# 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.**

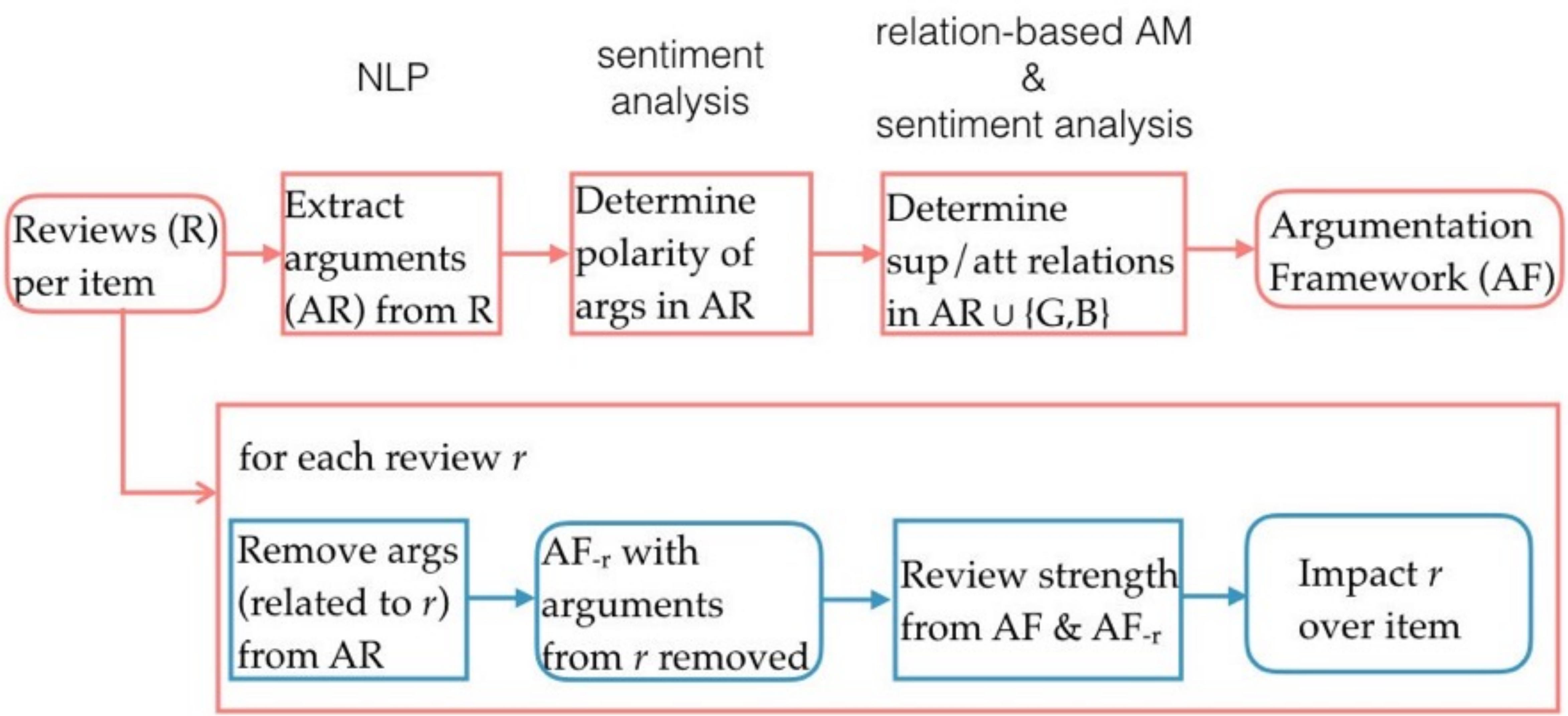
# Approach

- 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
sentiment analysis
relation-based AM & sentiment analysis
Reviews (R) per item
Extract arguments (AR) from R
Determine polarity of args in AR
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.'*

***$r_2$:*** *'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...'*

***$a_{11}$:*** *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 argument polarity

***$a_{11}$:*** *It had nice rooms (+)*

***$a_{12}$:*** *(It had) terrible food (-)*

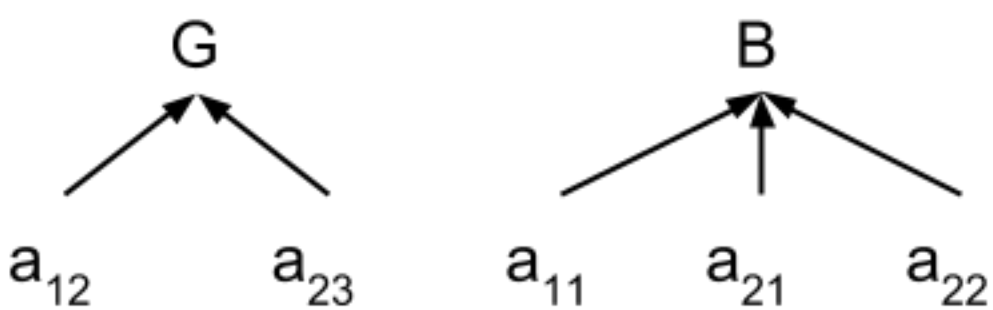
***$a_{21}$:*** *service was amazing (+)*

***$a_{22}$:*** *absolutely loved the room (+)*

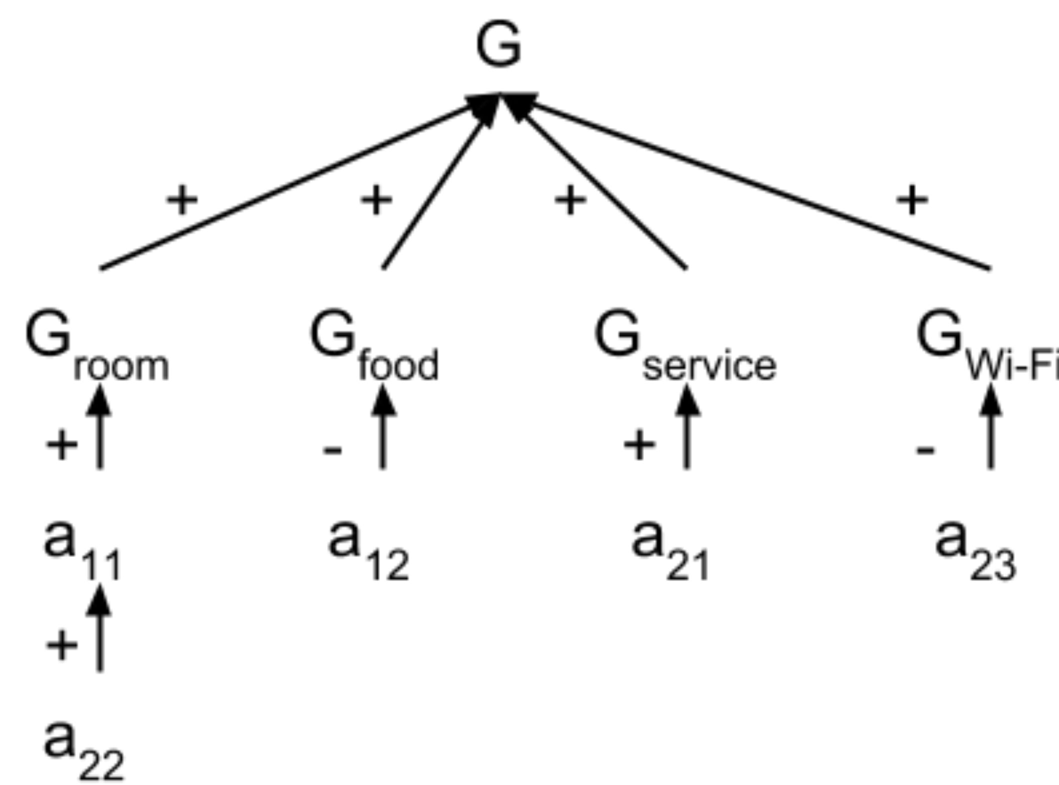
***$a_{23}$:*** *they do not offer free Wi-Fi so they expect you to pay to get Wi-Fi (-)*

# 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: $\boldsymbol{G_t}$

# Topic-independent AAF

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

# AAF from example

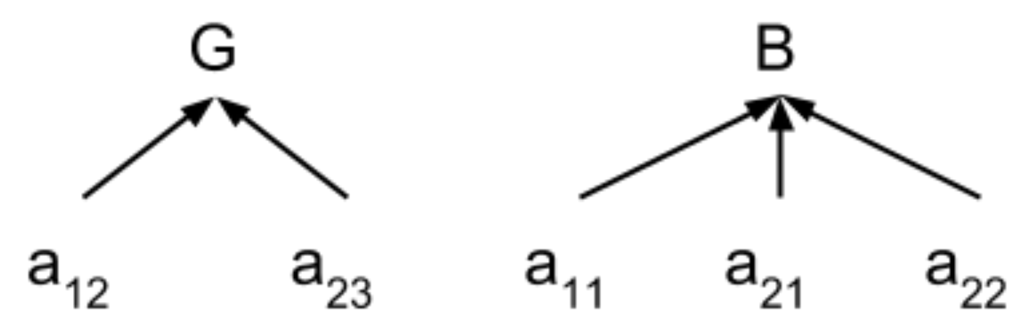
***$a_{11}$***: *It had nice rooms (+)*

***$a_{12}$***: *(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 (-)*



# Topic-dependent BAF

- identify topics (and related arguments)
- arguments extracted from reviews
- 1 special argument: $\boldsymbol{G}$
- 1 special argument per topic: $\boldsymbol{G_t}$
- relations determined using relation-based AM

# Topic-dependent BAF

***$a_{11}$***: *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** (-)*

**Topics**

- room
- food
- service
- Wi-Fi

# Topic-dependent BAF - Determining relations

| **Feature** | **Detail** |
|---|---|
| number of words | for each sentence |
| avg word length | for each sentence |
| sentiment polarity | for each sentence |
| Jaccard similarity | size of the intersection of words in sentences compared to the size of union of words in sentences |
| Levenshtein distance | count of replace and delete operations required to transform one sentence into the other |
| word order | normalized difference of word order between the sentences |
| Malik | sum of maximum word similarity scores of words in same POS class normalized by sum of sentence's lengths (path and lch) |
| combined semantic and syntactic | linear combination of semantic vector similarity and word order similarity (path and lch) |

# BAF from example

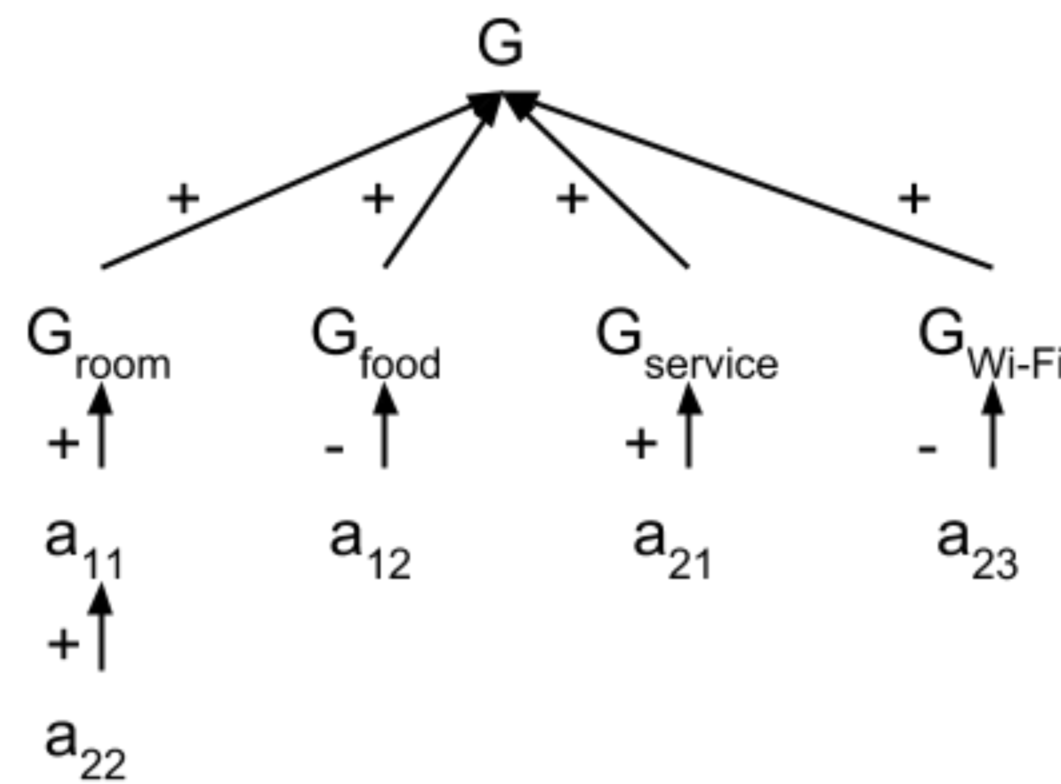
***$a_{11}$***: *It had nice* ***rooms*** *(+)*

***$a_{12}$***: *(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*** **=** $$\begin{cases} v_0 & if\ v_a = v_s \\ v_0 - log(v_0 * |v_s - v_a|) & if\ v_a > v_s \\ v_0 + log((1 - v_0) * |v_s - v_a|) & if\ v_a < v_s \end{cases}$$

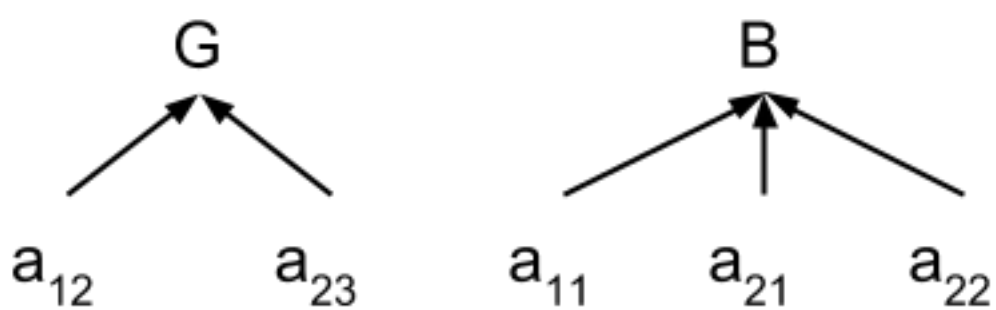
# Argumentative features

impact of review ***r*** :

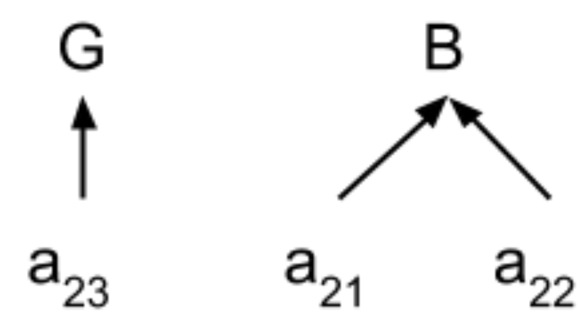
|strength given **R** - strength given **R\{r}**|

# Argumentative features in AAF

***$r_1$*** *- argumentative features*



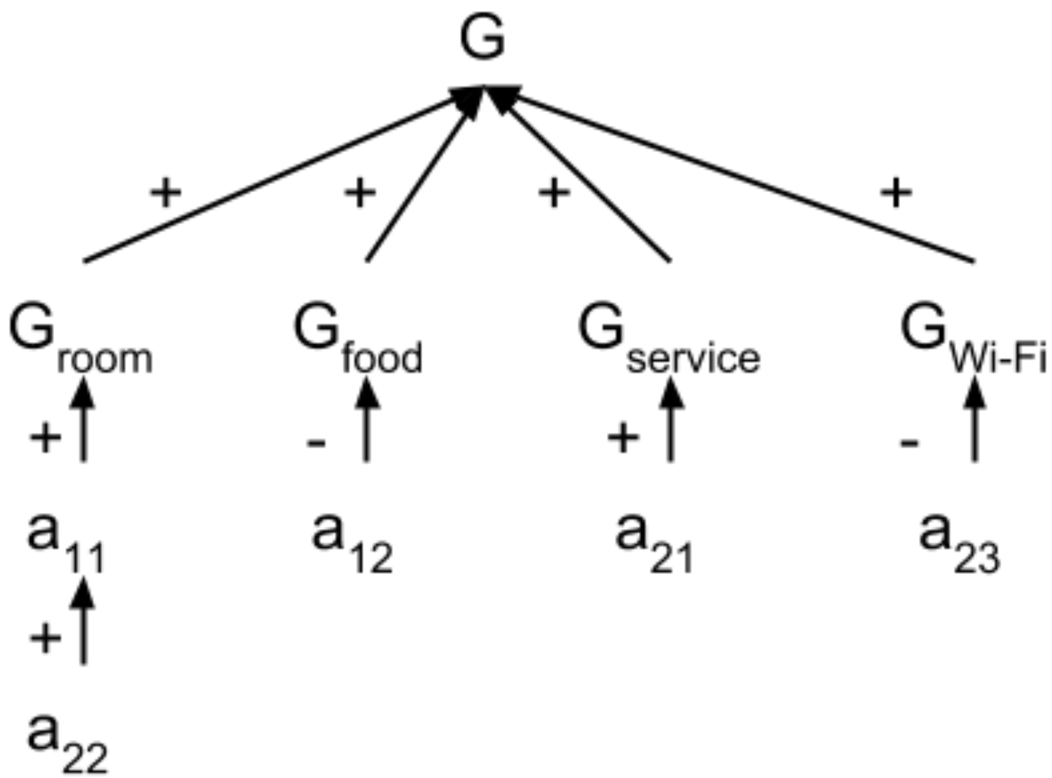

AF given **R**



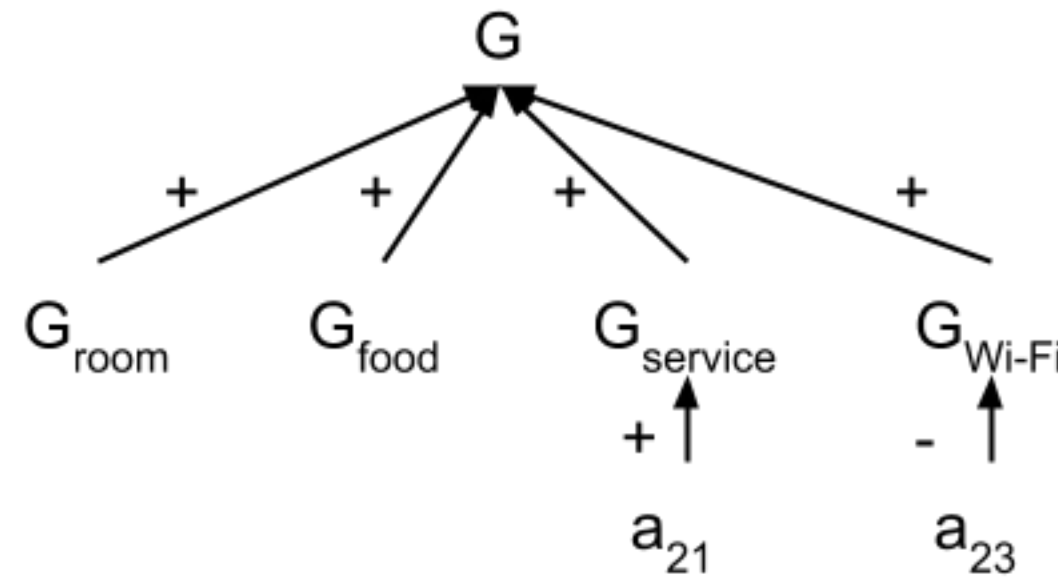

AF given **R\{$r_1$}**

# Argumentative features in BAF

***$r_1$** - argumentative features*




AF given **R**




AF given **$R \setminus \{r_1\}$**

# Deceptive reviews - standard NLP features

| Category | Features |
|---|---|
| Personalization | nr self references<br>nr 2nd person pronouns<br>nr other references<br>nr group pronouns |
| Quantity | nr sentences<br>nr words<br>nr nouns<br>nr verbs |
| Complexity | avg sentence length<br>avg word length |
| Diversity | lexical |
| Uncertainty | nr modal verbs<br>nr modifiers |

# Results

| Random Forests | Hotel | Restaurant |
|---|---|---|
| Baseline | 76.25% | 69% |
| AAF | 77.75% | 71.25% |
| BAF | 79.81% | 73% |

# Future work

- other AM techniques
- semi-supervised approach
- compute argument strength

Thank you!