Extracting Patterns of Harmful Expressions for Cyberbullying Detection

Michal Ptaszynski, Fumito Masui, Yasutomo Kimura, Rafal Rzepka and Kenji Araki
Outline

1. Cyberbullying as social problem
2. Previous research
3. Proposed method
4. Experiments
5. Future work
Cyberbullying

- Slandering and humiliating people on the Internet.
- Recently noticed social problem.

HELP by ICT

INTERNET PATROL

- Internet monitoring by PTA.
- Request site admin to remove harmful entries.
- High cost of time and fatigue for net-patrol members.
Previous Research

- **2009**
  - Affect analysis of cyberbullying data

- **2010**
  - SVM / optimization
  - SO-PMI-IR / phrases

- **2011**

- **2012**

- **2013**
  - Patents

- **2014**
  - Language Combinatorics
  - Language Combinatorics
  - Extra-Extracting Patterns of Harmful Expressions for Cyberbullying Detection, 7th Language & Technology Conference (LTC’15), 2015.11.27-29.

- **2015**
Previous Research

SO-PMI-IR / phrases

2010


2011


2012


2013

Language Combinatorics


Language Combinatorics / Preprocessing


2014

Affect analysis of cyberbullying data

Automatic acquisition of harmful words

2015

Brute Force Works Best Against Bullying, IJCAI 2015 Workshop on Intelligent Personalization (IP 2015), Buenos Aires, 2015.07.25-31


Previous Research

SO-PMI-IR / phrases


Language Combinatorics


Language Combinatorics / Preprocessing

2013 PATENT


Automatic acquisition of harmful words


Affect analysis of cyberbullying data


SVM / optimization


Previous Research

2009

Affect analysis of cyberbullying data

SVM / optimization


2010

SO-PMI-IR / phrases

2011

Language Combinatorics / Preprocessing


2012

Language Combinatorics

Category Relevance Optimization

Automatic acquisition of harmful words


Patent name: An Apparatus and Method for Detection of Harmful Entries on Internet


2013

SO-PMI-IR / phrases

Brute Force Works Best Against Bullying, IJCAI 2015 Workshop on Intelligent Personalization (IP 2015), Buenos Aires, 2015.07.25-31


2014

2015
Previous Research


Brute Force Works Best Against Bullying, IJCAI 2015 Workshop on Intelligent Personalization (IP 2015), Buenos Aires, 2015.07.25-31

Language Combinatorics / Preprocessing

Language Combinatorics

Automatic acquisition of harmful words

Category Relevance Optimization

Affect analysis of cyberbullying data

SVM / optimization

SO-PMI-IR / phrases

2013 PATENT
Previous Research

2009

Affect analysis of cyberbullying data

2010

SO-PMI-IR / phrases

2011

SVM / optimization

2012

Category Relevance Optimization

2013

2014

Language Combinatorics / Preprocessing

2015

Language Combinatorics

Automatic acquisition of harmful words


Patent name: An Apparatus and Method for Detection of Harmful Entries on Internet


Previous Research

- **Affect analysis of cyberbullying data**

- **SO-PMI-IR / phrases**

- **SVM / optimization**

- **2013 PATENT**

- **Language Combinatorics / Preprocessing**

- **Category Relevance Optimization**

- **Automatic acquisition of harmful words**
Previous Research

2009

Affect analysis of cyberbullying data

SO-PMI-IR / phrases

2010

SVM / optimization

2011

Language Combinatorics / Preprocessing

2012

Language Combinatorics

2013

Category Relevance Optimization

Automatic acquisition of harmful words

2014

2015


Dataset

- Actual data collected by Internet Patrol (annotated by experts)
- From unofficial school forums (BBS)
- Provided by Human Right Center in Japan (Mie Prefecture)
- According to the Definition by Japanese Ministry of Education (MEXT)
- 1,490 harmful and 1,508 non-harmful entries
Proposed Method
Sentence patterns = ordered non-repeated combinations of sentence elements.

for $1 \leq k \leq n$, there is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

all possible $k$-long patterns, and

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + \ldots + \frac{n!}{n!(n-n)!} = 2^n - 1$$

Extract patterns from all sentences and calculate occurrence.
## Language Combinatorics

**Example:** What a nice day!

### 5-element pattern: What a nice day! (1)

<table>
<thead>
<tr>
<th>4-el. patterns:</th>
<th>3-el. patterns:</th>
<th>2-el. patterns:</th>
<th>1-el. patterns:</th>
</tr>
</thead>
<tbody>
<tr>
<td>What a nice *!</td>
<td>a nice *!</td>
<td>What a</td>
<td>What</td>
</tr>
<tr>
<td>What a nice day</td>
<td>What a nice</td>
<td>What *!</td>
<td>a</td>
</tr>
<tr>
<td>What a * day!</td>
<td>What a *!</td>
<td>nice *!</td>
<td>nice</td>
</tr>
<tr>
<td>(5)</td>
<td>(10)</td>
<td>(10)</td>
<td>(5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Language Combinatorics

SPEC – Sentence Pattern Extraction arChitecture

Sentence patterns = ordered non-repeated combinations of sentence elements.

for $1 \leq k \leq n$, there is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

all possible $k$-long patterns, and

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + \ldots + \frac{n!}{n!(n-n)!} = 2^n - 1$$

Normalized pattern weight

$$w_j = \left( \frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5 \right) \times 2$$

Score for one sentence

$$score = \sum w_j, \ (1 \geq w_j \geq -1)$$

Experiment setup

Preprocessing
- 1. Tokenization
- 2. POS
- 3. Tokens+POS

Pattern List Modification
- 1. All patterns
- 2. Zero-patterns deleted
- 3. Ambiguous patterns deleted

Weight Calculation Modifications
- 1. Normalized
- 2. Award length
- 3. Award length and occurrence

All patterns vs. only n-grams

Automatic threshold setting

10-fold Cross Validation

Is it worth the time?

One experiment = 420 runs

Data is never perfectly balanced.
Results

Tokens+POS

Best F-score
F=0.8
P=0.76
R=0.84
Results

Tokens+POS

specific elements are more effective than generalized ones

POS

Best F-score
F=0.8
P=0.76
R=0.84
Results

Best BEP

Unmodified Tokens+POS
P=0.79
R=0.79
Results

Comparison with state-of-the-art
Results

Comparison with state-of-the-art

- More efficient (user does almost nothing)
- Applicable to other languages
- Can point out non-harmful elements
- Pattern lists contained all Nitta et al.’s seed words → could improve Nitta with patterns
Conclusions

• Presented research on cyberbullying detection.
• Proposed novel method.
  • Combinatorial algorithm applied in automatic extraction of sentence patterns.
• Used patterns in classification of cyberbullying.
• Tested on actual data obtained by Internet patrol.
• Outperformed previous methods.
• Requires minimal human effort.
Future work

• Apply different preprocessing and classifiers for further improvement.
• Test on new data
• Obtain new data by applying in practice.
• Verify the actual amount of CB information on the Internet and reevaluate in more realistic conditions.
Thank you for your kind attention!

Michal Ptaszynski
paszynski@ieee.org