1.6M Tweets Sentiment Analysis

CS542 Course Project

Jinye Cai, Xichao Geng, Tianmu Hu, Yu Sang

Sentiment Analysis

Predict the polarity, negative or positive of a sentence

Donald J. Trump
@realDonaldTrump

MAKE AMERICA GREAT AGAIN!

3:53 PM - 1 Sep 2018

- 98.1% positive

Snighdha Choudhury
@snighdhalicious

I hate the rain

4:34 AM - 23 Apr 2019

- 0.45% positive

Outline

1. Problem definition
2. Analysis
3. Model and Verification
4. Discussion
5. Conclusion

Problem Definition

Kaggle’s Sentiment140 Dataset

	1467810369 id	Mon Apr 06 22:19:45 PD date	NO_QUERY flag	_TheSpecialOne_ user	@switchfoot http://twit text
0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.

Problem Definition

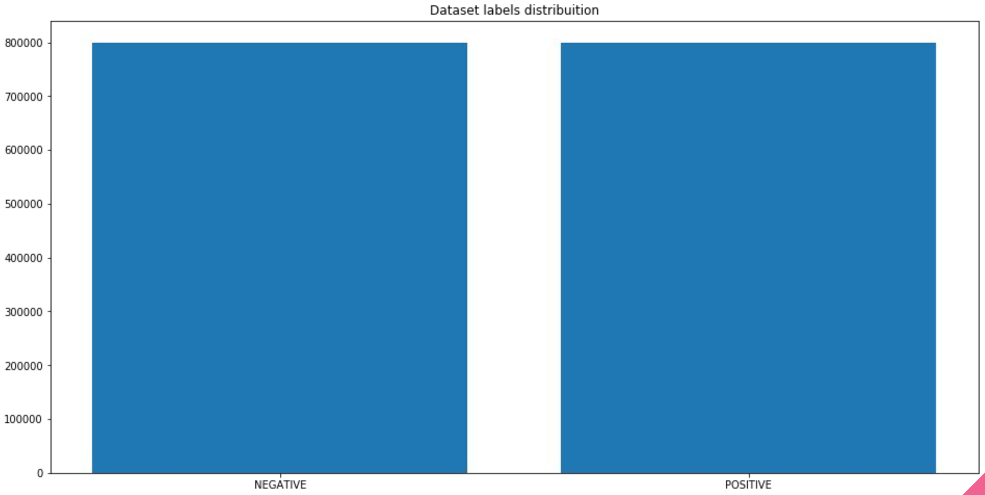
1. Text categorization of two classes: Positive or Negative
2. Natural Language Processing
3. Usage: Opinion mining/Reputation management/Brand monitoring etc.

Analysis

1. Train Set: 128 million tweets;

2. Test Set: 32 million tweets;

3. Output: 2 dimensions, probabilities of N/P;



How do human recognize sentiment?

1. **Words,**

   e.g. Perfect, it is a nice solution!

2. **Terms,**

   e.g. I do not like this person.

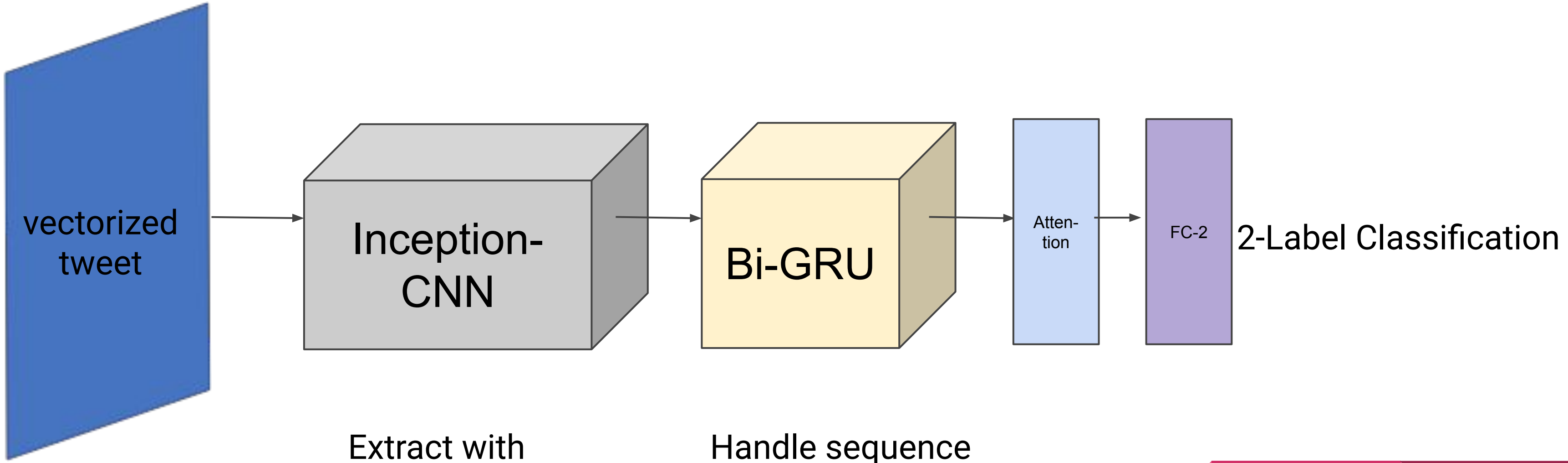
3. **Short sentences,**

   e.g. Although something negative, I feel positive.

How do human recognize sentiment?

1. Capability to extract feature with **different sizes**.
2. Memorizing and understanding the **sequence** of feature words.

Solution



Extract with different kernel sizes

Handle sequence pattern

Preprocess

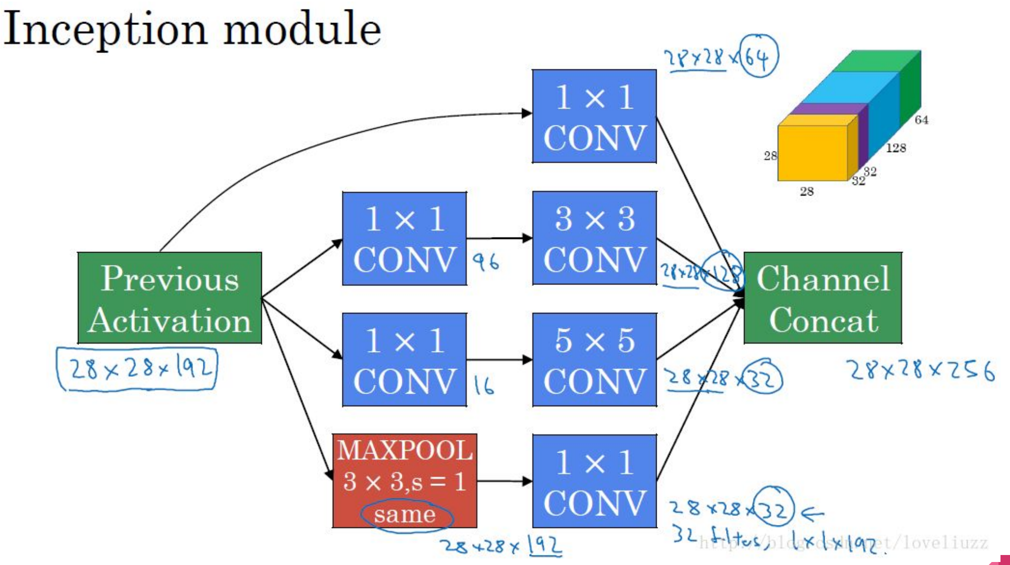
1. Tokenize words in data set

2. Word embedding with GLoVe

Unique Tokens	335,650
Word Vectors	400,000
Average words	11
Average chars	60

Model

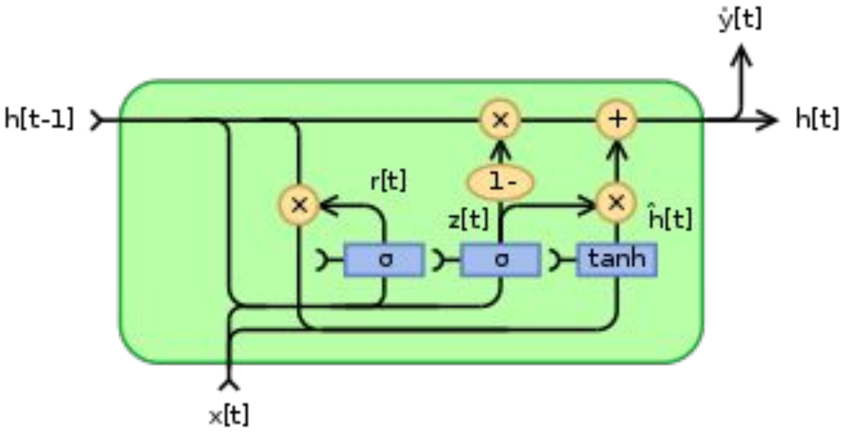
Inception CNN*



* Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

Model

Gated recurrent unit (GRU)*



* Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Model

Attention*

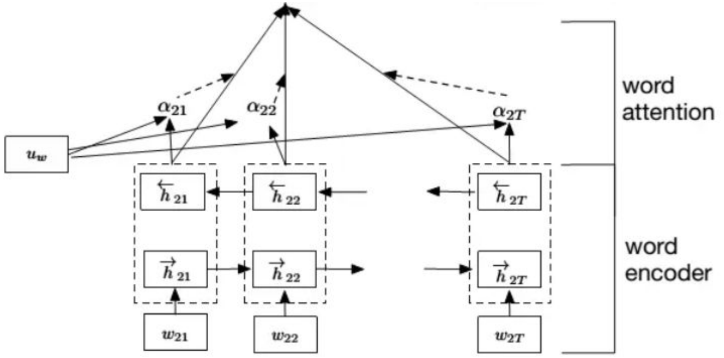


Figure 2: Hierarchical Attention Network.

$$u_{it} = \tanh(W_w h_{it} + b_w)$$
$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}$$
$$s_i = \sum_t \alpha_{it} h_{it}.$$

* Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480-1489).

Model

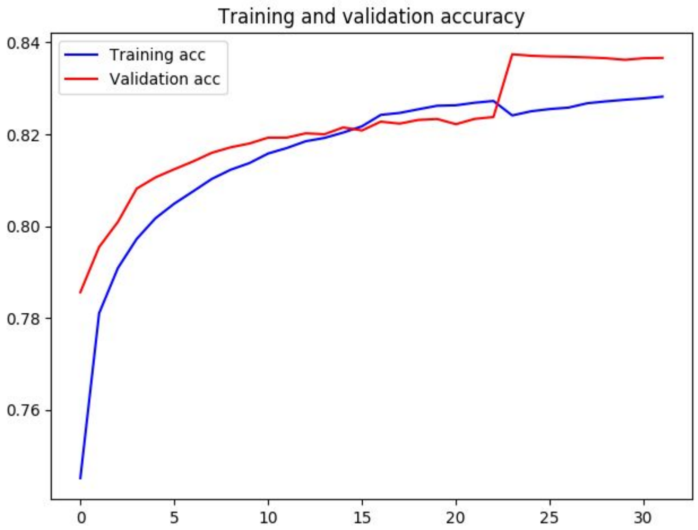
Summary

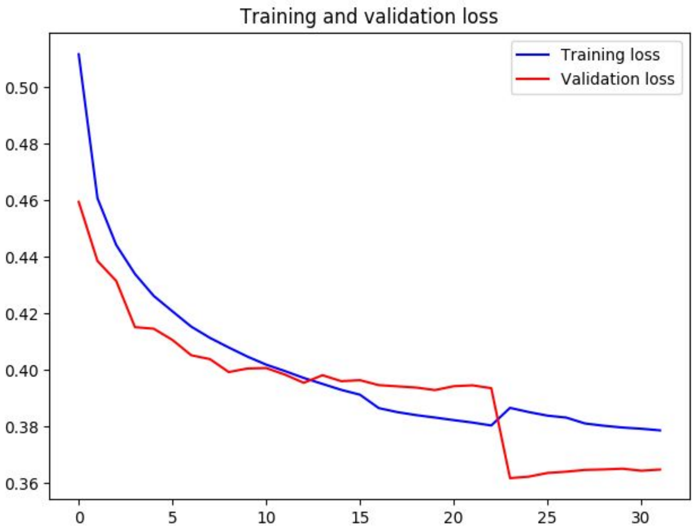
Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 150)	0	
embedding_1 (Embedding)	(None, 150, 200)	8000000	input_1[0][0]
dropout_1 (Dropout)	(None, 150, 200)	0	embedding_1[0][0]
conv1d_1 (Conv1D)	(None, 150, 64)	38464	dropout_1[0][0]
conv1d_3 (Conv1D)	(None, 150, 64)	12864	dropout_1[0][0]
conv1d_2 (Conv1D)	(None, 150, 64)	20544	conv1d_1[0][0]
conv1d_4 (Conv1D)	(None, 150, 64)	12352	conv1d_3[0][0]
conv1d_5 (Conv1D)	(None, 150, 64)	38464	dropout_1[0][0]
conv1d_6 (Conv1D)	(None, 150, 64)	12864	dropout_1[0][0]
mix0 (Concatenate)	(None, 150, 256)	0	conv1d_2[0][0] conv1d_4[0][0] conv1d_5[0][0] conv1d_6[0][0]

max_pooling1d_1 (MaxPooling1D)	(None, 75, 256)	0	mix0[0][0]
batch_normalization_1 (BatchNor	(None, 75, 256)	1024	max_pooling1d_1[0][0]
bidirectional_1 (Bidirectional)	(None, 75, 256)	295680	batch_normalization_1[0][0]
bidirectional_2 (Bidirectional)	(None, 75, 256)	295680	bidirectional_1[0][0]
attention_1 (Attention)	(None, 256)	331	bidirectional_2[0][0]
dropout_2 (Dropout)	(None, 256)	0	attention_1[0][0]
dense_1 (Dense)	(None, 2)	514	dropout_2[0][0]

Total params: 8,728,781
Trainable params: 728,269
Non-trainable params: 8,000,512

Model Verification





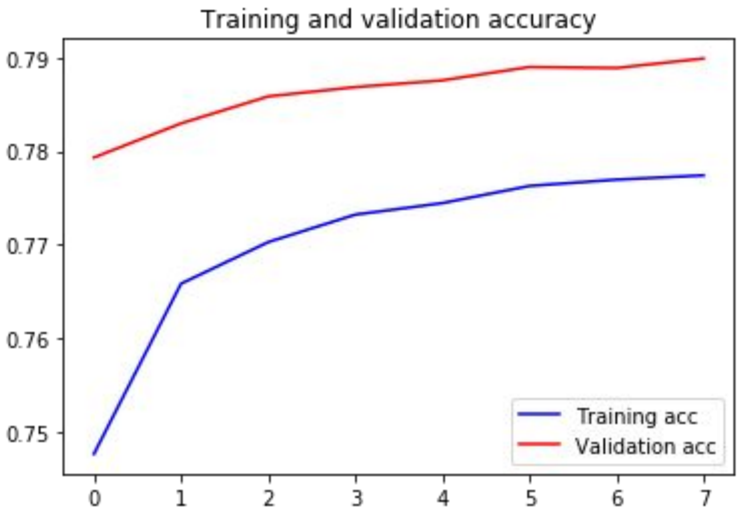
	Training Set	Validation Set
Accuracy (%)	82.82	83.66
Loss	0.3768	0.3648

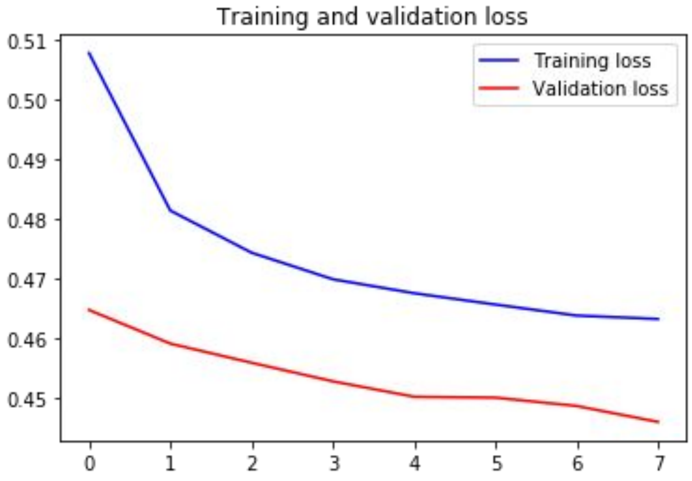
Model Verification

Baseline: 1-layer LSTM

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 300, 300)          87125700
_________________________________________________________________
dropout_1 (Dropout)          (None, 300, 300)          0
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               160400
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 101
=================================================================
Total params: 87,286,201
Trainable params: 160,501
Non-trainable params: 87,125,700
_________________________________________________________________
```

Model Verification





	Our Model	Baseline
Accuracy (%)	83.66	79.12
Loss	0.3648	0.4442

Demo

Extracting insights from Mr. President's tweets

Donald J. Trump
@realDonaldTrump

As I have been saying all along, NO COLLUSION - NO OBSTRUCTION!
https://t.co/BnMB5mvHAM

Apr. 18, 2019, 12:59 p.m.

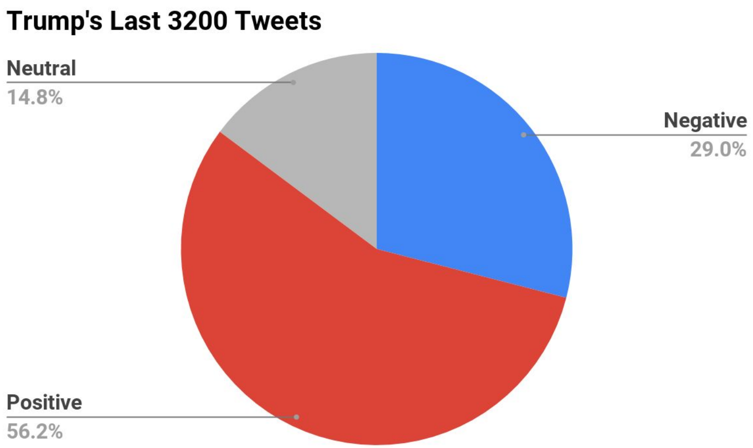
Donald J. Trump
@realDonaldTrump

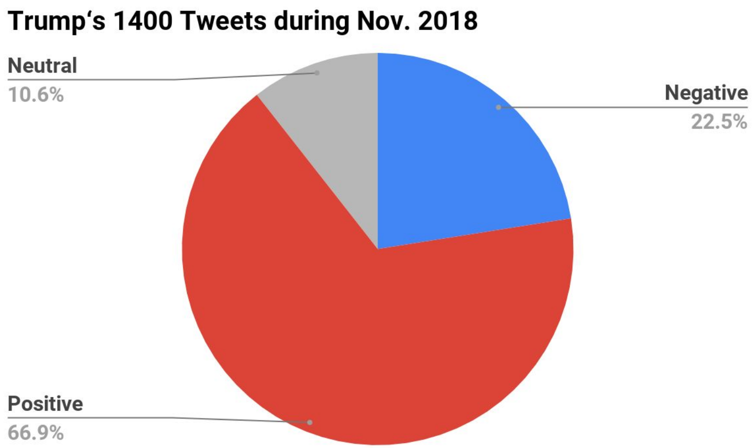
God bless the people of France!

Apr. 15, 2019, 5:58 p.m.

Demo

Extracting insights from Mr. President's tweets





Mr. President fought on his midterm!

Discussions

1. Reasonable to predict three labels:

   negative, positive, and neutral;

2. Can we do better?

   Multimodal, new structures, etc.

Conclusions

1. Represented texts in vectors;
2. Designed the **inception block of CNN** that creatively imitate human understanding of sentence, and **Bidirectional GRU** to handle sequence patterns;
3. Implemented **Hierarchical Attention** to memorize features of long sentence.
4. Improved the acc **~4%**, wrt baseline;
5. Demonstration our model by extracting insights from Mr. President's tweets.

References

[1] Kaggle.com. (2019). Sentiment140 dataset with 1.6 million tweets. [online] Available at: https://www.kaggle.com/kazanova/sentiment140 [Accessed 30 Apr. 2019].

[2] Nltk.org. (2019). Natural Language Toolkit — NLTK 3.4.1 documentation. [online] Available at: https://www.nltk.org/ [Accessed 30 Apr. 2019].

[3] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[4] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[5] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[6] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480-1489).

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[8] Twitter.com. (2019). Twitter. [online] Available at: https://twitter.com/ [Accessed 30 Apr. 2019].

Thank you !