Sentiment Analysis of Twitter Posts

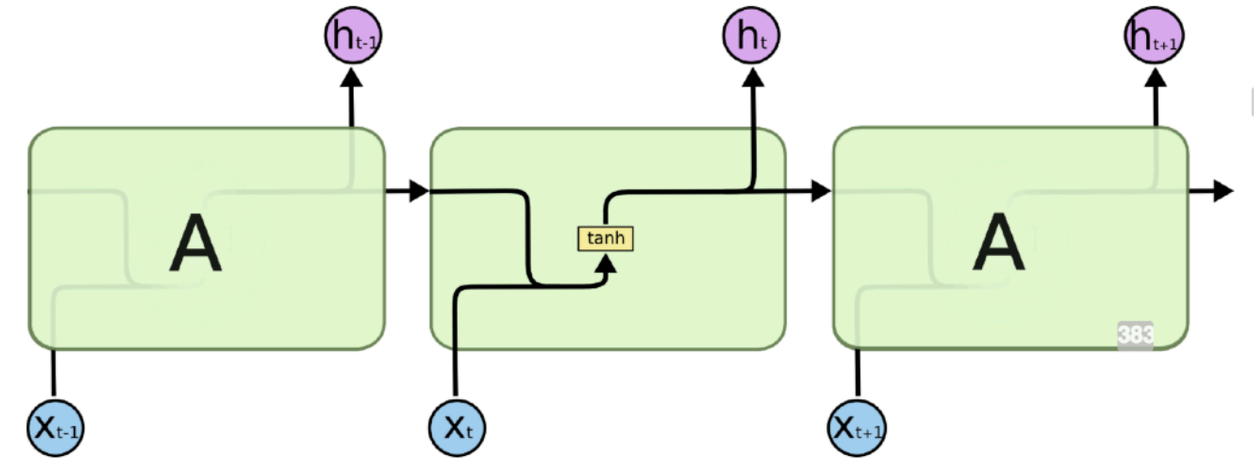
Michael Scheid
CECS 590

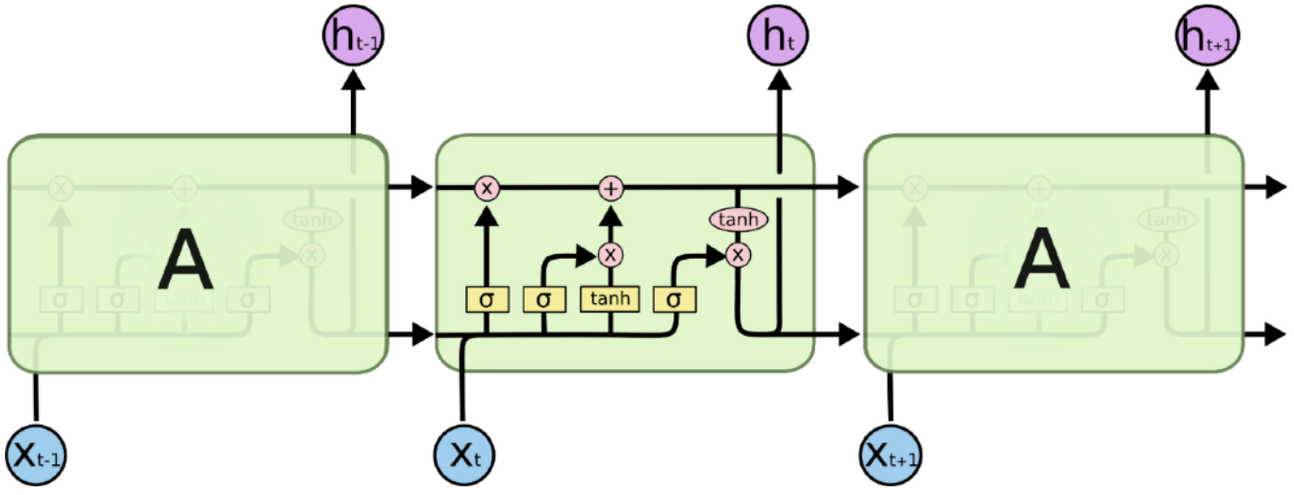
Contents

What is Sentiment Analysis?

- Sentiment is a general feeling, view or opinion.
- Sentiment = { Negative, Neutral, Positive }
- "Sunday is National Ice Cream Day, here's a sweet treat you can make at home http://t.co/IIbvqxGFMe http://t.co/vw7jzMQbCn" - positive
- “@sidrakhan222 I am very seriously injerd in a road Accident 2nd day of Eid with my sweet sis :-(“ -negative

RNN to LSTM





$$f_t = \sigma(W_f \times x_t + U_f \times h_{t-1} + b_f), \quad (3)$$

$$i_t = \sigma(W_i \times x_t + U_i \times h_{t-1} + b_i), \quad (4)$$

$$\tilde{C}_t = tanh(W_C \times x_t + U_C \times h_{t-1} + b_C), \quad (5)$$

$$C_t = i_t \times \tilde{C}_t + f_t \times C_{t-1}, \quad (6)$$

$$o_t = \sigma(W_o \times x_t + U_o \times h_{t-1} + b_o), \quad (7)$$

$$h_t = o_t \times tanh(C_t). \quad (8)$$

- Long-Short Term Memory (LSTM)
  - The memory cell contains 3 Gates: Input, Forget, Output

Natural Language Processing (NLP)

- Challenges
  - How do we represent statements in a form suitable for a NN?
  - Preserving information contained within a statement.
    - Example: “Ice”, “Cream”, “Ice Cream”
- Techniques
  - Vector space representation of words
  - NL Input -> Tokenization -> Numeric Vectors

Natural Language Processing cont.

- word2Vec: Predict the surrounding words in the radius of every word.
- P(context word | center word)

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m \le j \le m \\ j \neq 0}} \log P(w_{t+j} | w_t)$$

- GloVe: Count the overall frequency instead of sliding a window over the document.

$$J(\theta) = \frac{1}{2}\sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

Training LSTM

- Training Dataset: 20,000 Tweets from SemEval-2017_task_4_A
- Tokenization and Preprocessing:

```
It has been an absolute honor to serve our community and represent the Central Valley in
Congress over the past 8 yâ¦ https://t.co/nciU5P1yCK
  tokenizedDocument:

   10 tokens: absolute honor serve community represent central valley congress past yâ

    [100×14 single]
```

Training LSTM cont.

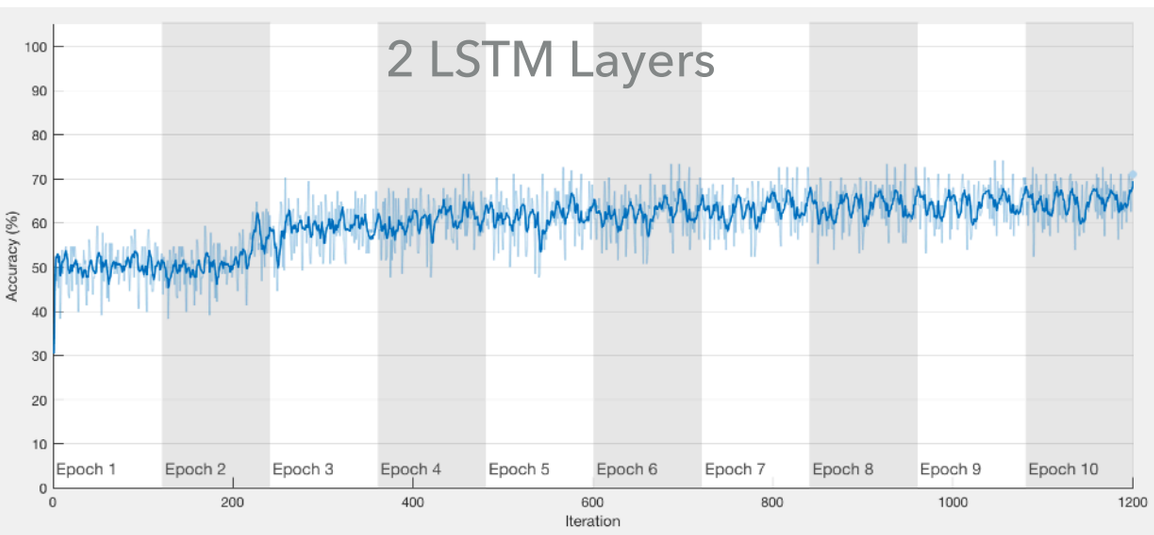
- Varied: LSTM layers, nodes per layer, epochs, and learning rate.
- Test dataset 7124 tweets
- Adding more LSTM layers did not improve the accuracy.

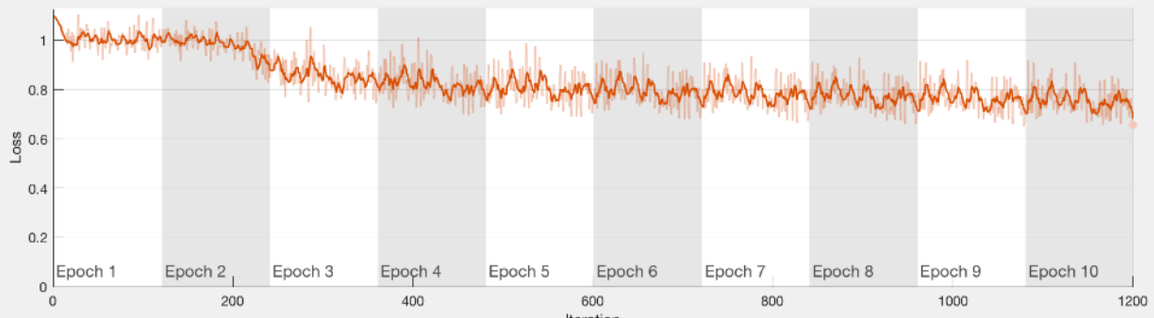
```
output size: dataTesting.rating
         7124                1

accuracy
    0.6465
```

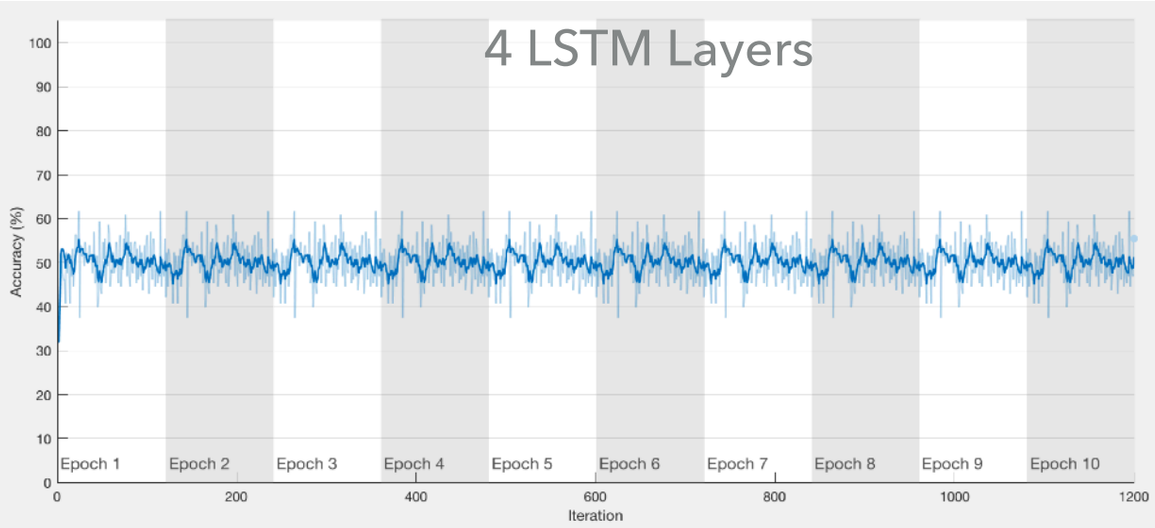
1 LSTM Layer

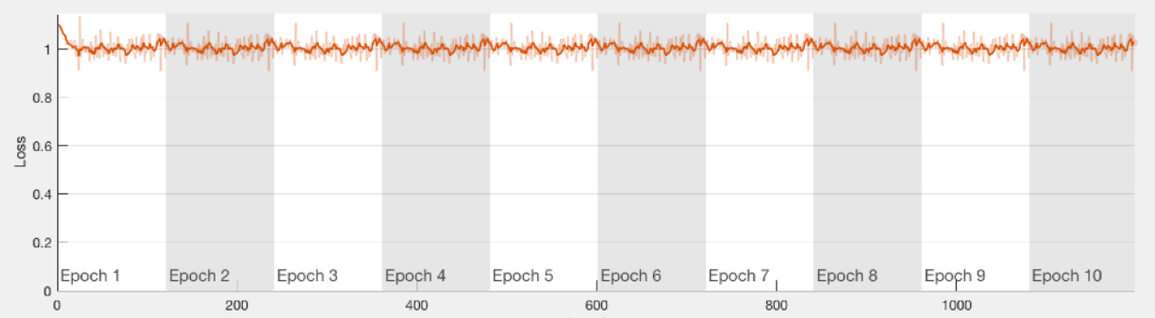
2 LSTM Layers





4 LSTM Layers





Applying the Model

- Data gathered from the California 2018 General Election.
- New York Times and Twitter were used to gather names and usernames.
- Twitter API
  - Register and wait for access tokens
  - Python wrapper: github.com/bear/python-twitter
  - Download last 200 tweets from twitter timeline by username. Saved to files: username.txt

Data collection methods

- For each candidate: Tweet -> Tokenize ->word vector -> pertained LSTM -> Sentiment

```
Reading data from /Users/mjscheid/PycharmProjects/twitter_1/rawTwitterData/kdeleon.txt
Today we officially launched @CalSavers, which makes saving up to leave the workforce as easy↙
as getting paid.

Prâ¦ https://t.co/RObHLvv6o8
  tokenizedDocument:

   12 tokens: today officially launched makes saving up leave workforce easy getting paid prâ

    [100×12 single]

     negative

Reading data from /Users/mjscheid/PycharmProjects/twitter_1/rawTwitterData/joshua_harder.txt
Time for 'the room where it happens'! https://t.co/MfABKeeVaP
  tokenizedDocument:

   3 tokens: time room happens

    [100×11 single]

     neutral

Reading data from /Users/mjscheid/PycharmProjects/twitter_1/rawTwitterData/RepJeffDenham.txt
It has been an absolute honor to serve our community and represent the Central Valley in↙
Congress over the past 8 yâ¦ https://t.co/nciU5P1yCK
  tokenizedDocument:

   10 tokens: absolute honor serve community represent central valley congress past yâ

    [100×14 single]

     positive
```

Results

name	screen_name	party	incumbent	state	outcome	percent	type_of_office	name_of_office
"Dianne Feinstein"	"DianneFeinstein"	0	0	"ca"	1	54.5	1	"us_senator"
"Kevin de Leon"	"kdeleon"	0	0	"ca"	0	45.5	1	"us_senator"
"Josh Harder"	"joshua_harder"	0	0	"ca"	1	51.4	0	"us_representative"
"Jeff Denham"	"RepJeffDenham"	1	0	"ca"	0	48.6	0	"us_representative"
"Gavin Newsom"	"GavinNewsom"	0	0	"ca"	1	61.8	1	"governor"
"John Cox"	"TheRealJohnHCox"	1	0	"ca"	0	38.2	1	"governor"
"Eleni Kounalakis"	"EleniForCA"	0	0	"ca"	1	56.5	1	"lieutenant_governor"
"Ed Hernandez"	"dredhernandez"	0	0	"ca"	0	43.5	1	"lieutenant_governor"

References

- [1]: Olah, C. (2015, August 27). Understanding LSTM Networks. Retrieved from http://colah.github.io/posts/2015-08-Understanding-LSTMs/
- [2]: Miedema, F., Bhulai, S, (2018, August 1). Sentiment Analysis with Long Short-Term Memory Networks. Retrieved from https://beta.vu.nl/nl/Images/werkstuk-miedema_tcm235-895557.pdf
- [3]: Rosenthal, S., Farra, N., Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. Retrieved from http://www.aclweb.org/anthology/S17-2088
- [4]: NA. (nd). MATLAB Documentation: Long Short-Term Memory Networks. Retrieved from https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html
- [5]: NA. (nd). MATLAB Documentation: Classify Text Data Using Deep Learning. Retrieved from https://www.mathworks.com/help/textanalytics/ug/classify-text-data-using-deep-learning.html;jsessionid=198f40443de068e7330b1e76e04c
- [6]: Manning, C., Raghavan, P,. Schütze, H.. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, (2008). Introduction to Information Retrieval. Retrieved from https://nlp.stanford.edu/IR-book/
- [7]: Harper, R. (2011, Spring). CS224n: Natural Language Processing with Deep Learning. Retrieved from http://web.stanford.edu/class/cs224n/
- [8]: Manning, C., Socher, R. (2017, April 3). Lecture 2 Word Vector Representations: word2vec. Retrieved from https://www.youtube.com/watch?v=ERibwqs9p38
- [9]: Socher, R. (2017, April 3). Lecture 3 | GloVe: Global Vectors for Word Representation. Retrieved from https://www.youtube.com/watch?v=ERibwqs9p38