Sentiment Analysis of Social Media Texts

WESST Tutorial

July 19, 2017

Kishaloy Halder
kishaloy@comp.nus.edu.sg

National University of Singapore

Web Information Retrieval / Natural Language Processing Group

Sentiment Analysis

- Is a given piece of text positive, negative, or neutral?
  - The text may be a sentence, a tweet, an SMS message, a customer review, a document, and so on.

Emotion Analysis

- What emotion is being expressed in a given piece of text?
  - Basic emotions: joy, sadness, fear, anger,…
  - Other emotions: guilt, pride, optimism, frustration,…

Slide adapted from [13]

Sentiment Analysis

- Is a given piece of text positive, negative, or neutral?
  - The text may be a sentence, a tweet, an SMS message, a customer review, a document, and so on.

Emotion Analysis

Not in the scope of this tutorial

- What emotion is being expressed in a given piece of text?
  - Basic emotions: joy, sadness, fear, anger,…
  - Other emotions: guilt, pride, optimism, frustration,…

Slide adapted from [13]

Sentiment Analysis: Domains

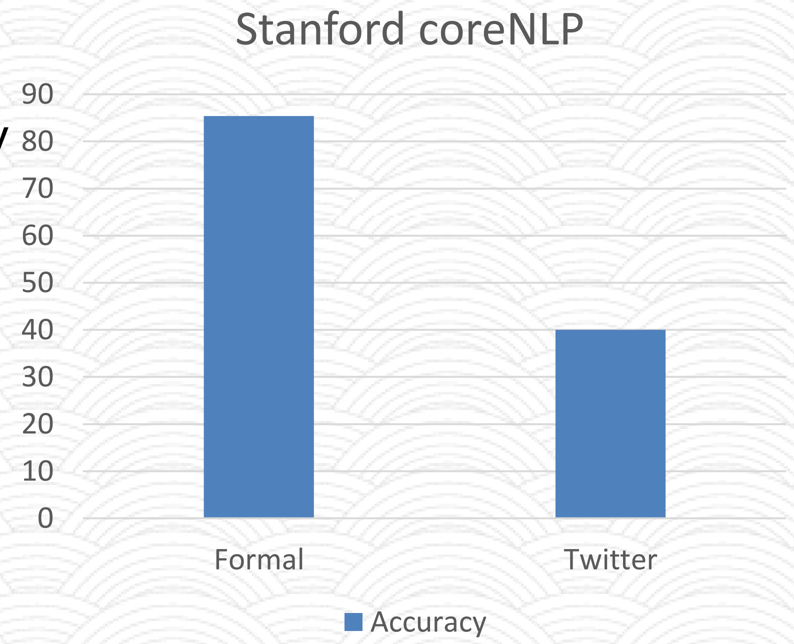
- News
- Legal
- Novels
- E-mails
- SMS
- Customer reviews
- Blog posts
- Tweets
- Facebook posts
- …

Formal text

Short informal text – collectively called Social Media texts

How Social Media text is different?

- Informal
- Short
  - 140 characters for tweets
- Abbreviations and shortenings
- Wide array of topics and large vocabulary
- Spelling mistakes and creative spellings
- Special strings
  - hashtags, emoticons, conjoined words
- High volume
  - 500 million tweets posted every day
- Often come with meta-information
  - date, links, likes, location
- Often express sentiment



Model trained on formal domain doesn't work on Twitter!

Outline

Data Collection

- APIs
- Python - Tweepy

Pre-processing

Models

Data Collection (Twitter)

- Twitter provides public APIs
  - https://dev.twitter.com/rest/public
- Register your app
  - https://apps.twitter.com/
- Obtain authentication key

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	[redacted]
Consumer Secret (API Secret)	[redacted]
Access Level	Read and write (modify app permissions)
Owner	[redacted]
Owner ID	[redacted]

Using Twitter APIs in Python

- Twitter provides REST APIs
- Install tweepy[1]
  - pip install tweepy
- Setup OAuth interface[2]

```python
import tweepy
from tweepy import OAuthHandler

consumer_key = 'YOUR-CONSUMER-KEY'
consumer_secret = 'YOUR-CONSUMER-SECRET'
access_token = 'YOUR-ACCESS-TOKEN'
access_secret = 'YOUR-ACCESS-SECRET'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)
```


1 http://docs.tweepy.org

2 https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/

Using Twitter APIs in Python: Streaming

- Setup stream of tweets based on filters[1]

```python
from tweepy import Stream
from tweepy.streaming import StreamListener

class MyListener(StreamListener):

    def on_data(self, data):
        try:
            with open('python.json', 'a') as f:
                f.write(data)
                return True
        except BaseException as e:
            print(&quot;Error on_data: %s&quot; % str(e))
        return True

    def on_error(self, status):
        print(status)
        return True

twitter_stream = Stream(auth, MyListener())
twitter_stream.filter(track=['#python'])
```

- Makes all the tweets available in json format in *python.json* file
  - Filtered with #python hashtag
  - To use multiple filters append them in the track array

1 https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/

Outline

Data Collection

Pre-processing

- Noisy elements
- #tags
- Normalization

Models

Pre-processing Social Media Text

- Social Media Text is noisy
  - Informal e.g., slangs
  - Misspellings e.g., *covfefe*
  - Elongated words e.g., *can't waittt*
  - Hashtags e.g., *#wesst2017*
  - Emoticons e.g., ☺ ☹
  - Urls
  - Random capitalization e.g., *NOT COOL!*
  - …
- Word coverage with standard dictionaries can be low (50-70%)

Pre-processing: Hashtags

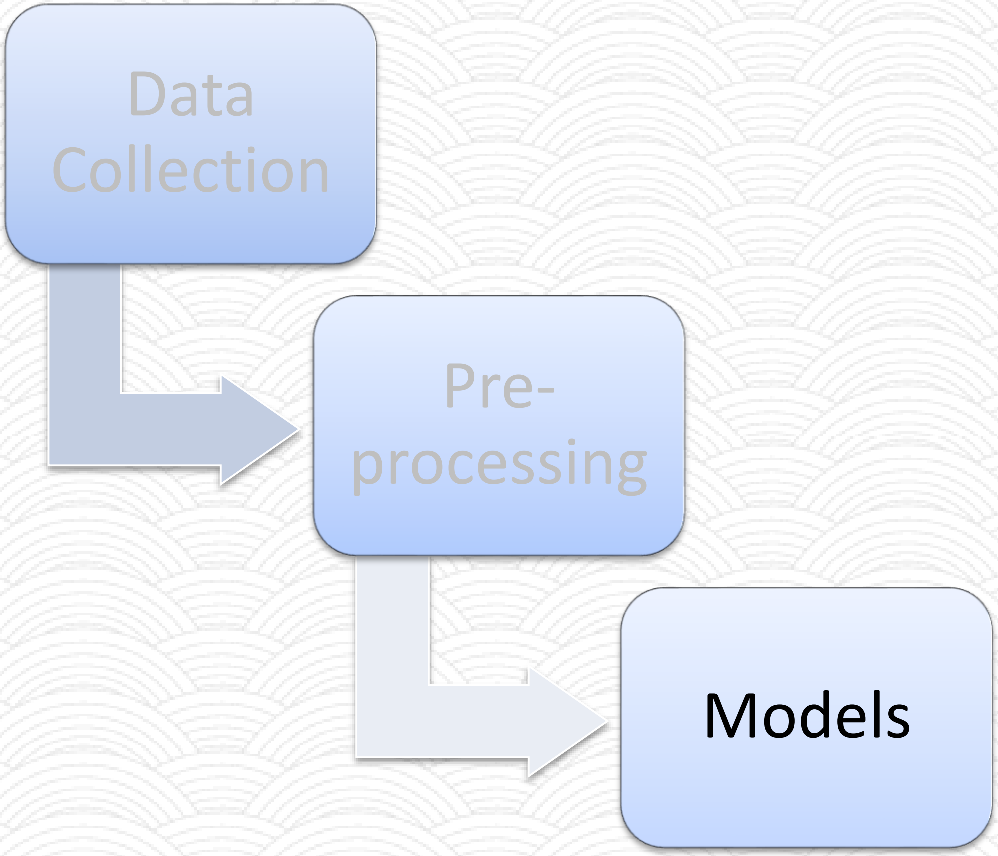
- Hashtagged words are good labels of sentiments and emotions
  - Can’t wait to have my own Google glasses #awesome
  - Some jerk just stole my photo on #tumblr. #grr #anger
- Hashtag Sentiment Lexicon
  - created from a large collection of hashtagged tweets
  - has entries for ~215,000 unigrams
- New hashtags are being generated every minute
- Breaking long hashtags into smaller instances [1]
  - #killthebill → kill the bill

Pre-processing: Normalization

- Remove patterns like ’RT’, ’@user name’, url
- Rectify informal/misspelled words using normalization dictionary [2]
  - “foundation” → “foudation”
  - “forgot” → “forgt”
- Expand abbreviations using slang dictionary[1]
- Removing emoticons
- Handling negation [3]
  - Presence of ‘not’ can negate the target polarity

1 Slang Dictionary - Text Slang & Internet Slang Words. http://www.noslang.com/dictionary/

Outline



- Rule Based
- Machine Learning
- Deep Learning

Rule Based Models

- Lexicalized hand-written rules:
  - Each rule is a pattern that matches words or sequences of words
  - Used in Teragram [4]
- Background data: use blogs, forums, news, and tweets to develop the rules
- Advantages:
  - explicit knowledge representation, so intuitive to develop and maintain.
- Disadvantages:
  - Coverage: often limited coverage → low recall
  - Extensibility: poor for new data/domains

Rule Based Models

- Lexicalized hand-written rules:
  - Each rule is a pattern that matches words or sequences of words
  - Used in Teragram [4]
- Background data: use blogs, forums, news, and tweets to develop the rules
- Advantages:
  - explicit knowledge representation, so intuitive to develop and maintain.
- Disadvantages:
  - Coverage: often limited coverage → low recall
  - Extensibility: poor for new data/domains

Knowledge acquired by applying rules can often be translated as features into statistical approaches

Conventional Machine Learning

- Standard Features

Features	Examples
N-grams	happy, am_very_happy, am_*_happy
Char n-grams	un, unh, unha, unhap
Emoticons	:D, >:(
hashtags	#excited, #NowPlaying
capitalizations	YES, COOL
Part of Speech	N: 5, V: 2, A:1
Negation	Neg:1

- Augmented Features [1]
  - Sentiment of the content of the associated URL, words from hashtags
- Classifier:
  - Linear SVM, Multinomial Naïve Bayes

Deep Learning Based Models

- General Word Embedding: representation of lexical items as points in a real-valued (low-dimensional) vector space.
- It is often computed by compressing a larger matrix to smaller one.

new		1		2	6			9		3	...
old	1	1			2	1		4		2	...
good	1		6	3		1		7	1		...
bad	2	1	4			2			3		...
…											...

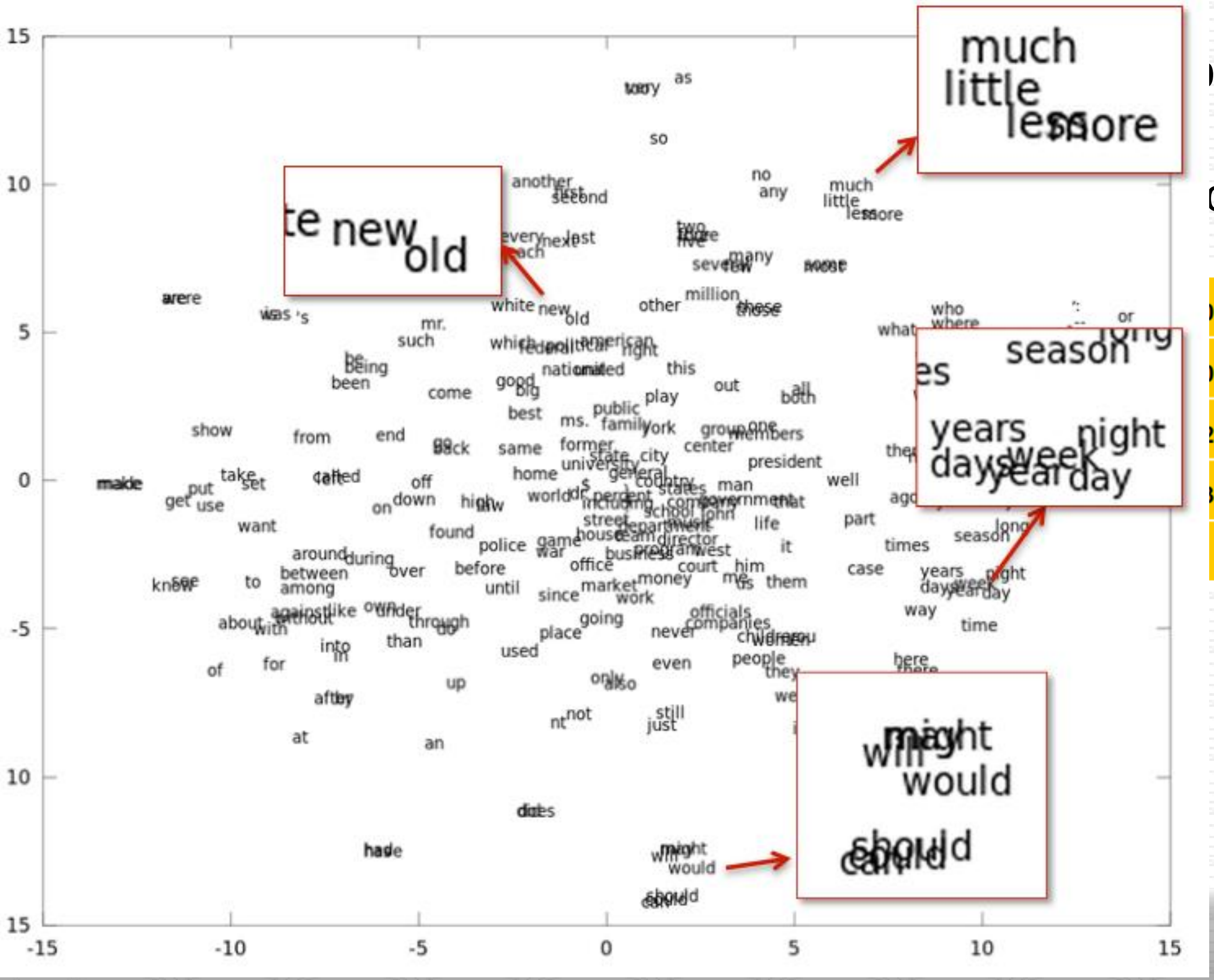
new	-0.03	0.5	0
old	-0.04	0.3	0
good	1.4	0	2.5
bad	1.3	0	3.6
…			

Keep (semantically or syntactically) close items in the original matrix/space to be close in the embedding space.

Deep Learning Based Models

- Gen … oints in a
- It is … one.

n
o
g
b
..



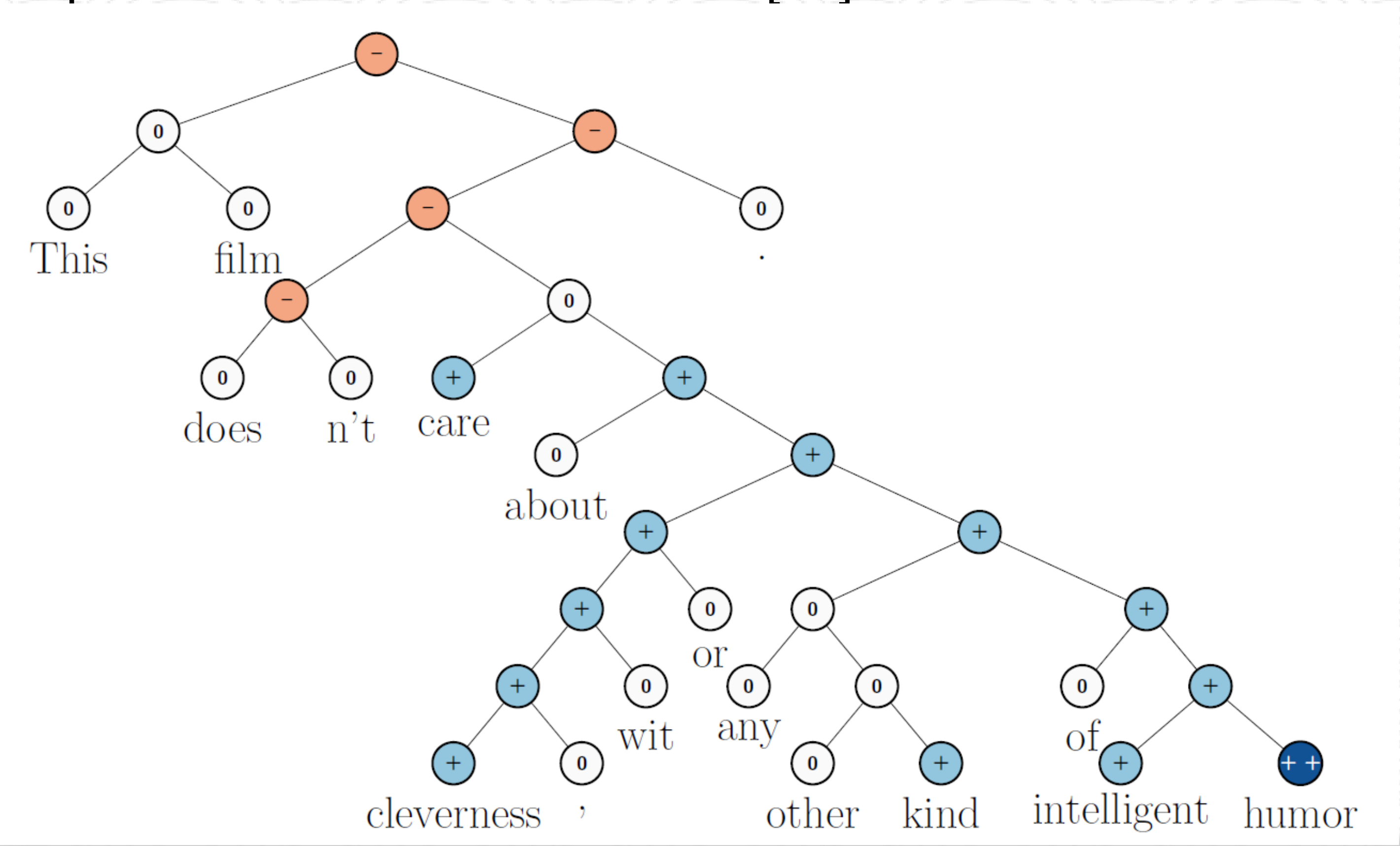
)
)
2.5
3.6

Sentiment Composition

- In addition to obtaining sentiment embedding, composing word sentiment to analyze larger pieces of text (e.g., sentences) is another important problem.
- Most work we have discussed so far is based on bag-of-words or bag-of-ngrams assumption.
- More principled models…
  - Convolution, LSTM in general

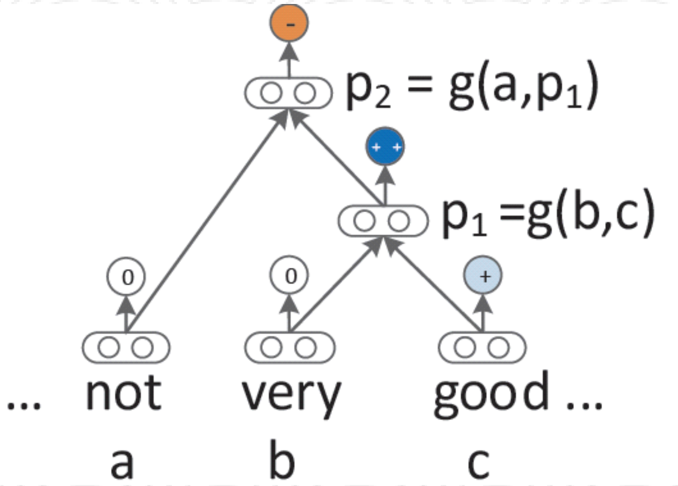
Sentiment Composition: Illustration

- Socher et al. (2013) proposed a recursive neural network to compose sentiment of a sentence [14].



Sentiment Composition: Training

- Tensors are critical in capturing interaction between two words/phrases being composed (e.g., a negator and the phrase it modifies.)



- Standard forward/backward propagation was adapted to learn the weights/parameters

Variations of Sentiment Analysis
&
Emerging Research

Opinion Mining

- What is an Opinion?
- An *opinion* is a quintuple

  $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$
  - o_j is a target object.
  - f_{jk} is a feature of the object o_j.
  - so_{ijkl} is the sentiment value of the opinion of the opinion holder h_i on feature f_{jk} of object o_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed

- Objective: Given an opinionated document,
  - Discover all quintuples $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$,
    - i.e., mine the five corresponding pieces of information in each quintuple, and

Aspect Based Sentiment Analysis

- Determine the polarity (positive, negative, neutral, or conflict) of each aspect category discussed in a given sentence extracted from a restaurant review

  "*To be completely fair, the only redeeming factor was the food, which was above average, but couldn't make up for all the other deficiencies of Teodora*."

- Aspect categories: food (positive), miscellaneous (negative)

Aspect Based Sentiment: Models

- Standard features for Supervised Models
  - ngrams, character ngrams
  - word cluster ngrams
  - sentiment lexicon features
  - Negation
- Task-specific features
  - find terms associated with a given aspect category using Yelp Restaurant Word – Aspect Association Lexicon
  - Add standard features generated just for those terms

food

Service

“The pizza was delicious, but the waiter was rude”

- Unsupervised methods use topic models [5]
  - Seed words to initialize the polarity classes
- Deep Learning based models [9]

Sentiment Analysis in Health Forums

- Emerging direction of research on Consumer Health Forums
  - Users share their clinical experience with others in the community[1]
- Critical for well being of patients with mental issues e.g., depression, Anxiety
- Mental Health Forums are getting popular[2]
  - Provides a platform for emotional support from others in the community
- Sentiment Analysis in Mental Health Forums
  - Can detect early symptoms of depression[7]
  - Track a patients emotional state over time[6]
  - Can help us prevent life-threatening situations
- Standard Features for Depression Detection
  - Increased negativity in user posts
  - Withdrawal from Social interactions

1 www.patientslikeme.com, www.healthboards.com
2 www.dailystrength.org

Summary

- Social Media Text varies widely from formal domain
  - Text normalization, cleaning is necessary for traditional lexical dictionary to work
- Discussed ways to collect Social Media Data (e.g., twitter)
- Discussed features for state-of-the-art models
  - Conventional Machine Learning, Deep Learning
- Variations of Sentiment Analysis
  - Opinion Mining, Aspect Based Sentiment Analysis
- Implication of sentiment analysis on Health Forums and emerging research directions

Thanks for listening!

Questions?

Email: kishaloy@comp.nus.edu.sg

References

1. Lahari Poddar, Kishaloy Halder, and Xianyan Jia. "Sentiment Analysis for Twitter: Going Beyond Tweet Text." *arXiv preprint arXiv:1611.09441* (2016).
2. Bo Han, Paul Cook, and Timothy Baldwin. “Automatically constructing a normalisation dictionary for microblogs.” In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 421–432, Jeju Island, Korea, July 2012. Association for Computational Linguistics.
3. Lapponi, Emanuele, Jonathon Read, and Lilja Ovrelid. "Representing and resolving negation for sentiment analysis." *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*. IEEE, 2012.
4. Reckman, Hilke, et al. “Teragram: Rule-based detection of sentiment phrases using sas sentiment analysis." (2013).
5. Lahari Poddar, Wynne Hsu, Mong Li Lee. “Author aware Aspect Topic Sentiment Model to Retrieve Supporting Opinions from Reviews.” EMNLP 2017.
6. Kishaloy Halder, Lahari Poddar, Min-Yen Kan, “Modeling Temporal Progression of Emotional Status in Mental Health Forum: A Recurrent Neural Net Approach”, WASSA, EMNLP 2017.
7. Nicolas Rey-Villamizar, Prasha Shrestha, Farig Sadeque, Steven Bethard, Ted Pedersen, Arjun Mukherjee, and Thamar Solorio. 2016. Analysis of anxious word usage on online health forums. EMNLP 2016.
8. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. AAAI.

References

9. Lakkaraju, Himabindu, Richard Socher, and Chris Manning. "Aspect specific sentiment analysis using hierarchical deep learning." *NIPS Workshop on Deep Learning and Representation Learning*. 2014.
10. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM'11, pp. 30--38, Portland, Oregon.
11. Becker, L., Erhart, G., Skiba, D., & Matula, V. (2013). Avaya: Sentiment analysis on Twitter with selftraining and polarity lexicon expansion. In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pp. 333{340, Atlanta, Georgia, USA.
12. Brody, S., & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 562--570, Stroudsburg, PA, USA.
13. Saif M. Mohammad and Xiaodan Zhu. "Sentiment Analysis of Social Media Texts!", A tutorial presented at the 2014 Conference on Empirical Methods on Natural Language Processing, October 2014, Doha, Qatar.
14. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '13. Association for Computational Linguistics.