Filtering Offensive Language in Online Communities using Grammatical Relations
Zhi Xu and Sencun Zhu

Introduction and Motivation

Online Communities
- People in the online social networking websites create social aggregations, called Online Communities
- Offensive language has become a big issue of online communities

Offensive language has spread into almost every corner of online communities

To the community
- Undermine the community’s reputation
- Drive users away

To the user
- Bring negative influence to user’s mental health, especially for youth and children

This work focuses on how to remove offensive language within user messages

The Offensive Language Filtering Problem

Existing automatic filtering approaches

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>“it is aston martin and you are a crying pig”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword Censoring Approach</td>
<td>“it is aston martin and you are a c**** p**”</td>
</tr>
<tr>
<td>Content Control Approach</td>
<td>“” (blocked)</td>
</tr>
</tbody>
</table>

- Break the readability of text
- Readers can easily guess the removed words

- Too coarse-grained
- Easy to bypass
- Inoffensive part may be removed falsely

Manual filtering (outputs the BEST filtering result)

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>“it is aston martin”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual Filtering Approach</td>
<td></td>
</tr>
</tbody>
</table>

- Remove offensive part precisely
- Inoffensive part remains
- The text after filtering is still readable

The “Filtering instead of blocking” philosophy

- Precisely identify all offensive contents and remove them semantically, so that viewers will not notice the existence of offensive language in the original sentence;
- Keep the readability and inoffensive content in the sentence, so that the author will still be allowed to express his opinion freely as long as it is not offensive;

A Sentence-level Semantic Filtering Approach

Step 1: Grammatical Analysis

- Parse Tree
- Root
- NP
- VP
- RelTree
- Clause Level
- Typed Dependency Relations

Step 2: Bottom-up Estimation

- Parse Tree
- Root
- NP
- VP
- Clause Level
- Word Level

Applications and Evaluations

Administrator side applications
- YouTube Dataset
  - 11670 text comments collected YouTube
  - 2063 sentences containing offensive words
  - Compare the proposed semantic filtering approach with manual filtering approach
    - Correct Filtering: 90.94%
    - Insufficient Filtering: 2.81%
    - Excessive Filtering: 6.25%

Browser side applications
- Firefox extension for parental control

Reference: This poster is based on the paper “Filtering Offensive Language in Online Communities using Grammatical Relations,” Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, July 13-14, 2010, Redmond, Washington, US

More information is available: http://www.cse.psu.edu/~zux103