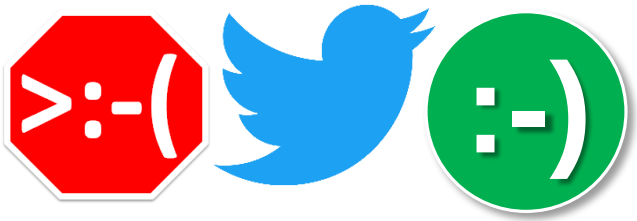
UNIWERSYTET JAGIELLOŃSKI W KRAKOWIE

KITAMI Institute of Technology

ai & nlp workshop day

PJATK 学大科工 ポーランド 報情本日

PAN




Sponsoring:
samurai LABS

# PolEval 2019 Task 6: First Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter

(Pierwsze Zadanie dot. Automatycznego Wykrywania Cyberagresji w Polskim Twitterze)

Michal Ptaszynski
KIT, Japan
ptaszynski@ieee.org

Agata Pieciukiewicz
PJATK, Poland
agata.niescieruk@pjwstk.edu.pl

Pawel Dybala
UJ, Poland
pawel.dybala@uj.edu.pl

# Outline

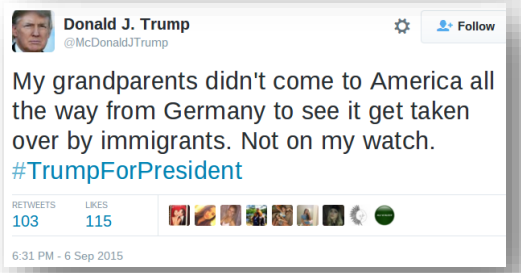
- Why this task today?
- Task description
  - Task 6-1: harmful vs. non-harmful
  - Task 6-2: cyberbullying vs. hate-speech vs. non-harmful
- Cyberbullying: a working definition
- Participants
- Winners
- Future Plans

# Thanks to our sponsor


- SamuraiLabs
- https://www.samurailabs.ai/

# Why this task today?

- Since 2007
  - Awareness in Japan:
    Takikawa suicide incident

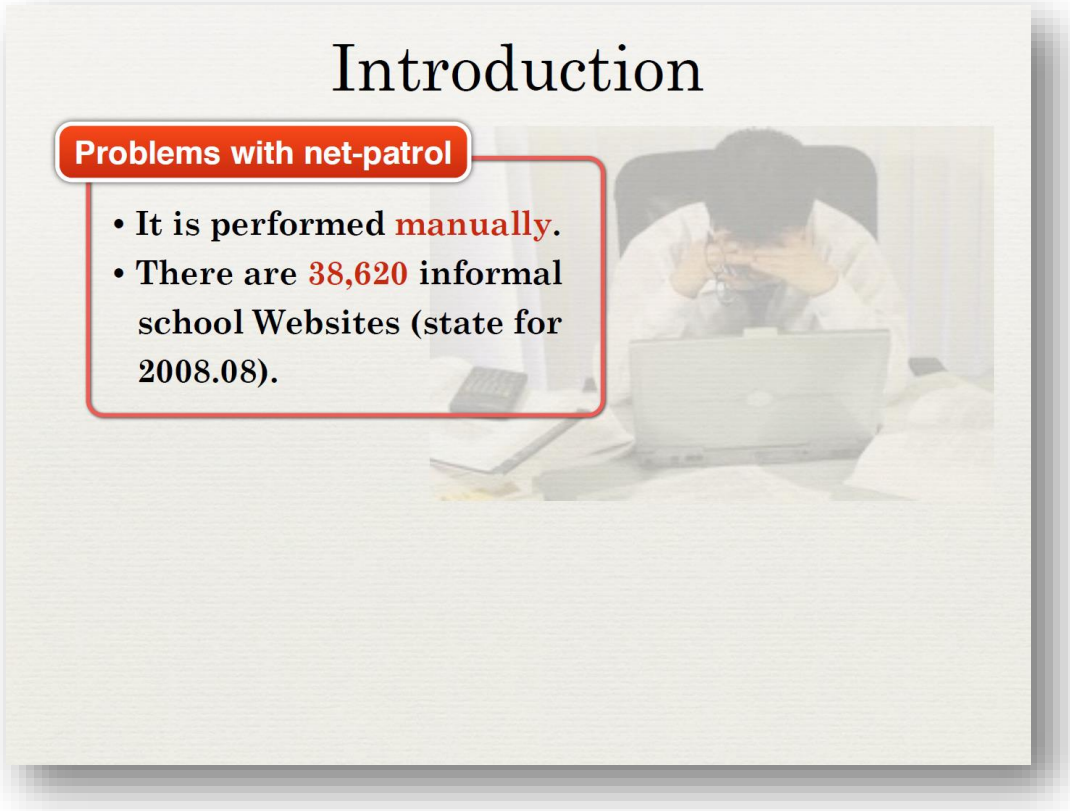



- Since 2015/2016
  - Post-truth reality: USA, UK, Hungary, etc.


- Increase of abuse on Internet
  - Harassment
  - Cyberbullying
  - Fake-news
  - Large-scale manipulations
    (Cambridge Analytica, etc.)

# Why this task today?

**In 2008:**



# Why this task today?


Introduction

**Problems with net-patrol**

- It is performed manually.
- There are 38,620 informal school Websites (state for 2008.08).

# Why this task today?

- World 10 years ago:
- Twitter, Facebook, Youtube just started
- No Instagram
- Cyberbullying / cyber-bullying / cyber bullying ?

# Why this task today?

- First studies in social sciences:
  - **Hinduja, S., & Patchin, J. W.** (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant behavior*, *29*(2), 129-156.
  - **Olweus, D.** (2012). Cyberbullying: An overrated phenomenon?. *European Journal of Developmental Psychology*, *9*(5), 520-538.
  - **Pyżalski, J.** (2012). From cyberbullying to electronic aggression: Typology of the phenomenon. *Emotional and behavioural difficulties*, *17*(3-4), 305-317.

# Why this task today?

- 2008: First studies in Automatic Cyberbullying Detection
- 2009: First application-ready methods
- Nobody interested in global solution

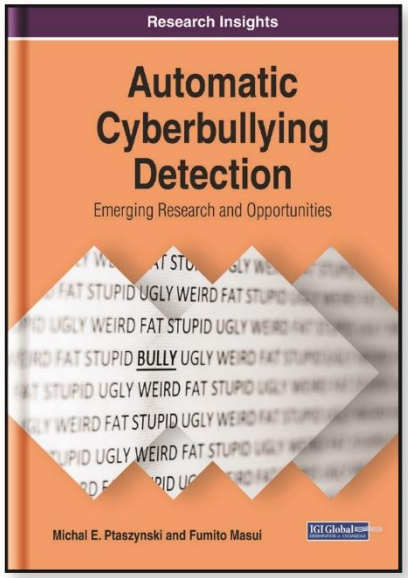
# Why this task today?

- 2008: First studies in Automatic Cyberbullying Detection
- 2009: First application-ready methods
- Nobody interested in global solution

- 2018: Problem too large
- 2019: Companies, authorities in Japan asking from help:
  - Japanese Police
  - Several companies
- One person/lab/company cannot solve all problems
- Foundation of the field
- **Need army of people → This is why this task today!**

https://www.igi-global.com/book/automatic-cyberbullying-detection/187225

# Automatic Cyberbullying Detection: Emerging Research and Opportunities




Michal E. Ptaszynski (Kitami Institute of Technology, Japan) and
Fumito Masui (Kitami Institute of Technology, Japan)

## Description:

Due to the prevalence of social network service and social media, the problem of cyberbullying has risen to the forefront as a major social issue over the last decade. Internet hate, harassment, cyberstalking, cyberbullying—these terms, which were almost unknown 10 years ago—are in the everyday lexicon of all internet users. Unfortunately, it is becoming increasingly difficult to undertake continuous surveillance of websites as new ones are appearing daily. Methods for automatic detection and mitigation for online bullying have become necessary in order to protect the online user experience.

**Automatic Cyberbullying Detection: Emerging Research and Opportunities** provides innovative insights into online bullying and methods of early identification, mitigation, and prevention of harassing speech and activity. Explanations and reasoning for each of these applied methods are provided as well as their pros and cons when applied to the language of online bullying. Also included are some generalizations of cyberbullying as a phenomenon and how to approach the problem from a practical technology-backed point of view. The content within this publication represents the work of deep learning, language modeling, and web mining. It is designed for academicians, social media moderators, IT consultants, programmers, education administrators, researchers, and professionals and covers topics centered on identification methods and mitigation of internet hate and online harassment.

**ISBN:** 9781522552499 **Release Date:** November, 2018 **Copyright:** 2019 **Pages:** 190

# Task Descriptions

# Task 6-1: harmful vs. non-harmful

- Goal:

  Distinguish between
- normal (non-harmful) tweets (class: 0)
- any kind of harmful information (class: 1).

# Task 6-2: cyberbullying vs. hate-speech vs. non-harmful

- Goal:

  Distinguish between
- Non-harmful (0)
- Cyberbullying:
  addressed to a private person/s (1)
- Hate-speech:
  public person/entity/larger group (2)

# Task/dataset preparation:

- Data acquisition:
  **Agata Pieciukiewicz**, Polish-Japanese Academy of Information Technology,

- Tweet tagging:
  Students of Jagiellonian University, employed by Kotoken Language Laboratory under the supervision of dr. **Pawel Dybala**

- General oversight and data check:
  **Michal Ptaszynski**, Kitami Institute of Technology, Japan

# Cyberbullying: a working definition (1)

**General characteristics:**

- Power imbalance (Nierównowaga sił) *
- Repetitiveness (Powtarzalność) *
- Peer group (Grupa rówieśnicza) *

* Difficult to grasp on the Internet

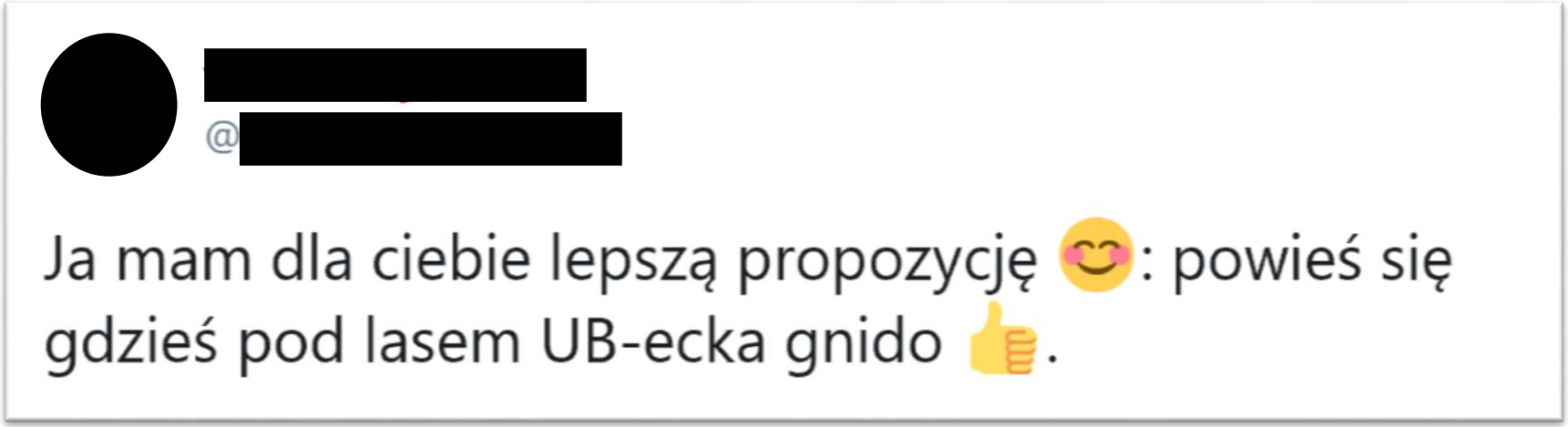
# Cyberbullying: a working definition (2)

**Content (what to look for when annotating)**

- Disclosure of private information (Tel. Number, e-mail, address, school name / number, class, PESEL/Private ID, credit card no.)
- Personal attack ("Hang yourself, bitch!")
- Threats ("I will find you and I will kill you")
- Blackmail ("I will tell everyone where you live if you do not pay me")
- Ridiculing ("Look at that fat guy")
- Gossip / insinuations ("Hey, did you know he’s gay?")
- Accumulation of profanity (longer "bundle" of profanity)
- Various combinations of all of the above

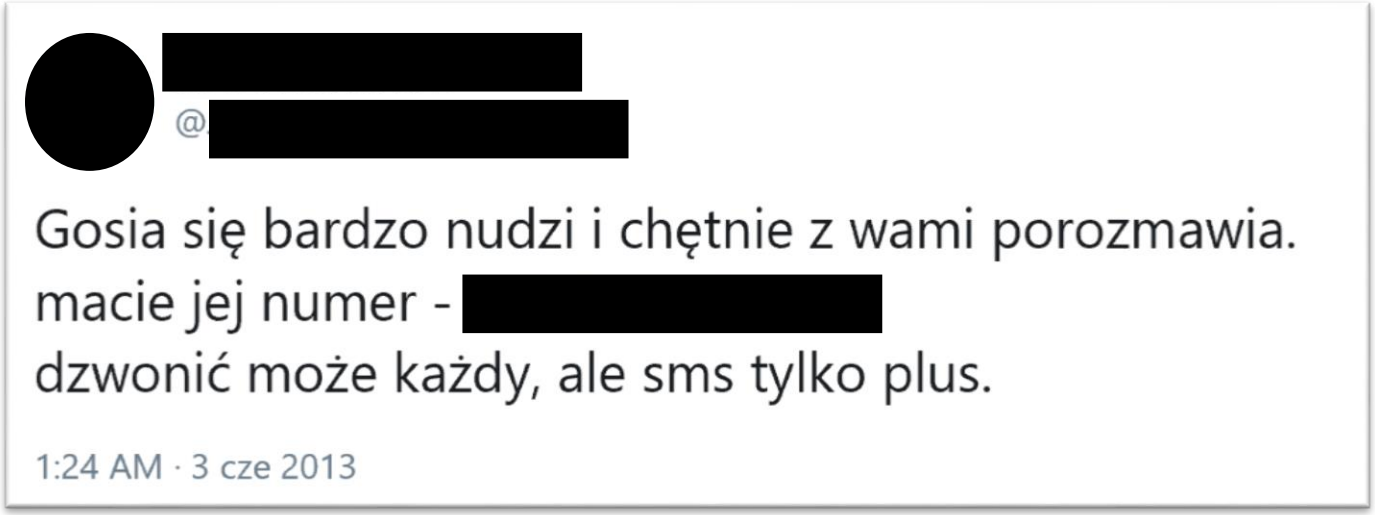
# Examples

# Example 1



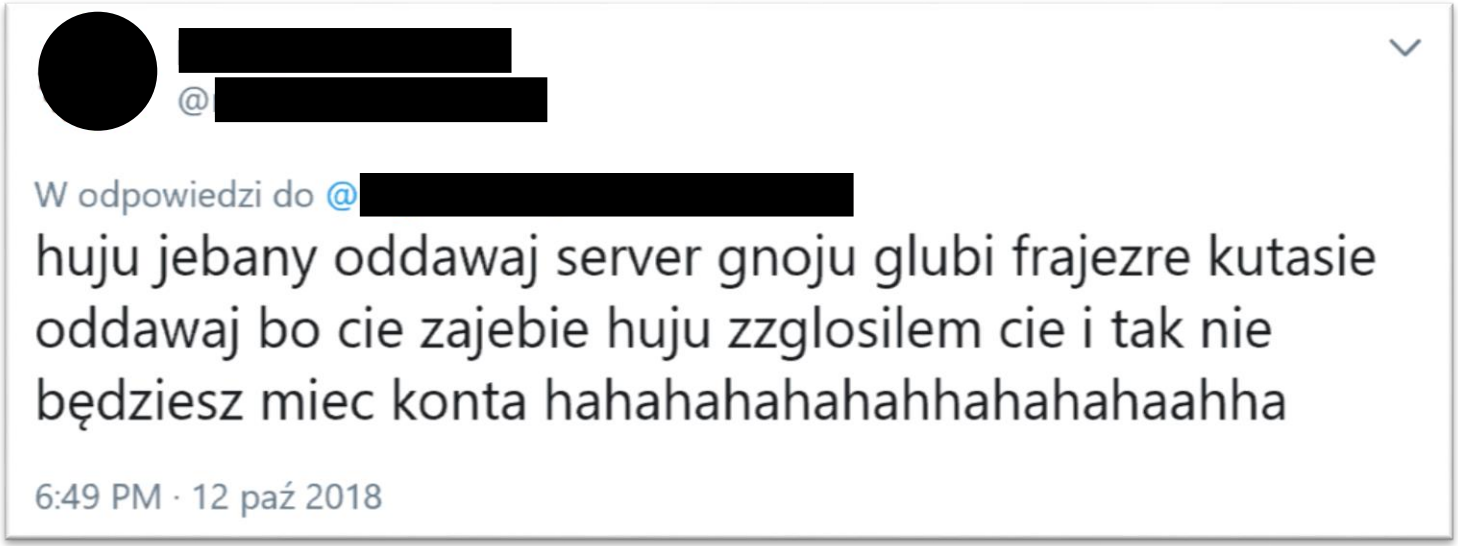

- Aggressive vocabulary
- Personal attack

# Example 2




- Disclosure of personal information:
  - Phone number

# Example 3




- Accumulation of profanity
- Threats
- Ridiculing

# Dataset

https://github.com/ptaszynski/cyberbullying-Polish

Michal Ptaszynski, Agata Pieciukiewicz, Pawel Dybala. (2019). **Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter**, In *Proceedings of the PolEval 2019 Workshop*, Warsaw. Institute of Computer Science, Polish Academy of Sciences

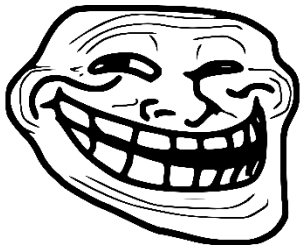
# Participants

# Participants

- 16 submissions

# Participants

- 16 submissions
- 14 valid submissions
  - One double submission (uploading accident)
  - One troll



# Participants

- 16 submissions
- 14 valid submissions
  - One double submission (uploading accident)
  - One troll
- 9 unique teams

- Task 6-1:  14 submissions
- Task 6-2:  8 submissions

# Participants

- Application of widely available solutions
  - Fast.AI NLP
  - BERT
  - FastText
  - Flair
  - SentencePiece
  - SVM
- Original solutions
  - Przetak

# Task 6-1 Results

| Submission author(s) | Affiliation | Submitted system | Precision | Recall | F-score | Accuracy |
|---|---|---|---|---|---|---|
| **Piotr Czapla, Marcin Kardas** | **n-waves** | **n-waves ULMFiT** | **66.67%** | **52.24%** | **58.58%** | **90.10%** |
| Marcin Ciura | independent | Przetak | 66.35% | 51.49% | 57.98% | 90.00% |
| Tomasz Pietruszka | Warsaw University of Technology | ULMFiT + SentencePiece + BranchingAttention | 52.90% | 54.48% | 53.68% | 87.40% |
| Sigmoidal Team (Renard Korzeniowski, Przemysław Sadowski, Rafał Rolczyński, Tomasz Korbak, Marcin Możejko, Krystyna Gajczyk) | Sigmoidal | ensamble spacy + tpot + BERT | 52.71% | 50.75% | 51.71% | 87.30% |
| Sigmoidal Team | Sigmoidal | ensamble + fastai | 52.71% | 50.75% | 51.71% | 87.30% |
| Sigmoidal Team | Sigmoidal | ensenble spacy + tpot | 43.09% | 58.21% | 49.52% | 84.10% |
| Rafał Prońko | CVTimeline | Rafal | 41.08% | 56.72% | 47.65% | 83.30% |
| Rafał Prońko | CVTimeline | Rafal | 41.38% | 53.73% | 46.75% | 83.60% |
| Maciej Biesek | independent | model1-svm | 60.49% | 36.57% | 45.58% | 88.30% |
| Krzysztof Wróbel | AGH, UJ | fasttext | 58.11% | 32.09% | 41.35% | 87.80% |
| Katarzyna Krasnowska-Kieraś, Alina Wróblewska | IPI PAN | SCWAD-CB | 51.90% | 30.60% | 38.50% | 86.90% |
| Maciej Biesek | independent | model2-gru | 63.83% | 22.39% | 33.15% | 87.90% |
| Maciej Biesek | independent | model3-flair | 81.82% | 13.43% | 23.08% | 88.00% |
| Jakub Kuczkowiak | UWr | Task 6: Automatic cyberbullying detection | 17.41% | 32.09% | 22.57% | 70.50% |

# Task 6-2 Results

| Submission author(s) | Affiliation | Name of the submitted system | Micro-Average F-score | Macro-Average F-score |
|---|---|---|---|---|
| **Maciej Biesek** | independent | **model1-svm** | **87.60%** | **51.75%** |
| Sigmoidal Team (Renard Korzeniowski, Przemysław Sadowski, Rafał Rolczynski, Tomasz Korbak, Marcin Możejko, Krystyna Gajczyk) | Sigmoidal | ensamble spacy + tpot + BERT | 87.10% | 46.45% |
| Krzysztof Wróbel | AGH, UJ | fasttext | 86.80% | 47.22% |
| Maciej Biesek | | model3-flair | 86.80% | 45.05% |
| Katarzyna Krasnowska-Kieraś, Alina Wróblewska | IPI PAN | SCWAD-CB | 83.70% | 49.47% |
| Maciej Biesek | independent | model2-gru | 78.80% | 49.15% |
| Jakub Kuczkowiak | UWr | Task 6: Automatic cyberbullying detection | 70.40% | 37.59% |
| Sigmoidal Team | Sigmoidal | ensamble + fastai | 61.60% | 39.64% |

# Winners

- Task 6-1:
  1. n-waves (based on fast.AI)
  2. (runner-up) Przetak (original solution)

- Task 6-2:
  1. Maciej Biesek (SVM)
  2. (runner-up) Sigmoidal

# Winners

- All participants are winners
- Contributed to the creation of the field

but

- Do not feel satisfied
- Go and use your solutions in practice

# Future plans

- Dataset
  - ~10x larger
  - Twitter conversation threads
  - More expert annotations
- Solutions
  - More feature engineering
  - More error analysis
- Implementations
  - Most released on GitHub
  - More practical implementations = less cyberbullying

UNIWERSYTET JAGIELLOŃSKI W KRAKOWIE

KITAMI Institute of Technology

ai & nlp workshop day

PJATK 学大科工 ポーランド 報情本日

PAN


Sponsoring:
samurai LABS

# PolEval 2019 Task 6: First Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter

(Pierwsze Zadanie dot. Automatycznego Wykrywania Cyberagresji w Polskim Twitterze)

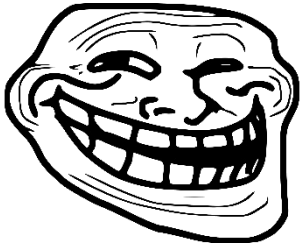
## Let's stay in touch!


Michal Ptaszynski
KIT, Japan
ptaszynski@ieee.org

Agata Pieciukiewicz
PJATK, Poland
agata.niescieruk@pjwstk.edu.pl

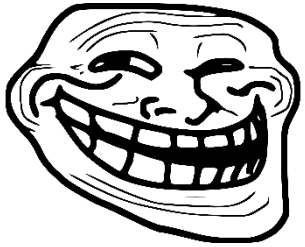
Pawel Dybala
UJ, Poland
pawel.dybala@uj.edu.pl

# Our Private Troll



- 100% Accuracy, 100% F-score
- Submitted pre-released test set answers

# Our Private Troll

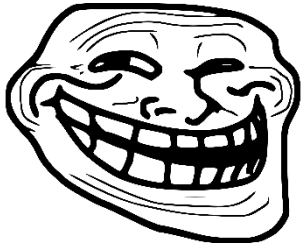


- 100% Accuracy, 100% F-score
- Submitted pre-released test set answers

czw., 11 kwi 2019 o 21:31 Michal Ptaszynski <michal.ptaszynski@gmail.com> napisał(a):

> Szanowny Panie,
>
> Z tej strony kłania się Michal Ptaszynski, organizator zadania 6 na PolEval 2019.
>
> Właśnie podsumowuje wyniki tego zadania i widzę, ze osiągnął pan niezwykle wysoki wynik 100% wszystkiego...
>
> Niestety, w przeciwieństwie do wszystkich innych uczestników nie udostępnił pan kodu swojego systemu, wiec nie jesteśmy w stanie obiektywnie ocenić czy rzeczywiście wykonał pan zadanie.
>
> Jeśli chciałby pan, aby pana system był wzięty pod uwagę przy podliczaniu wyników, prosiłbym o udostępnienie kodu.
>
> Pozdrawiam,
> --
> Michal Ptaszynski

# Our Private Troll



- 100% Accuracy, 100% F-score
- Submitted pre-released test set answers

czw., 11 kwi 2019 o 21:31 Michal Ptaszynski <michal.ptaszynski@gmail.com> napisał(a):

Szanowny Panie,

Z tej strony kłania się Michal Ptaszynski, organizator zadania 6 na PolEval 2019.

Właśnie podsumowuje wyniki tego zadania i widzę, ze osiągnął pan niezwykle wysoki wynik 100% wszystkiego...

Niestety, w prz
udostępnił pan
obiektywnie oc

Jeśli chciałby
wyników, prosi

Pozdrawiam,
--
Michal Ptaszyn

**Od:** Jan Złoty <janzlotyk@gmail.com>
**Data:** Sat, 13 Apr 2019 05:55:43 +0900
**Temat:** Re: Zgloszenie na PolEval
**Do:** "Michal Ptaszynski" <michal.ptaszynski@gmail.com>

---

Do I have to release the source code of my system to participate in PolEval?
No, but you are strongly encouraged to do so. Releasing the source code is a strict condition for receiving the award.