Cyberbullying Blocker Application for Android

Pawel LEMPA, Michal PTASZYNSKI, Fumito MASUI

Kitami Institute of Technology

Department of Computer Science

Agenda

1. Cyberbullying
2. Platform selection
3. Existing solutions
4. Application assumptions
5. Cyberbullying detection methods
6. Application interface
7. Harmful Content Detection Process
8. Preliminary tests
9. Conclusion
10. Future works

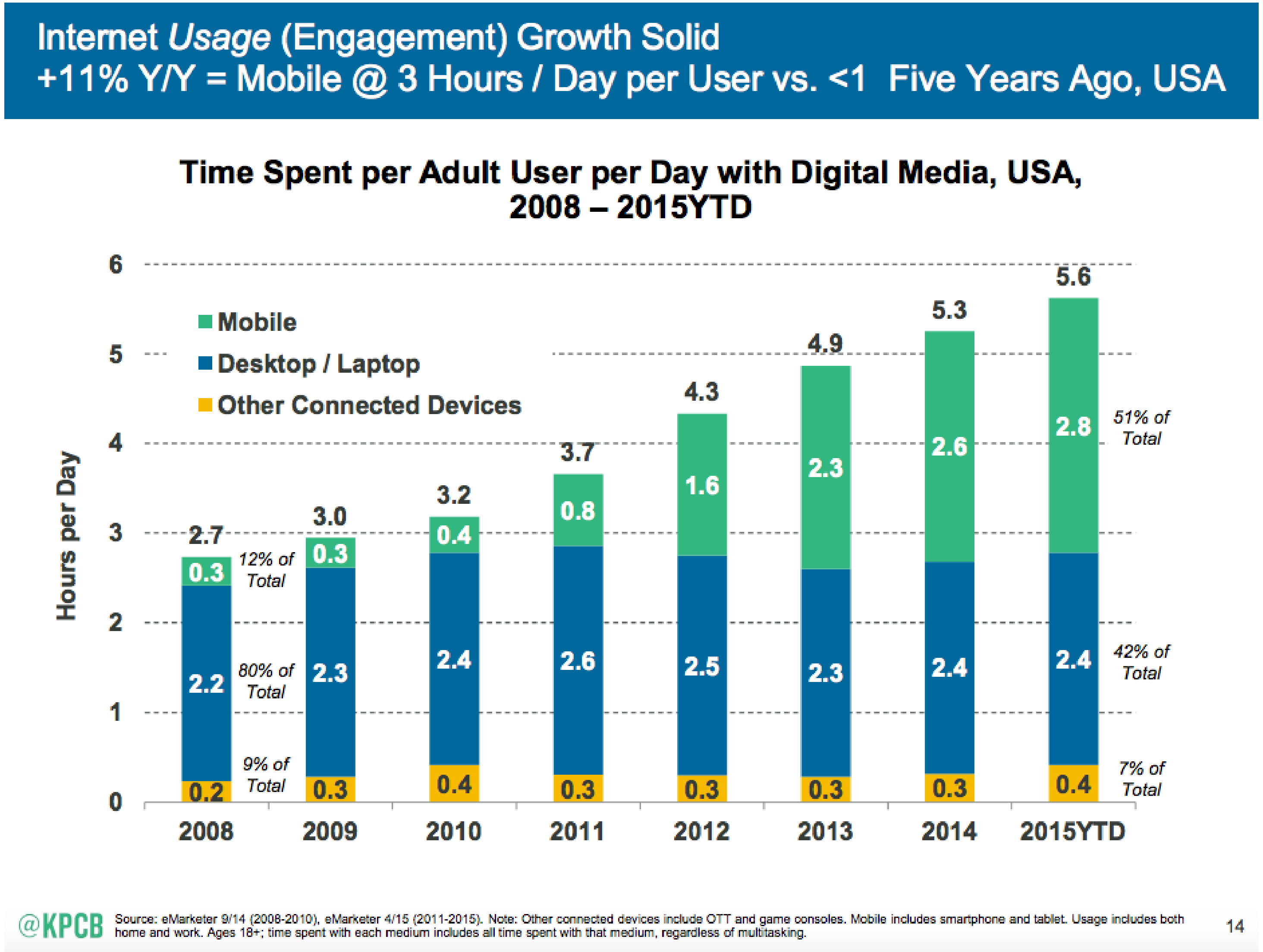
Cyberbullying

Slandering and humiliating people on the Internet.

By using Internet services and mobile technologies, such as:

- web pages
- discussion groups
- instant messaging
- SMS text messaging

Platform selection

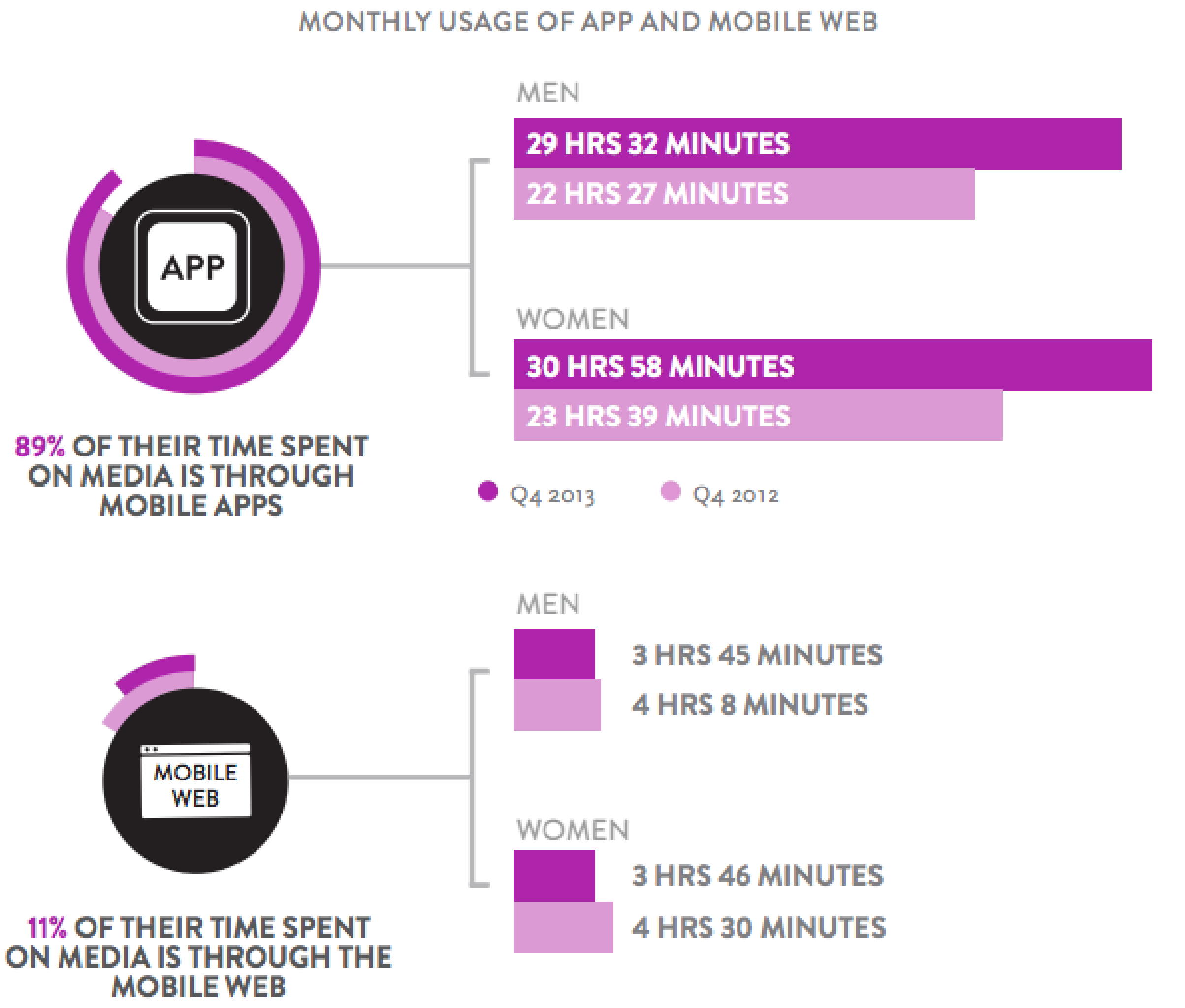


Internet *Usage* (Engagement) Growth Solid
+11% Y/Y = Mobile @ 3 Hours / Day per User vs. <1 Five Years Ago, USA

Time Spent per Adult User per Day with Digital Media, USA, 2008 – 2015YTD

Source: eMarketer 9/14 (2008-2010), eMarketer 4/15 (2011-2015). Note: Other connected devices include OTT and game consoles. Mobile includes smartphone and tablet. Usage includes both home and work. Ages 18+; time spent with each medium includes all time spent with that medium, regardless of multitasking.

Mobile Apps or mobile Web



MONTHLY USAGE OF APP AND MOBILE WEB

Existing solutions

Stopped development

FearNot!

Not relayed yet

BullyGuardPro

Closed

Samaritans Radar

In development state

Uonevu

Too simple

ReThink

Ethics problem

PocketGuardian

Application assumptions

- Detecting harmful words on mobile devices.
- Possibility to test different approaches for detecting cyberbullying (different methods).
- Easy to test, easy to modify (add or change detection methods).

Cyberbullying detection methods

Two detection methods for first version of application.

Both previously created to work on computers

Both methods adapted to the Java language and Android environment

Plans to add more methods in future

Method A

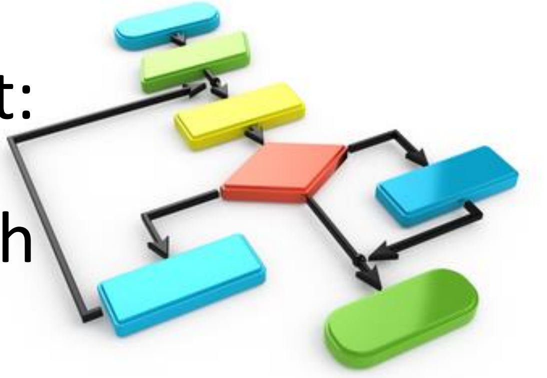
- Developed by Ptaszynski et al. 2015
- The method classifies messages as harmful or not by using a classifier trained with a language modelling method based on a brute force search algorithm applied to language modelling.
- The method uses sophisticated sentence patterns with disjoint elements automatically extracted with a novel language modelling method developed by (Ptaszynski et al.2011).
- The patterns are defined as ordered combinations of sentence elements which are used for brute-force searching within input sentences.

Michal Ptaszynski, Fumito Masui, Yasutomo Kimura, Rafal Rzepka, Kenji Araki. 2015. "Extracting Patterns of Harmful Expressions for Cyberbullying Detection", 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'15), November 27-29, 2015, Poznan, Poland.
Michal Ptaszynski, Rafal Rzepka, Kenji Araki, Yoshio Momouchi. 2011. Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion. In International Journal of Computational Linguistics (IJCL), Vol. 2, No. 1, pp. 24-36.

Method B

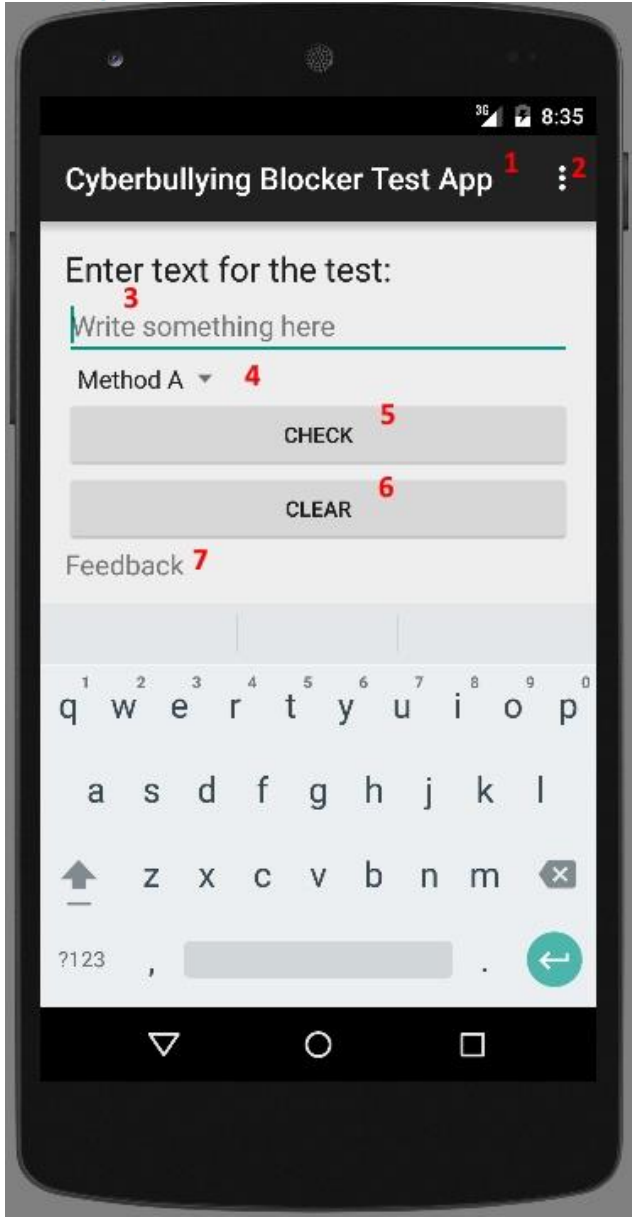
Uses a list of seed words from three categories to calculate semantic orientation score SO-PMI-IR and then maximize the relevance of categories with input sentence according to a method developed by (Nitta et al. 2013).

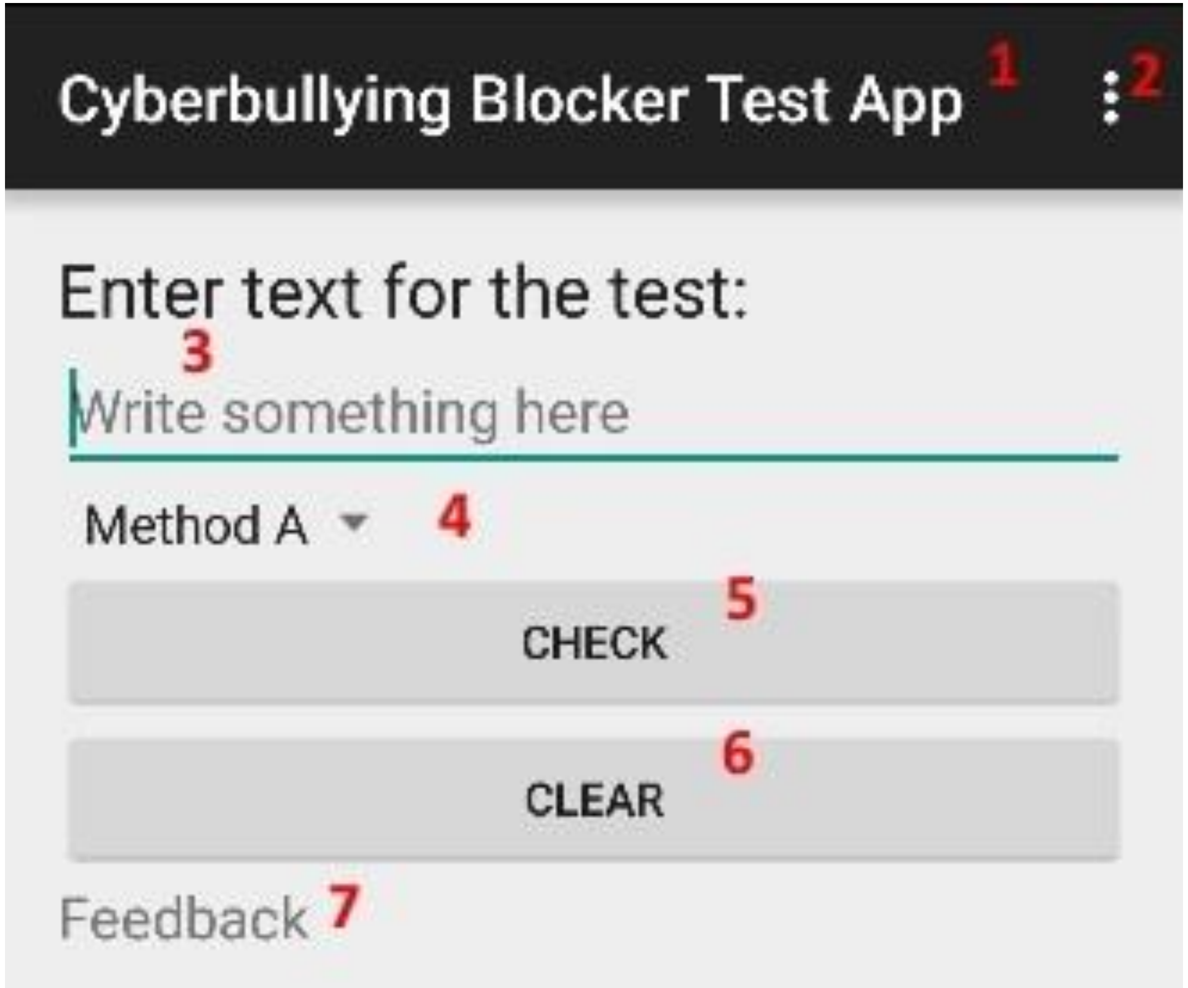
Three steps in the classification of the harmfulness of input:
1. Phrase extraction.
2. Categorization and harmful word detection together with harmfulness polarity determination.
3. Relevance maximization.



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 Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), pp. 579-586, Nagoya, Japan, October 14-18, 2013.

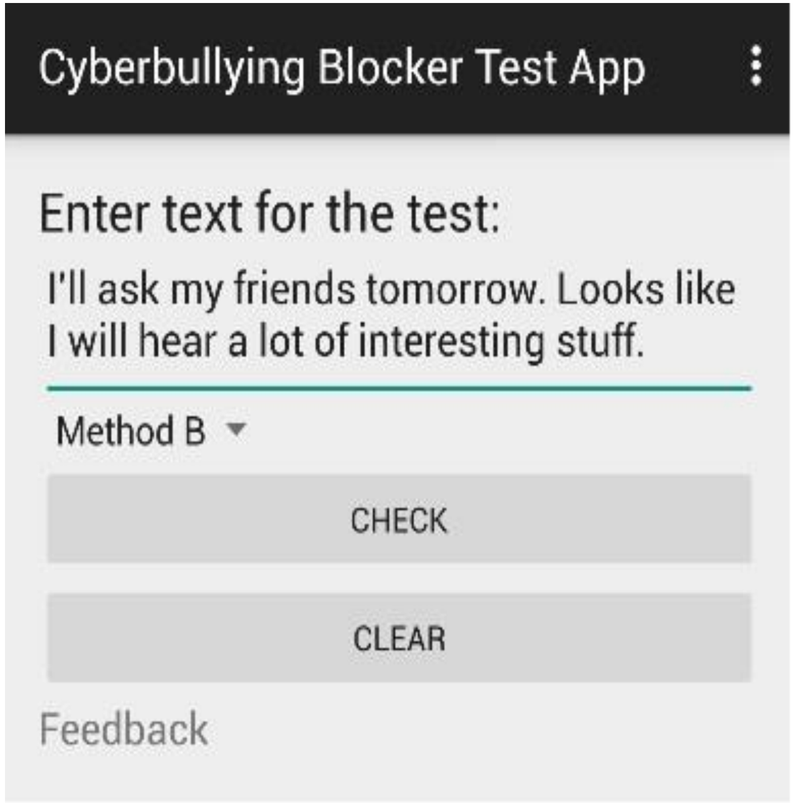
Application interface





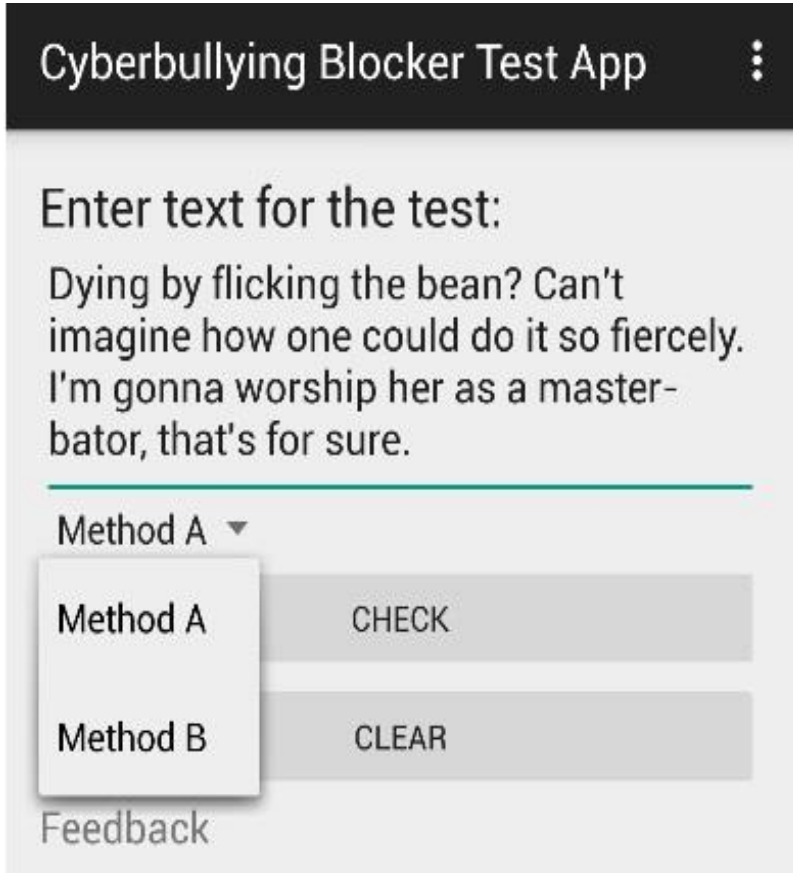
Harmful Content Detection Process

1. Sentence input.
2. Selection of detection method.
3. Launch of the checking process.
4. Feedback.



I'll ask my friends tomorrow. Looks like I will hear a lot of interesting stuff.

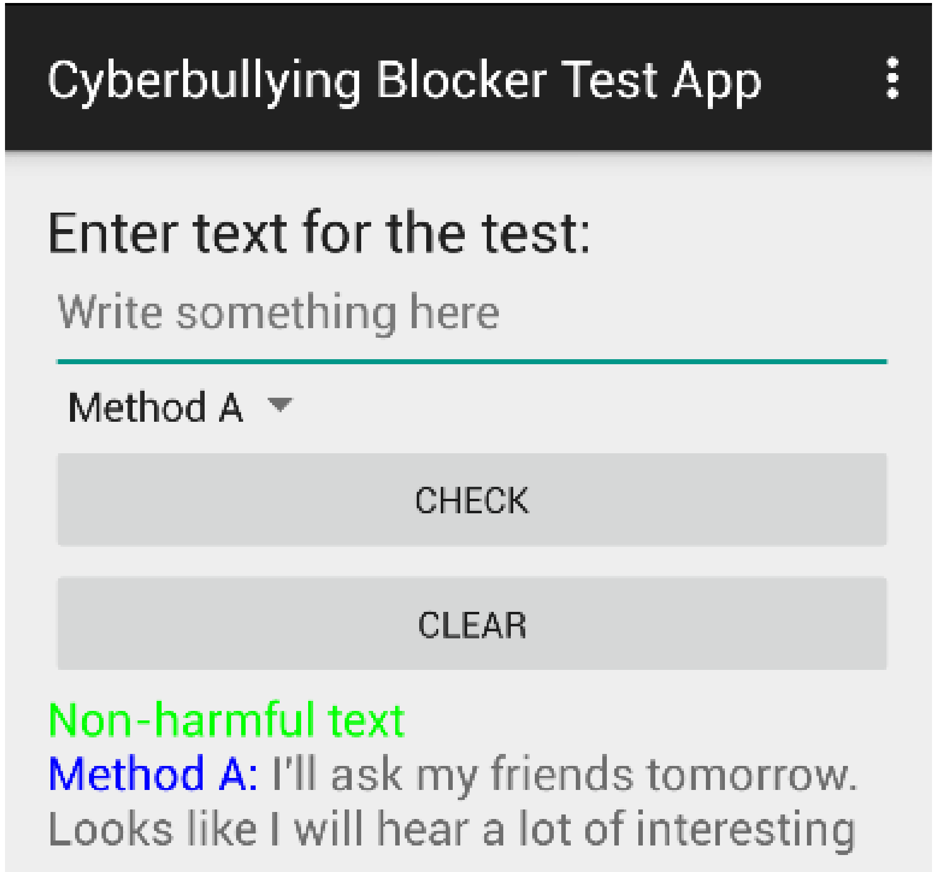
Method B



Dying by flicking the bean? Can't imagine how one could do it so fiercely. I'm gonna worship her as a master-bator, that's for sure.

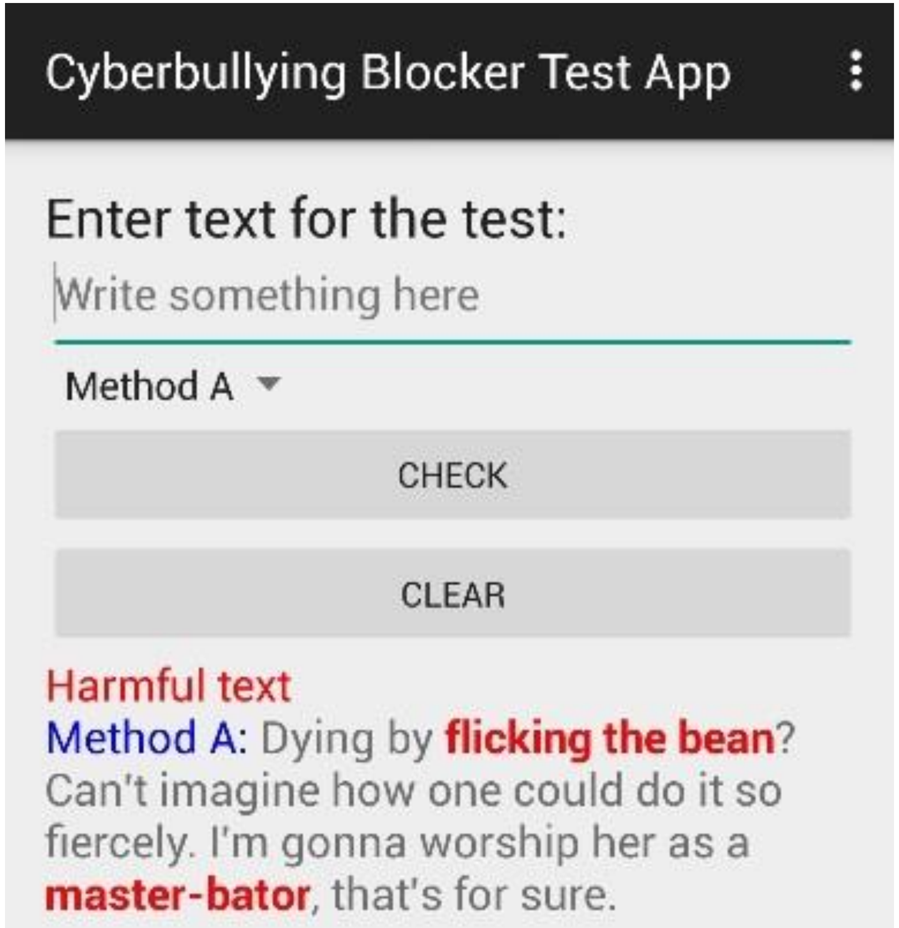
Method A

Feedback



Enter text for the test:

Non-harmful text
Method A: I'll ask my friends tomorrow. Looks like I will hear a lot of interesting



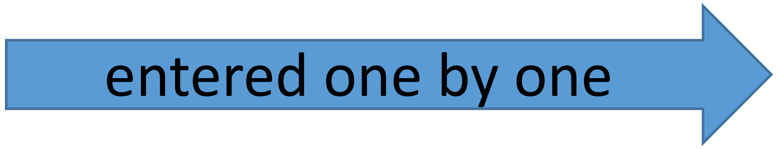
Enter text for the test:

Harmful text
Method A: Dying by **flicking the bean**? Can't imagine how one could do it so fiercely. I'm gonna worship her as a **master-bator**, that's for sure.

Preliminary tests

Verification whether the applied classification methods performed correctly under the new environment, and if they returned a proper feedback.

Set of sentences with harmful and not harmful words.



entered one by one

Tested by both methods, results compared to each other.

Both of the applied cyberbullying detection methods work presently for the Japanese language.

Tests were performed on virtual devices emulated by Genymotion engine:

- Sony Xperia with Android 4.2.2.
- Google Nexus 10 with Android 5.0 .

And on Smartphone LG G2 with Android 5.0.2.

Tests were focused on performance of used algorithms on mobile devices.

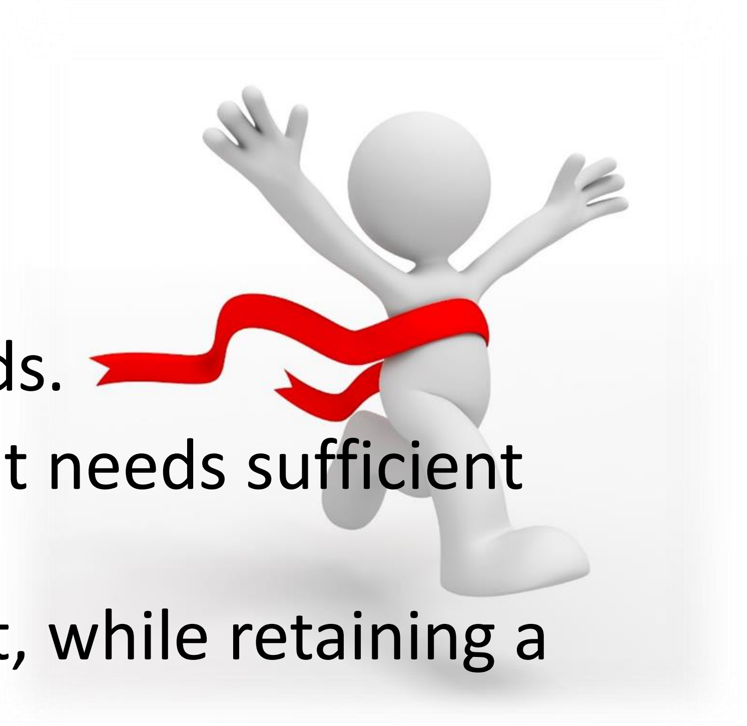
Depending on the classification method, the detected harmful words can differ.

Processing speed of detection is associated with the type of device used.

Results

	Method A	Method B
Avg. # of words / patterns used in classification	11,832,430	9
Internet connection required	NO	YES
Avg. time for one sentence (in seconds)	88.46 s	8.41 s
Avg. time for one sentence (in minutes)	(1.47 m)	(0.14 m)
Best Precision	89%	91%
Recall at Best Precision	34%	9%
F-score for Best Precission	49%	16%

Conclusion



- Applied two methods to find potential harmful words.
- Method A, does not require Internet connection, but needs sufficient computing power.
- Method B to work requires an access to the Internet, while retaining a need for low computing power.
- Future use of this application is linked with communication via the Internet, and each new generation of smartphones represents a major technological leap so both differences do not affect the developed application.

Future works

- Improve a performance of detection of the harmful words by using another methods or by optimizing the existing ones.
- Implement our software to communication applications (Facebook, Twitter, etc.):
  - Plugins
  - Virtual keyboard
- Creating an app for other dominant mobile systems.

Thank you for your attention

Pawel LEMPA, Michal PTASZYNSKI, Fumito MASUI