Determining the Polarity of Postings for Discussion Search

Ingo Frommholz and Marc Lechtenfeld

University of Duisburg-Essen
{ingo.frommholz|marc.lechtenfeld}@uni-due.de

Diss-Stadl
1. Oktober 2008
Outline

1. Polarity in Discussion Search
2. Using SVM for Polarity Detection
3. Experiments and Results
4. Conclusion
Polarity in Discussion Search
Example Discussion Thread

Vista is a very secure operating system.

I disagree. They found lots of security holes shortly after Vista was released.

Vista is Microsoft's new operating system. It is the successor of XP.

Nonsense!

Very true! Vista is secure.

C1

C2

C3

C4

C5
POLAR

- Probabilistic Object-oriented Logics for Annotation-based Retrieval
- Model and query discussion threads
- Discussion search
- Integrates content, structure and external knowledge
- Supports polarity of comments
Example Representation in POLAR

c1[ 0.7 vista ... # tf
   0.6 -*c2 # negative
   0.6 -*c3 # negative
   0.6 *c4 # positive
 ]
c2[ 0.6 vista ...]
c3[ ... ]
c4[ 0.8 vista ...]
c5[ 0.45 vista ...]

0.5 °vista # idf

Vista is a very secure operating system.

Vista is Microsoft's new operating system. It is the successor of XP.

I disagree. They found lots of security holes shortly after Vista was released.

Nonsense!

Very true! Vista is secure.
Content-only Query for “vista”

tf × idf-based probability

q[ vista ]  # the query
?- D->q    # D implies q

# content-based ranking
0.350 c1
0.300 c2
0.225 c5
Content-only Query for “vista”

tf × idf-based probability

q[ vista ]  # the query
?- D→q  # D implies q

# content-based ranking
0.350 c1
0.300 c2
0.225 c5

$c_1$ ranked first due to its content. But is $c_1$ a “good” comment?

I disagree. They found lots of security holes shortly after Vista was released.

Vista is a very secure operating system.

Nonsense!

Very true! Vista is secure.

Vista is Microsoft's new operating system. It is the successor of XP.
# A priori unconditional trust
0.5 uncond_trust(c1)
0.5 uncond_trust(c2)
0.5 uncond_trust(c3)
0.5 uncond_trust(c4)
0.5 uncond_trust(c5)

# Trust and polarity
trust(C) :- uncond_trust(C)
trust(C) :- C[*A] & trust(A)
!trust(C) :- C[-*A] & trust(A)
Combination of Trustworthiness and Content

\[ \neg D \rightarrow q \quad \& \quad \text{trust}(D) \]

0.1500 (c2)
0.1125 (c5)
0.1115 (c1)

I. Frommholz, M. Lechtenfeld
Determining the Polarity of Postings for Discussion Search

University of Duisburg-Essen
Combination of Trustworthiness and Content

?- D->q & trust(D)
0.1500  (c2)
0.1125  (c5)
0.1115  (c1)

\[ c_1 \text{ now ranked on 3rd position due to negative replies!} \]
The Problem

Polarity of postings usually not known, so the question is:
The Problem

Polarity of postings usually not known, so the question is:

**How (and how good) can we determine the polarity of postings automatically?**
The Problem

Polarity of postings usually not known, so the question is:

**How (and how good) can we determine the polarity of postings automatically?**

One possible answer: Try and evaluate a machine learning approach!
Using SVM for Polarity Detection
Classification Task

- **Object**
  - Discussion posting

**Criterion**

Judgemental character of the statements regarding the previous posting

**Categories**

- Positive (supporting) posting
- Negative (deprecatory) posting
- Neutral (non-judgemental) posting
Classification Task

- **Object**
  - Discussion posting

- **Criterion**
  - Judgemental character of the statements regarding the *previous* posting

---

**Text Classifier**

---

Polarity in Discussion Search

Using SVM for Polarity Detection

Experiments and Results

Conclusion

I. Frommholz, M. Lechtenfeld

Determining the Polarity of Postings for Discussion Search

University of Duisburg-Essen
Classification Task

Object
- Discussion posting

Criterion
- Judgemental character of the statements regarding the *previous* posting

Categories
- Positive (supporting) posting
- Negative (deprecatory) posting
- Neutral (non-judgemental) posting
Automatic Classification of Postings

Feature Sources

- Content of the Posting: Text
  ⇒ Non-Content-Oriented Text Classification
  ≈ Sentiment Classification [Pang et al., 2002]

Support Vector Machine [Vapnik, 1998]
(Supervised Machine Learning Technique)
Automatic Classification of Postings

Feature Sources

- Content of the Posting: Text
  ⇒ Non-Content-Oriented Text Classification
  ≈ Sentiment Classification [Pang et al., 2002]
- Context of the Posting: Thread
Automatic Classification of Postings

- Feature Sources
  - Content of the Posting: Text
    - Non-Content-Oriented Text Classification
      ≈ Sentiment Classification [Pang et al., 2002]
  - Context of the Posting: Thread

- Classifier
  - Support Vector Machine [Vapnik, 1998]
    (Supervised Machine Learning Technique)
Collection: ZDNet News

I. Frommholz, M. Lechtenfeld
Determining the Polarity of Postings for Discussion Search
University of Duisburg-Essen

- Articles and comments about IT-related topics
- Tagging, thread structure, metadata like author and creation time

Sun begins open-source Solaris era
By Stephen Shankland, News.com
Published on ZDNet News: Jun 14, 2005 4:00:00 AM

Sun Microsystems has released Solaris as open-source software, a move that's central to the company's plan to regain lost relevance and fend off rivals Red Hat, IBM and Microsoft.

The company on Tuesday posted more than 5 million lines of source code for the heart of the operating system—kernel and networking code—at the OpenSolaris Web site. The source code components, such as file systems, device drivers and networking protocols, are available under the Apache license. Sun will release the rest of the source code in increments over the coming months. The company says it has made the code available through the OpenSolaris Web site in response to the growing demand for open-source operating systems.

Re: Sun begins open-source Solaris era
This will be a very important step. Detailed documentation will be crucial to support this effort given that Solaris is arriving a bit late to the Open Source game.

And when will we see Purple Cat Solaris?
For this to work, we will need non-Sun Solaris distributions. For that matter will they permit the use of the name Solaris? If not, they need to create a name that all can use (like Linux) that all can use to describe distributions.

OpenSolaris is unique...
OpenSolaris is going to be a great addition to the current open source OS offerings. I myself primarily run Linux, but I've been patiently waiting for the open-sourcing of Solaris. I'm very curious to see how well it will be received by the OSS community. Though Linux is great in its own right, here is why I think an open-source Solaris is more attractive for most businesses:

Excellent post.
Well thought out and articulated.
Term and Document Features

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Presence of tokens (unigrams),</td>
</tr>
</tbody>
</table>

Author: B
Title: Never

>> I think it is wrong! ...
I agree with you!
It is the best ...!
I think it is not the worst ...!
See you ... :-)
## Term and Document Features

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Presence of tokens (unigrams), Punctuation</td>
</tr>
</tbody>
</table>

**Author:** B  
**Title:** Never

>> I think it is wrong! ...

I agree with you!
It is the best ... !
I think it is not the worst ... !
See you ... :-)

---

### Polarity in Discussion Search

<table>
<thead>
<tr>
<th>I. Frommholz, M. Lechtenfeld</th>
<th>Determining the Polarity of Postings for Discussion Search</th>
<th>University of Duisburg-Essen</th>
<th>Conclusion</th>
</tr>
</thead>
</table>
Term and Document Features

Author: B
Title: Never

>> I think it is wrong! ...
I agree with you! It is the best ... !
I think it is not the worst ... !
See you ... :-)

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Presence of tokens (unigrams), Punctuation</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Emoticon (EMOTICON_),</td>
</tr>
</tbody>
</table>
Term and Document Features

Author: B
Title: Never

>> I think it is wrong! ...
I agree with you!
It is the best ... !
I think it is not the worst ... !
See you ... :-)

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Presence of tokens (unigrams), Punctuation</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Emoticon (EMOTICON_), Negation (NOT_)</td>
</tr>
</tbody>
</table>

I. Frommholz, M. Lechtenfeld
Determining the Polarity of Postings for Discussion Search
University of Duisburg-Essen
## Term and Document Features

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Presence of tokens (unigrams),</td>
</tr>
<tr>
<td></td>
<td>Punctuation</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Emoticon (EMOTICON_),</td>
</tr>
<tr>
<td></td>
<td>Negation (NOT_)</td>
</tr>
<tr>
<td>Source</td>
<td>Title terms (TITLE_),</td>
</tr>
</tbody>
</table>

Author: B
Title: Never

>> I think it is wrong! ...
I agree with you!
It is the best ...!
I think it is not the worst ...!
See you ... :−)
## Term and Document Features

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Token</strong></td>
<td>Presence of tokens (unigrams), Punctuation</td>
</tr>
<tr>
<td><strong>Sentiment</strong></td>
<td>Emoticon (EMOTICON_), Negation (NOT_)</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>Title terms (TITLE_), Quotation terms (QUOT_),</td>
</tr>
</tbody>
</table>

**Author:** B  
**Title:** Never

>> I think it is wrong! ...

I agree with you! It is the best ... !
I think it is not the worst ... !
See you ... :-)

---

**Polarity in Discussion Search**

- Frommholz, I.
- Lechtenfeld, M.

**Using SVM for Polarity Detection**

- Determining the Polarity of Postings for Discussion Search

**Experiments and Results**

- University of Duisburg-Essen

**Conclusion**
## Term and Document Features

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Token</strong></td>
<td>Presence of tokens (unigrams), Punctuation</td>
</tr>
<tr>
<td><strong>Sentiment</strong></td>
<td>Emoticon (EMOTICON_), Negation (NOT_)</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>Title terms (TITLE_), Quotation terms (QUOT_), Terms of prev. posting (PREV_)</td>
</tr>
</tbody>
</table>

**Author: B**
**Title: Never**

>> I think it is wrong! ...
I agree with you!
It is the best ... !
I think it is not the worst ... !
See you ... :-)

---

**I. Frommholz, M. Lechtenfeld**
Determining the Polarity of Postings for Discussion Search
University of Duisburg-Essen
## Term and Document Features

**Author:** B  
**Title:** Never

- ">> I think it is wrong! ...  
I agree with you!  
It is the best ... !  
I think it is not the worst ... !
See you ... :-)

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Token</strong></td>
<td>Presence of tokens (unigrams), Punctuation</td>
</tr>
<tr>
<td><strong>Sentiment</strong></td>
<td>Emoticon (EMOTICON_), Negation (NOT_)</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>Title terms (TITLE_), Quotation terms (QUOT_), Terms of prev. posting (PREV_)</td>
</tr>
<tr>
<td><strong>Posting</strong></td>
<td>Posting length</td>
</tr>
</tbody>
</table>
## Term and Document Features

**Author:** B  
**Title:** Never

>> I think it is wrong! ...  
I agree with you!  
It is the best ... !  
I think it is not the worst ... !  
See you ... :-)

<table>
<thead>
<tr>
<th>Group</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Token</strong></td>
<td>Presence of tokens (unigrams), Punctuation</td>
</tr>
<tr>
<td><strong>Sentiment</strong></td>
<td>Emoticon (EMOTICON_), Negation (NOT_)</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>Title terms (TITLE_), Quotation terms (QUOT_), Terms of prev. posting (PREV_)</td>
</tr>
<tr>
<td><strong>Posting</strong></td>
<td>Posting length</td>
</tr>
</tbody>
</table>
Term and Document Features

Preprocessing

>> I think it is wrong! ...
I agree with you!
It is the best ... !
I think it is not the worst ... !
See you ... :-}
Term and Document Features

Author: B
Title: Never

>> I think it is wrong! ...
I agree with you!
It is the best ... !
I think it is not the worst ... !
See you ... :-)
Term and Document Features

Author: B
Title: Never

>> I think it is wrong! ...
I agree with you! It is the best ... !
I think it is not the worst ... !
See you ... :-)
Term and Document Features

Preprocessing
- No usage of Stop Word Lists
- Stemming

Feature Reduction
- $\chi^2$ Test
- Minimal Occurrences

I. Frommholz, M. Lechtenfeld
Determining the Polarity of Postings for Discussion Search
University of Duisburg-Essen
Context Features

Structure Features
Context Features

Structure Features

Author Sequences
Context Features

Posting of Author A

Structure Features

Author Sequences
Context Features

Structure Features

Author Sequences

I. Frommholz, M. Lechtenfeld
Determining the Polarity of Postings for Discussion Search
University of Duisburg-Essen
Context Features

Structure Features

- Author Sequences

Polarity in Discussion Search

Using SVM for Polarity Detection

Experiments and Results

Conclusion

I. Frommholz, M. Lechtenfeld

Determining the Polarity of Postings for Discussion Search

University of Duisburg-Essen
Context Features

Structure Features
- Author Sequences
  - A – B – A

I. Frommholz, M. Lechtenfeld
Determining the Polarity of Postings for Discussion Search
University of Duisburg-Essen
Context Features

Structure Features
- Author Sequences
  - A – B – A
  - A – A

Polarity in Discussion Search

Using SVM for Polarity Detection

Experiments and Results

Conclusion
Context Features

Structure Features

- Author Sequences
  - A – B – A
  - A – A
- Quotations
Context Features

Structure Features

- Author Sequences
  - A – B – A
  - A – A

- Quotations
  - Existence/Number
Context Features

Structure Features

- Author Sequences
  - A – B – A
  - A – A

- Quotations
  - Existence/Number
  - Proportion (Quotation; current/previous Posting)
Context Features

Structure Features

- Author Sequences
  - A – B – A
  - A – A

- Quotations
  - Existence/Number
  - Proportion (Quotation; current/previous Posting)

- Annotations
  - Alternating Quotations and Annotations
Context Features

Response Behaviour
Context Features

Response Behaviour

- Time
- Response Time
Context Features

Response Behaviour
- Time
  - Response Time
  - Response Duration
Context Features

Response Behaviour

- Time
  - Response Time
  - Response Duration
  - Creation Time (day of the week, hour, time of the day)
Context Features

Response Behaviour

- **Time**
  - Response Time
  - Response Duration
  - Creation Time (day of the week, hour, time of the day)
Context Features

Response Behaviour

- Time
  - Response Time
  - Response Duration
  - Creation Time (day of the week, hour, time of the day)
Context Features

Response Behaviour

- Time
- Response Time
- Response Duration
- Creation Time (day of the week, hour, time of the day)
- Responses
Context Features

Response Behaviour

- Time
  - Response Time
  - Response Duration
  - Creation Time (day of the week, hour, time of the day)

- Responses
  - Number of Responses (direct, total, to the previous posting)
Context Features

Response Behaviour

- **Time**
  - Response Time
  - Response Duration
  - Creation Time (day of the week, hour, time of the day)

- **Responses**
  - Number of Responses (direct, total, to the previous posting)
Context Features

References Features

Topic Features
Context Features

For further information see:  
http://www.wikipedia.org/

References Features

- Hyperlinks

Topic Features
Context Features

For further information see:
http://www.wikipedia.org/

References Features

- Hyperlinks
- Mentioned Author

Topic Features
Context Features

For further information see:
http://www.wikipedia.org/

If you are interested in information about the topic mentioned by Charlie ...

Title: Microsoft: No antivirus product yet
ZDNet Tags: Antivirus software, Microsoft, ...

References Features

- Hyperlinks
- Mentioned Author
- Article Posting

Topic Features

- Article Topics (ZDNet Tags)
For further information see:
http://www.wikipedia.org/

If you are interested in information about the topic mentioned by Charlie...

Title:
Microsoft: No antivirus product yet

ZDNet Tags:
Antivirus software, Microsoft,...

References Features
- Hyperlinks
- Mentioned Author
- Article Posting

Topic Features
- Article Topics (ZDNet Tags)
Context Features

For further information see:
http://www.wikipedia.org/

If you are interested in information about the topic mentioned by Charlie ...

Title:
Microsoft: No antivirus product yet

ZDNet Tags:
Antivirus software, Microsoft, ...

Posting A

TOPIC_Antivirus...
TOPIC_Microsoft ...

References Features

- Hyperlinks
- Mentioned Author
- Article Posting

Topic Features

- Article Topics (ZDNet Tags)
Experiments and Results
Experimental Setup

Cross-Evaluation

Stratified ten-fold cross-validations

Example:
3-fold cross-val.

- Training set
- Test set

<table>
<thead>
<tr>
<th>1. Run</th>
<th>Partition 1</th>
<th>Partition 2</th>
<th>Partition 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Run</td>
<td>Partition 1</td>
<td>Partition 2</td>
<td>Partition 3</td>
</tr>
<tr>
<td>3. Run</td>
<td>Partition 1</td>
<td>Partition 2</td>
<td>Partition 3</td>
</tr>
</tbody>
</table>
Experimental Setup

- **Cross-Evaluation**
  - Stratified ten-fold cross-validations

  - Example:
    - 3-fold cross-val.
    - | 1. Run | Partition 1 | Partition 2 | Partition 3 |
      | Training set | Partition 1 | Partition 2 | Partition 3 |
      | Test set | Partition 1 | Partition 2 | Partition 3 |

- **Evaluation Measure & Baseline**
  - Accuracy (percentage of postings classified correctly)

---

I. Frommholz, M. Lechtenfeld
Determining the Polarity of Postings for Discussion Search
University of Duisburg-Essen
Experimental Setup

- Cross-Evaluation
  - Stratified ten-fold cross-validations

  Example:
  3-fold cross-val.
  
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition 1</td>
<td>Partition 2</td>
<td>Partition 3</td>
<td>Partition 1</td>
<td>Partition 2</td>
<td>Partition 3</td>
</tr>
</tbody>
</table>

- Evaluation Measure & Baseline
  - Accuracy (percentage of postings classified correctly)
  - Random-Choice-Baseline (0.5)
Experimental Setup

Cross-Evaluation

- Stratified ten-fold cross-validations

Example:

<table>
<thead>
<tr>
<th>1. Run</th>
<th>2. Run</th>
<th>3. Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition 1</td>
<td>Partition 2</td>
<td>Partition 3</td>
</tr>
<tr>
<td>Training set</td>
<td>Test set</td>
<td></td>
</tr>
</tbody>
</table>

Evaluation Measure & Baseline

- Accuracy (percentage of postings classified correctly)
- Random-Choice-Baseline (0.5)

Data

- 150 positive and 150 negative postings
## Experiments and Results

### Term Features

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5</td>
</tr>
<tr>
<td>Init. term feature set</td>
<td>0.69</td>
</tr>
<tr>
<td>Best term feature set</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Best feature set contains:
- Tagged quotation terms
- Tagged terms of the previous posting
Experiments and Results

Term Features

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5</td>
</tr>
<tr>
<td>Init. term feature set</td>
<td>0.69</td>
</tr>
<tr>
<td>Best term feature set</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Best feature set contains:

- Tagged quotation terms
- Tagged terms of the previous posting
## Experiments and Results

### Context Features

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5</td>
</tr>
<tr>
<td>Init. context feature set</td>
<td>0.67</td>
</tr>
<tr>
<td>Best context feature set</td>
<td>0.7733</td>
</tr>
</tbody>
</table>

**Best features:**
- Response behaviour: Response time and creation time
- Topic: Article topic tags

---

I. Frommholz, M. Lechtenfeld  
Determining the Polarity of Postings for Discussion Search  
University of Duisburg-Essen
Experiments and Results

Context Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5</td>
</tr>
<tr>
<td>Init. context feature set</td>
<td>0.67</td>
</tr>
<tr>
<td>Best context feature set</td>
<td>0.7733</td>
</tr>
</tbody>
</table>

Best features:

- Response behaviour: Response time and creation time
- Topic: Article topic tags
Experiments and Results

Polarity Detection with Term and Context Features

Baseline 0.5
## Experiments and Results

Polarity Detection with Term and Context Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5</td>
</tr>
<tr>
<td>Term Features</td>
<td>0.76</td>
</tr>
<tr>
<td>Context Features</td>
<td>0.7733</td>
</tr>
</tbody>
</table>

Polarity in Discussion Search

Using SVM for Polarity Detection

Experiments and Results

Conclusion
Experiments and Results

Polarity Detection with Term and Context Features

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5</td>
</tr>
<tr>
<td>Term Features</td>
<td>0.76</td>
</tr>
<tr>
<td>Context Features</td>
<td>0.7733</td>
</tr>
<tr>
<td>Term &amp; Context Features</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Conclusion
Polarity (and trustworthiness) is a valuable kind of evidence for discussion search.
Conclusion

- Polarity (and trustworthiness) is a valuable kind of evidence for discussion search
- Support Vector Machines with appropriate term and context features can be applied for polarity detection
Conclusion

- Polarity (and trustworthiness) is a valuable kind of evidence for discussion search
- Support Vector Machines with appropriate term and context features can be applied for polarity detection
- Our first experiments gained an accuracy around 79% for this task
Thanks for your attention!

...any further questions?