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

Original Sentence	***"it is aston martin and you are a crying pig"***
Keyword Censoring Approach	***"it is aston martin and you are a c**** p**"*** •Break the readability of text •Readers can easily guess the removed words
Content Control Approach (thld=2)	***" " (blocked)*** •Too coarse-grained •Easy to bypass •Inoffensive part may be removed falsely

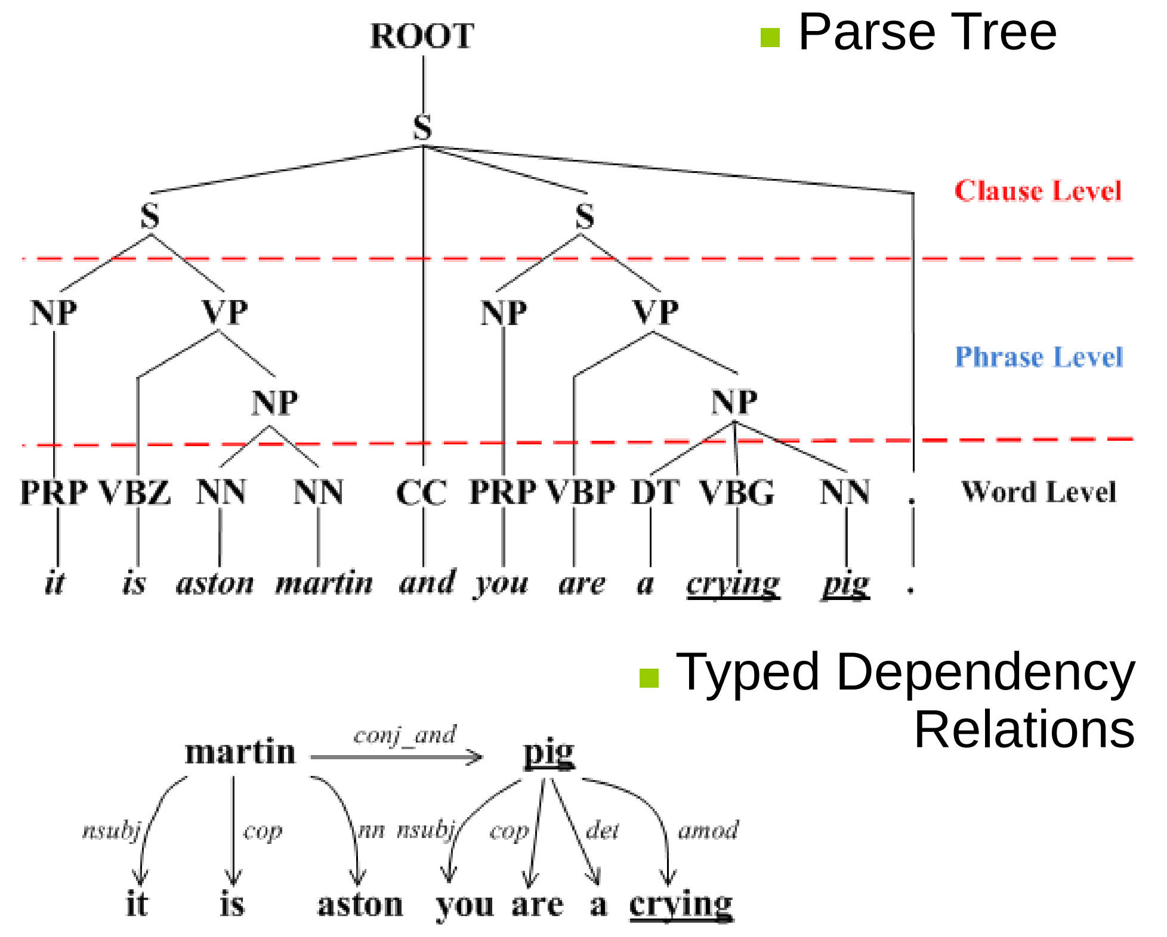
- Manual filtering (outputs the BEST filtering result)

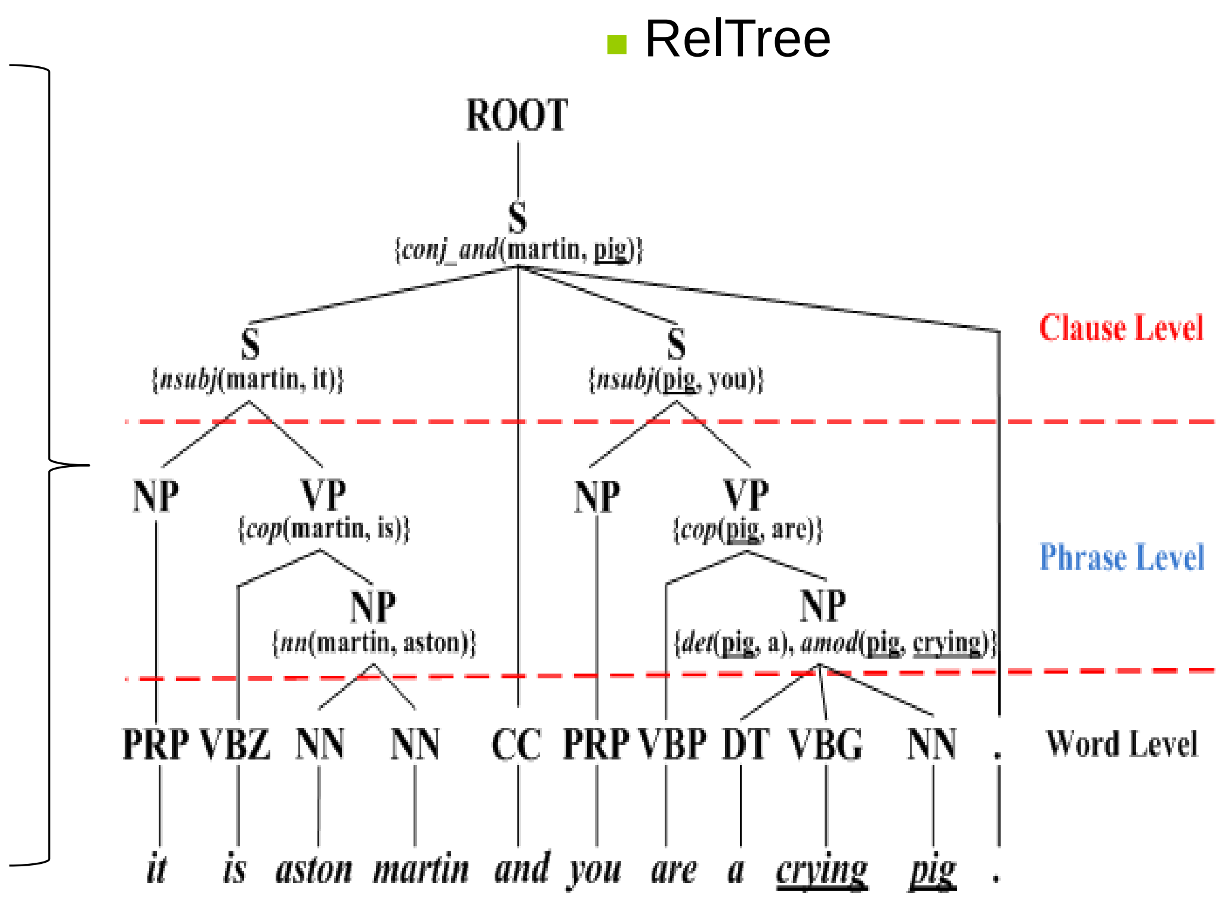
Manual Filtering Approach	***"it is aston martin"*** •Remove offensive part precisely •Inoffensive part remains •The text after filtering is still readable

- 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;

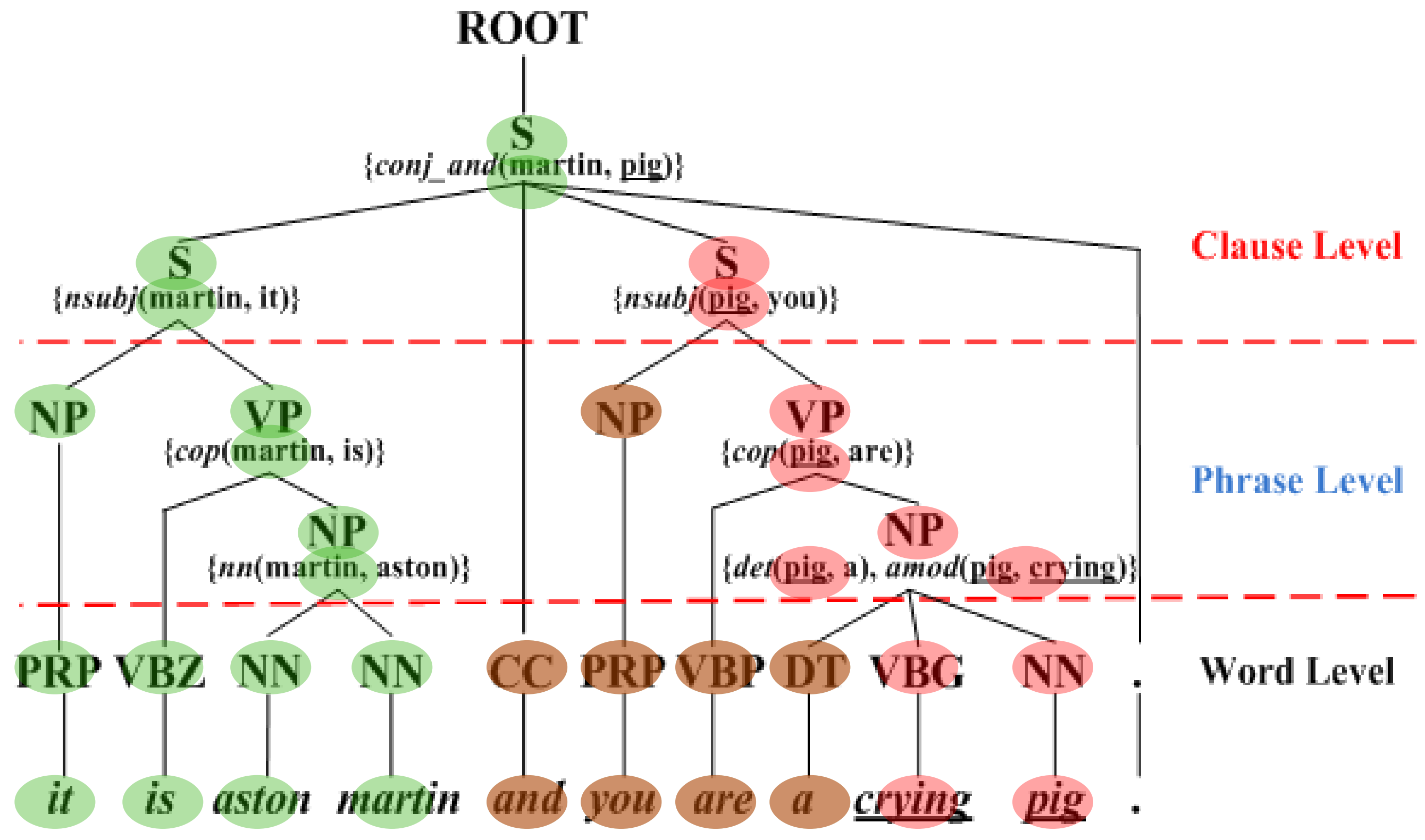
A Sentence-level Semantic Filtering Approach

- Step 1: Grammatical Analysis


Parse Tree
ROOT
S
S
S
Clause Level
NP
VP
NP
VP
Phrase Level
NP
NP
PRP VBZ NN NN CC PRP VBP DT VBG NN .
Word Level
it is aston martin and you are a crying pig .
Typed Dependency Relations
martin
conj_and
pig
nsubj
cop
nn
nsubj
cop
det
amod
it is aston you are a crying



- Step 2: Bottom-up Estimation

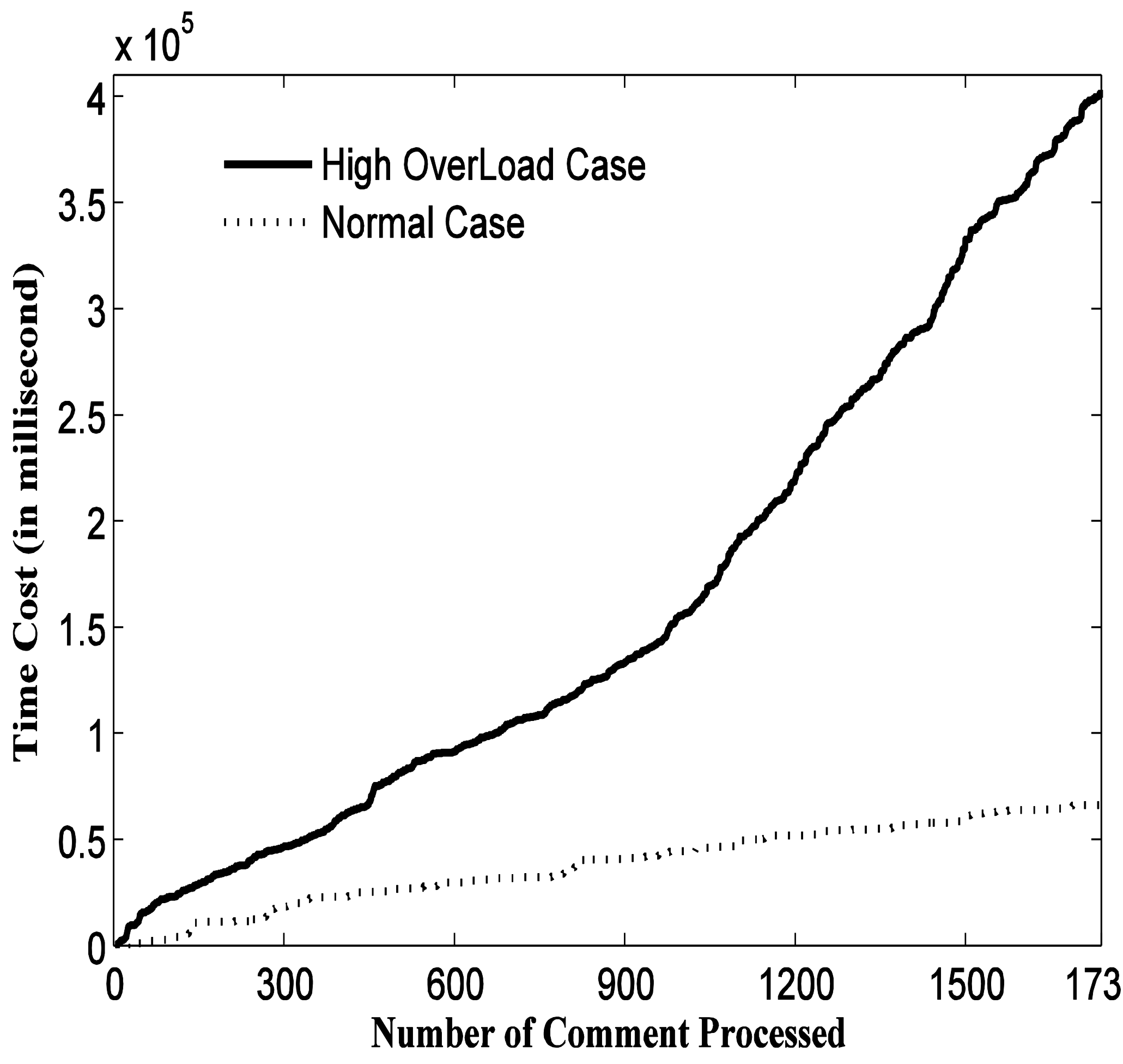


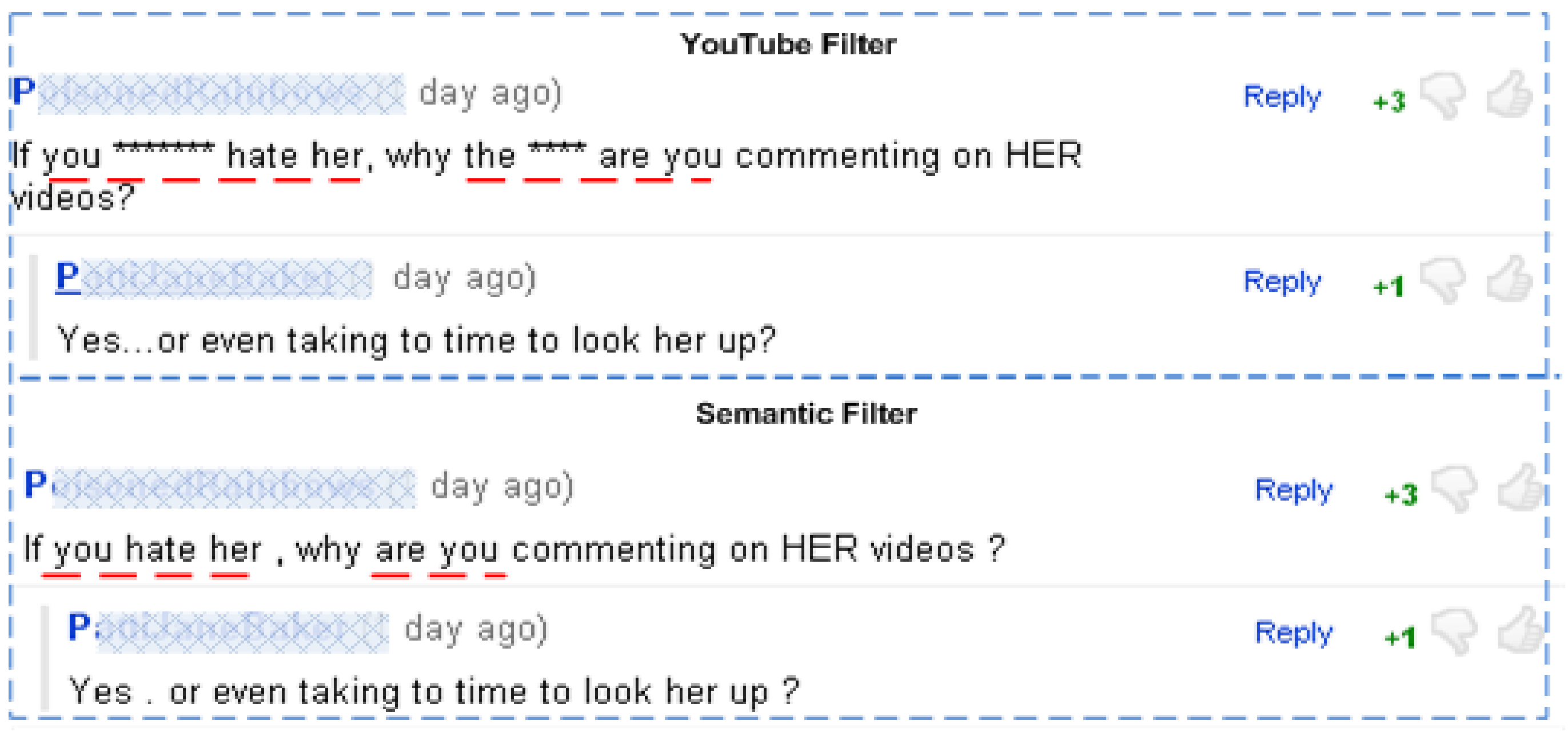
Applications and Evaluations

- 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





YouTube Filter

P[redacted] day ago) Reply +3

If you ******* hate her, why the **** are you commenting on HER videos?

P[redacted] day ago) Reply +1

Yes…or even taking to time to look her up?

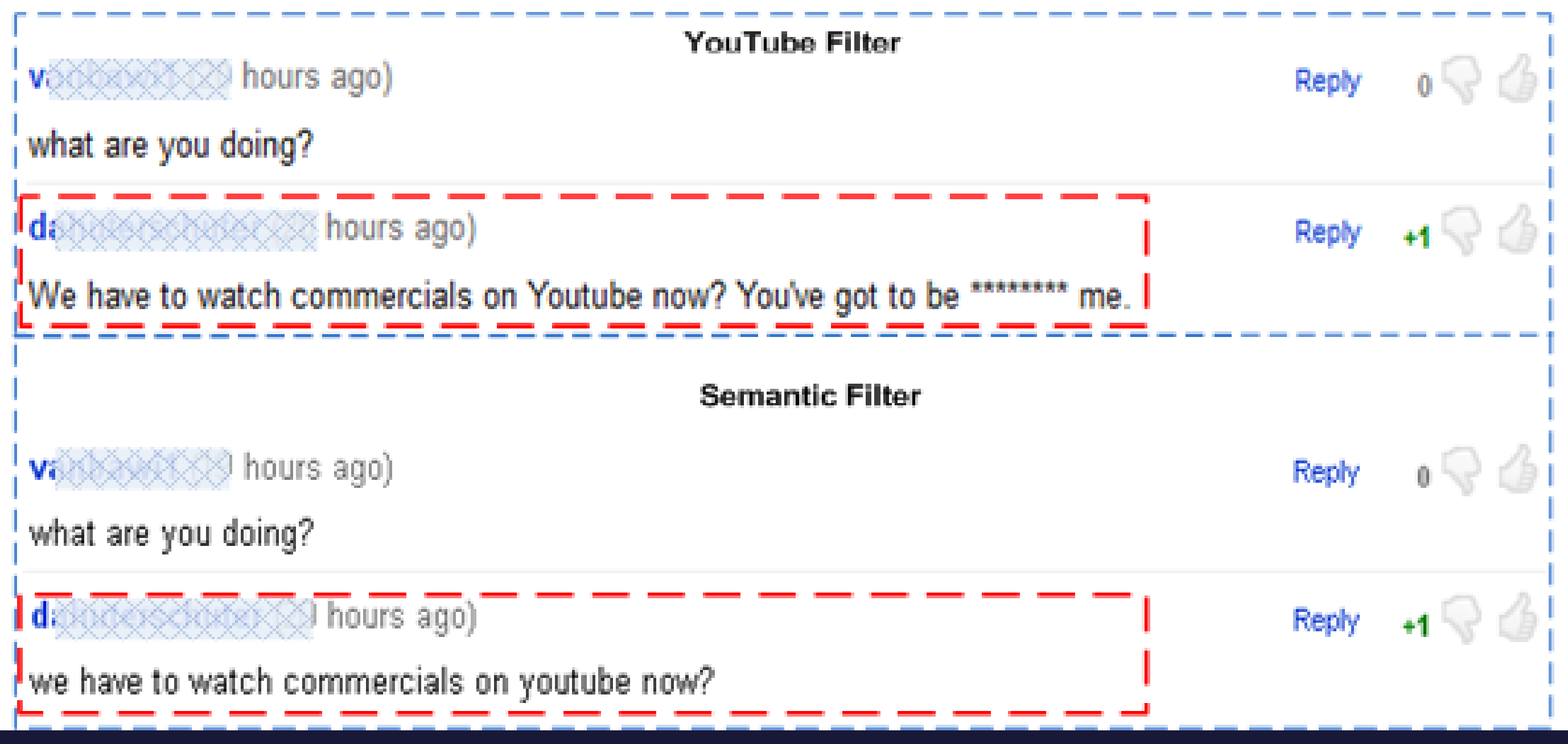
Semantic Filter

P[redacted] day ago) Reply +3

If you hate her , why are you commenting on HER videos ?

P[redacted] day ago) Reply +1

Yes . or even taking to time to look her up ?



YouTube Filter

v[redacted] hours ago) Reply 0

what are you doing?

d[redacted] hours ago) Reply +1

We have to watch commercials on Youtube now? You've got to be ******** me.

Semantic Filter

v[redacted] hours ago) Reply 0

what are you doing?

d[redacted] hours ago) Reply +1

we have to watch commercials on youtube now?

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