Detecting deceptive reviews using Argument Mining

Oana Cocarascu
Imperial College London
Truthful or deceptive?

⭐⭐⭐⭐⭐
Their service was amazing, and we absolutely loved the beautiful indoor pool. I would recommend staying here.

⭐⭐⭐⭐⭐
The staff was super friendly and helpful and the location was fantastic. Highly recommended!

⭐
Pathetic and rude. Hotel better find some better employees for their guests to truly enjoy their stay.
Approach

• mine Argumentation Frameworks (AFs)

• argumentative features for classifiers from dialectical strength
Related work

• opinion spam (Ott et al. 2011, Shojaee et al. 2013, Fusilier et al. 2015)

• opinion spammers (Lim et al. 2010, Mukherjee et al. 2012)

• Argument Mining (Palau & Moens 2011, Lippi & Torroni 2016)
  • argumentative sentence
  • argument components
  • relations between arguments
Overview

NLP

Reviews (R) per item

Extract arguments (AR) from R

Determine polarity of args in AR

sentiment analysis

Relation-based AM & sentiment analysis

Determine sup/att relations in AR ∪ {G,B}

Argumentation Framework (AF)

for each review r

Remove args (related to r) from AR

AF_r with arguments from r removed

Review strength from AF & AF_r

Impact r over item
Argumentation Frameworks (AFs)

- Abstract Argumentation Framework (AAF)
- Abstract Bipolar Argumentation Framework (BAF)
Example

\( r_1 \): ‘It had nice rooms but terrible food.’

\( r_2 \): ‘Their service was amazing and we absolutely loved the room. They do not offer free Wi-Fi so they expect you to pay to get Wi-Fi...’
Extracting arguments

$r_1$: ‘It had nice rooms but terrible food.’

$a_{11}$: It had nice rooms
$a_{12}$: (It had) terrible food

$r_2$: ‘Their service was amazing and we absolutely loved the room. They do not offer free Wi-Fi so they expect you to pay to get Wi-Fi...’

$a_{21}$: service was amazing
$a_{22}$: absolutely loved the room
$a_{23}$: they do not offer free Wi-Fi so they expect you to pay to get Wi-Fi
Determine argument polarity

\( a_{11} \): It had nice rooms (+)
\( a_{12} \): (It had) terrible food (-)

\( a_{21} \): service was amazing (+)
\( a_{22} \): absolutely loved the room (+)
\( a_{23} \): they do not offer free Wi-Fi so they expect you to pay to get Wi-Fi (-)
Determine support/attack relations

sentiment analysis -> AAF

relation-based Argument Mining
+ sentiment analysis -> BAF
Mining AFs for detecting deception

- topic-independent AAF
  - 2 special arguments: $G$ and $B$

- (noun-level) topic-dependent BAF
  - 1 special argument: $G$
  - 1 special argument per topic: $G_t$
Topic-independent AAF

- arguments extracted from reviews
- 2 special arguments: $G$ and $B$
- attack relation determined by argument polarity
**AAF from example**

\[ a_{11}: \text{It had nice rooms (+)} \]
\[ a_{12}: \text{(It had) terrible food (-)} \]
\[ a_{21}: \text{service was amazing (+)} \]
\[ a_{22}: \text{absolutely loved the room (+)} \]
\[ a_{23}: \text{they do not offer free Wi-Fi so they expect you to pay to get Wi-Fi (-)} \]
Topic-dependent BAF

- identify topics (and related arguments)
- arguments extracted from reviews
- 1 special argument: $G$
- 1 special argument per topic: $G_t$
- relations determined using relation-based AM
Topic-dependent BAF

\[ a_{11} : \text{It had nice } \textbf{rooms} \ (+) \]
\[ a_{12} : \text{(It had) terrible } \textbf{food} \ (-) \]
\[ a_{21} : \text{service was amazing} \ (+) \]
\[ a_{22} : \text{absolutely loved the } \textbf{room} \ (+) \]
\[ a_{23} : \text{they do not offer free } \textbf{Wi-Fi} \text{ so they expect you to pay to get } \textbf{Wi-Fi} \ (-) \]

Topics

- room
- food
- service
- Wi-Fi
### Topic-dependent BAF - Determining relations

<table>
<thead>
<tr>
<th><strong>Feature</strong></th>
<th><strong>Detail</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>number of words</td>
<td>for each sentence</td>
</tr>
<tr>
<td>avg word length</td>
<td>for each sentence</td>
</tr>
<tr>
<td>sentiment polarity</td>
<td>for each sentence</td>
</tr>
<tr>
<td>Jaccard similarity</td>
<td>size of the intersection of words in sentences compared to the size of union of words in sentences</td>
</tr>
<tr>
<td>Levenshtein distance</td>
<td>count of replace and delete operations required to transform one sentence into the other</td>
</tr>
<tr>
<td>word order</td>
<td>normalized difference of word order between the sentences</td>
</tr>
<tr>
<td>Malik</td>
<td>sum of maximum word similarity scores of words in same POS class normalized by sum of sentence’s lengths (path and lch)</td>
</tr>
<tr>
<td>combined semantic and syntactic</td>
<td>linear combination of semantic vector similarity and word order similarity (path and lch)</td>
</tr>
</tbody>
</table>
BAF from example

\( a_{11} \): It had nice rooms (+)
\( a_{12} \): (It had) terrible food (-)
\( a_{21} \): service was amazing (+)
\( a_{22} \): absolutely loved the room (+)
\( a_{23} \): they do not offer free Wi-Fi so they expect you to pay to get Wi-Fi (-)
Calculating argument strength

*base score* of arguments

*F - aggregating* the argument strength

*C - combining* base score with the aggregated score of attackers/supporters
Calculating argument strength

**Base score** of arguments: 0.5

\[
F = \begin{cases} 
0 & n = 0 \\
1 - \log \prod_{i=1}^{n} (|1 - v_i|) & n > 0
\end{cases}
\]

\[
C = \begin{cases} 
v_0 & if \ v_a = v_s \\
v_0 - \log(v_0 \times |v_s - v_a|) & if \ v_a > v_s \\
v_0 + \log((1 - v_0) \times |v_s - v_a|) & if \ v_a < v_s
\end{cases}
\]
Argumentative features

impact of review $r$ :

$|\text{strength given } R - \text{strength given } R\setminus\{r\}|$
Argumentative features in AAF

$r_1$ - argumentative features

AF given $R$

AF given $R\backslash\{r_1\}$
Argumentative features in BAF

$r_1$ - argumentative features

AF given $R$  

AF given $R \setminus \{r_1\}$
Deceptive reviews - standard NLP features

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personalization</td>
<td>nr self references</td>
</tr>
<tr>
<td></td>
<td>nr 2nd person pronouns</td>
</tr>
<tr>
<td></td>
<td>nr other references</td>
</tr>
<tr>
<td></td>
<td>nr group pronouns</td>
</tr>
<tr>
<td>Quantity</td>
<td>nr sentences</td>
</tr>
<tr>
<td></td>
<td>nr words</td>
</tr>
<tr>
<td></td>
<td>nr nouns</td>
</tr>
<tr>
<td></td>
<td>nr verbs</td>
</tr>
<tr>
<td>Complexity</td>
<td>avg sentence length</td>
</tr>
<tr>
<td></td>
<td>avg word length</td>
</tr>
<tr>
<td>Diversity</td>
<td>lexical</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>nr modal verbs</td>
</tr>
<tr>
<td></td>
<td>nr modifiers</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Random Forests</th>
<th>Hotel</th>
<th>Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>76.25%</td>
<td>69%</td>
</tr>
<tr>
<td>AAF</td>
<td>77.75%</td>
<td>71.25%</td>
</tr>
<tr>
<td>BAF</td>
<td>79.81%</td>
<td>73%</td>
</tr>
</tbody>
</table>
Future work

• other AM techniques
• semi-supervised approach
• compute argument strength
Thank you!