Understanding Neural Networks through Representation Erasure

Authors: Jiwei Li, Will Monroe and Dan Jurfsky

Anupam Basu

CISC850
Cyber Analytics

Abstract

- A general methodology to understanding a neural model by using erasure
- To provide a way to conduct an error analysis on a neural model

Introduction

- Cons of Neural Networks
  - Poor interpretability of its components
  - Hard to pinpoint when it makes mistakes

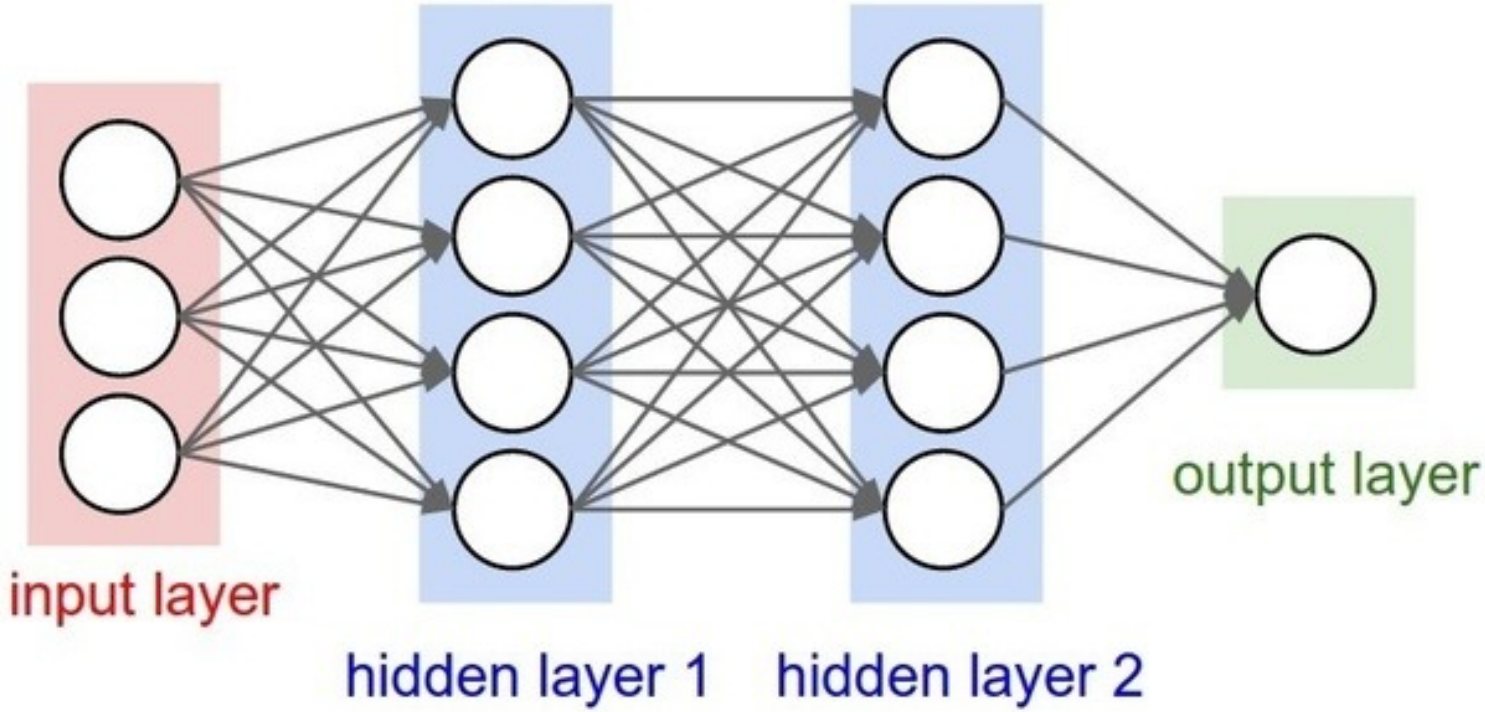
Introduction

- Erasing components :
  - Can improve performance
  - Show importance of components

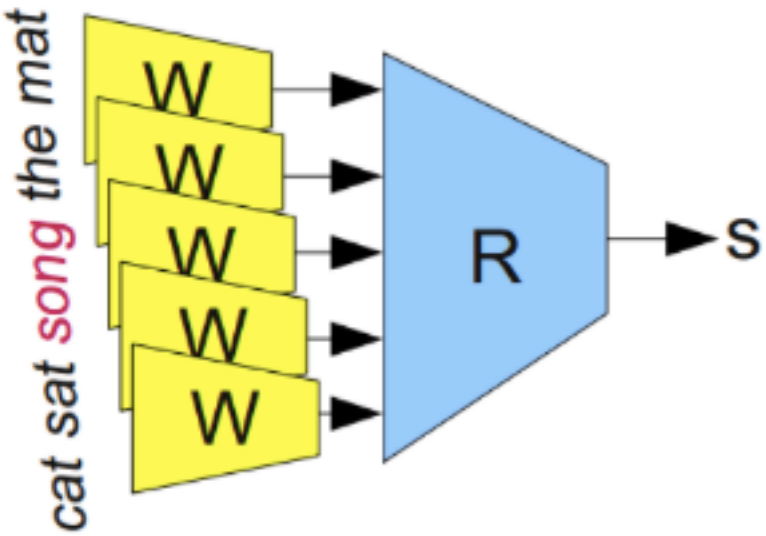
Introduction

- Erasing of features are done on these levels:
  - Input-word vector dimensions
  - Intermediate hidden units
  - Input words

Introduction



Introduction



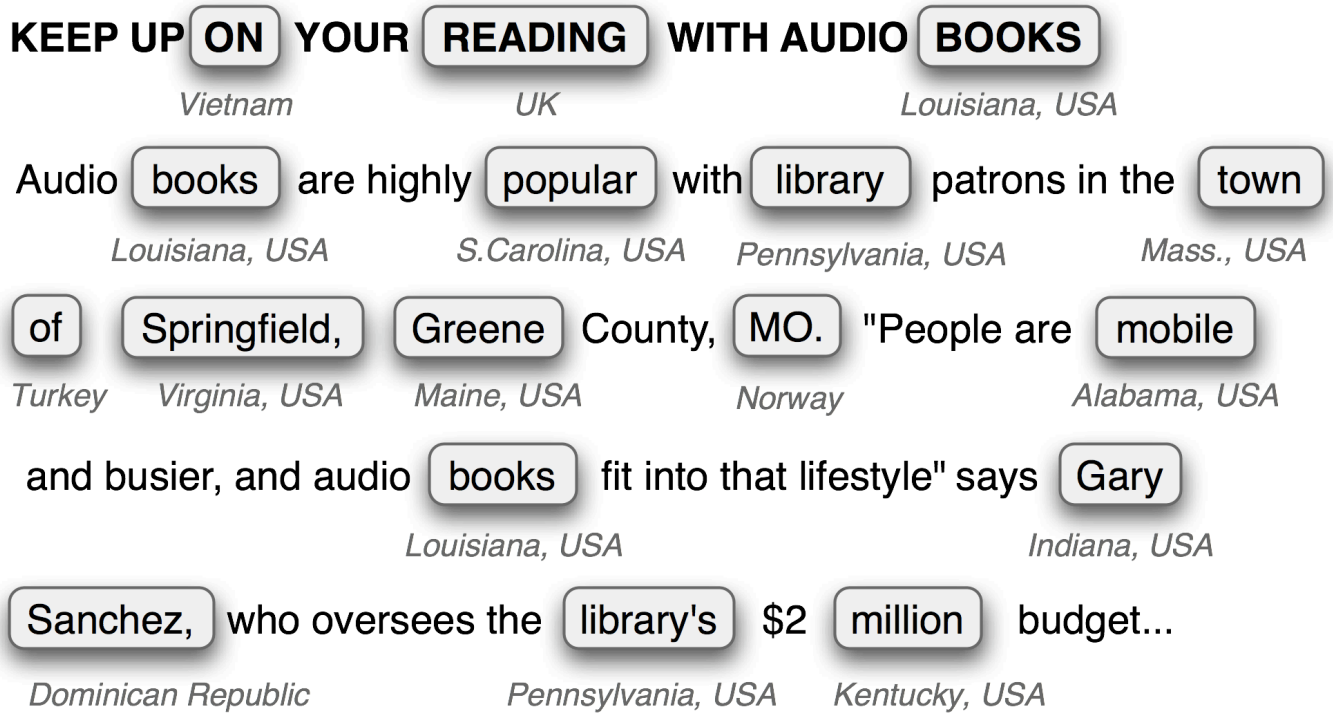
Linking Word Vector Dimensions to Linguistic Features

- Using the model to understand neural models at word vector dimensions:
  - Visualization Model
  - Tasks and Training
  - Results

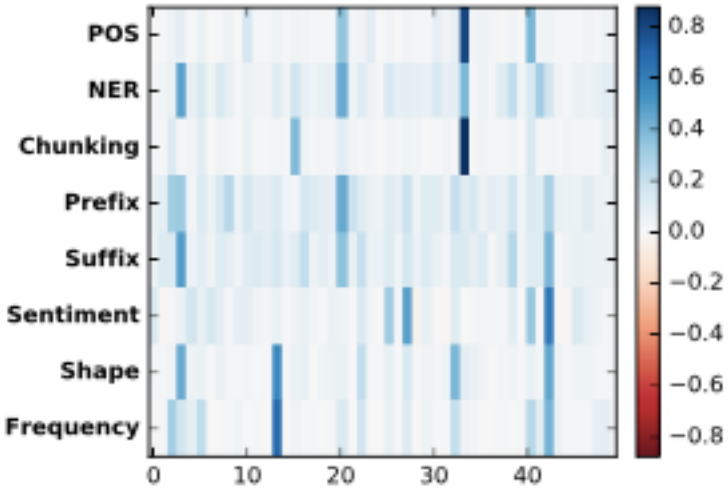
Linking Word Vector Dimensions to Linguistic Features

- Linguistic Features which can be used :
  - Parts of Speech
  - Named Entity class
  - word frequency
  - word-shape

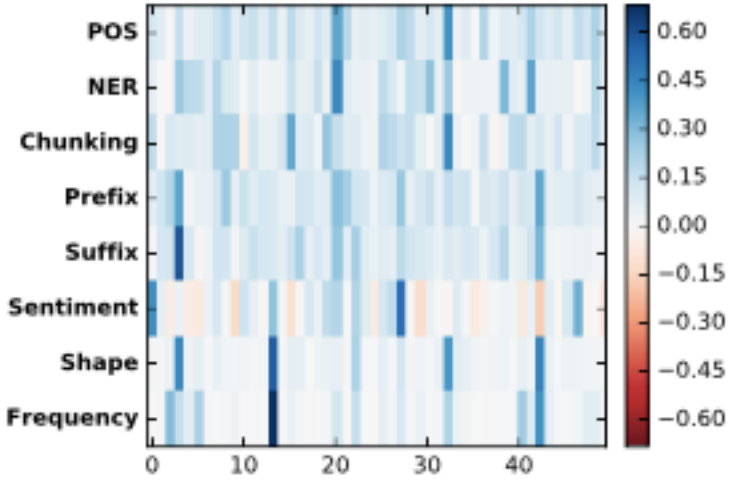
Linking Word Vector Dimensions to Linguistic Features



Linking Word Vector Dimensions to Linguistic Features

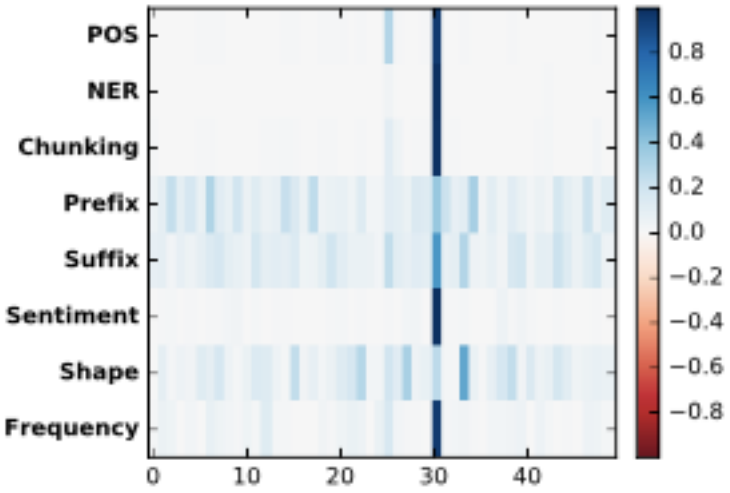


(a) Word2vec, no dropout.



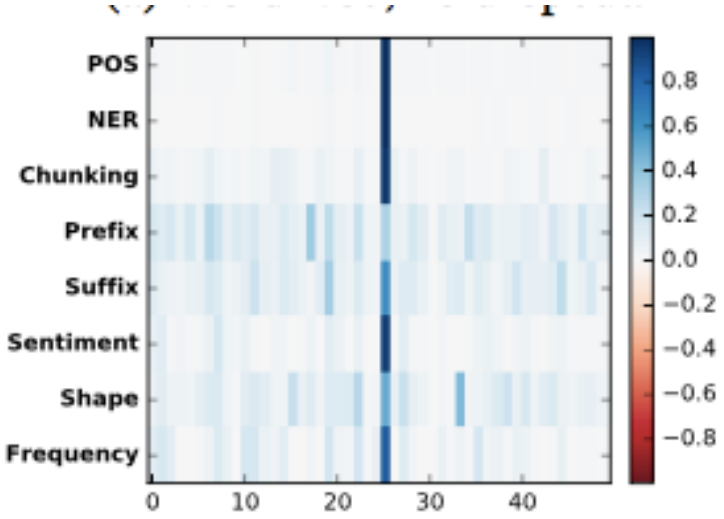
(b) Word2vec, with dropout.

Linking Word Vector Dimensions to Linguistic Features

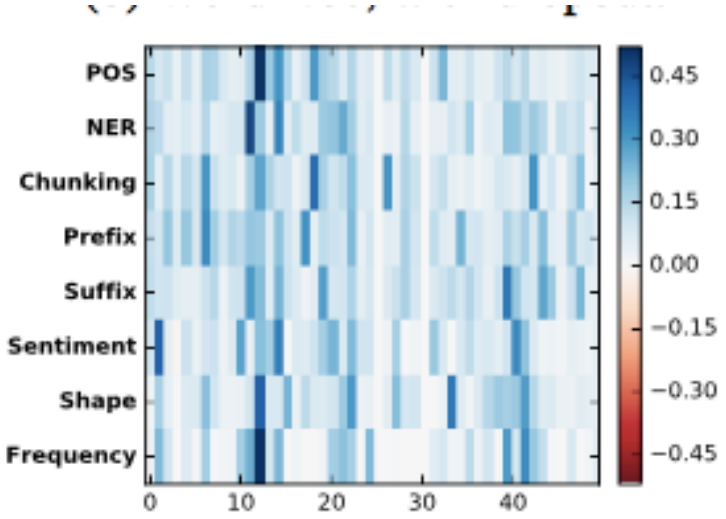


(c) GloVe, no dropout.

Linking Word Vector Dimensions to Linguistic Features

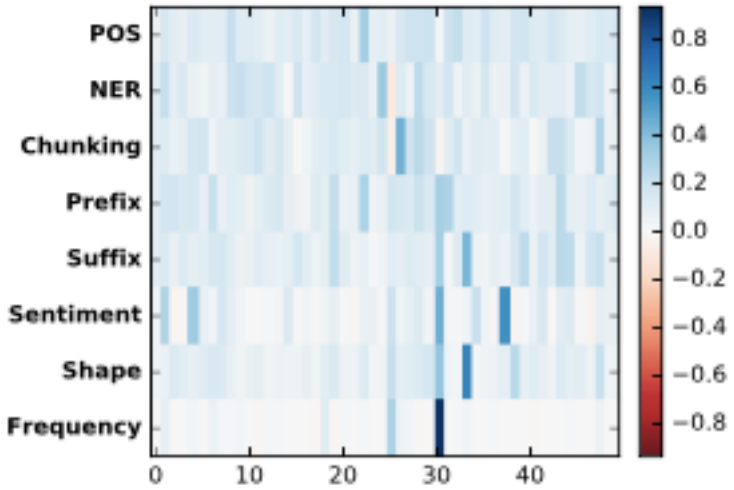


(d) GloVe, no dropout; 31rd dimension removed.



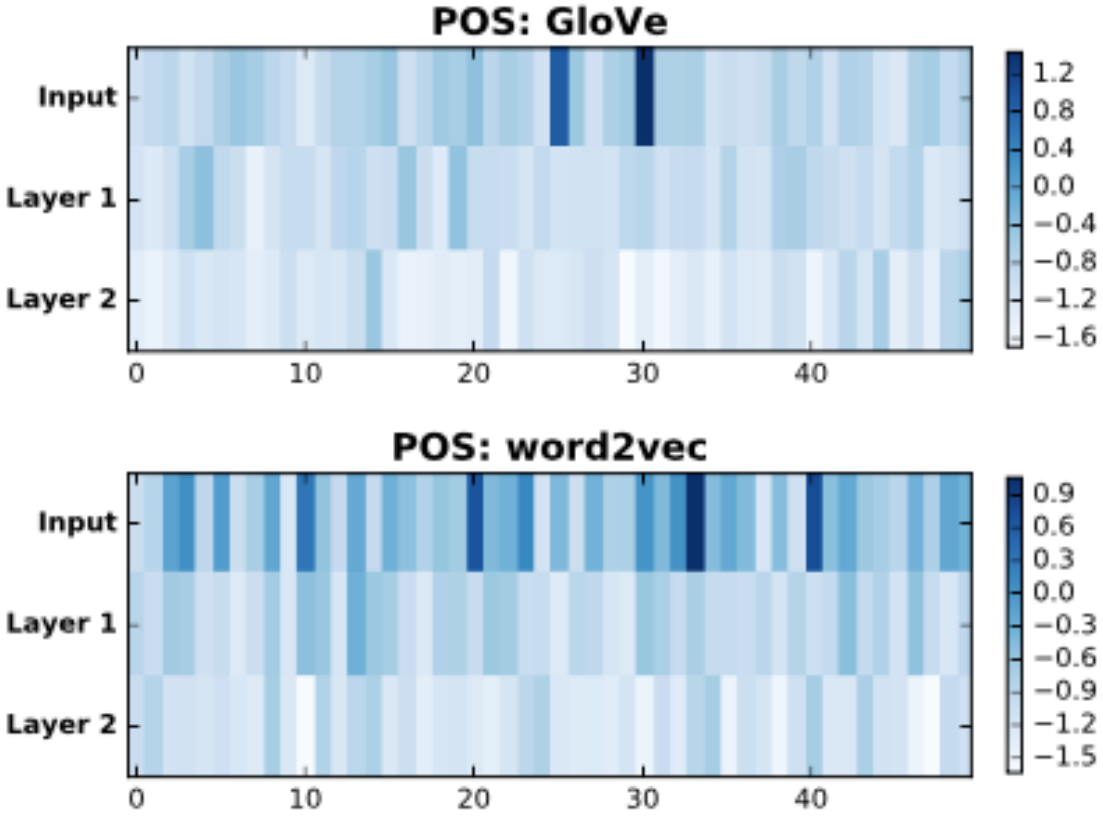
(e) GloVe, no dropout; 31rd, 26th dimensions removed.

Linking Word Vector Dimensions to Linguistic Features



(f) GloVe, with dropout.

Linking Word Vector Dimensions to Linguistic Features



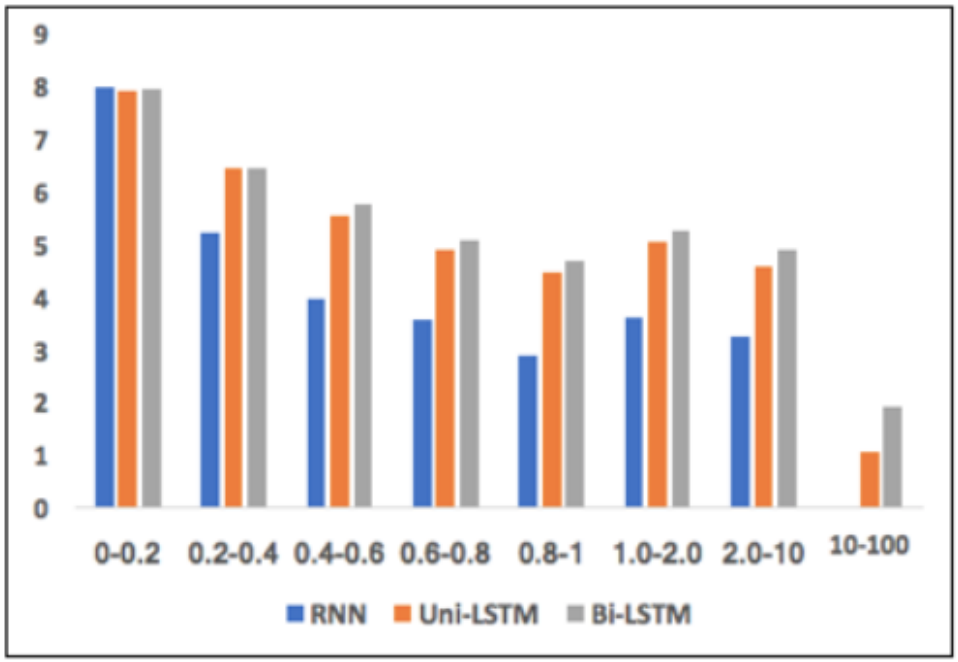
Finding Important Words in Sentiment Analysis

- Using the model to understand neural models at word level:
  - Computing log likelihood of correct sentiment when a particular word is erased

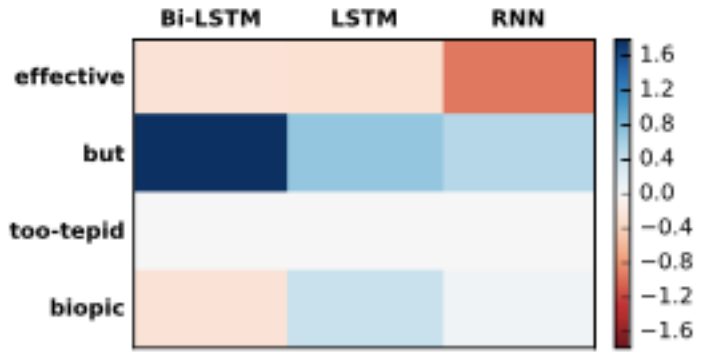
Finding Important Words in Sentiment Analysis

word	Bi-LSTMs	Uni-LSTMs	RNN
greatest	9.463	5.593	0.742
wonderful	9.521	3.292	0.704
worst	7.739	4.698	0.967
excellent	6.835	4.883	1.859
best	4.916	2.448	0.548
hated	6.557	3.512	4.338
love	1.678	1.786	0.999
unforgettable	2.286	1.648	1.482
waste	4.579	3.600	2.342
disaster	3.728	3.362	0.021

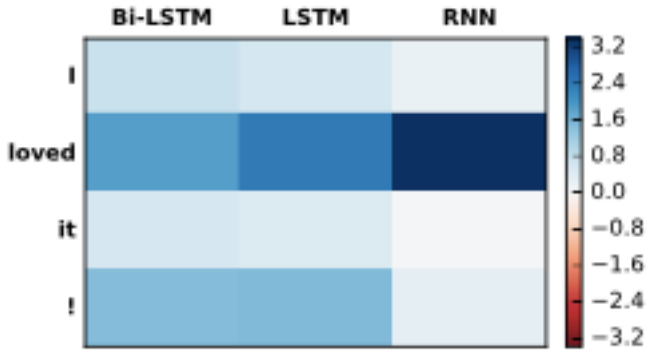
Finding Important Words in Sentiment Analysis



Finding Important Words in Sentiment Analysis

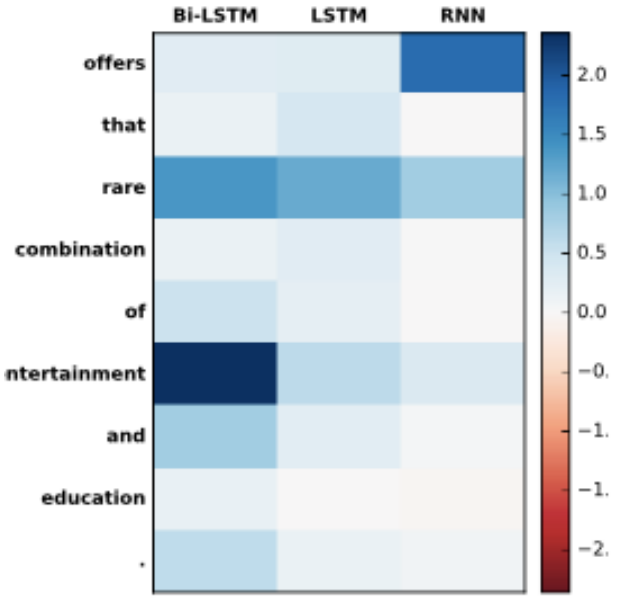


(a) Neutral

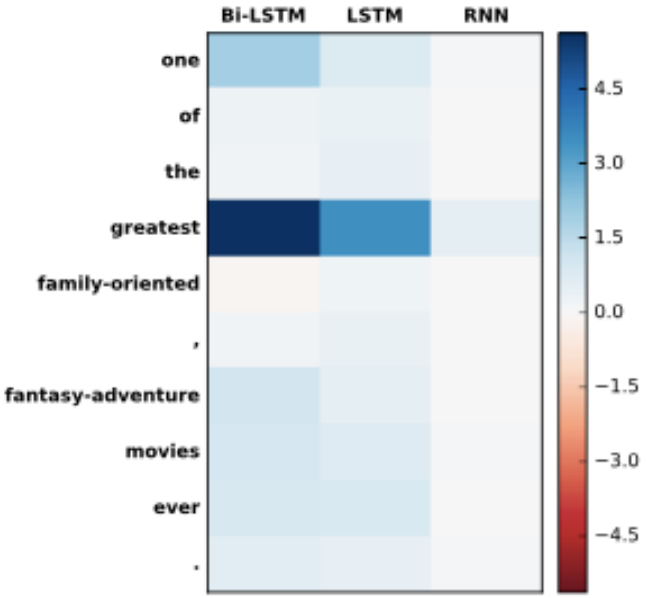


(b) Strong positive

Finding Important Words in Sentiment Analysis

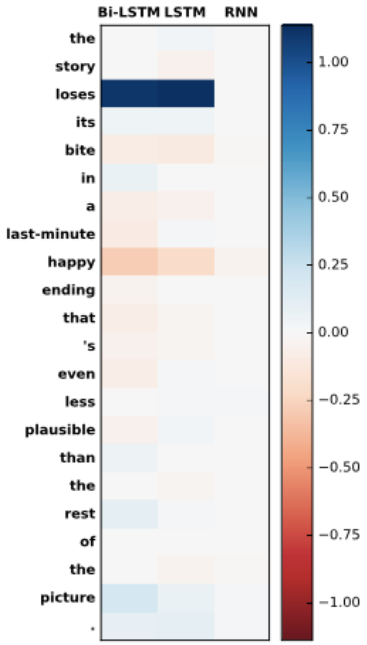


(c) Strong positive



(d) Strong positive

Finding Important Words in Sentiment Analysis



(e) Strong negative

Reinforcement Learning for Finding Decision-Changing Phrases

- Using the model to understand neural models at sentence level:
  - Task, Dataset and Training
  - Results

Reinforcement Learning for Finding Decision-Changing Phrases

(1) clean updated room. friendly efficient staff . rate was too high 199 plus they charged 10 day for internet access in the room .
(2) the location is fantastic. the staff are helpful and service oriented . sleeping rooms meeting rooms and public lavatories not cleaned on a daily basis . the hotel seems a bit old and a bit tired overall . trolley noise outside can go into the wee hours . if you get a great price for a few nights this hotel may be a good choice . breakfast is very nice remember if you just stick to the cold buffet it is cheaper .
(3) location is nice . but goes from bad to worse once you walk through the door . staff very surly and unhelpful . room and hallway had a very strange smell . rooms very run down . so bad that i checked out immediately and went to another hotel . intercontinental chain should be ashamed .
(4) i took my daughter and her step sister to see a show at webster hall . it is so overpriced i 'm in awe . i felt safe . the rooms were tiny . lots of street noise all night from the partiers at the ale house below .

Reinforcement Learning for Finding Decision-Changing Phrases

(1) clean updated room. friendly efficient staff . rate was too high 199 plus they charged 10 day for internet access in the room .
the location is fantastic. the staff are helpful and service oriented . (2) sleeping rooms meeting rooms and public lavatories not cleaned on a daily basis . the hotel seems a bit old and a bit tired overall . trolley noise outside can go into the wee hours . if you get a great price for a few nights this hotel may be a good choice . breakfast is very nice remember if you just stick to the cold buffet it is cheaper .
(3) location is nice . but goes from bad to worse once you walk through the door . staff very surly and unhelpful . room and hallway had a very strange smell . rooms very run down . so bad that i checked out immediately and went to another hotel . intercontinental chain should be ashamed .
(4) i took my daughter and her step sister to see a show at webster hall . it is so overpriced i 'm in awe . i felt safe . the rooms were tiny . lots of street noise all night from the partiers at the ale house below .

Conclusion

- This methodology shows the benefits and harms in erasing representation, helps in the error analysis of neural networks.
- This has the potential to benefit a wide variety of models and tasks.

Thank You