Sentiment Analysis

Improving results with complex generated features

Lucas Silva & Dr. Pearl
Agenda

The field: Sentiment Analysis

Silva & Rocha: research

Evolution! Improving sentiment analysis

Methodology

Expected results
The field: Sentiment analysis
"Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text"

(Liu, Bing)
Or...
I loved the new Spider Man movie, but I hated the poor performance of Andrew Garfield as Peter Parker.
I loved the new Spider Man movie, but I hated the poor performance of Andrew Garfield as Peter Parker.
I loved the new Spider Man movie, but I hated the poor performance of Andrew Garfield as Peter Parker.
Organizing the field

SA

ML

LB
Organizing the field

- SA
- ML
- LB
Machine Learning

Training Docs → Feature Vectors → ML Alg → Predictive Model → Positive, Negative
Machine Learning

Training Docs → Feature Vectors → ML Alg → Predictive Model → Positive, Negative

New Doc → Predictive Model → Positive, Negative
Machine Learning

- Training Docs → Feature Vectors → ML Alg → Predictive Model → Positive, Negative
ngrams

PoS tag

Prior polarity

Polarity shifters

Specific sets of words

[‘love’, ‘hate’]

[‘“I love’, ‘love you’]

[‘I love you’]
ngrams

PoS tag → ['I_PRP', 'love_VBP', 'you_PRP']

Prior polarity

Polarity shifters

Specific sets of words
ngrams

PoS tag

Prior polarity → ['love_pos']

Polarity shifters

Specific sets of words
ngrams
PoS tag
Prior polarity
Polarity shifters
Specific sets of words
  intensifiers
diminishers
negation
ngrams
PoS tag
Prior polarity
Polarity shifters
Specific sets of words
  adjectives
  adverbs
Main algorithms are...

Naive Bayes
Assumes that all the features are independent and equally important

Support Vector Machines
Has the ability to define the most important features
Organizing the field
Lexicon Based

<table>
<thead>
<tr>
<th></th>
<th>word_token_01</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.75</td>
<td>...</td>
</tr>
<tr>
<td>Negative</td>
<td>0.25</td>
<td>...</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.05</td>
<td>...</td>
</tr>
</tbody>
</table>

Document -> Lexicon -> Positive
Document -> Lexicon -> Negative
Lexicon Based

<table>
<thead>
<tr>
<th>Word Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.75</td>
</tr>
<tr>
<td>Negative</td>
<td>0.25</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Document → Lexicon → Positive, Negative, Neutral
Main lexicons are...

General Inquirer
WordNet
SentiWordNet
Valence Shifters

Modifiers: intensifiers and diminishers

This is really good!
Valence Shifters

Modifiers: intensifiers and diminishers

This is really good!

+ 0.78
Valence Shifters

Modifiers: intensifiers and diminishers

This is \textit{really} \textbf{good}!

\[ X 1.5 \]

\[ + 1.17 \]
Valence Shifters

Negation: switch

This is not good!
Valence Shifters

Negation: switch

This is not good!

+ 0.78
Valence Shifters

Negation: switch

This is not good!

-0.78
Valence Shifters

Negation: shift

This is not good!

- $0.156$
Valence Shifters

window of words x window of structure

This is not really good!
Valence Shifters

Window: words \textbf{X} structure

This is not really good!
Valence Shifters

Window: words X structure

```
This is not really good
```

Diagram:
```
S
  NP
    DT
    This
  VP
    RB
    is
    RB
    not
    ADJP
      RB
      really
      JJ
      good
```
Valence Shifters

Window: words X structure

- **DT**: This
- **NP**: is
- **S**: not
- **VP**: really
- **ADJP**: good
- **RB**: good
Silva & Rocha: research
Dataset

Topic: entrepreneurship/startups

Automatically acquired using crawler

180 articles with 4563 sentences
Goal

Show that it's possible to create accurate SA classifiers using data provided by untrained annotators

Pipelines

  Initial data classification
  Annotation and final classification
Methodology

Pipelines

Initial data classification
Annotation and final classification

Classifier: naive Bayes

Features: unigrams

Approach: Generated different training data and compare results of the classifiers trained with them
Evolution!
Improving sentiment analysis
Goal

To study how the process to generate word prior polarity feature impacts the accuracy of the sentiment classification task
Methodology

1. Framework to generate complex features
2. Sentiment classification task
3. Experiment
Methodology

1. Framework to define polarity of words
2. Sentiment classification task
3. Experiment
Docs

Pre processing

Define polarity score of words

Apply SentiWordNet

Apply valence shifter’s rules

Valence shifters

Define final words polarity
Docs

Preprocessing

Define polarity score of words

Apply SentiWordNet

Apply valence shifter's rules

Valence shifters

Define final words polarity
<table>
<thead>
<tr>
<th>This is <strong>good</strong></th>
<th>.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is <strong>really good</strong></td>
<td>.1 + .75</td>
</tr>
<tr>
<td>This is <strong>not good</strong></td>
<td>1 -</td>
</tr>
<tr>
<td>This is <strong>not really good</strong></td>
<td>1 −</td>
</tr>
<tr>
<td>This is <strong>bad</strong></td>
<td>-.625</td>
</tr>
</tbody>
</table>
Methodology

1. Framework to define polarity of words
2. Sentiment classification task
3. Experiment
## Classification

### Features

<table>
<thead>
<tr>
<th></th>
<th>this</th>
<th>is</th>
<th>good</th>
<th>really</th>
<th>not</th>
<th>bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>.75</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>.85</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>-.25</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>-.15</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-.625</td>
</tr>
</tbody>
</table>
Classification

Classifier: Support Vector Machines

10-fold cross validation to define best params
Methodology

1. Framework to define polarity of words
2. Sentiment classification task
3. Experiment
Experiment

Compare:

Baseline (Silva&Rocha)

Support Vector Machines with different windows for shifters

Negation shift

Negation switch

Precision, recall and F-score to define accuracy
Expected results
We believe that...

Negation shift $\gg$ Negation switch

Window of structure $\gg$ Window of words

SVM $\gg$ naive Bayes
Thank you!