SENTIMENT LABELLING ANALYSIS

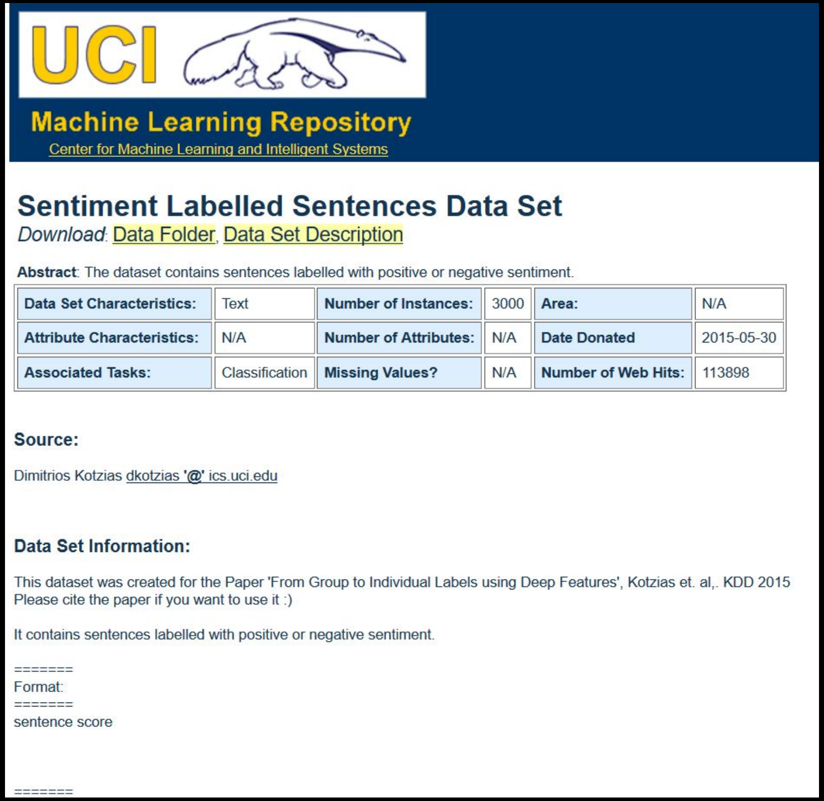
By Lixiong Feng, Qihao Huang, Chenhao Li,
Yongheng Zhang, Zihua Zhang

WHY IS IT IMPORTANT?

- Customer reviews
- User feedback
- Elections
- In general

OUR DATA

- Reviews on phones and accessories on Amazon
- Original paper published in 2015, using this dataset and two more from Yelp and IMDb



UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Sentiment Labelled Sentences Data Set

Download: Data Folder, Data Set Description

Abstract: The dataset contains sentences labelled with positive or negative sentiment.

Data Set Characteristics:	Text	**Number of Instances:**	3000	**Area:**	N/A
Attribute Characteristics:	N/A	**Number of Attributes:**	N/A	**Date Donated**	2015-05-30
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	113898

Source:

Dimitrios Kotzias dkotzias '@' ics.uci.edu

Data Set Information:

This dataset was created for the Paper 'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015
Please cite the paper if you want to use it :)

It contains sentences labelled with positive or negative sentiment.

=======
Format:
=======
sentence score

=======

'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015
https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

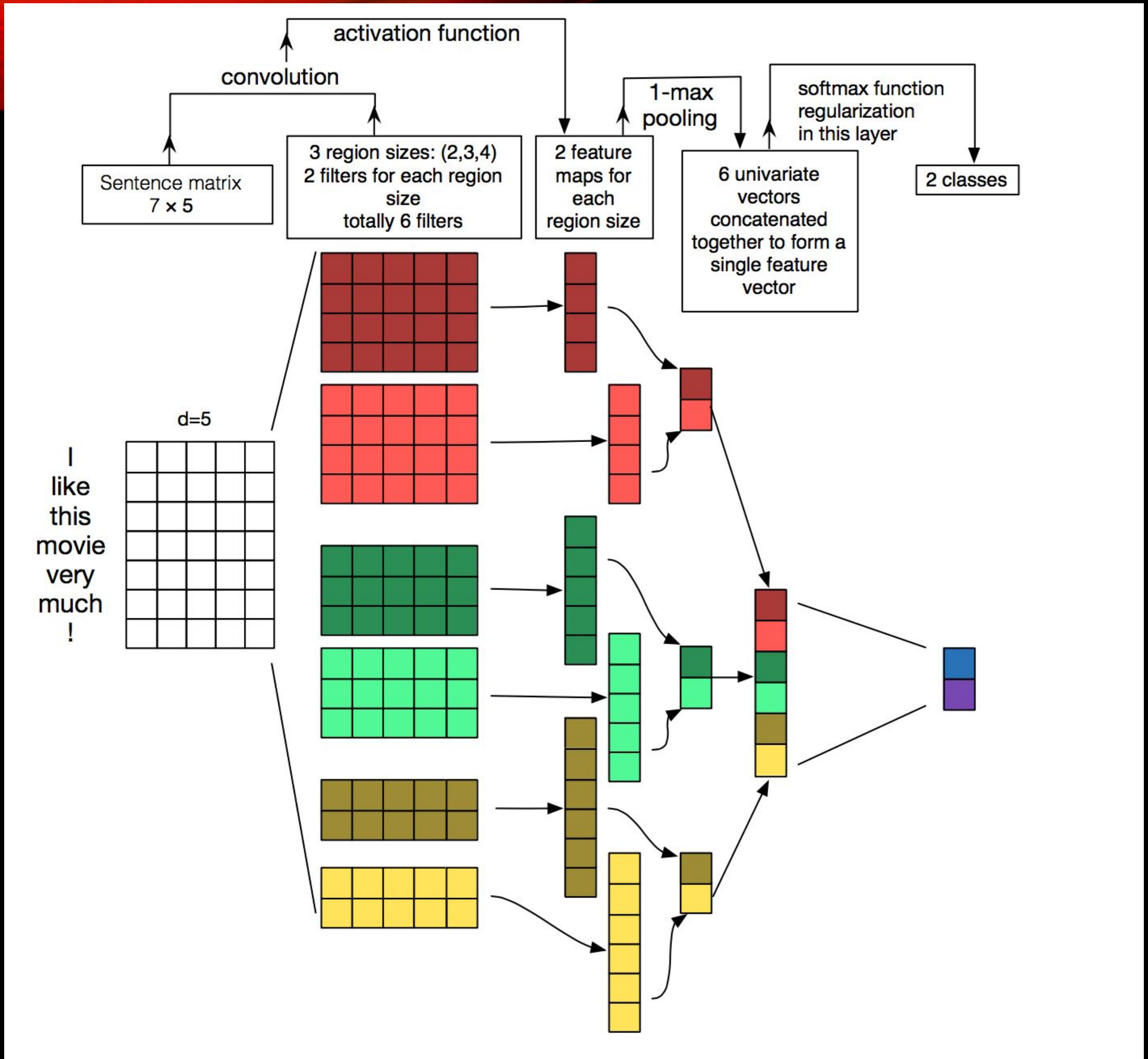
PREPROCESSING

- Remove most punctuations; not remove .!?
- Lowercase all words
- And, most importantly...

Word2Vec

- Words appear near each other has close vector
- good for finding relative positive/negative words
- drop words appears less than 5 times
- end up with 340 vocabularies

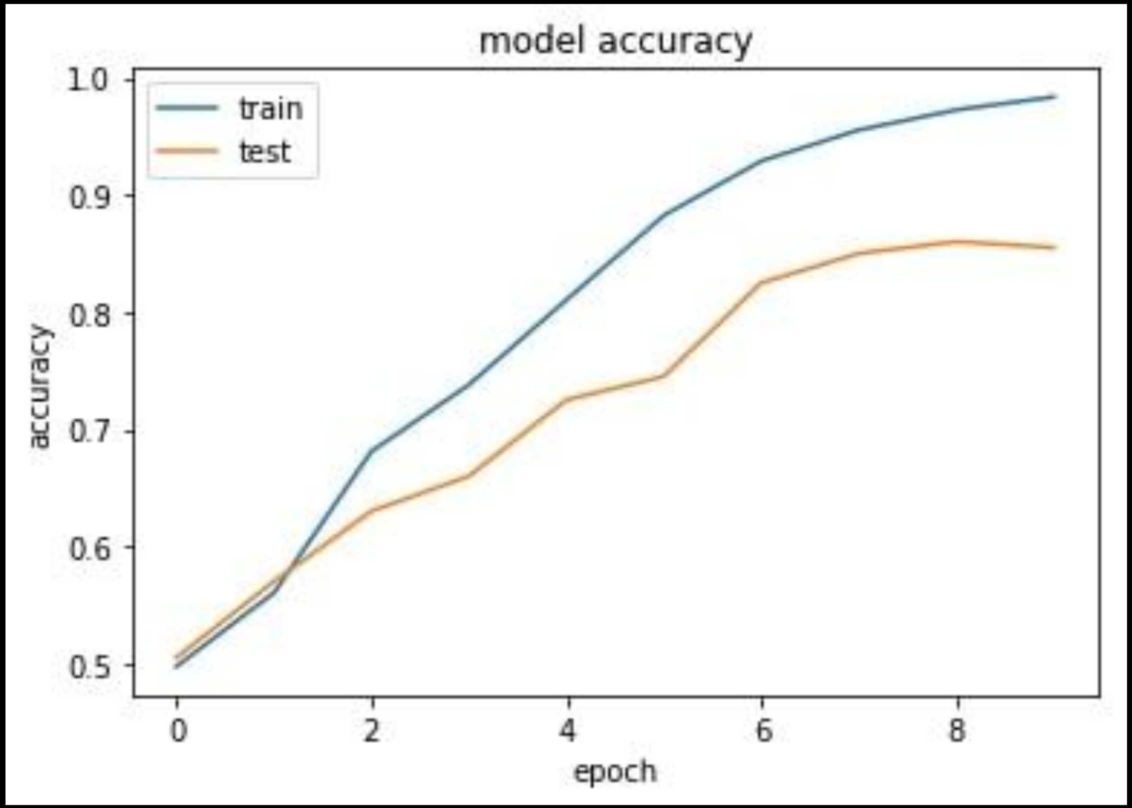
CNN

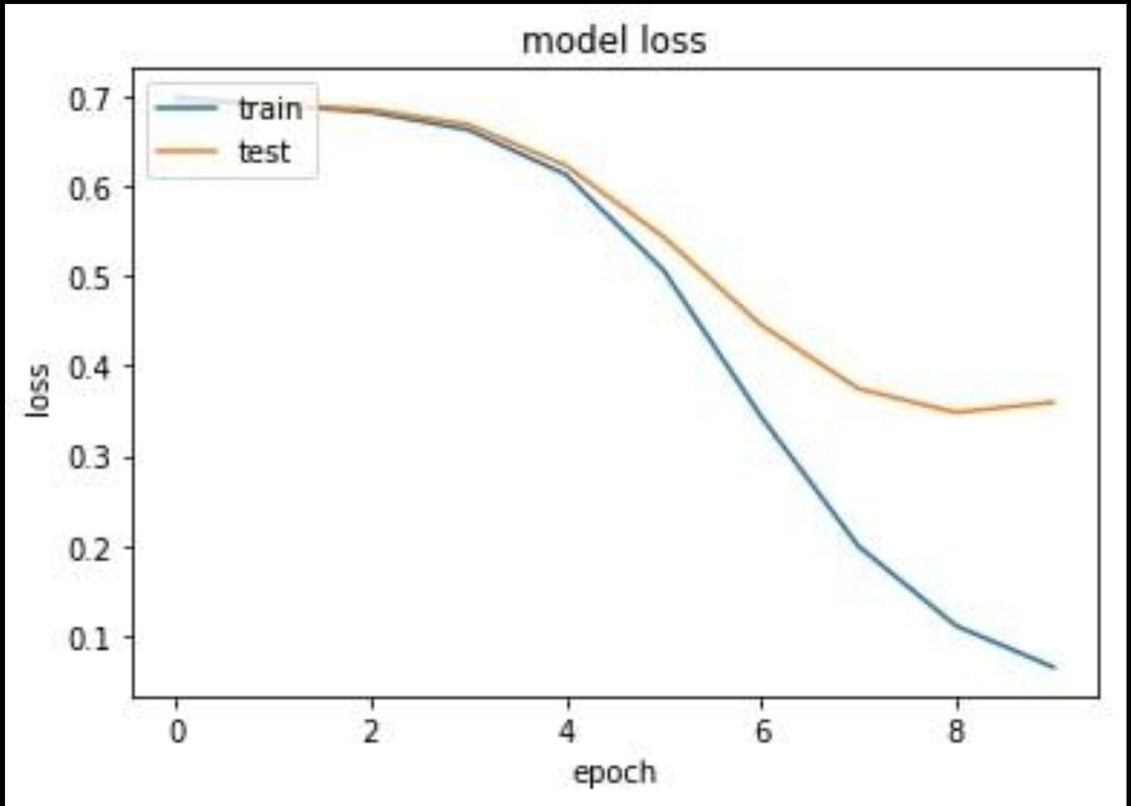


- Advantage:
  - It requires fewer parameters.
  - It is good at extracting relevant information.
  - It has been used in solving many image and text based problems.
- Architecture:
  - 128 filters with kernel size of 3
  - Batch size is 128

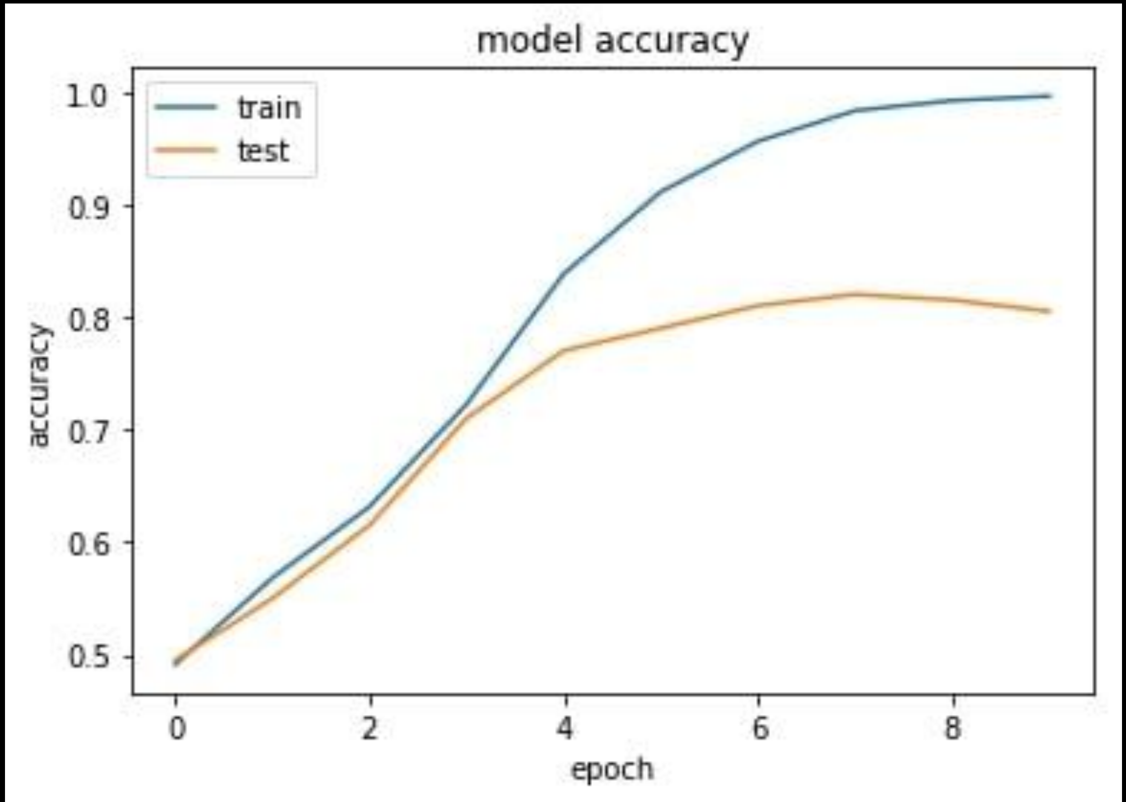
Image Reference : http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

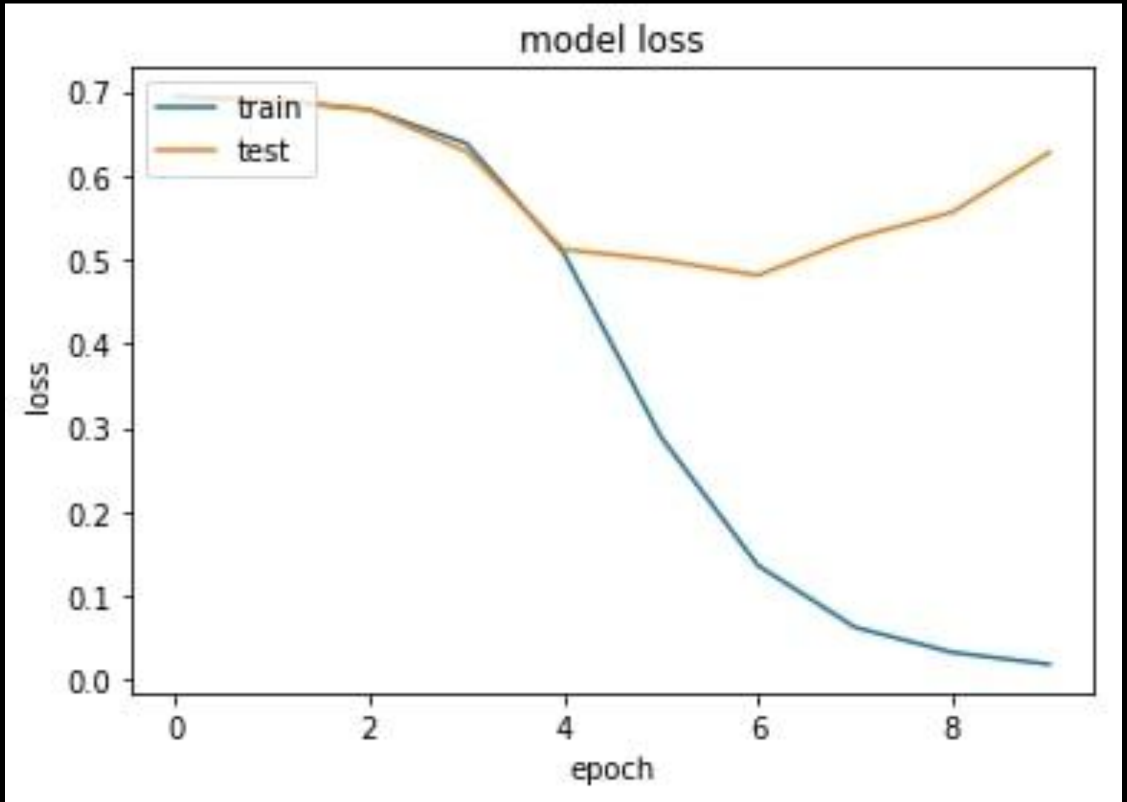
CNN-1 Convolutional Layer



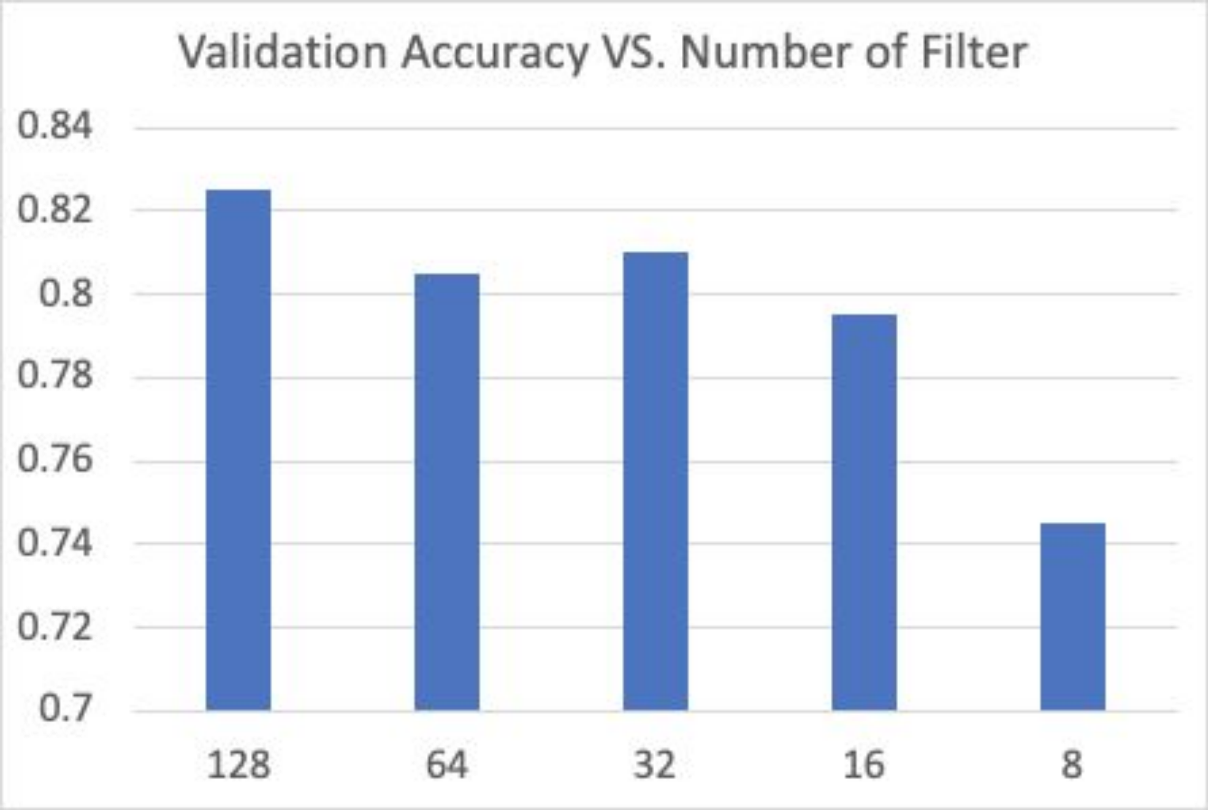


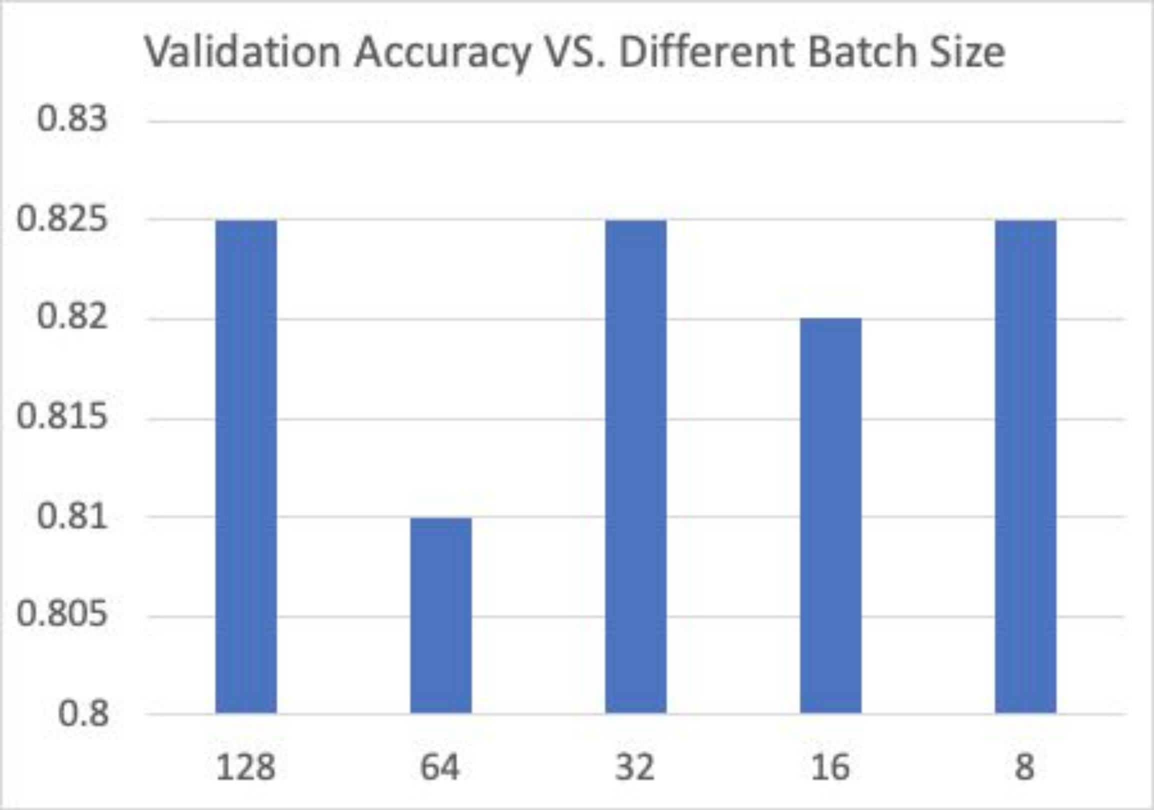
CNN- 2 Convolutional Layers





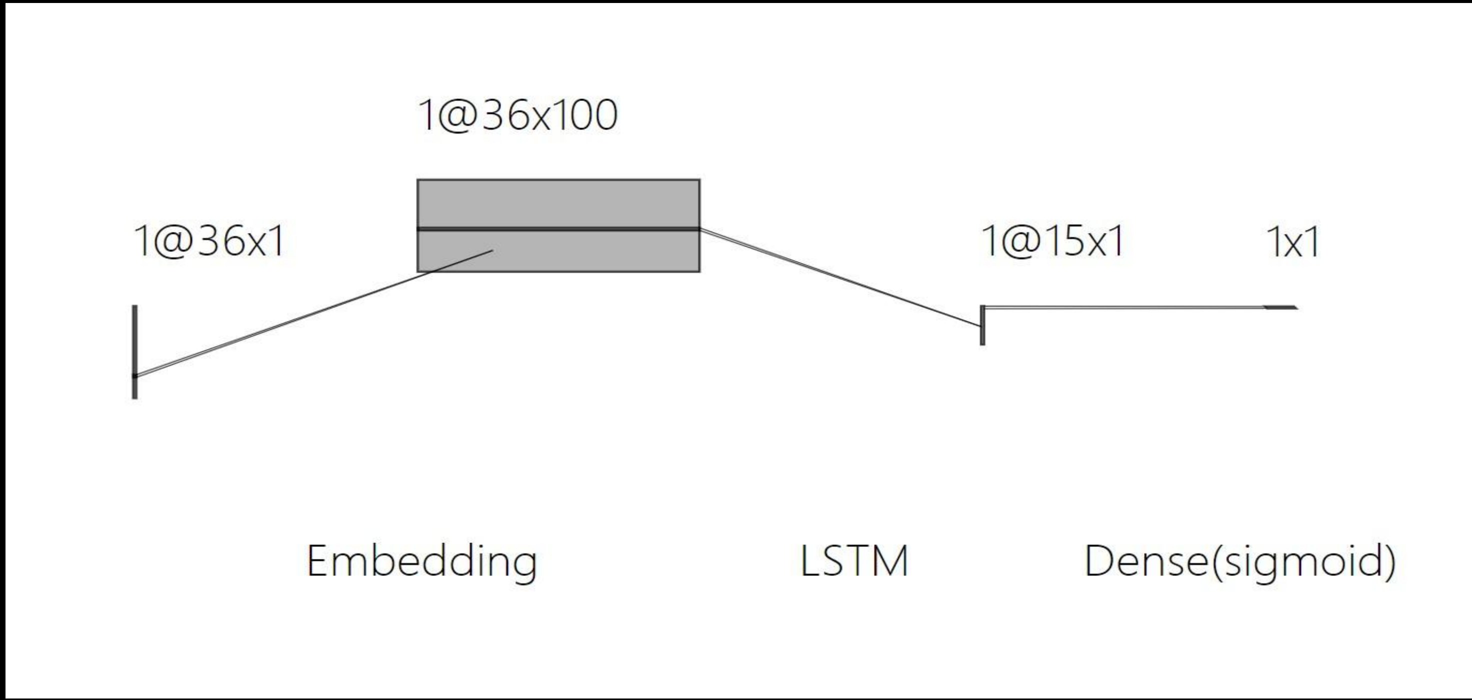
Adjusting Hyper Parameter



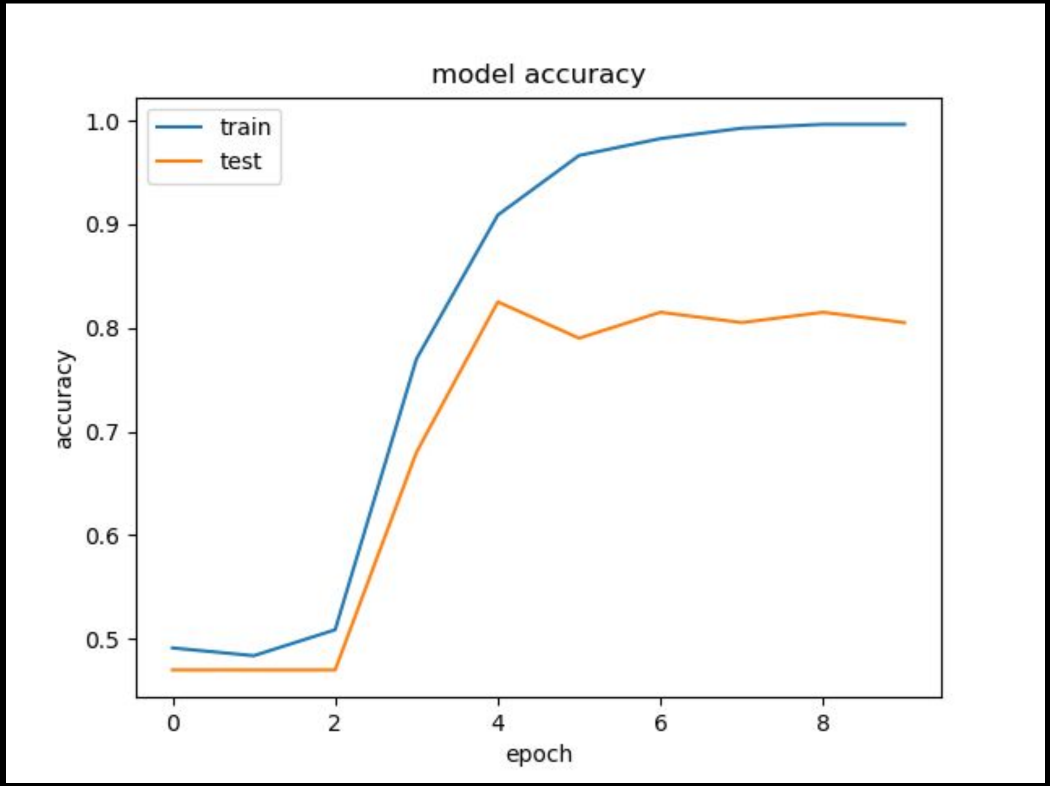


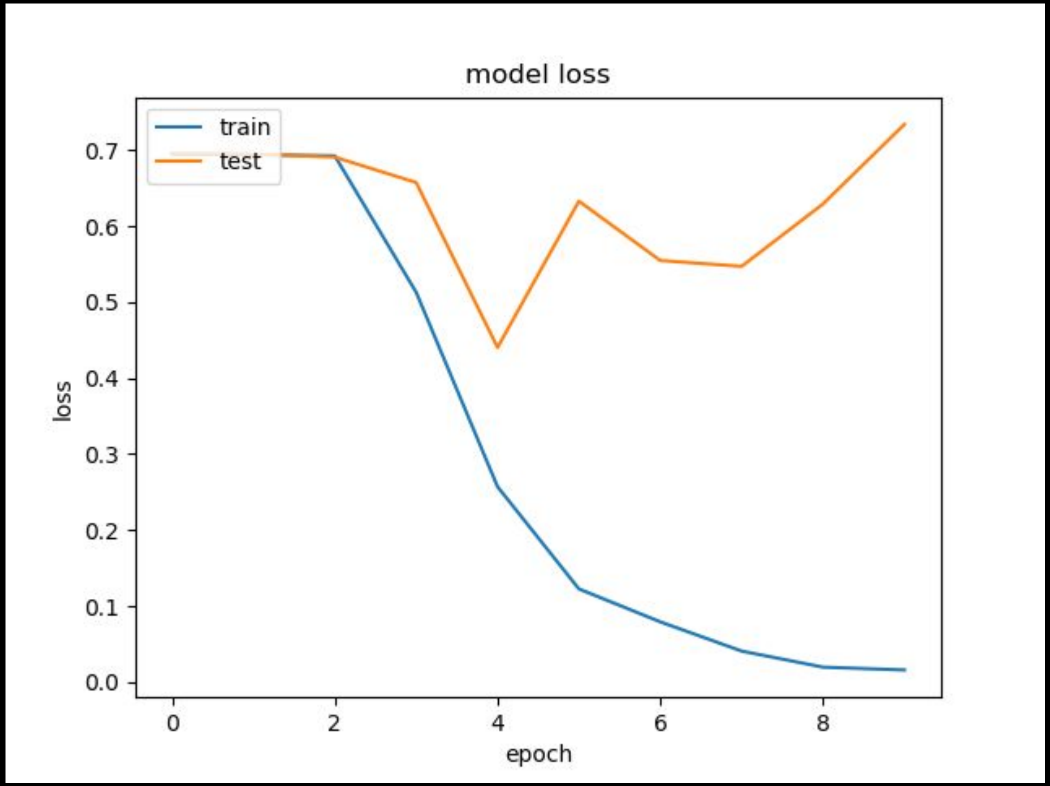
LSTM

- Analysis on sentence, better to look at the one sentence as a whole
- LSTM is a good tool for us to achieve the purpose

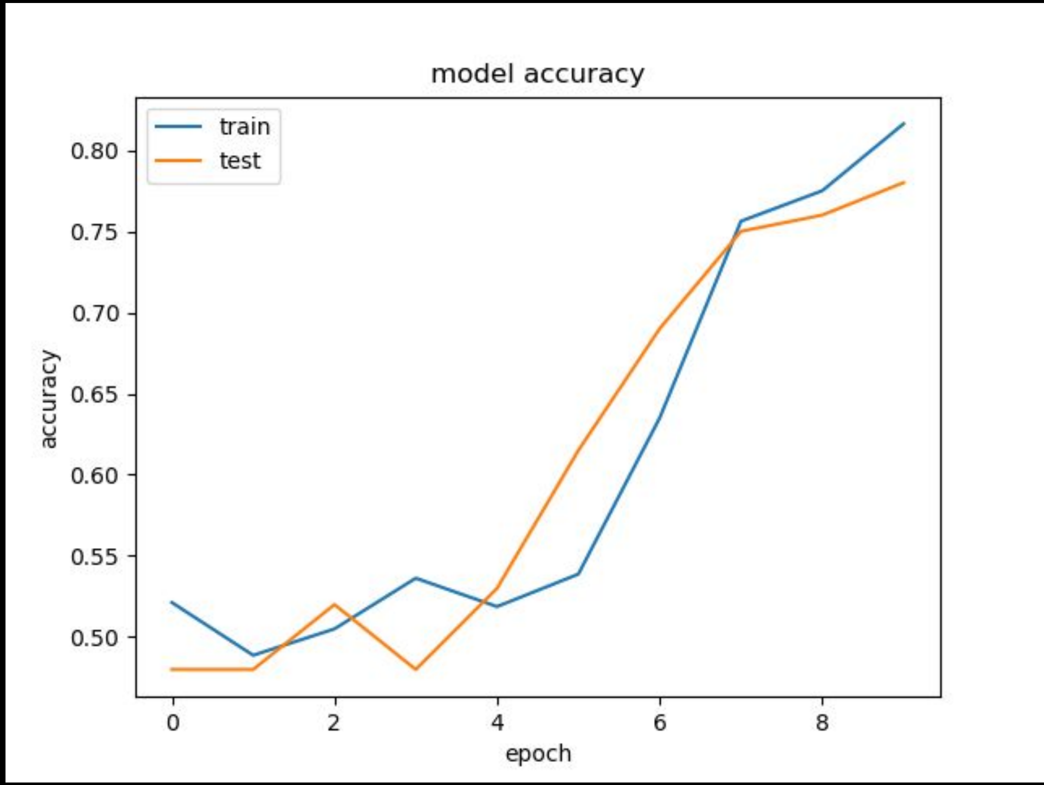


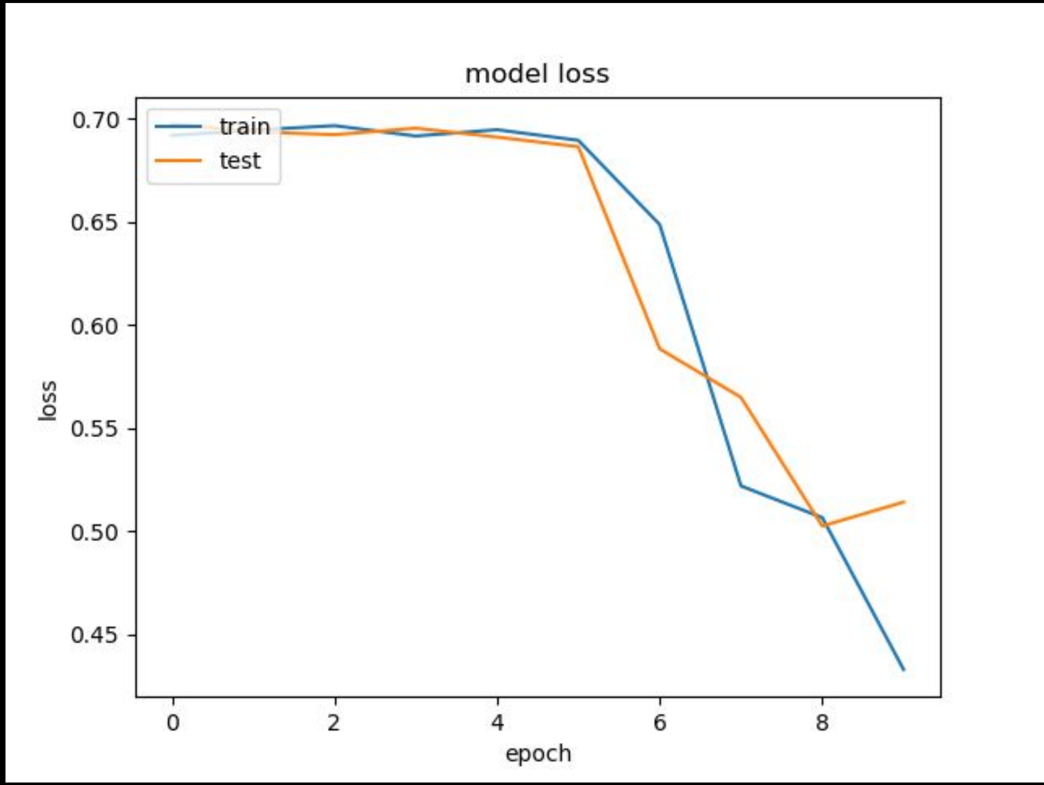
LSTM Result


model accuracy
train
test
accuracy
epoch
1.0
0.9
0.8
0.7
0.6
0.5
0
2
4
6
8


model loss
train
test
loss
epoch
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8

LSTM with Word2Vec Result



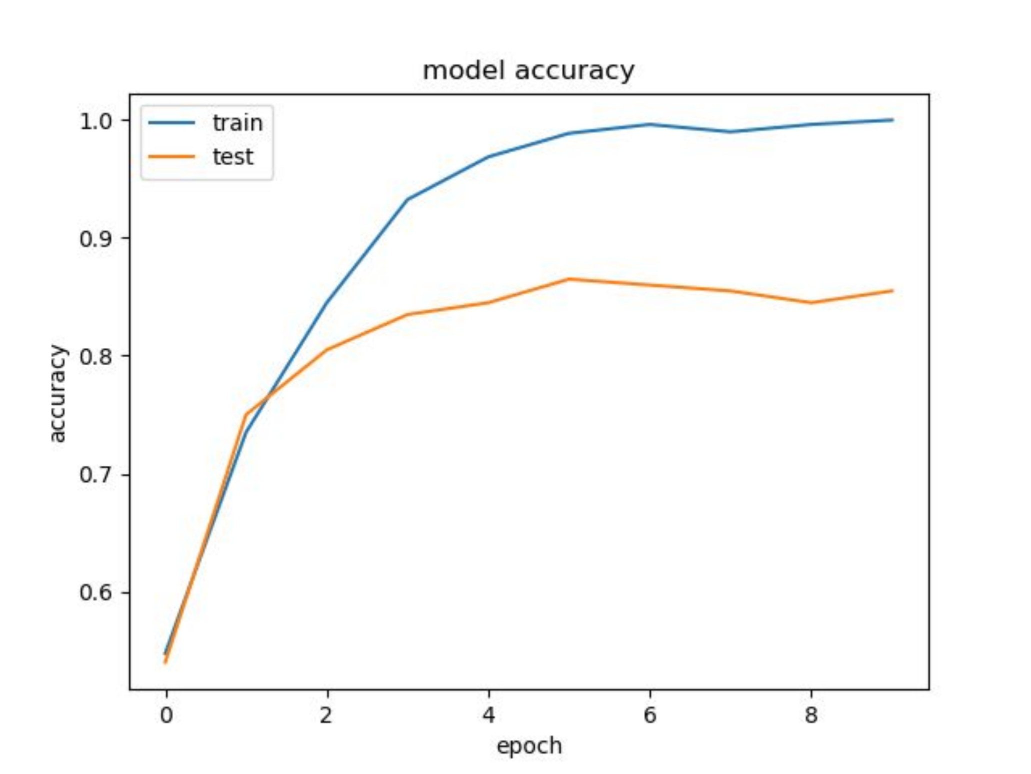


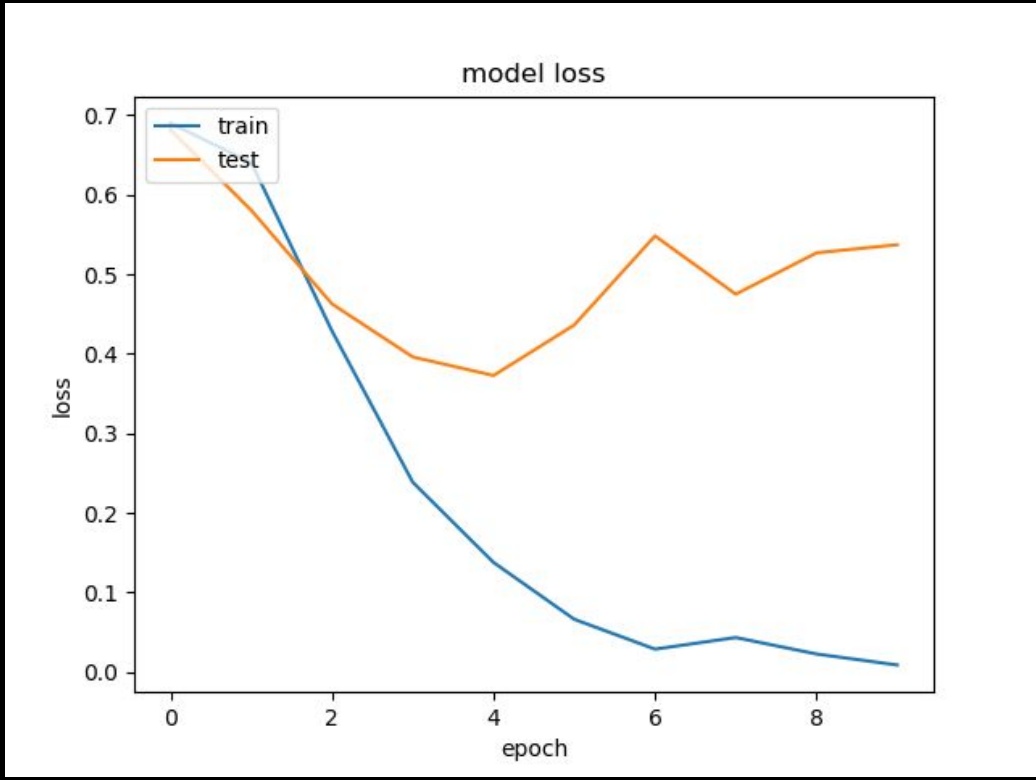
Bidirectional LSTM

- overfitting since the third epoch, not too accurate
- LSTM is not stable when using Word2Vec
- And it does not train perfectly in the beginning epochs
- May use bidirectional since we also need to take consider if following words are negative while previous are positive, and vise-versa

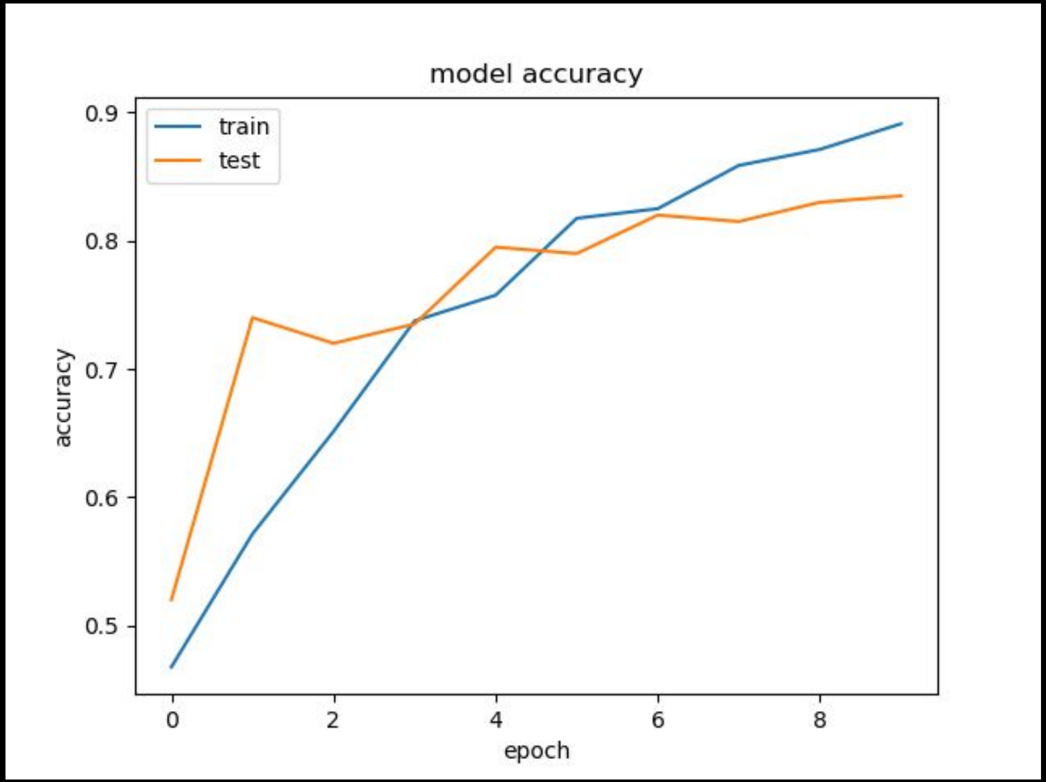
reference: A. Graves, A. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 6645-6649. doi: 10.1109/ICASSP.2013.6638947

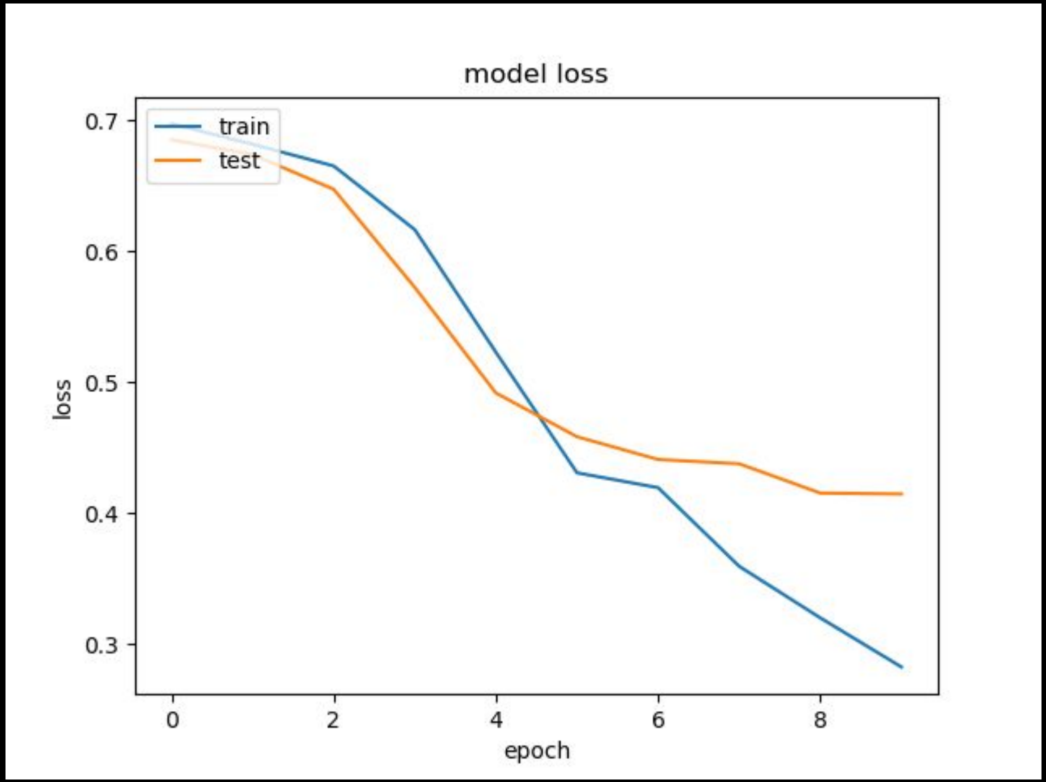
Bidirectional LSTM Result





Bidirectional LSTM with Word2Vec Result

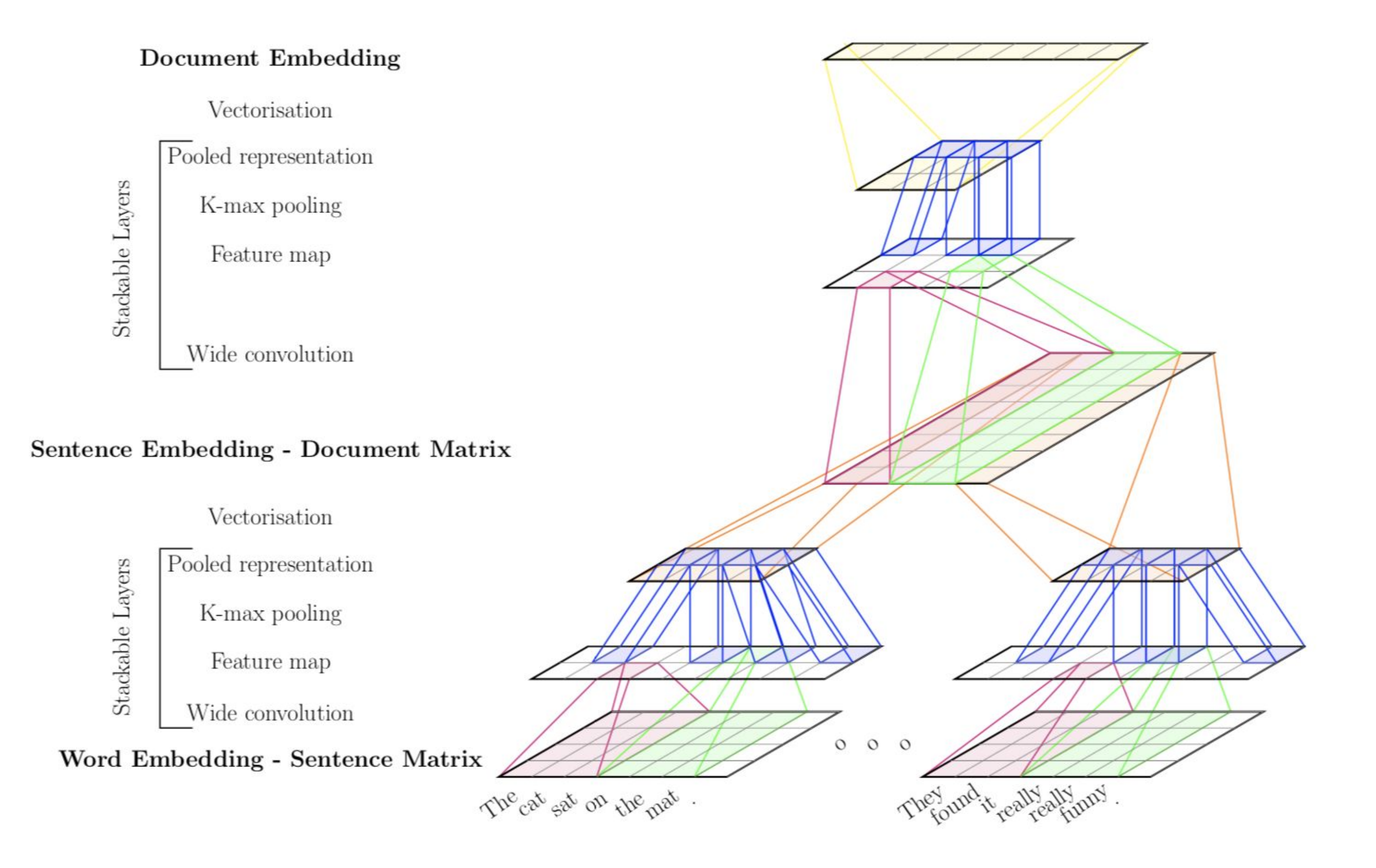




ANALYSIS OF RESULT

- Accuracy ~83%
- Word2Vec reduces overfitting
- Overfitting still exists
- A more complex model can help
- how complex?

STATE-OF-ART



Used convolutional neural network for embedding

Get vector-based representation for sentence (Intermediate representation)

Transform the sentence embedding to full document representation (review)

M. Denil, A. Demiraj, and N. de Freitas. Extraction of salient sentences from labelled documents. Technical report, University of Oxford, 2014.

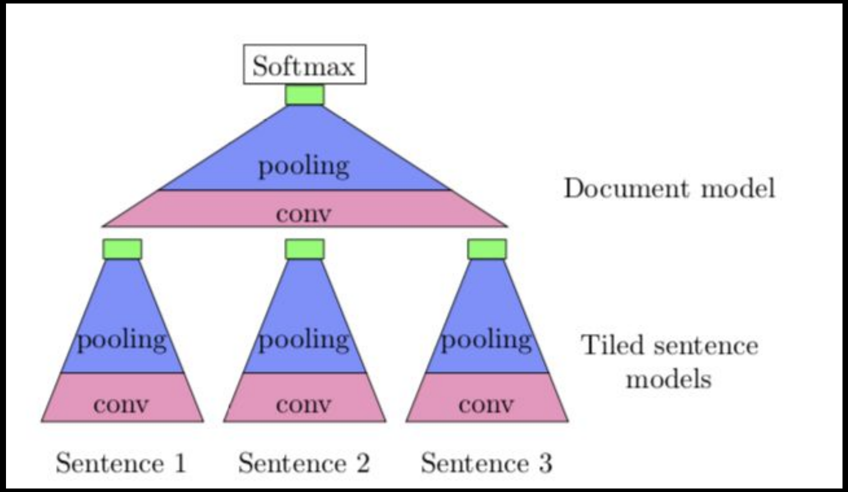
'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015
https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015
https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

STATE-OF-ART

	Accuracy			AUC		
	Amazon	IMDb	Yelp	Amazon	IMDb	Yelp
Logistic w/ BOW on Documents	85.8%	86.20%	**91.25%**	88.08%	88.32	**94.41**
Logistic w/ BOW on Sentences	88.3%	81.81%	78.16%	87.19%	82.67	67.87
Logistic w/ Embeddings on Documents	67.82%	58.23%	81.00%	61.24%	60.77	82.59
GICF w/ Embeddings on Sentences	**92.8%**	**88.56%**	88.73 %	**91.73%**	**88.36%**	92.36%

Table 3: Accuracy and Area-Under-the-Curve (AUC) scores for predicting labels at the group (document) level for the baselines and our proposed method (GICF). Training is always done at the group level. Testing on sentences corresponds to scoring each sentence separately and aggregating the results. BOW or embeddings corresponds to the features used.



Sentence Level:

20-Dimensional word embedding, convolved with 10 feature maps with width 15

Followed by 7-max pooling layer and a tanh nonlinearity+Dropout 0.2

Document Level:

Convolves inputs with 30 feature maps with width 9

Followed by 5-max pooling layer and a tanh nonlinearity+Dropout 0.5

FUTURE IMPROVEMENTS

Change the number of hidden layers
Change the activation function
Change the number of epoch
Change the number of neurons
Use cross validation
Combine CNN with LSTM

QUESTIONS?