Machine Learning and Affect Analysis Against Cyber-Bullying

Michal Ptaszynski, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka, Kenji Araki
Presentation Outline

• Introduction
• What is Cyber-Bullying?
• Machine Learning Method for Cyber-bullying Detection
• Affect Analysis of Cyber-Bullying Data
• Conclusions & Future Work
Introduction

- Dialogue agent–companion needs to be aware of undesirable activities (in or around the user)
- Application: Web security
  - Could these be stopped with Web-mining?
  - Bus-jack case, 2000, Japan
  - 9.11
Introduction

• Dialogue agent–companion needs to be aware of undesirable activities in or around the user

• Applications: Web security
  • Could these be stopped with Web-mining?

Bus-jack case, 2000, Japan

9/11

Demi Moore Saves A Teen Through Twitter 03.19
Introduction

• Dialogue agent–companion needs to be aware of undesirable activities (in or around the user)

• Application: Web security
  • Could these be stopped with Web-mining?

  • Bus-jack case, 2000, Japan
  • 9.11

We need an artificial Demi Moore!

• Applications:
  • Could these be stopped with Web-mining?
  • Bus-jack case, 2000, Japan
  • 9.11

Demi Moore Saves A Teen Through Twitter 03.19

A Teen's Dark Tweet To Demi

R U rly asking 4 help?RT @: i'm going 2 send a live feed of me hanging myself. No 1 cares if I die or not. about 14 hours ago via TweetDeck
New Threat: Cyber-Bullying

- cyber-bullying (or cyber-harassment, cyber-stalking)
  - Cyber-bullying happens ”when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person.”
    - The National Crime Prevention Council in America
  - cyber-bullying ”involves the use of information and communication technologies to support deliberate, repeated, and hostile behavior by an individual or group, that is intended to harm others.”
New Threat: Cyber-Bullying

• In Japan:
  – several suicide cases of cyber-bullying victims
  – Ministry of Education officially considers cyber-bullying a problem and produces a manual for spotting and handling the cyber-bullying cases.
    • Ministry of Education, Culture, Sports, Science and Technology, 2008:
      – 'Netto jou no ijime' ni kansuru taiou manyuaru jirei shuu (gakkou, kyouin muke)
      – ["Bullying on the net" Manual for handling and the collection of cases (directed to school teachers)] (in Japanese).
New Threat: Cyber-Bullying

• In Japan:
  – Volunteers (teachers, PTA members) started Online Patrol (OP) to spot CB cases...
  – But there is too much of it! (impossible deal with all of it manually)
New Threat: Cyber-Bullying

• In Japan:
  – Volunteers (teachers, PTA members) started Online Patrol (OP) to spot CB cases...
  – But there is too much of it! (impossible deal with all of it manually)

Need to help OP automatically spot Cyber-bullying
Cyber-bullying Detection Method

- Construction of lexicon of words distinguishable for cyber-bullying
- Estimation of word similarity (due to slang modifications of words)
- Classification of entries into harmful/non-harmful
- Ranking according to harmfulness
Lexicon Construction

• Words distinguishable for cyber-bullying = vulgarities
  – In English: f**ck, b*tch, sh*t, c*nt, etc..
  – In Japanese: uzai (freaking annoying), kimo (freaking ugly), etc.

↓

• Usually not recognized by parsers
Lexicon Construction

• Obtained Cyber-bullying data (from Online Patrol of Japanese secondary school sites)*

• Read and manually specified 216 distinguishable vulgar words.

• Added to parser dictionary:

<table>
<thead>
<tr>
<th>Example:</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>kimoi</em> (freaking ugly)</td>
</tr>
<tr>
<td>POS: Adjective;</td>
</tr>
<tr>
<td>Headword: <em>kimoi</em> (hit-rate: 294);</td>
</tr>
<tr>
<td>Reading: kimo;</td>
</tr>
<tr>
<td>Pronunciation: kimo;</td>
</tr>
<tr>
<td>Conjugated form: uninflected;</td>
</tr>
</tbody>
</table>

*) From Human Rights Research Institute Against All Forms for Discrimination and Racism-MIE, Japan
Similarity Estimation

- Jargonization (online slang)
  - English: “CU” (see you [later]), “brah” (bro[ther], friend)
  - Japanese:

<table>
<thead>
<tr>
<th>original word</th>
<th>colloquial transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>kimoi (freaking ugly, gross)</td>
<td>kimosu, kishoi, kisho, ...</td>
</tr>
<tr>
<td>uzai (freaking annoying)</td>
<td>uzee, UZAI, uzakko, ...</td>
</tr>
<tr>
<td>busaiku (ugly bitch)</td>
<td>buchaiku, bussaiku, ...</td>
</tr>
</tbody>
</table>

*Problem: The same words will not be recognized or will be recognized as separate words.*
Similarity Estimation

• Use Levenshtein Distance
  
  “The Levenshtein Distance between two strings is calculated as the minimum number of operations required to transform one string into another, where the available operations are only deletion, insertion or substitution of a single character.”

<table>
<thead>
<tr>
<th>transformed word</th>
<th>performed operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>kimosu</td>
<td>substitution of ’s’ to ’i’; distance = 1;</td>
</tr>
<tr>
<td>→ kimoiu</td>
<td>deletion of final ’u’; distance = 2;</td>
</tr>
<tr>
<td>→ kimoii</td>
<td></td>
</tr>
</tbody>
</table>

Similarity Estimation

- Add heuristic rules for optimization

<table>
<thead>
<tr>
<th>Rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. deletion syllable prolongations</td>
<td><em>kimoooi → kimo</em></td>
</tr>
<tr>
<td>2. unification of word first letter</td>
<td><em>In case of uzai we will consider only the words beginning with u</em></td>
</tr>
</tbody>
</table>

- With the threshold set on 2, the Precision before applying the rules was 58.9% and was improved to 85.0%.
SVM Classification

• Support Vector Machines (SVM) are a method of supervised machine learning developed by Vapnik [14] and used for classification of data.
• Training data: 966 entries (750 hamful, 216 non-harmful)
• Calculate result as balanced F-score (with Precision and Recall)
• Perform 10-fold cross validation on all data
SVM Classification

• 10-fold cross validation on all data
  – Divide data to 10 parts
  – Use 9 for training and 1 for test
  – Perform 10 times and take an approximation.

Precision=79.9%, Recall=98.3% \rightarrow F=88.2%
Affect Analysis of Cyber-Bullying Data

• The Affect Analysis system used:
  – ML-Ask:
    1. Determines Emotiveness
    2. Determines the types of emotions expressed

Affect Analysis of Cyber-Bullying Data

• The Affect Analysis system used:
  – ML-Ask:
    1. Emotiveness:
       1. Determine whether utterance is emotive (0/1)
       2. Calculate emotive value of an utterance (0-5)
       3. Number of emotive utterances in conversation
       4. Approx emotive value for all utterances
       5. Determine number of emotiveness’ features:
          • Interjections
          • Exclamations
          • Vulgarities
          • Mimetic expressions
Affect Analysis of Cyber-Bullying Data

- The Affect Analysis system used:
  - ML-Ask:
    2. Determines the types of emotions expressed:
       One of 10 emotion types said to be the most appropriate for the Japanese language:
       `ki/yorokobi` (joy, delight), `do/ikari` (anger), `ai/aware` (sadness, gloom), `fu/kowagari` (fear), `chi/haji` (shame, shyness), `ko/suki` (liking, fondness), `en/iya` (dislike), `ko/takaburi` (excitement), `an/yasuragi` (relief) and `kyo/odoroki` (surprise, amazement)

       Based on an emotive expression database

Affect Analysis of Cyber-Bullying Data

• Results
Affect Analysis of Cyber-Bullying Data

• Results

  1. Emotiveness:
     1. Determine whether utterance is emotive
     2. Calculate emotive value of an utterance
     3. Number of emotive utterances in conversation
     4. Approx emotive value for all utterances
     5. Determine number of emotiveness’ features:
        • Interjections
        • Exclamations
        • Vulgarities
        • Mimetic expressions

Many moderate proofs:
Harmful data is less emotive

There are two distinctive features
Affect Analysis of Cyber-Bullying Data

• Results

  2. Emotion types

  – More positive emotions in non-harmful data
  – Slightly more negative emotions in harmful data
  – Detailed analysis: fondness is often used in irony
Conclusions

• New problem: Cyber-Bullying
• Prototype Machine Learning Method for Cyber-bullying Detection
  – Results not ideal, but somewhat encouraging
• Affect Analysis of Cyber-Bullying Data
  – Cyber-bullying is less “emotive” (cold irony)
  – Distinctive features of CB: vulgarities, mimetic expressions
  – Expressions of emotions considered as positive are often used in ironic meaning
Future Works

• New vulgarities are created everyday
  – Create a method for extraction of vulgarities
  – Find a syntactic model of vulgar expression

• Implement in to a web crawler automatically performing online patrol (e.g. for school web sites)
Thank you for your attention.