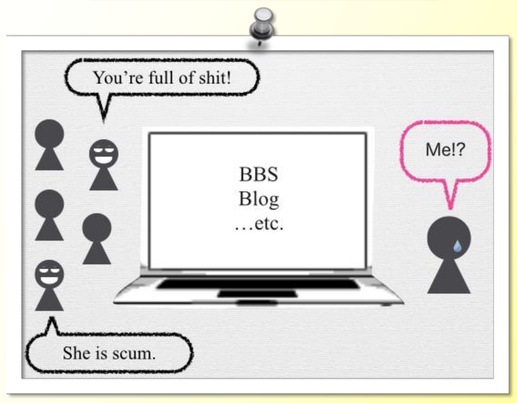
Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information

Suzuha Hatakeyama — Kitami Institute of Technology
Fumito Masui — Kitami Institute of Technology
Michal Ptaszynski — Kitami Institute of Technology
Kazuhide Yamamoto — Nagaoka University of Technology

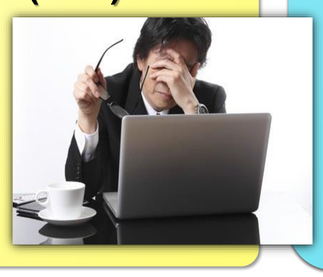
Background



CYBERBULLYING
Recently noticed social problem

INTERNET PATROL

- Internet monitoring by Parent-Teacher Association (PTA).
- Request site admin to remove harmful entries.
- High cost of time and fatigue for net-patrol members.



Category Relevance Maximization Method

Phrase Extraction
Extract *phrases* from sentence using dependency relations

(Noun, Noun)
(Noun, Verb)
(Noun, Adjective)

Ex. *"Cute girl, but bad personality."*
(cute, girl), (bad, personality)

Relevance Estimation
Calculate relevance of each phrase to seed words

Estimation Model (extended Turney's SO-PMI [3])

$$score = max\,(max\,(PMI(p_i, w_j)))$$

Maximize category relevance of phrase p_i to seed word w_j

Seed words
Typical words related to cyberbullying

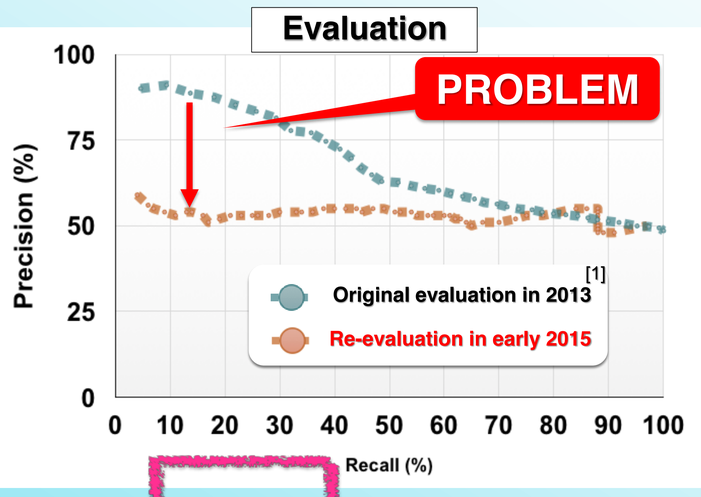
Category1	Category2	Category3
Obscene words	Violent words	Abusive words
Sex	Die	Annoying
Slut	Kill	Gross
Bl*wjob	Slap	Ugly

Defined by MEXT [4]

Our research

Help Internet Patrol with ICT	→ Automatic detection of cyberbullying entries
Performance improvement of method by Nitta et al. [1]	→ Automatic acquisition and update of seed words

Evaluation



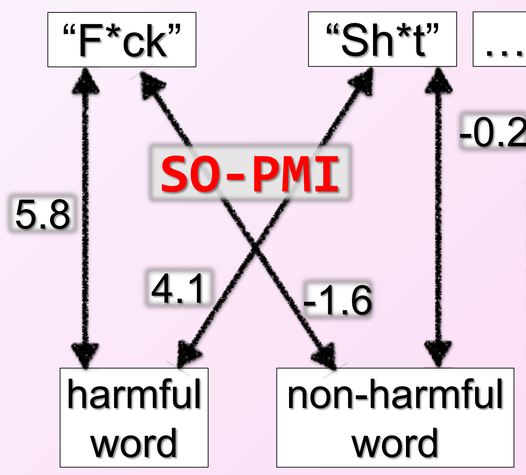
Possible reasons
(1) page rankings change
(2) net-patrol movement
(3) usage policies tightening

Automatic Acquisition of Seed Words

Primary Filtering (cleaning)

Nitta's [1] seed words ✕ Ishizaka's [2] seed words ✕ Non-harmful words

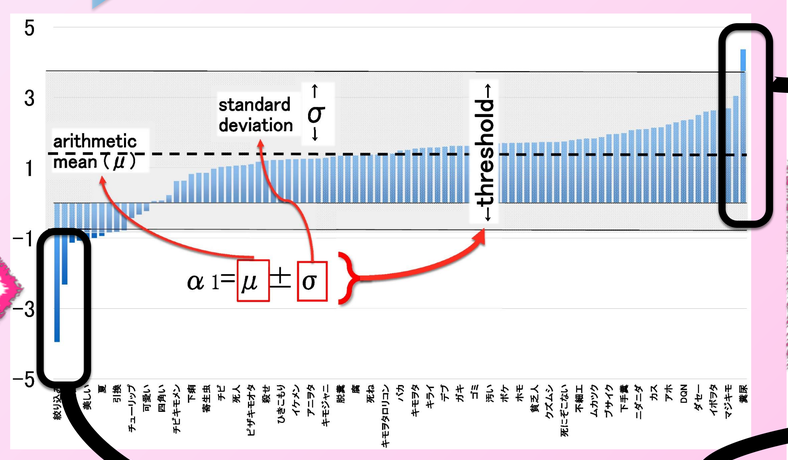
"F*ck" "Sh*t" …



SO-PMI>0 ↓ harmful

Secondary Filtering (optimizing)

Filtered seed words ✕ Harmful word candidates



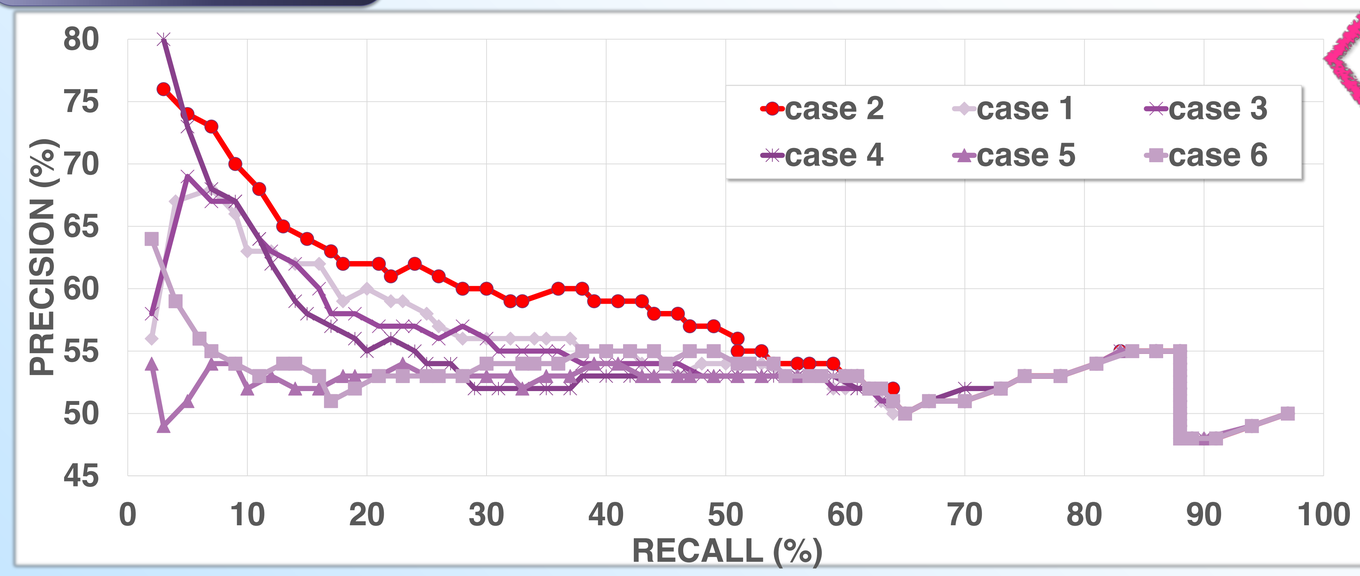
above threshold = harmful

Seed word candidate

case1 : 7 words	obtained after Primary Filtering with [2] 17 words
case2 : 12 words	obtained after Primary Filtering with [1] 9 words
case3 : 16 words	7 words from case1 + original 9 words [1]
case4 : 21 words	12 words from case2 + original 9 words [1]
case5 : 5 words (baseline 1)	originally used by [2]
case6 : 9 words (baseline 2)	originally used by [1]

Apply in Classification

Evaluation



Evaluation criteria:
① Highest F-score for longest threshold
② Highest break even point (BEP) of P&R
③ Highest Precision in general
④ Largest area under the curve (AUC) of P&R (same as in [1])
⑤ Always better to simply add words?

question	winner					
	case 1	case 2	case 3	case 4	case 5	case 6
①		●			●	●
②	●	●				●
③		●		●		
④		●				
⑤	●	●				

McNemar test

	case1	case2	case3	case4	case5	case6
case5	189.00 ***	26.88 ***	0.83	0.30	–	16.98 ***
case6	145.00 ***	5.80 *	9.47 **	10.29 **	18.56 ***	–

* p<0.5, ** p<0.1, *** p<0.01

Conclusions

→ **Best performance was achieved by filtering methods (case1 and case2)**
→ **Seed word filtering increases performance in general**
→ **But too much was no good (Only secondary filtering was better than Primary + Secondary)**
→ **Single filtering was also more time efficient**
→ **Simply adding more words does not increase performance**

REFERENCES

[1] Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, Kenji Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization**. In *Proc. of IJCNLP 2013*, pp. 579-586.
[2] Tatsuya Ishizaka, Kazuhide Yamamoto. 2011. **Automatic detection of sentences representing slandering on the Web** (In Japanese). *Proc. of NLP2011*, pp. 131-134.
[3] Peter D. Turney. 2002. **Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews**. In *Proc. of ACL-2002*, pp. 417-424, 2002.
[4] Ministry of Education, Culture, Sports, Science and Technology (MEXT). 2008. **"Bullying on the Net" Manual for handling and collection of cases (for schools and teachers)** (in Japanese). Published by *MEXT*.