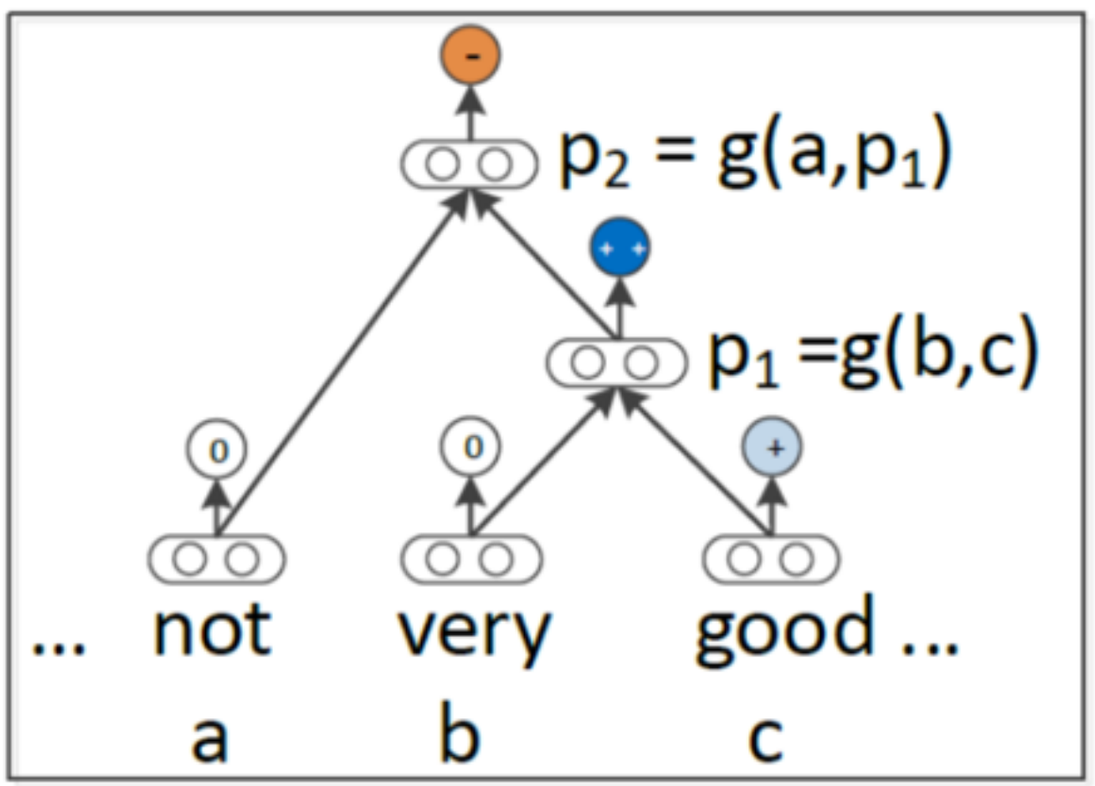
RECURSIVE NEURAL MODELS



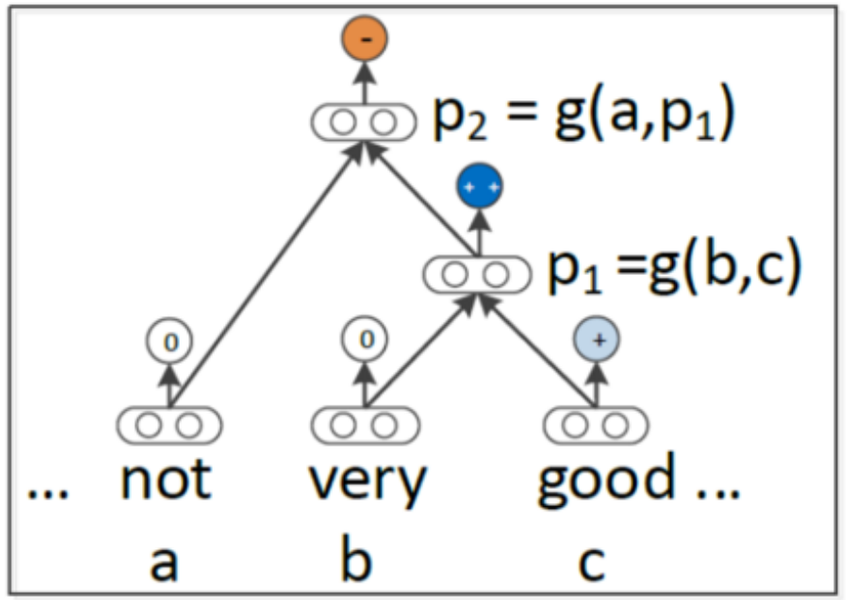
Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

RECURSIVE NEURAL MODELS

- RNN: Recursive Neural Network
- MV-RNN: Matrix-Vector RNN
- RNTN: Recursive Neural Tensor Network



OPERATIONS IN COMMON



- Word vector representations

  Word vectors: d-dimensional, initialized by randomly from a U(-r,r), r = 0.0001

  Word embedding Matrix $L \in \mathbb{R}^{d \times |V|}$, stacked by all the word vectors, trained jointly with compositionality models

- Classification

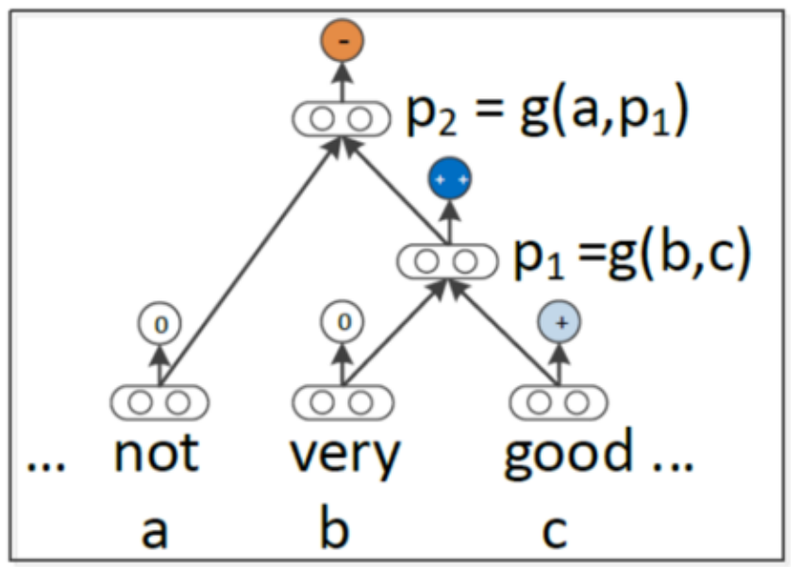
  Posterior probability over labels given the word vector:

  $$y^a = \text{softmax}(W_s a)$$

  $W_s \in \mathbb{R}^{5 \times d}$ – Sentiment classification matrix

RECURSIVE NEURAL MODELS[1]

- Focused on compositional representation learning of
  - Hierarchical structure, features and prediction
- Different combinations of
  - Training Objective
  - Composition Function
  - Tree Structure

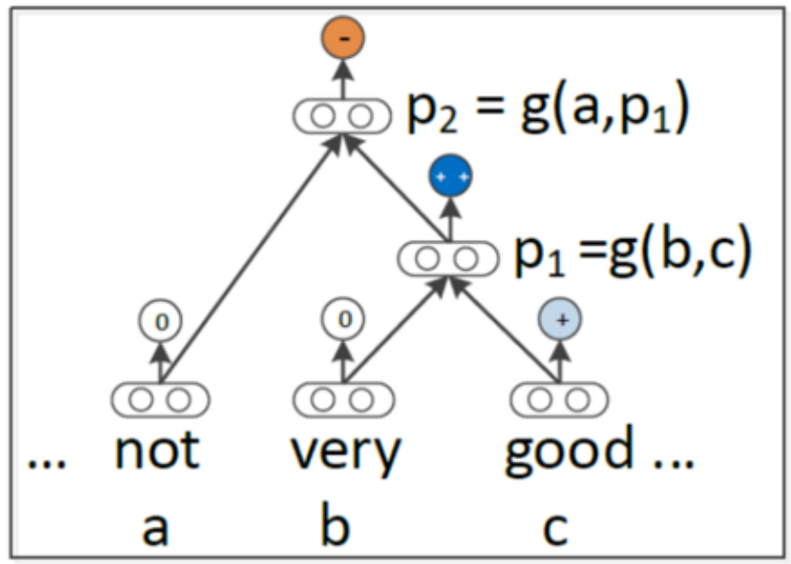


[1]Adapted from Richard Socher's slides: https://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

STANDARD RECURSIVE NEURAL NETWORK

- Compositionality Function:

$$p_1 = f\left(W\begin{bmatrix} b \\ c \end{bmatrix}\right), p_2 = f\left(W\begin{bmatrix} a \\ p_1 \end{bmatrix}\right)$$



$f = \tanh$ – standard element-wise nonlinearity

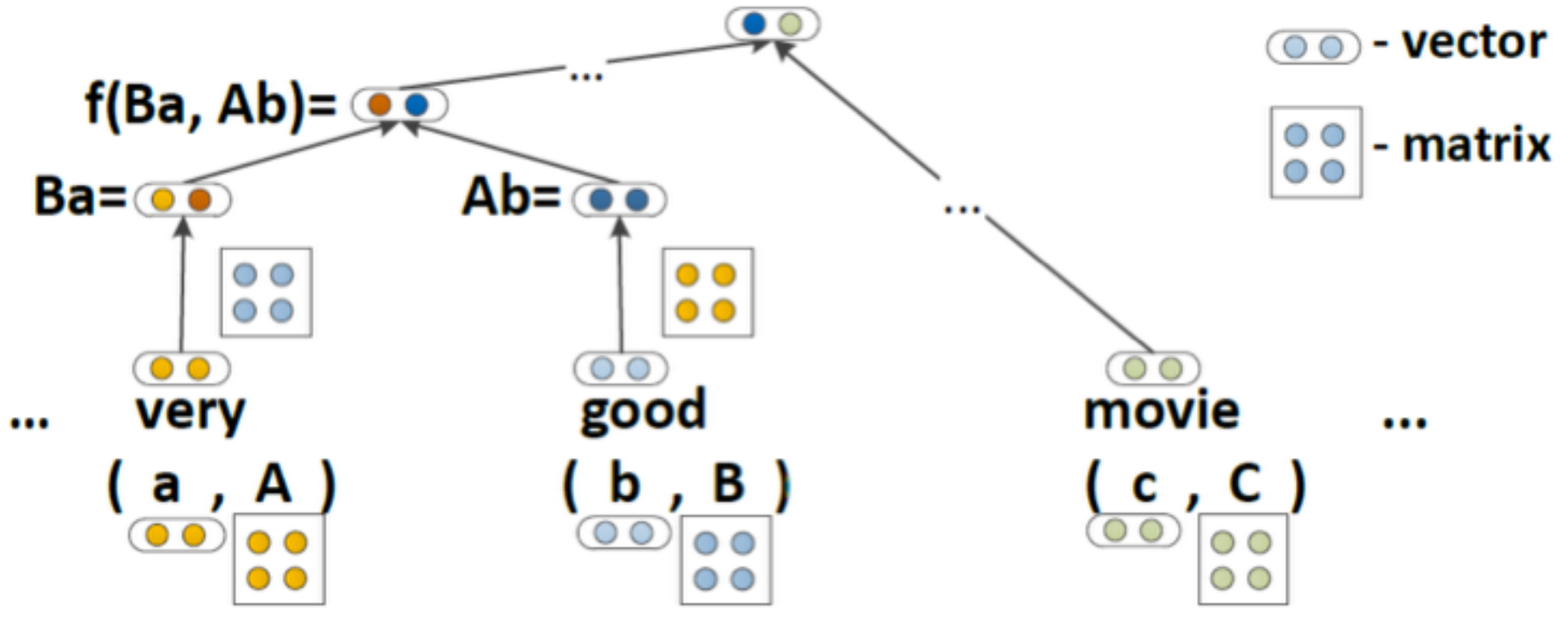
$W \in \mathbb{R}^{d \times 2d}$ – main parameter to learn

MV-RNN: MATRIX-VECTOR RNN

- Composition Function:

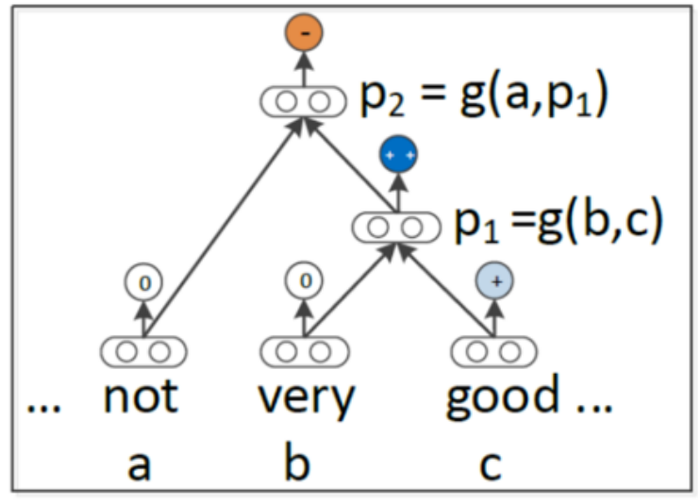
$$p_1 = f\left(W\begin{bmatrix} Ba \\ Ab \end{bmatrix}\right), P_1 = f\left(W_M\begin{bmatrix} A \\ B \end{bmatrix}\right)$$

$$W_M \in \mathbb{R}^{d\times 2d}$$

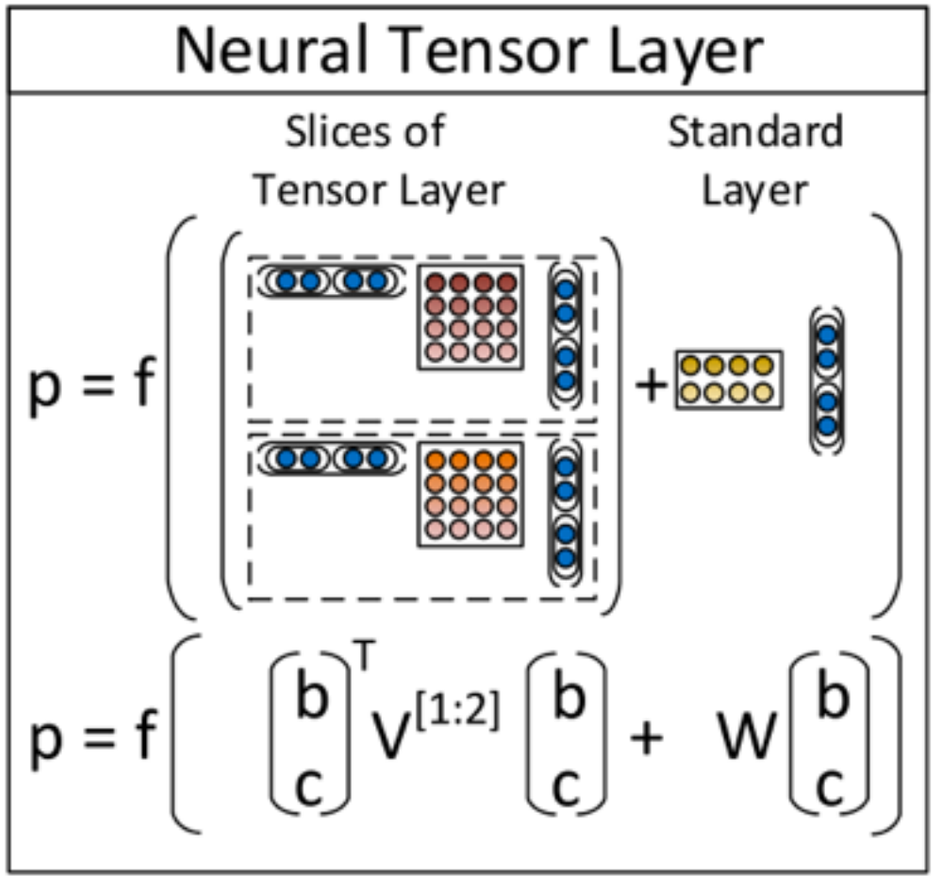


Adapted from Richard Socher's slides: https://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

RECURSIVE NEURAL TENSOR NETWORK



- More expressive than previous RNNs
- Basic idea: Allow more interactions of vectors



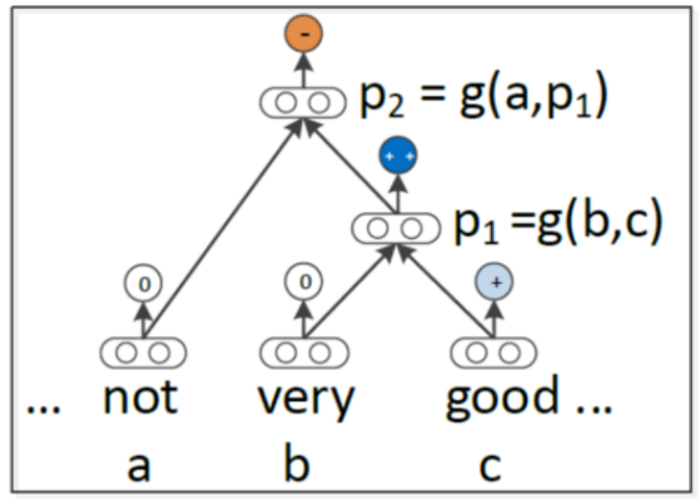
- Composition Function

$$p_1 = f\left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix}\right)$$

$$p_2 = f\left(\begin{bmatrix} a \\ p_1 \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} a \\ p_1 \end{bmatrix} + W \begin{bmatrix} a \\ p_1 \end{bmatrix}\right)$$

- The tensor can directly relate input vectors
- Each slice of the tensor captures a specific type of composition

TENSOR BACKPROP THROUGH STRUCTURE



- Minimizing cross entropy error:

$$E(\theta) = \sum_i \sum_j t_j^i \log y_j^i + \lambda \|\theta\|^2 \qquad \theta = (V, W, W_s, L)$$

- Standard softmax error vector:

$$\delta^{i,s} = \left(W_s^T (y^i - t^i)\right) \otimes f'(x^i),$$

- Update for each slice:

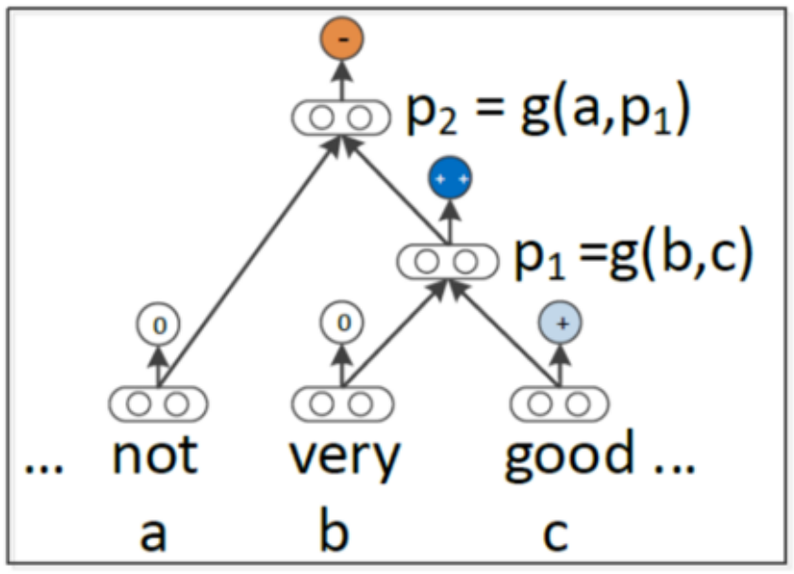
$$\frac{\partial E^{p_2}}{\partial V^{[k]}} = \delta_k^{p_2,com} \begin{bmatrix} a \\ p_1 \end{bmatrix} \begin{bmatrix} a \\ p_1 \end{bmatrix}^T$$

TENSOR BACKPROP THROUGH STRUCTURE

- Main backdrop rule to pass error down from parent:

$$\delta^{p_2,down} = \left(W^T \delta^{p_2,com} + S\right) \otimes f'\left(\begin{bmatrix} a \\ p_1 \end{bmatrix}\right)$$

$$S = \sum_{k=1}^{d} \delta_k^{p_2,com} \left(V^{[k]} + \left(V^{[k]}\right)^T\right) \begin{bmatrix} a \\ p_1 \end{bmatrix}$$



- Add errors from parent and current softmax

$$\delta^{p_1,com} = \delta^{p_1,s} + \delta^{p_2,down}[d+1:2d]$$

- Full derivative for slice $V^{[k]}$

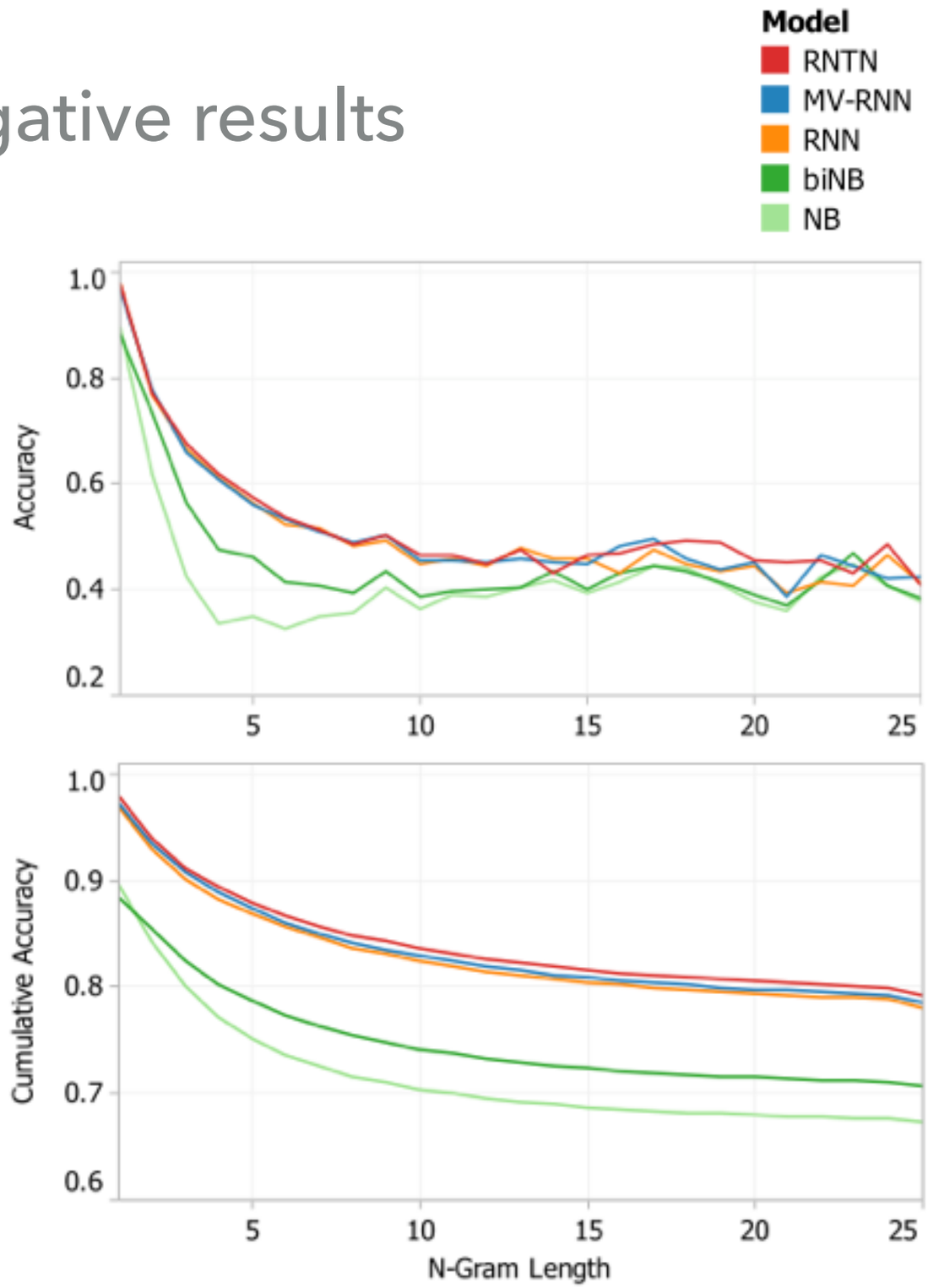
$$\frac{\partial E}{\partial V^{[k]}} = \frac{E^{p_2}}{\partial V^{[k]}} + \delta_k^{p_1,com} \begin{bmatrix} b \\ c \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix}^T$$

RESULTS ON TREEBANK

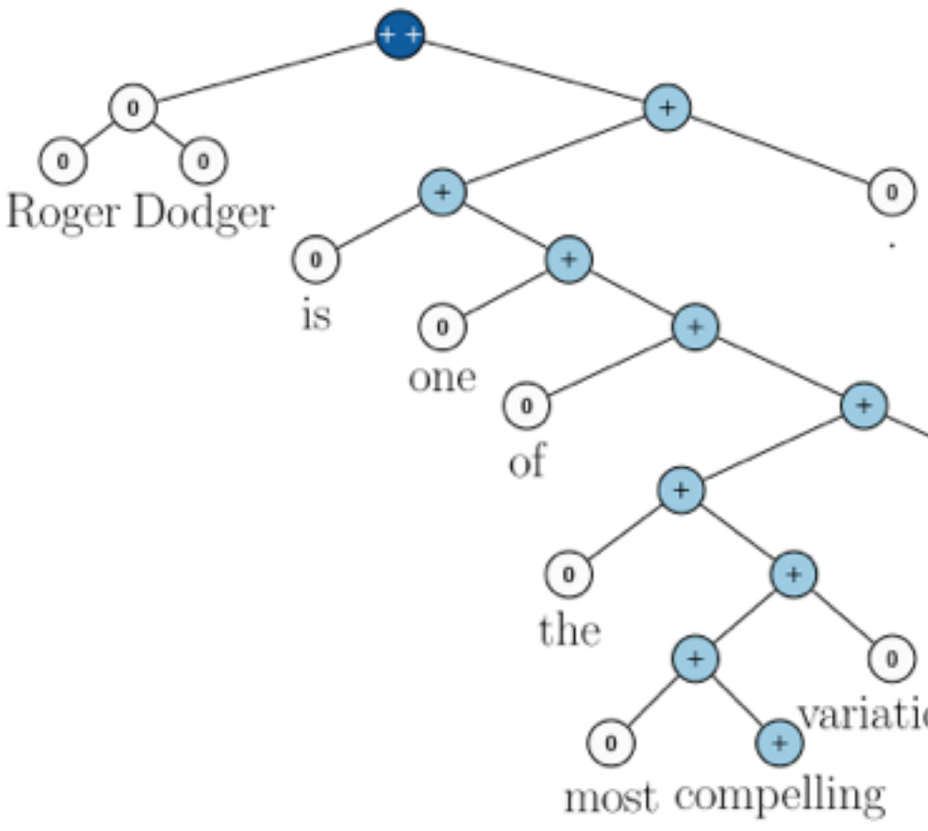
- Fine-grained and Positive/Negative results

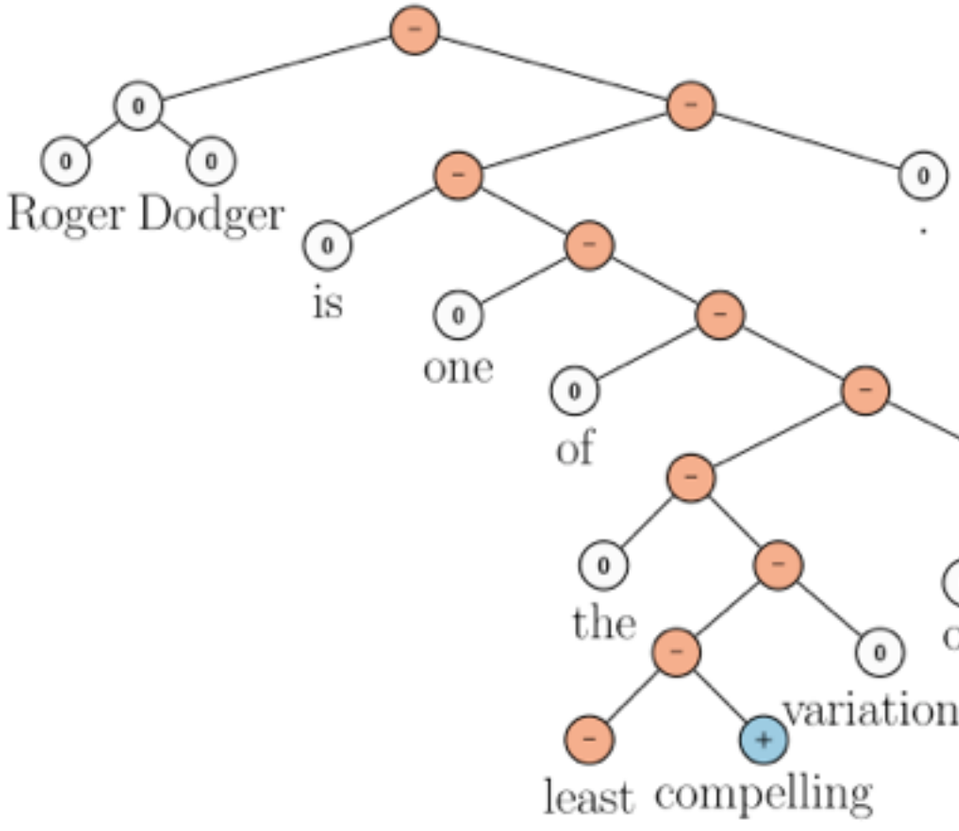
Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.7**	**87.6**	**85.4**

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



NEGATION RESULTS

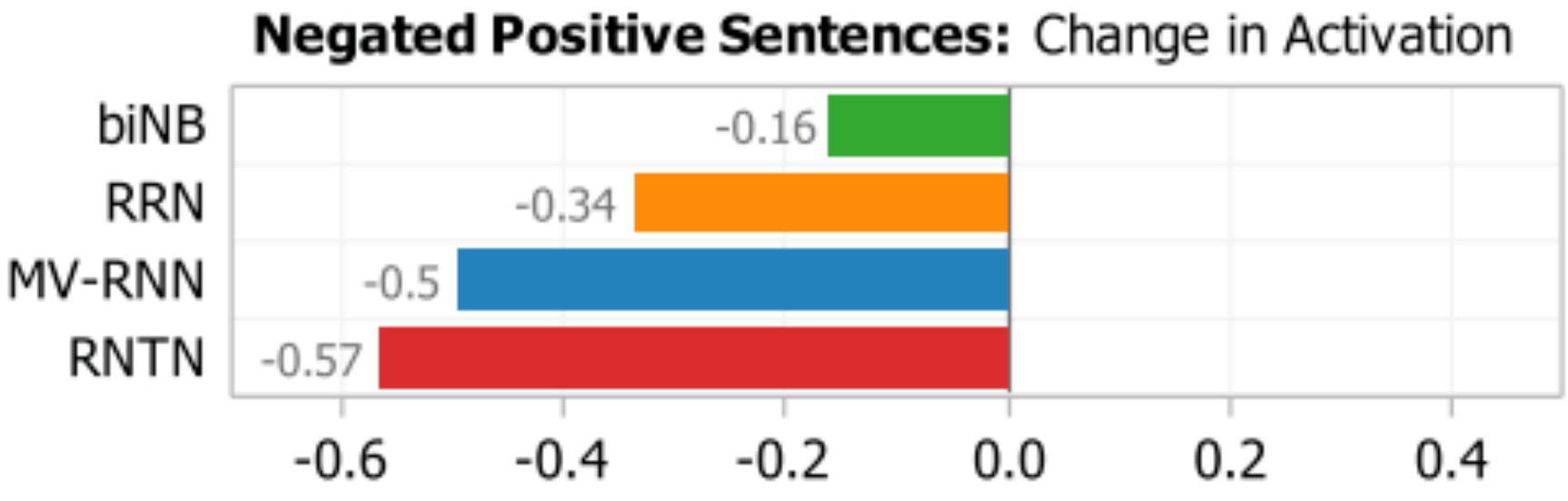




NEGATION RESULTS

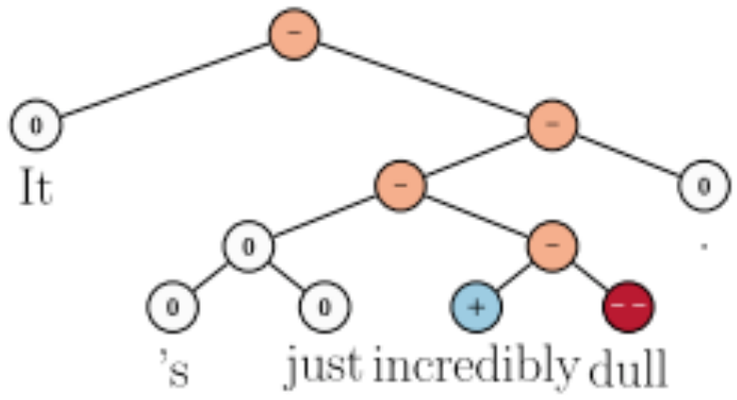
- Negating Positive

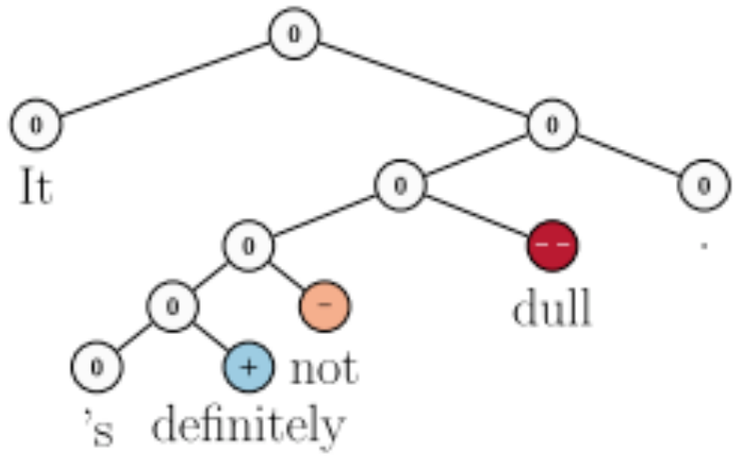
Model	Accuracy	
	Negated Positive	Negated Negative
biNB	19.0	27.3
RNN	33.3	45.5
MV-RNN	52.4	54.6
RNTN	**71.4**	**81.8**

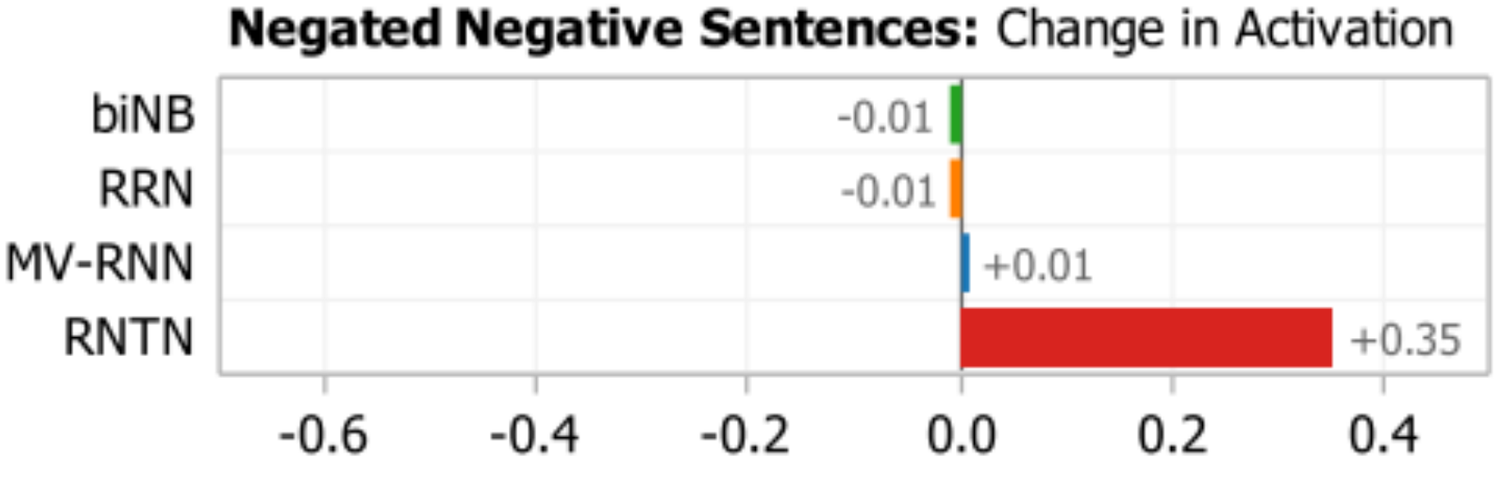


NEGATION RESULTS

- Negating Negative
  - When negative sentences are negated, the overall sentiment should become less negative, but not necessarily positive
  - – Positive activation should increase



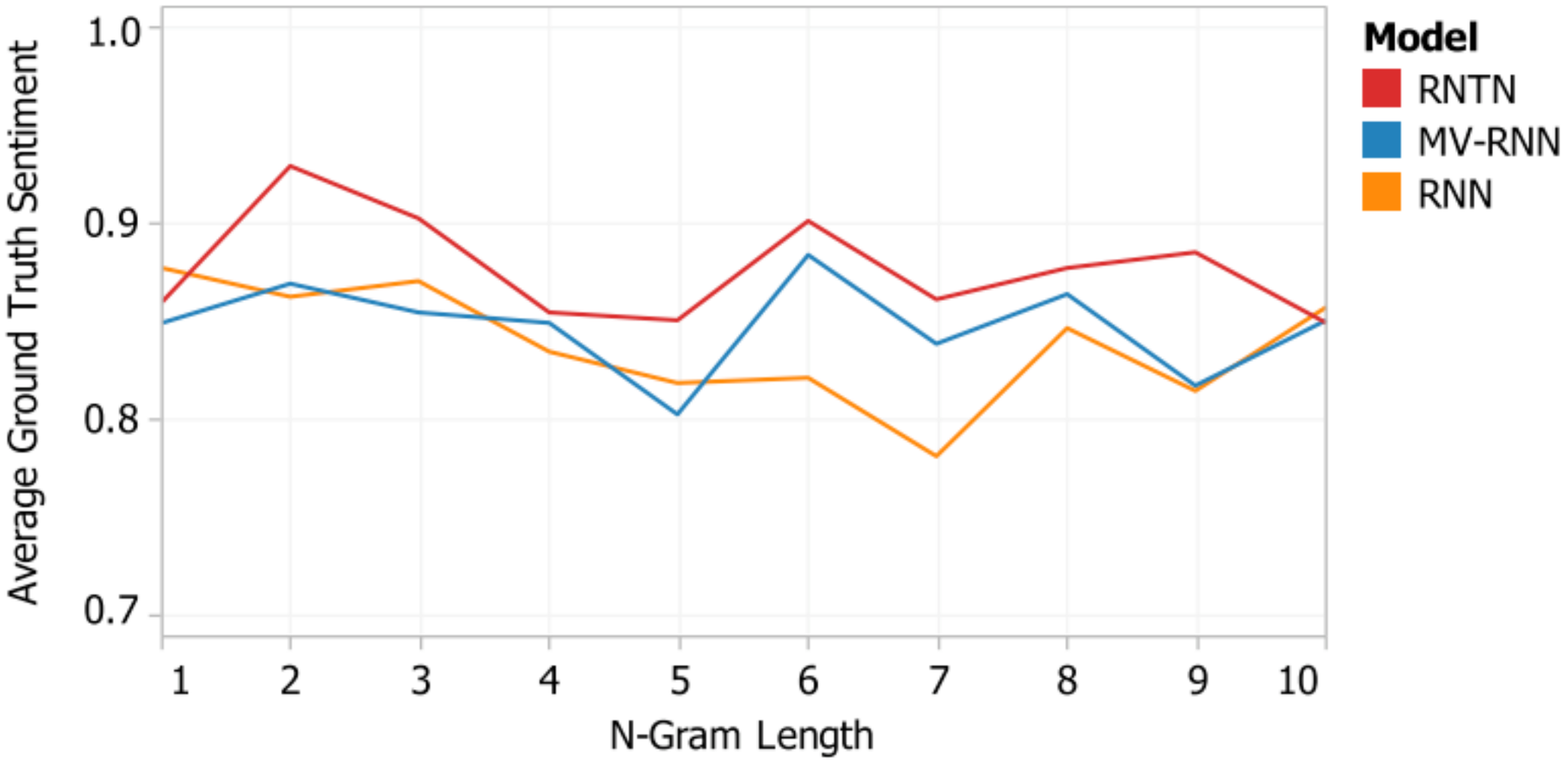




Model	Accuracy	
	Negated Positive	Negated Negative
biNB	19.0	27.3
RNN	33.3	45.5
MV-RNN	52.4	54.6
RNTN	**71.4**	**81.8**

n	Most positive n-grams	Most negative n-grams
1	engaging; best; powerful; love; beautiful	bad; dull; boring; fails; worst; stupid; painfully
2	excellent performances; A masterpiece; masterful film; wonderful movie; marvelous performances	worst movie; very bad; shapeless mess; worst thing; instantly forgettable; complete failure
3	an amazing performance; wonderful all-ages triumph; a wonderful movie; most visually stunning	for worst movie; A lousy movie; a complete failure; most painfully marginal; very bad sign
5	nicely acted and beautifully shot; gorgeous imagery, effective performances; the best of the year; a terrific American sports movie; refreshingly honest and ultimately touching	silliest and most incoherent movie; completely crass and forgettable movie; just another bad movie. A cumbersome and cliche-ridden movie; a humorless, disjointed mess
8	one of the best films of the year; A love for films shines through each frame; created a masterful piece of artistry right here; A masterful film from a master filmmaker,	A trashy, exploitative, thoroughly unpleasant experience ; this sloppy drama is an empty vessel.; quickly drags on becoming boring and predictable.; be the worst special-effects creation of the year

Examples of n-grams for which the RNTN predicted the most positive and most negative responses



Average ground truth sentiment of top 10 most positive n-grams at various n. RNTN selects more strongly positive phrases at most n-gram lengths compared to other models.

DEMO

- http://nlp.stanford.edu:8080/sentiment/rntnDemo.html
- Stanford CoreNLP