Sentiment Analysis of Social Media Texts

WESST Tutorial
July 19, 2017
Kishaloy Halder
kishaloy@comp.nus.edu.sg
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,…
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
- …
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](https://dev.twitter.com/rest/public)
- Register your app
  - [https://apps.twitter.com/](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.*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Key (API Key)</td>
<td>[Redacted]</td>
</tr>
<tr>
<td>Consumer Secret (API Secret)</td>
<td>[Redacted]</td>
</tr>
<tr>
<td>Access Level</td>
<td>Read and write (modify app permissions)</td>
</tr>
<tr>
<td>Owner</td>
<td>[Redacted]</td>
</tr>
<tr>
<td>Owner ID</td>
<td>[Redacted]</td>
</tr>
</tbody>
</table>
Using Twitter APIs in Python

- Twitter provides REST APIs
- Install tweepy
  - pip install tweepy
- Setup OAuth interface

```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

```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("Error on_data: {}".format(e))
            return True

    def on_error(self, 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 dictionary¹

- Removing emoticons

- Handling negation [3]
  - Presence of ‘not’ can negate the target polarity
Outline

Data Collection

Pre-processing

Models

- 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 \(\rightarrow\) 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

<table>
<thead>
<tr>
<th>Features</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-grams</td>
<td>happy, am_very_happy, am_*_happy</td>
</tr>
<tr>
<td>Char n-grams</td>
<td>un, unh, unha, unhap</td>
</tr>
<tr>
<td>Emoticons</td>
<td>:D, &gt;:(</td>
</tr>
<tr>
<td>hashtags</td>
<td>#excited, #NowPlaying</td>
</tr>
<tr>
<td>capitalizations</td>
<td>YES, COOL</td>
</tr>
<tr>
<td>Part of Speech</td>
<td>N: 5, V: 2, A:1</td>
</tr>
<tr>
<td>Negation</td>
<td>Neg:1</td>
</tr>
</tbody>
</table>

- 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.

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

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

\[ p_2 = g(a, p_1) \]

\[ p_1 = g(b, c) \]

- 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

“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


8. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. AAAI.
References


