Back up your Stance:

Recognizing Arguments in Online Discussions

Filip Boltužić and Jan Šnajder
Argument based opinion mining

Argument:

one or more premises leading to exactly one conclusion.

Why Analyze Arguments?

- Basis of users’ opinions.
- Brand analysis.
- Political opinion mining.
- Legal
- Scientific
- Social

Argument-based opinion mining.
used to back up opinion

Manually annotated corpus for argument recognition in online discussion forums.
A discussion on the topic
Should gay marriage be legal?

Gay marriages must be legal in all 50 states. A marriage is covenant between 2 people regardless of their genders. Discrimination against gay marriage is unconstitutional and biased. Tolerance, education and social justice make our world a better place.

Argument in favour
“It is discriminatory to refuse gay couples the right to marry”.

Novel:
Argument based opinion mining as compared to opinion mining in online discussion forums.

Assumption:
Topic-dependent set of arguments exist and have been prepared in advance. User arguments belong to this predefined set of arguments.

Task: What arguments, from predefined set of arguments are used in user comments?
COMARG Corpus

**Websites**
- Procon.org: User comments
- Idebate.org: Arguments

**Procon for arguments?**
**Two Topics:** Gay Marriage (GM)
                     Under God in Pledge (UGIP)

More than 300 comments, balanced between pro and con
Spam removal by skim read?

**UGIP**
175 comments and 6 arguments with stance.

**GM**
198 comments and 7 arguments with stance.
<table>
<thead>
<tr>
<th>Predefined Arguments</th>
<th>Example Gay Marriage</th>
</tr>
</thead>
<tbody>
<tr>
<td>It is discriminatory to refuse gay couples the right to marry.</td>
<td>Pro</td>
</tr>
<tr>
<td>Gay couples should be able to take advantage of the fiscal and legal benefits of marriage.</td>
<td>Pro</td>
</tr>
<tr>
<td>Marriage is about more than procreation, therefore gay couples should not be denied the right to marry due to their biology.</td>
<td>Pro</td>
</tr>
<tr>
<td>Gay couples can declare their union without resort to marriage.</td>
<td>Con</td>
</tr>
<tr>
<td>Gay marriage undermines the institution of marriage, leading to an increase in out of wedlock births and divorce rates.</td>
<td>Con</td>
</tr>
<tr>
<td>Major world religions are against gay marriages.</td>
<td>Con</td>
</tr>
<tr>
<td>Marriage should be between a man and a woman.</td>
<td>Con</td>
</tr>
</tbody>
</table>
Annotations

Comment Argument Pairs
Pair Arguments and Comments for a topic
- Gay Marriage 1386
- Under God in Pledge 1050

Annotation Process
Three Annotators
- Each annotator independently annotated the complete dataset of 2,436 comment-argument pairs.
- Filtering: Remove Problem Pairs
  - No agreement among the three annotators
  - Ordinal distance between labels assigned by the annotators > 1
- Revise problematic comment-argument pairs.
### Inter Annotator Agreements

<table>
<thead>
<tr>
<th>IAA</th>
<th>UGIP</th>
<th>GM</th>
<th>UGIP+GM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fleiss’ Kappa</td>
<td>0.46</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>Cohen’s Kappa</td>
<td>0.46</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>Weighted Kappa</td>
<td>0.45</td>
<td>0.51</td>
<td>0.50</td>
</tr>
<tr>
<td>Pearson’s r</td>
<td>0.68</td>
<td>0.74</td>
<td>0.71</td>
</tr>
</tbody>
</table>

### Gold Standard: Majority Label

#### Label Distribution

<table>
<thead>
<tr>
<th>TOPIC</th>
<th>A</th>
<th>a</th>
<th>N</th>
<th>s</th>
<th>S</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>UGIP</td>
<td>48</td>
<td>86</td>
<td>691</td>
<td>58</td>
<td>130</td>
<td>1,013</td>
</tr>
<tr>
<td>GM</td>
<td>89</td>
<td>73</td>
<td>849</td>
<td>98</td>
<td>176</td>
<td>1,285</td>
</tr>
<tr>
<td>UGIP+GM</td>
<td>137</td>
<td>159</td>
<td>1540</td>
<td>156</td>
<td>306</td>
<td>2,298</td>
</tr>
</tbody>
</table>

A: explicitly attacks the argument
a: vaguely/implicitly attacks the argument 12.9%
N: makes no use of the argument 67%
s: vaguely/implicitly supports the argument
S: explicitly supports the argument 20.1%
Argument Recognition Model

comment-argument pair $\rightarrow$ less domain dependent

- TE features
- STS features
- SA features

A (explicitly attacks)
a (implicitly attacks)
N (no argument)
s (implicitly supports)
S (explicitly supports)
Argument Recognition Model  
-- SA features

**Stance Alignment (SA) feature:**

$SA = 1$, if *pro* comment + *pro* argument,

or if *con* comment + *con* argument;

$SA = 0$, otherwise.

Use SA to display the argument polarity, and to demonstrate how important it is to known stance to distinguish S vs A, or s vs a.

Is it proper to use this feature for the classifier while it has strong inference to the class labels? (see Table 7 SA: A-N-S)
Argument Recognition Model
-- TE features

Text Entailment (TE) feature:
T(Text) entails H(Hypothesis) (T ⇒ H) if, typically, a human reading T would infer that H is most likely true.

example of positive entailment:
- Text: If you help the needy, God will reward you.
- Hypothesis: Giving money to a poor man has good consequences.
Example 1.
T1: Research shows that drivers speaking on a mobile phone have much slower reactions in braking tests than non-users, and are worse even than if they have been drinking.

H: The use of cell-phones while driving is a public hazard.

(Cabrio and Villata, 2012)
Argument Recognition Model
-- TE features

Example 2 (Continued).
T2: Regulation could negate the safety benefits of having a phone in the car. When you’re stuck in traffic, calling to say you’ll be late can reduce stress and make you less inclined to drive aggressively to make up lost time.

\[ T2 \text{ attacks } H \]

H: The use of cell-phones while driving is a public hazard.

(Cabrio and Villata, 2012)
Argument Recognition Model
-- TE tool

Excitement Open Platform (EOP)

Raw entailment problems → Linguistic Analysis Pipeline (LAP) → Annotated entailment problems

Linguistic Analysis Tools

UIMA Components

Entailment Core (EC)

Entailment Decision Algorithm (EDA)

Dynamic and Static Components (Algorithms and Knowledge)

Decision (Entailment or NonEntailment) & Confidences

(Padó et al., 2014)

EOP: http://hltfbk.github.io/Excitement-Open-Platform/
Argument Recognition Model
-- TE tool

EOP - Entailment Decision Algorithm (EDA) result type hierarchy

(1) **Binary EDA decision** (Entailment/NonEntailment); *Why not include “Abstain”? Why not use finer-grain for the 5 labels?*
(2) **Seven pre-trained EDAs**, some contain only syntactical features, others rely on resources such as WordNet and VerbOcean; *Why and how do they choose 7 EDAs?*
(3) Each EDA generates output **decisions & confidences**;

The features of the classifier: outputs of all seven EDAs (14 in total).
Argument Recognition Model
-- TE features

The ratio of positive entailment decision across labels (averaged over seven EDAs), scaled to a [0,1] interval.

Pro > Con; Explicitly > Implicitly.

The comment is expected to entail the argument phrase.

figure 1
Argument Recognition Model
-- STS features

Semantic Textual Similarity (STS) feature:
“the degree of semantic equivalence between two texts”.
(Agirre et al., 2012)

Given a comment and an argument, compute
(1) the continuous similarity score of argument & comment;
(2) the vector of similarity: argument & sentences of the comment,
   (vector length=largest # of sentences) 29
(3) the maximum & the mean of sentence-level similarities. 2

⇒ 31 STS features in total.
Argument Recognition Model  
-- STS features

**STS feature:**
Average similarity score on sentence and comment level across labels, scaled to a [0, 1] interval.

Attacked > Supported
N -> the smallest score

Is the high STS score of A (explicit attack) consistent with the low TE decision in figure 1?

Exp:
"He has a car" <=> "She has a car"
similarity=high, entailment=no
Experimental Evaluation - The Setup

- Three different settings for classifying comment argument pairs
  - A-a-N-s-S : Polarity + Degree of explicitness
  - Aa-N-sS : Only the polarity
  - A-N-S : Pairs in which arguments are either not used or used explicitly (only for comparison).

<table>
<thead>
<tr>
<th>Label</th>
<th>Description: Comment...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>...explicitly attacks the argument</td>
</tr>
<tr>
<td>a</td>
<td>...vaguely/implicitly attacks the argument</td>
</tr>
<tr>
<td>N</td>
<td>...makes no use of the argument</td>
</tr>
<tr>
<td>s</td>
<td>...vaguely/implicitly supports the argument</td>
</tr>
<tr>
<td>S</td>
<td>...explicitly supports the argument</td>
</tr>
</tbody>
</table>
The Setup - Cont

- For comparison we have two baselines:
  - Majority Class Classifier (MCC) - assigns label N to every instance
  - a bag-of-words overlap classifier (BoWO) - uses the word overlap between the comment and the argument as the only feature.

- For comparison: SVM (Support Vector Machine) algorithm with a Radial Basis Function kernel.

- For training and evaluation: using nested 5X3 cross-validation
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>A-a-N-s-S</th>
<th>Aa-N-sS</th>
<th>A-N-S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UGIP</td>
<td>GM</td>
<td>UGIP</td>
</tr>
<tr>
<td>MCC baseline</td>
<td>68.2</td>
<td>69.4</td>
<td>68.2</td>
</tr>
<tr>
<td>BoWO baseline</td>
<td>68.2</td>
<td>69.4</td>
<td>67.8</td>
</tr>
<tr>
<td>TE</td>
<td>69.1</td>
<td><strong>81.1</strong></td>
<td>69.6</td>
</tr>
<tr>
<td>STS</td>
<td>67.8</td>
<td>68.7</td>
<td>67.3</td>
</tr>
<tr>
<td>SA</td>
<td>68.2</td>
<td>69.4</td>
<td>68.2</td>
</tr>
<tr>
<td>STS+SA</td>
<td>68.2</td>
<td>69.5</td>
<td>67.5</td>
</tr>
<tr>
<td>TE+SA</td>
<td>68.9</td>
<td>72.4</td>
<td><strong>71.0</strong></td>
</tr>
<tr>
<td>TE+STS+SA</td>
<td><strong>70.5</strong></td>
<td>72.5</td>
<td>68.9</td>
</tr>
</tbody>
</table>

Table 7: Argument recognition F1-score (separate models for UGIP and GM topics)

- Baselines: Perform similar
- Better results
- Only the STS or the SA: perform similar to baseline
- No Benefit
- Does better in simpler setting
To Evaluate the Models on Unseen Topics -- Results (contd.)

- No change in performance of TE+SA, therefore, the models are topic independent!!

<table>
<thead>
<tr>
<th>Model</th>
<th>UGIP $\rightarrow$ GM</th>
<th>GM $\rightarrow$ UGIP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A-a-N-s-S</td>
<td>Aa-N-sS</td>
</tr>
<tr>
<td>STS+SA</td>
<td>69.4</td>
<td>69.4</td>
</tr>
<tr>
<td>TE+SA</td>
<td>72.6</td>
<td>73.5</td>
</tr>
<tr>
<td>STS+TE+SA</td>
<td>71.5</td>
<td>72.2</td>
</tr>
</tbody>
</table>

Table 8: Argument recognition F1-score on UGIP and GM topics (cross-topic setting)
The models perform less well on smaller classes (A, a, s, and S), hence the macro-averaged F1-scores are much lower than the micro-averaged F1-scores.

- The recall is lower than the precision.
- The STS+TE+SA model slightly outperforms the TE+SA model on the A-a-N-s-S problem, while on the other problem formulations the TE+SA model performs best.
Error Analysis

- Problems
  - Understanding users actual intention/meaning
  - Distinguishing between arguments that are mentioned and those that are not (s and a)

- The TE model in the majority of cases outperforms the STS model.

- Nonetheless, in some cases, the STS-based model outperformed it because the word overlap between the argument and the comment in quite high, although they completely differ in meaning.
Conclusion/Feedback

- Why take arguments and comments from different websites?
- Why use a sample space that small?
- Do we require the stance?