Kitami Institute of Technology

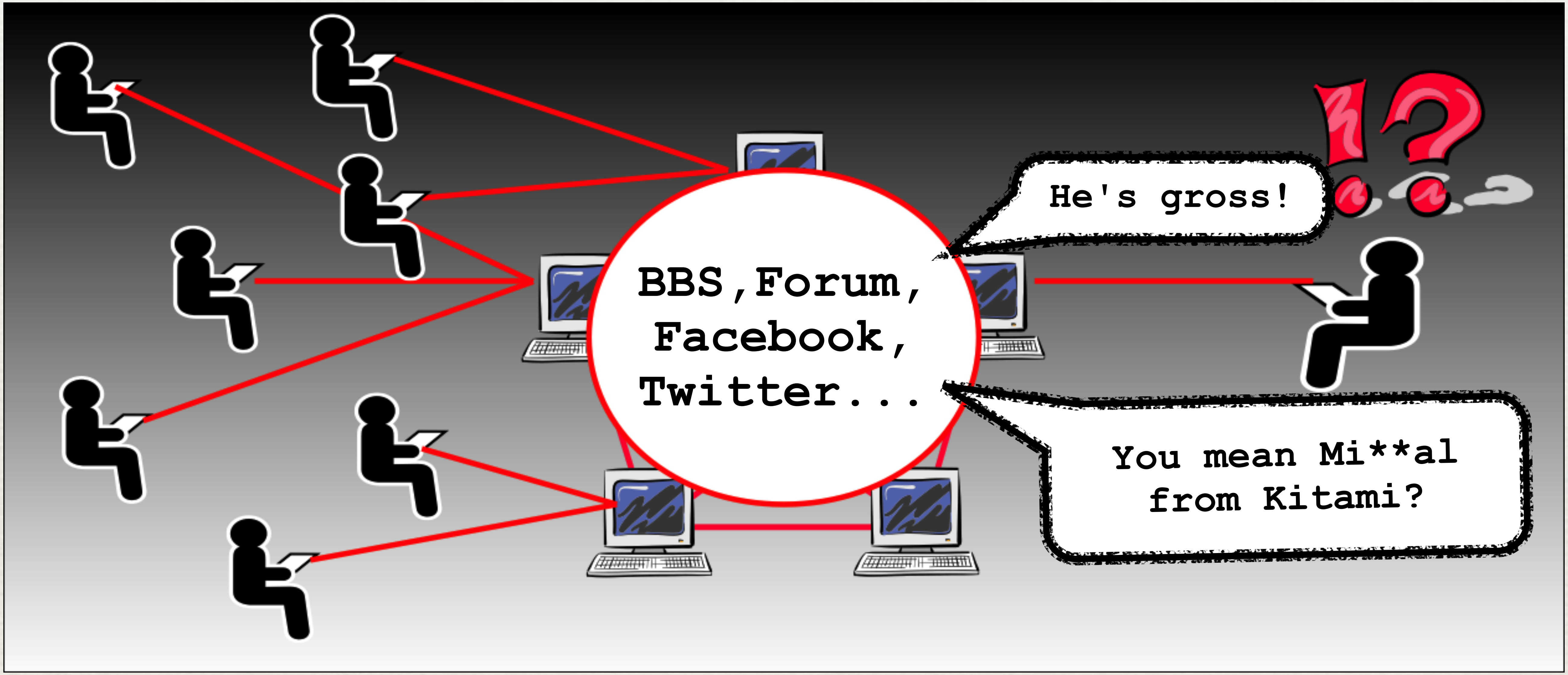
Automatically detecting cyberbullying on the Internet using methods from the fields of Artificial Intelligence and Natural Language Processing

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

Outline

1. Introduction
2. Affect analysis of cyberbullying data
3. Lexicon construction
4. Word similarity estimation
5. Classification
   - SVM-based method
   - PMI-IR-based method
6. Conclusions and Future work

Introduction



Cyberbullying (slandering and humiliating people on the Internet) is a new social problem.

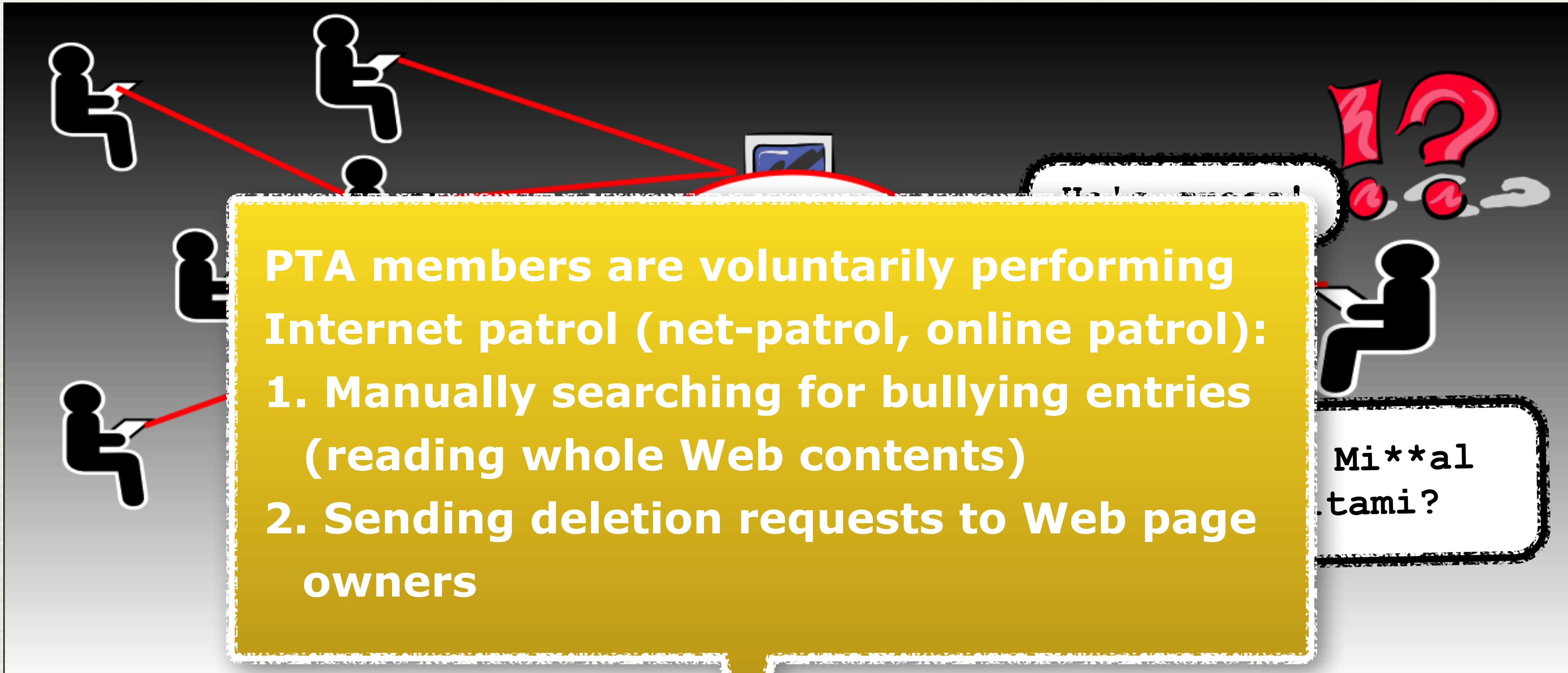
Introduction

- cyberbullying (or cyber-harassment, cyber-stalking)
  - Cyberbullying 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
  - cyberbullying ”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.”
    - B. Belsey. Cyberbullying: An Emerging Threat for the ”Always On” Generation, http://www.cyberbullying.ca/pdf/Cyberbullying Presentation Description.pdf

Introduction

- In Japan:
  - several suicide cases of cyberbullying victims
  - Ministry of Education officially considers cyberbullying a problem and produces a manual for spotting and handling the cyberbullying 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 Internet" Manual for handling and the collection of cases (directed to school teachers)] (in Japanese).

Introduction



PTA members are voluntarily performing Internet patrol (net-patrol, online patrol):

1. **Manually searching for bullying entries (reading whole Web contents)**
2. **Sending deletion requests to Web page owners**

Cyberbullying (slandering and humiliating people on the Internet) is a new social problem.

Introduction

Problems with net-patrol

- It is performed **manually**.
- There are **38,620** unofficial school Web sites (state for 2008.08).

Introduction

Problems with net-patrol

- It is performed **manually**.
- There are **38,620** unofficial school Web sites (state for 2008.08).

There is too much of it! (impossible to deal with all of it manually)

Introduction

Problems with net-patrol

- It is performed **manually**.
- There are **38,620** unofficial school Web sites (state for 2008.08).

There is too much of it! (impossible to deal with all of it manually)

Need to help net-patrol members by automatically spotting cyberbullying entries

Affect Analysis of Cyberbullying Data

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

M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -', In Proceedings of The Conference of the Pacific Association for Computational Linguistics 2009 (PACLING-09), pp. 223-228 (2009).

Affect Analysis of Cyberbullying 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 Cyberbullying 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:

       喜 ki/yorokobi (**joy**, delight), 怒 do/ikari (**anger**), 哀 ai/aware (**sadness**, gloom), 怖 fu/kowagari (**fear**), 恥 chi/haji (**shame**, shyness), 好 ko/suki (liking, **fondness**), 厭 en/iya (**dislike**), 昂 ko/takaburi (**excitement**), 安 an/yasuragi (**relief**) and 驚 kyo/odoroki (**surprise**, amazement)

       Based on an emotive expression database

A. Nakamura. Kanjo hyogen jiten [Dictionary of Emotive Expressions] (in Japanese), Tokyodo Publishing, Tokyo, 1993.

Affect Analysis of Cyberbullying Data

- Results

Affect Analysis of Cyberbullying Data

- Results

  1. Emotion types

  – More positive emotions in non-harmful data

  – Slightly more negative emotions in harmful data

  – Detailed analysis: fondness is often used in irony

*) results not significant

Affect Analysis of Cyberbullying Data

- Results

  2. Emotiveness:

Many moderate proofs:
Harmful data is less emotive

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

There are two distinctive features

Affect Analysis of Cyberbullying Data

- Results

  2. 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 fo
     5. Determine numbe
        - Interjections
        - Exclamations
        - **Vulgarities**
        - Mimetic expressions

Many moderate proofs:
Harmful data is less emotive

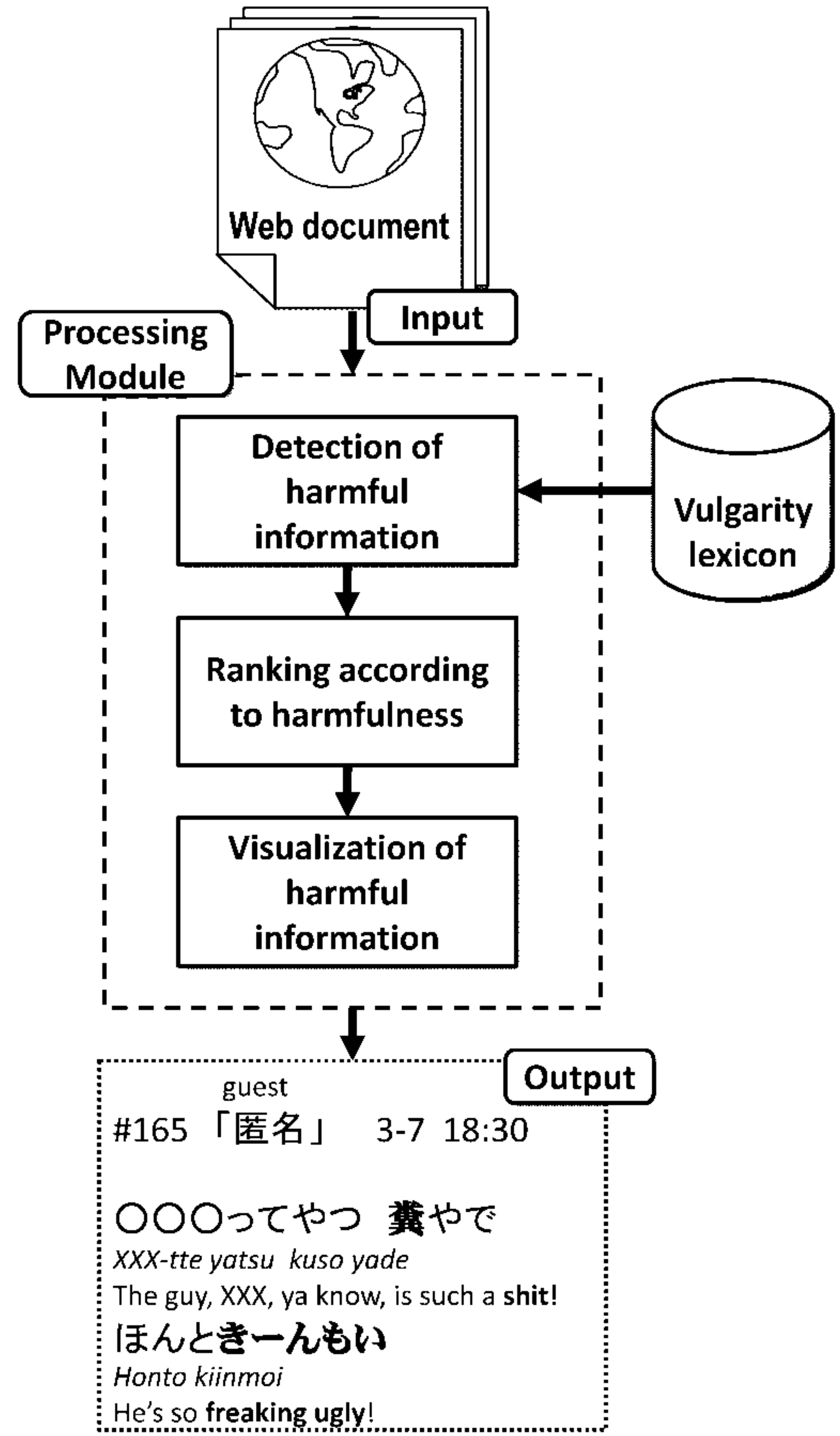
Focused of vulgarities (largest difference)

There are two distinctive features

Cyberbullying Detection

Cyberbullying Detection Method

1. Construction of lexicon of words distinguishable for cyber-bullying
2. Estimation of word similarity (due to slang modifications of words)
3. Classification of entries into harmful/non-harmful
4. 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), kimoi (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:

Example:

kimoi (freaking ugly)
POS: Adjective;
Headword: *kimoi* (hit-rate: 294);
Reading: kimoi;
Pronunciation: kimoi;
Conjugated form: uninflected;

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

Similarity Estimation

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

original word	**colloquial transformation**
kimoi (freaking ugly, gross)	*kimosu*, *kishoi*, *kisho*, ...
uzai (freaking annoying)	*uzee*, *UZAI*, *uzakkoi*, ...
busaiku (ugly bitch)	*buchaiku*, *bussaiku*, ...

*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."

transformed word		performed operation
	kimosu	
→	*kimoiu*	substitution of 's' to 'i'; distance = 1;
→	*kimoi*	deletion of final 'u'; distance = 2;

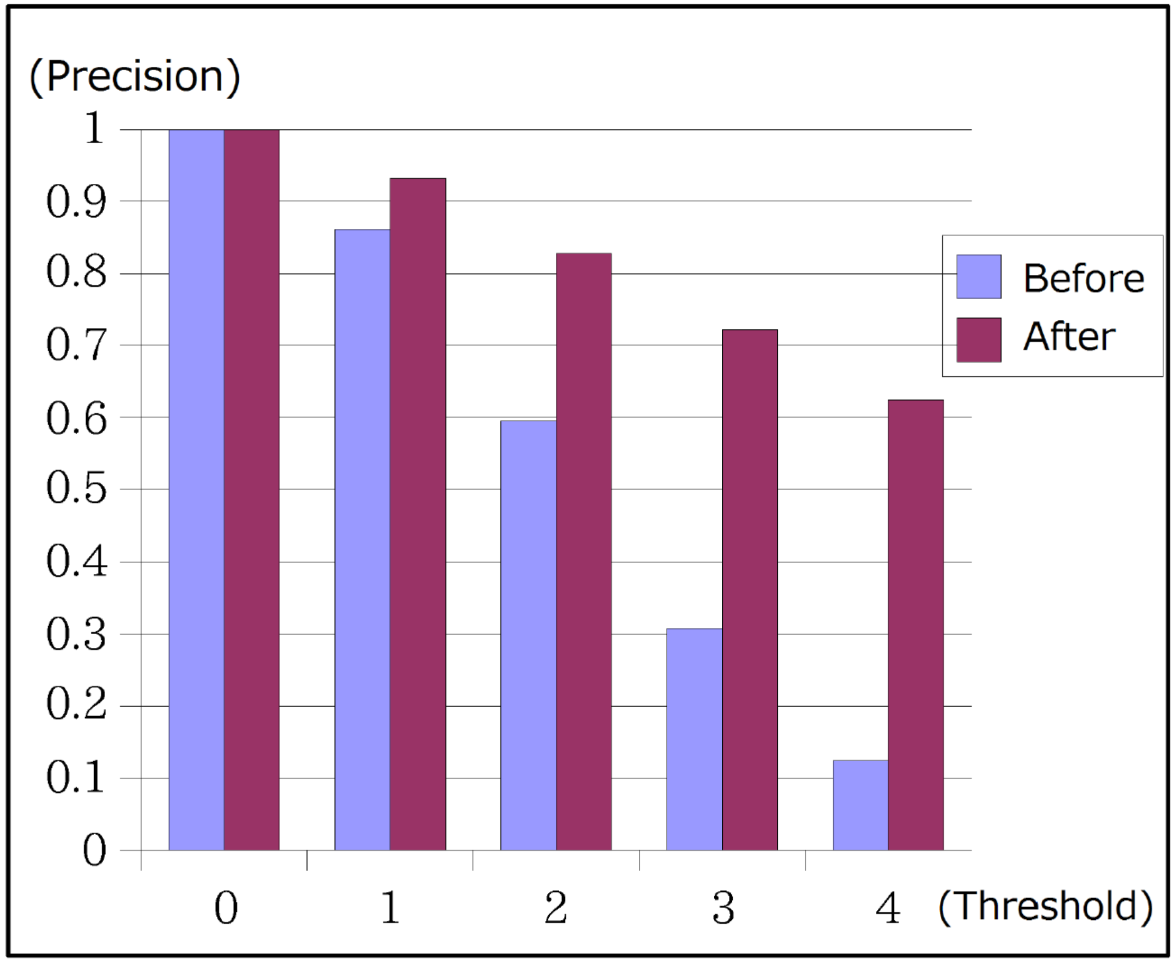
V. I. Levenshtein. Binary Code Capable of Correcting Deletions, Insertions and Reversals. Doklady Akademii Nauk SSSR, Vol. 163, No. 4, pp. 845-848 (1965).

Similarity Estimation

- Add heuristic rules for optimization

Rule	**Example**
1. deletion syllable prolongations	*kimoooi* → *kimoi*
2. unification of word first letter	In case of *uzai* we will consider only the words beginning with *u*

- 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 and used for classification of data.
- Training data: 966 entries (750 hamful, 216 non-harmful, *later added data to 1. have about 50/50 harmful and non-harmful, and 2. doubled the number of data)
- 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%, F=88.2%

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%, F=88.2%

Problem: Adding more data lowers results

PMI-IR Classification

Tatsuaki Matsuba, Fumito Masui, Atsuo Kawai, Naoki Isu. 2001. *Gakkou hikoushiki saito ni okeru yuugai jouhou kenshutsu wo mokuteki to shita kyokusei hantei moderu ni kansuru kenkyu* [**Study on the polarity classification model for the purpose of detecting harmful information on informal school sites**] (in Japanese), In *Proceedings of The Seventeenth Annual Meeting of The Association for Natural Language Processing (NLP2011)*, pp. 388-391.

PMI-IR Classification

$$\text{PMI-IR}(\boxed{word}) = \text{PMI}(\boxed{word}, \boxed{excellent}) - \text{PMI}(\boxed{word}, \boxed{poor})$$

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews [2002, Turney]

PMI-IR Classification

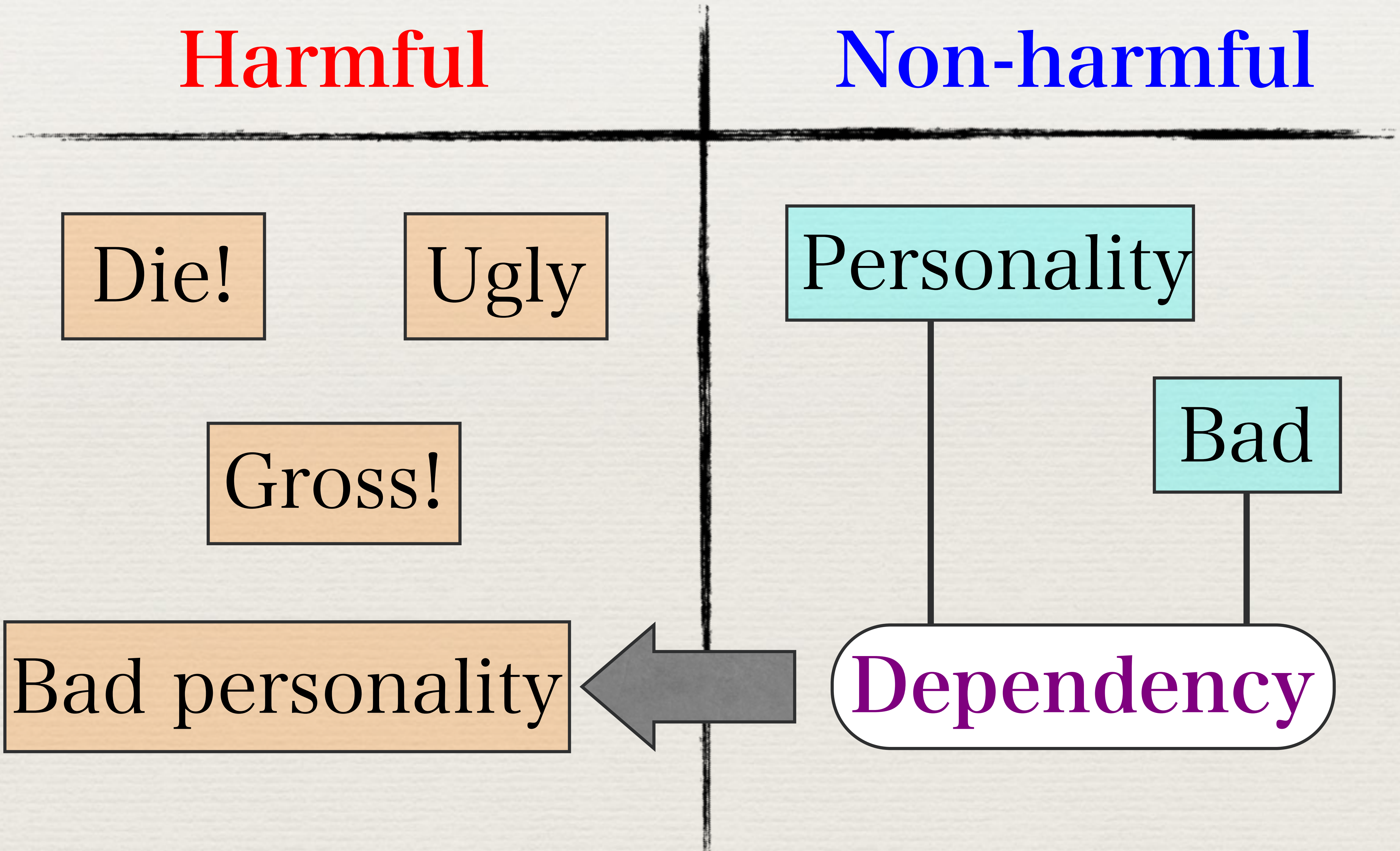
PMI-IR(word) = PMI(word,excellent) – PMI(word,poor)

PMI-IR(phrase) = max(PMI(phrase, "die","kill","slap"*) , …*
… ,
PMI(phrase,"annoying","gross","ugly"*))*

Phrase

Harmful polarity words

Phrase


Harmful
Non-harmful
Die!
Ugly
Gross!
Bad personality
Personality
Bad
Dependency

Phrase

Harmful entry

Phrase patterns

- **Noun**-**noun** ex.:**monkey face**
- **Noun**-**verb** ex.:**kare**wo**korosu** (**kill him**)
- **Noun**-**adjective** ex.:**seikaku** ga **warui** (**bad personality**)

PMI-IR Classification

$$\text{PMI-IR}(\boxed{word}) = \text{PMI}(\boxed{word}, \boxed{excellent}) - \text{PMI}(\boxed{word}, \boxed{poor})$$

$$\text{PMI-IR}(\boxed{phrase}) = \max(\text{PMI}(\boxed{phrase}, \boxed{\text{"die","kill","slap"}}), \ldots, \ldots, \text{PMI}(\boxed{phrase}, \boxed{\text{"annoying","gross","ugly"}}))$$

Phrase

Harmful polarity words

Harmful polarity words

Harmful words

- ugly hag
- gross
- b*tch

etc...

Add them to morphological analyzer (255 w)

Prevent morphological analysis Errors

Harmful polarity words

Harmful words

- ugly hag
- gross
- b*tch

etc…

Obscene

Violent

Abusive

Ministry of Education, Culture, Sports, Science and Technology (MEXT). 2008. ‘Netto-jou no ijime' ni kansuru taiou manyuaru jirei shuu (gakkou, kyouin muke) [”Bullying on the Net” Manual for handling and collection of cases (for schools and teachers)] (in Japanese). Published by MEXT.

Harmful polarity words

Harmful words

- ugly hag
- gross
- b*tch

etc…

Obscene

"sex", "slut", "fellatio"

Violent

"die [imp.]", "kill","slap"

Abusive

"annoy-ing","gross", "ugly"

3 most frequent words for each category

Harmful polarity words

PMI-IR Classification

Tested two versions of the method

1. Taking average of all scores
2. Taking maximized scores

Tested methods on two datasets

1. Similar distribution of data (cyb./non-cyb.)
2. Distribution similar to reality

PMI-IR Classification

PMI-IR(phrase) = max(PMI(phrase, "die","kill","slap"*) , …*
… ,
PMI(phrase,"annoying","gross","ugly"*))*

PMI-IR Classification

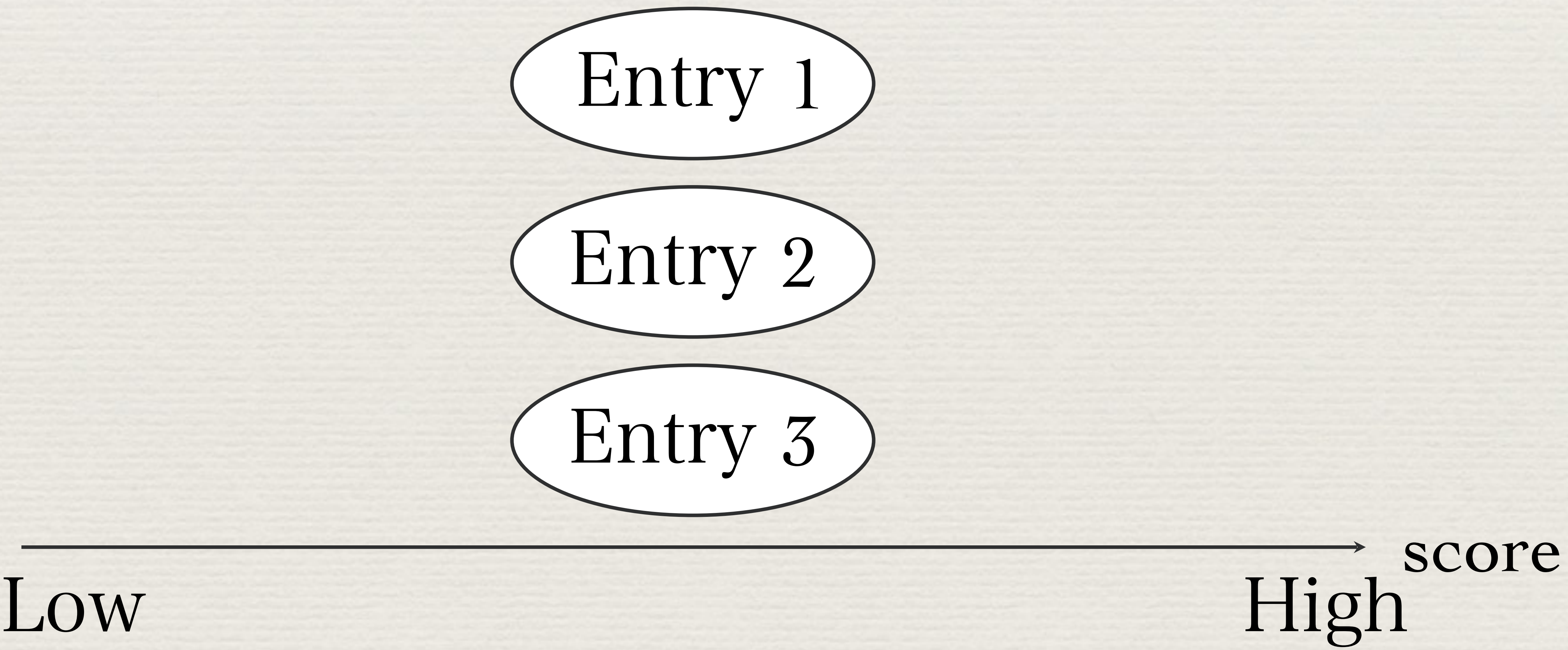
PMI-IR(phrase) = max(PMI(phrase, "die","kill","slap"*) , …*
… ,
PMI(phrase,"annoying","gross","ugly"*))*

score = max $\boxed{\textbf{\textit{(PMI-IR(phrase))}}}$

score : Harmful polarity score
of an entry

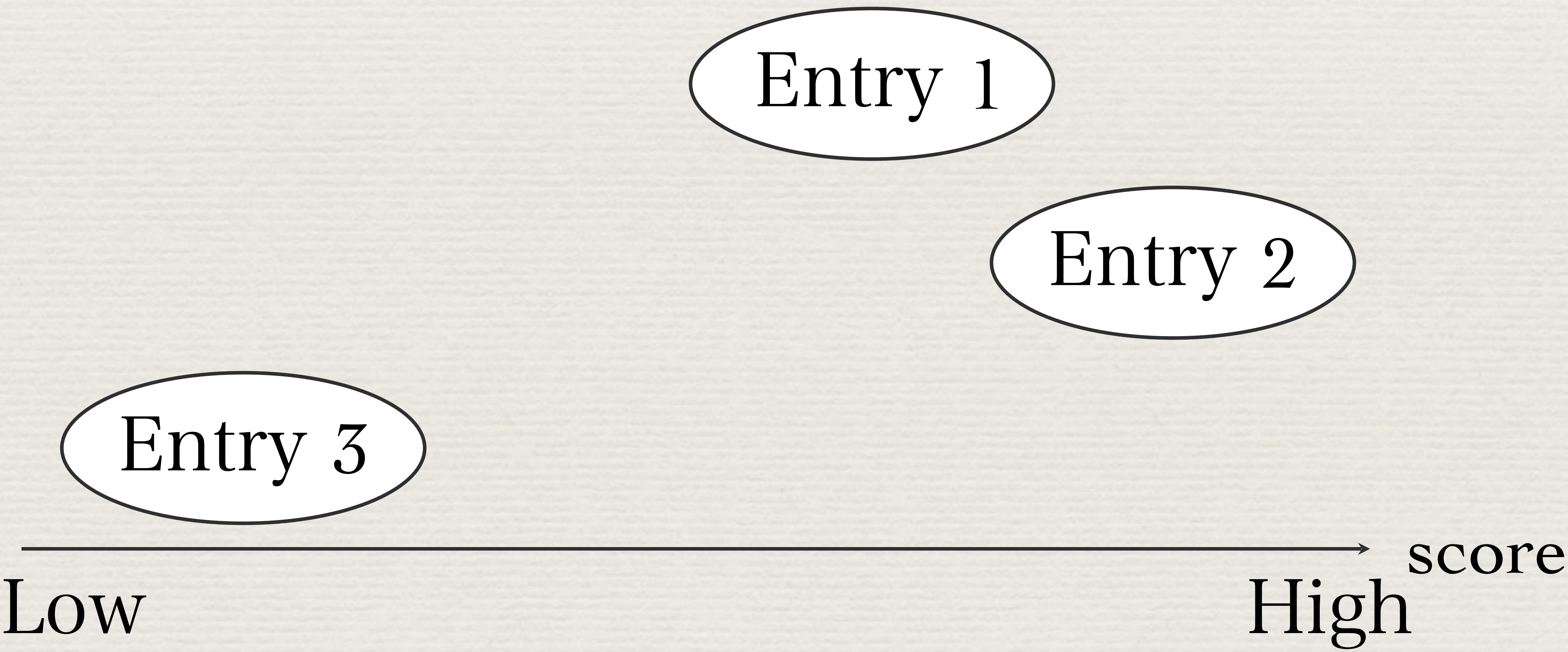
Harmful polarity estimation

score = max(PMI-IR(phrase))



Harmful polarity estimation

score = max(PMI-IR(phrase))



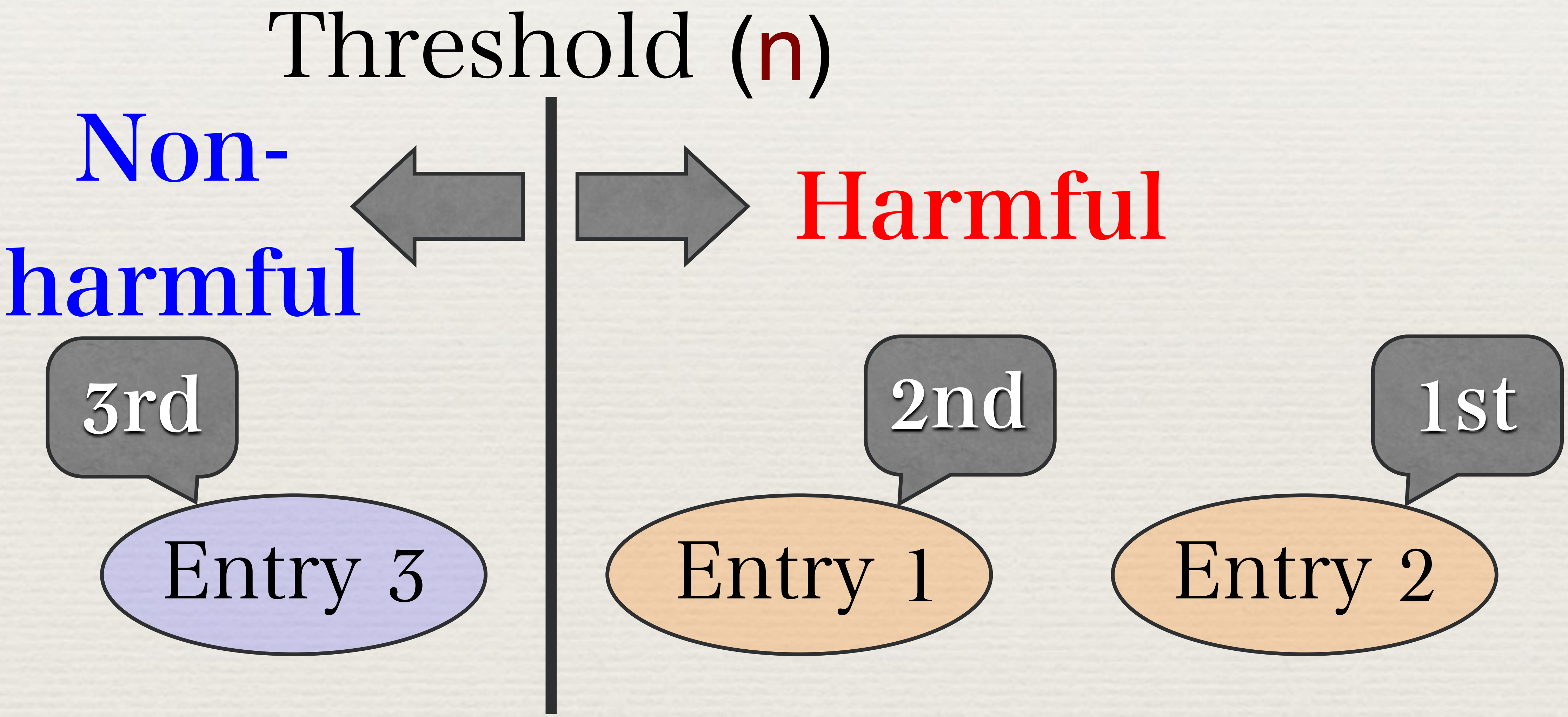
Harmful polarity estimation

Entry 3

Entry 1

Entry 2

Harmful polarity estimation



Test dataset - study

Goal

Find the actual amount of harmful entries of Web pages (school unofficial BBS)

Data

Three random school unofficial BBS

Time

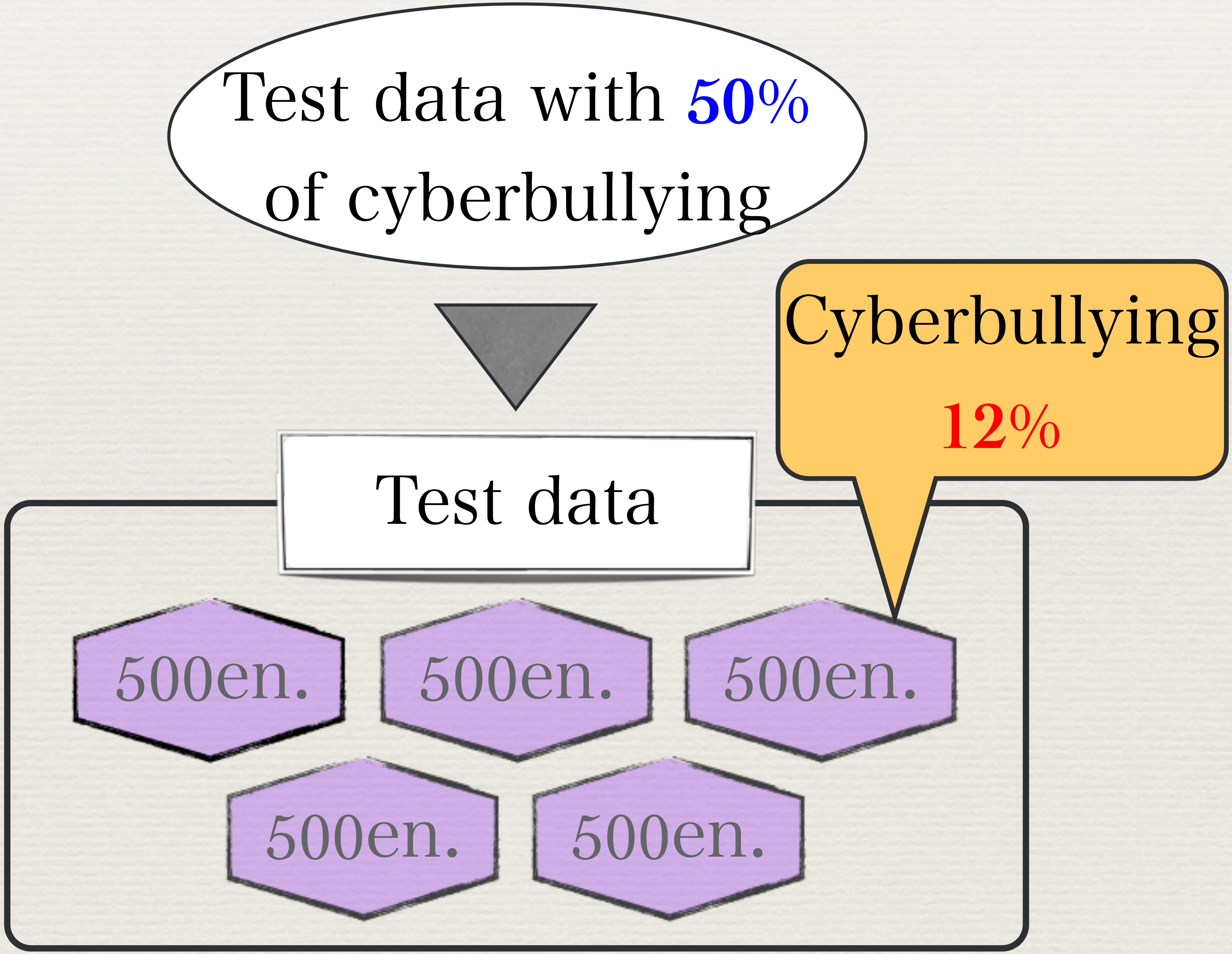
2012/01/27~2012/01/30

Test dataset - study

BBS	Overall number of entries	Cyberbullying entries	Percentage(%)
BBS(1)	600	75	12.5
BBS(2)	736	90	12.2
BBS(3)	886	100	11.3

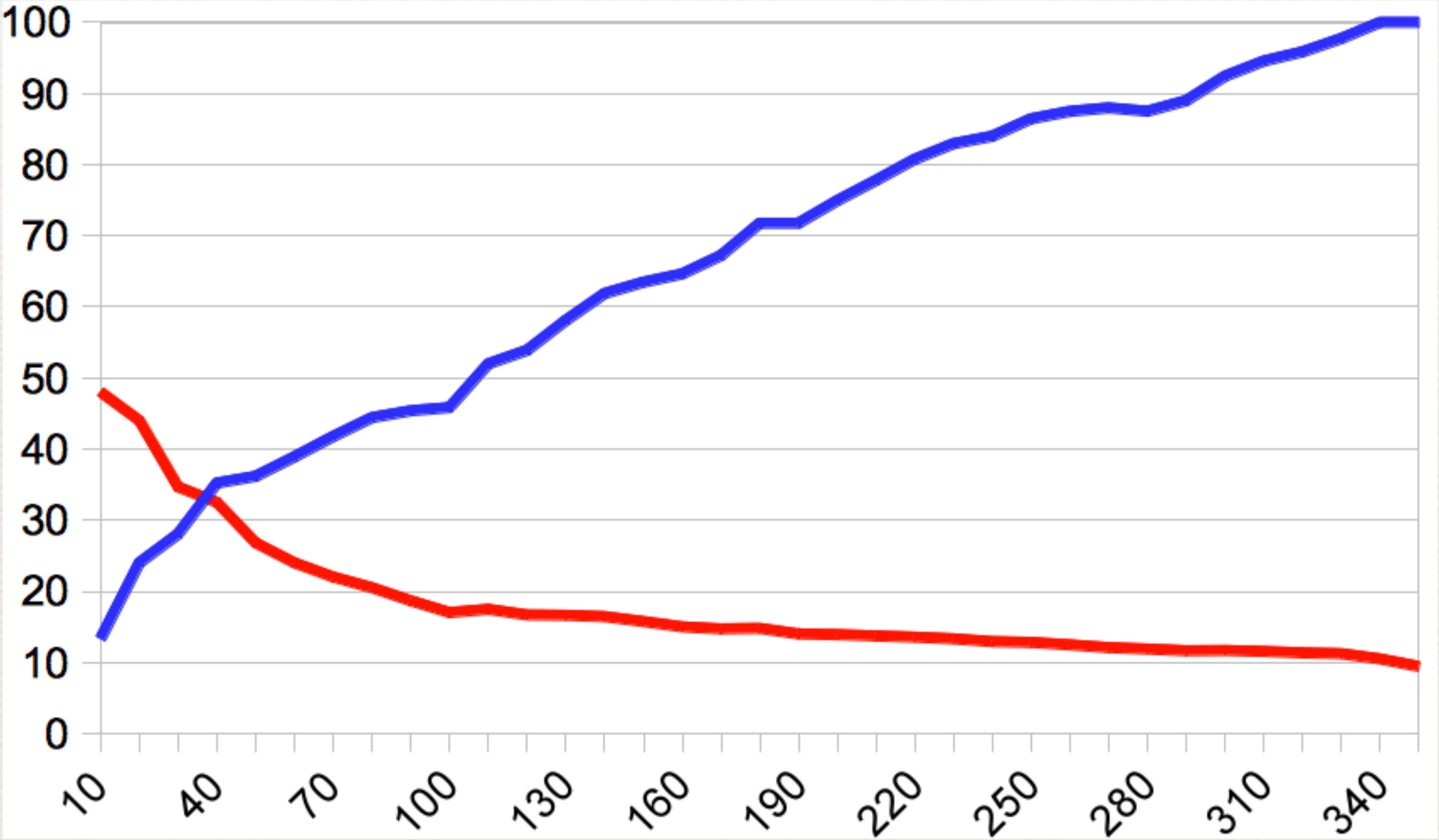
Harmful entries = **12%**

Preparation of test data



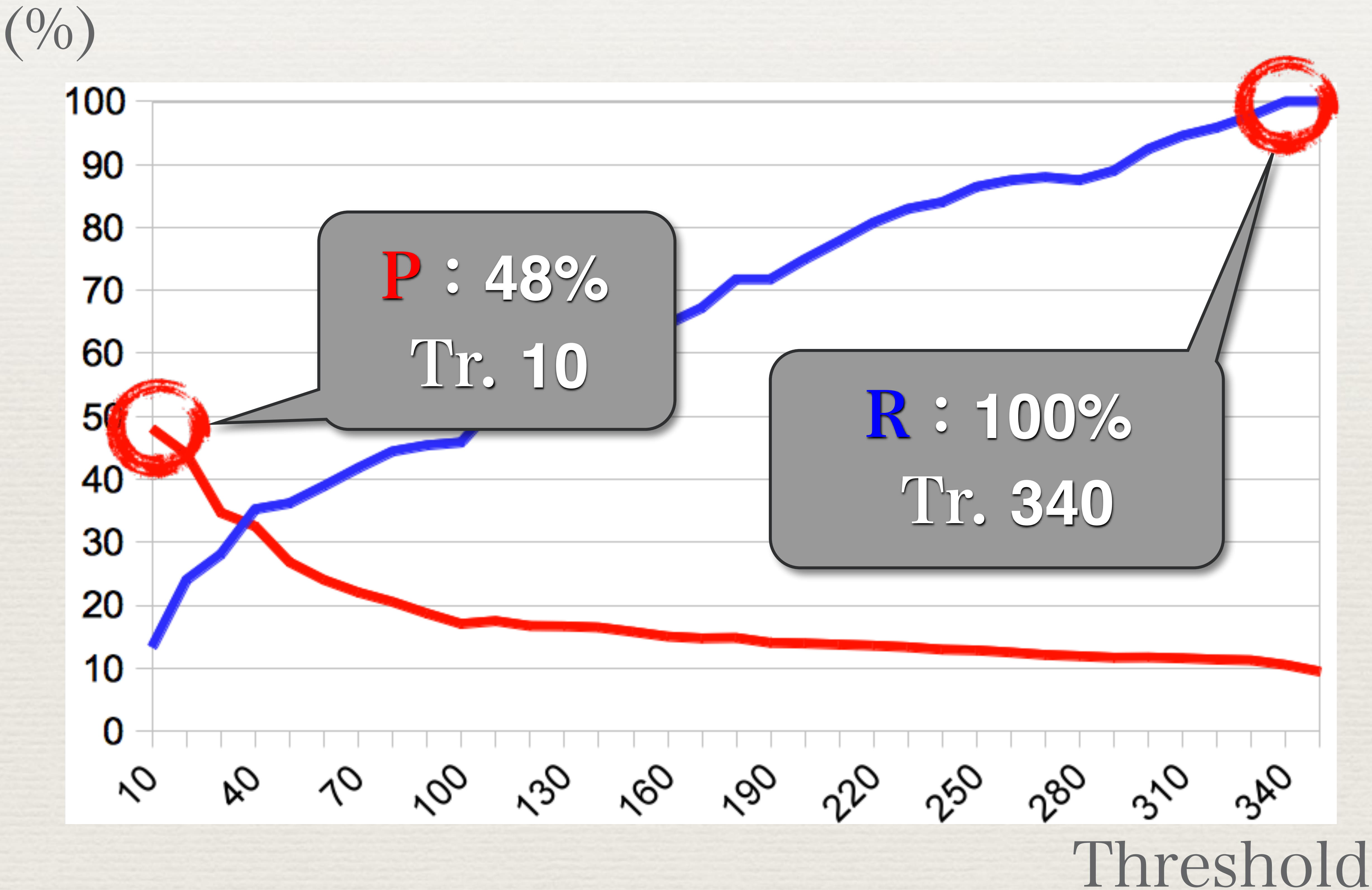
Results

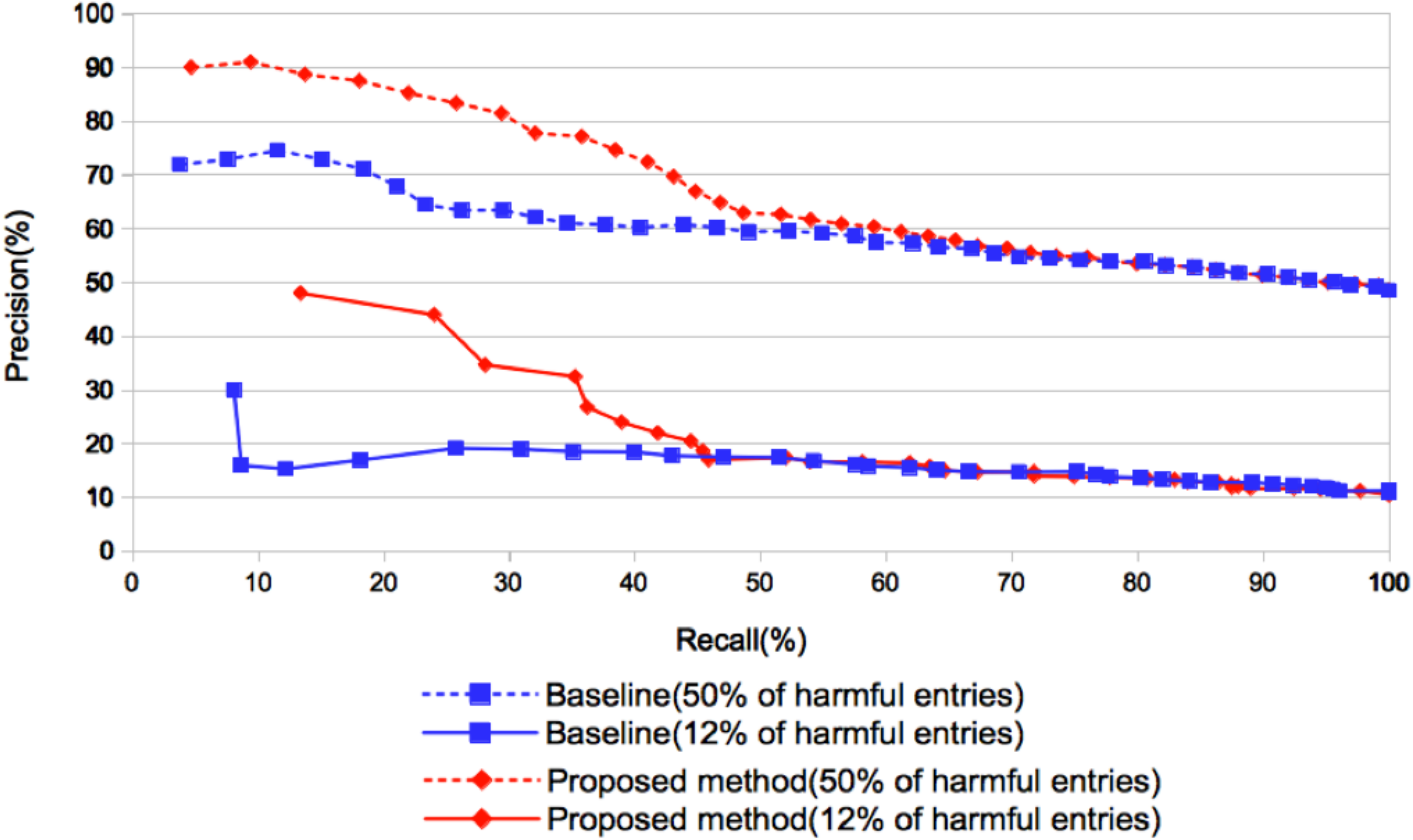
(%)



Threshold

Results


(%)
100
90
80
70
60
50
40
30
20
10
0
10
40
70
100
130
160
190
220
250
280
310
340
P : 48%
Tr. 10
R : 100%
Tr. 340
Threshold



100
90
80
70
60
50
40
30
20
10
0
Precision(%)
0
10
20
30
40
50
60
70
80
90
100
Recall(%)
Baseline(50% of harmful entries)
Baseline(12% of harmful entries)
Proposed method(50% of harmful entries)
Proposed method(12% of harmful entries)

Discussion

Entries with high scores

Harmful entry

- I hate that ugly b*tch → Hate_b*tch
- gotta kill that freak with atopy → Atopy_kill

Discussion

Entries with high scores

Harmful entry

- I hate that ugly b*tch
- gotta kill that freak with atopy

Hate-b*tch

Atopy-kill

Non-harm.

- I live outside of the prefecture

Outside-live

Discussion

Entries with high scores

Harmful entry

- I hate that ugly b*tch
- gotta kill that freak with atopy

Hate-b*tch

Atopy-kill

Non-harm.

- I live outside of

Outside-live

Some phrases are used both in harmful and non-harmful entries (ambiguous, neutral)

Discussion

Could use **non-harmful polarity words** to disambiguate such cases

- Check which words/phrases co-occur most often in non harmful entries

Discussion

Entries with **low** score

Ex. : Michal from Kitami Inst. of Tech.

- entries revealing personal information such as names of a person or school.

At the moment these have low relevance

Discussion

Entries with **low** score

Ex. : Michal from Kitami Inst. of Tech.

▶ entries revealing personal information such as names of a person or school.

Could gather words used for describing private information and implement a method for spotting it.

Conclusions

- New problem: Cyberbullying
- Affect Analysis of Cyberbullying Data
  - Cyberbullying is less “emotive” (cold irony)
  - Distinctive features of CB: vulgarities, mimetic expressions
  - Expressions of emotions considered as positive are often used in ironic meaning
- Proposed a Prototype Method for Cyberbullying Detection
  - First on SVM, then on PMI-IR

Conclusions

PMI-IR based method

- checked the amount of cyberbullying in reality
- tested methods on such data

What influenced the results:

▶**Neutral phrases**

▶**Entries containing private information**

Future work

- perform a study non-harmful words
- apply in processing private names

Future work

- 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)