

# #TwitterBuddy TB



@Adam Piróg

@Mateusz Stolarski

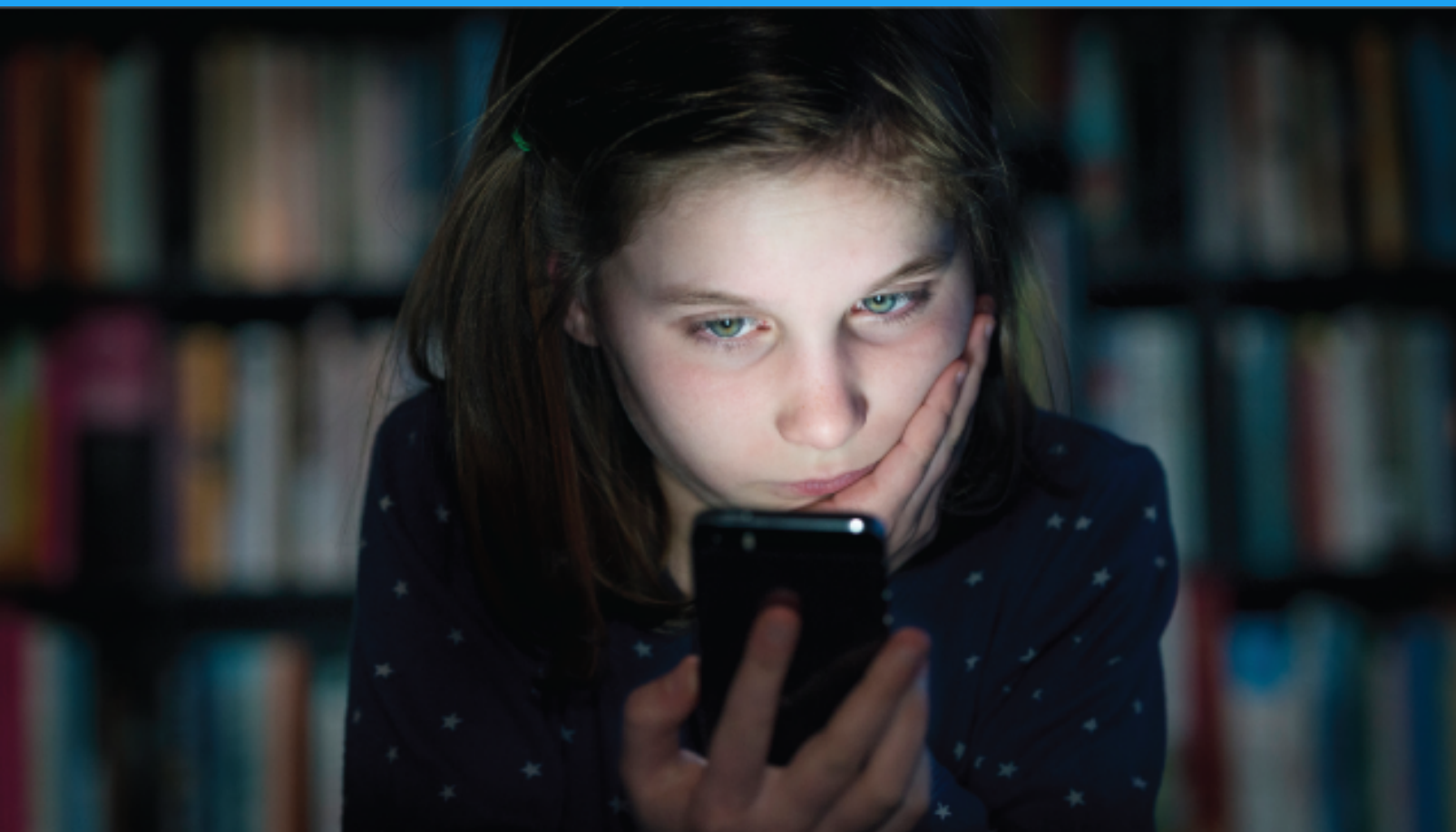
@Przemysław Mikluszka

@Jan Kulbiński

Opiekunowie projektu:  
dr hab. inż. Piotr Bródka  
dr hab. inż. Radosław Michalski

**"59% nastolatków w USA było  
zastraszanych lub nękanych w Internecie"**

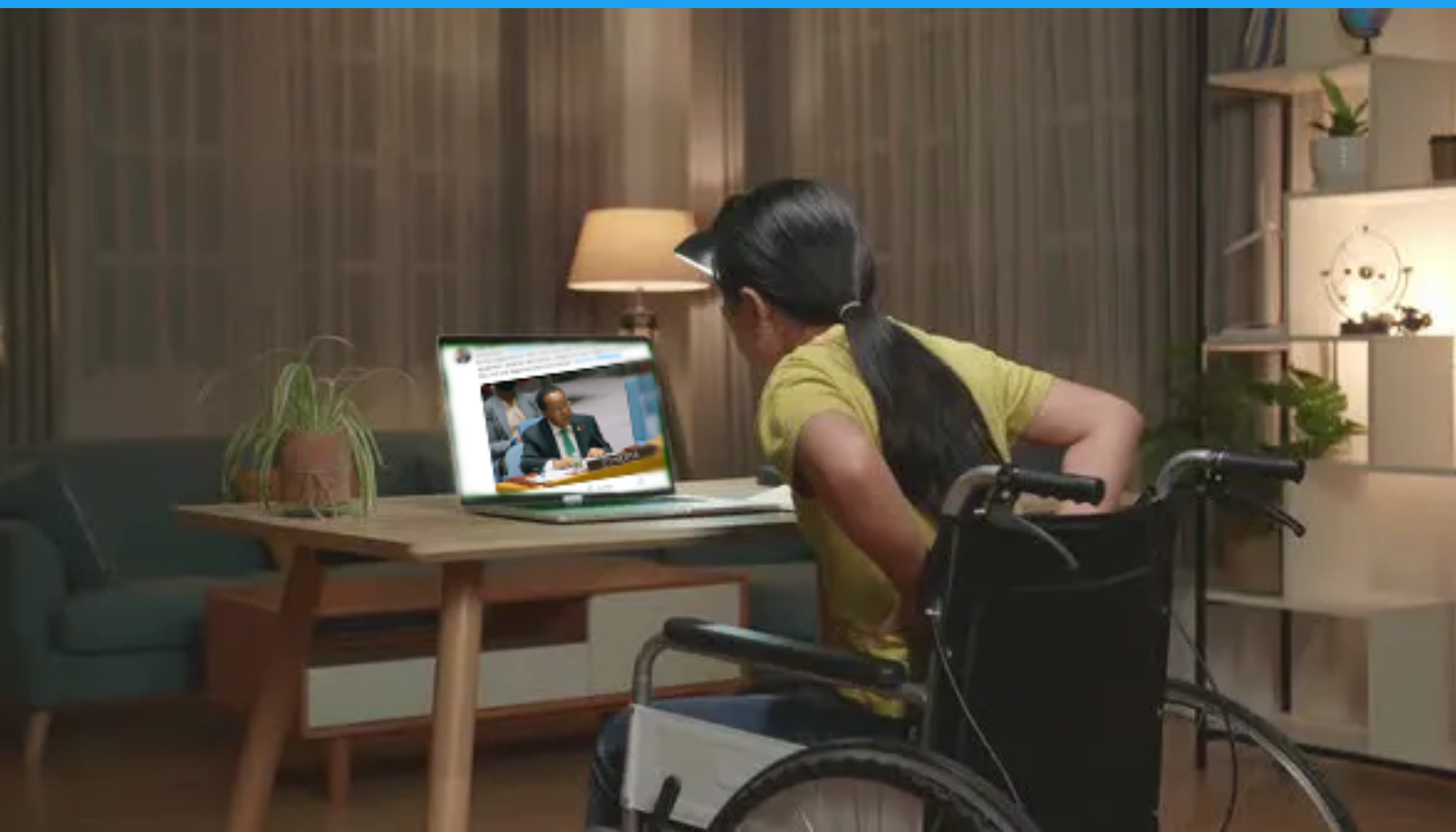
Pew Research Center [1]



[1] Pew Research Center - [www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/](http://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/)

**"73% osób z niepełnosprawnością doświadczyło cyberprzemocy"**

Toth et al. [2]



[2] Kowalski RM, Toth A. Cyberbullying among Youth with and without Disabilities. J Child Adolesc Trauma. 2017, 15 Mar 2017

**"Osoby LGBTQIA+ są niemal dwukrotnie bardziej  
narażone na nękanie w sieci"**

U.S. Department of Health and Human Services  
Centers for Disease Control and Prevention [3]

[3] [www.cdc.gov/healthyyouth/data/yrbs/pdf/2019/su6901-H.pdf](http://www.cdc.gov/healthyyouth/data/yrbs/pdf/2019/su6901-H.pdf)





**"Fałszywe informacje mogą służyć jako środek cyberprzemocy wykorzystywany do dyskredytowania określonych osób lub grup"**

**"Ostatnie badania podkreślają związek między rozprzestrzenianiem fałszywych informacji a nękaniami"**

Zizumbo-Colunga et al. [4]

# PROBLEM BIZNESOWY

- Subiektywna definicja mowy nienawiści
- Indywidualna wrażliwość na treści w internecie
- Moderacja nie uwzględnia potrzeb indywidualnych użytkowników
- Brak detekcji fałszywych informacji

# NASZE ROZWIĄZANIE

7

Interaktywna wtyczka przeglądarkowa

---

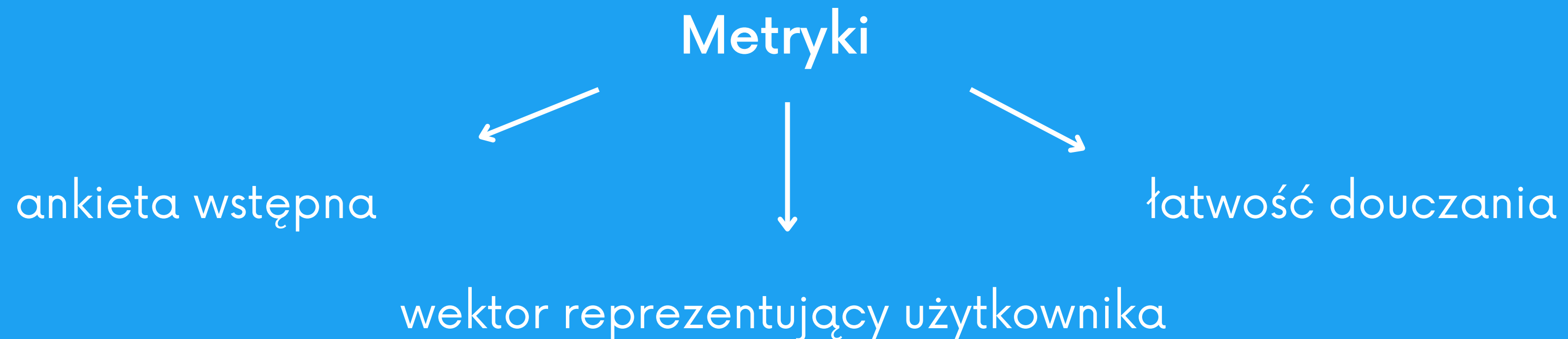
## WARTOŚĆ BIZNESOWA

- Automatyczna, spersonalizowana moderacja treści
- Wykrywanie mowy nienawiści
- Weryfikacja faktów w wypowiedziach użytkowników

# SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

9

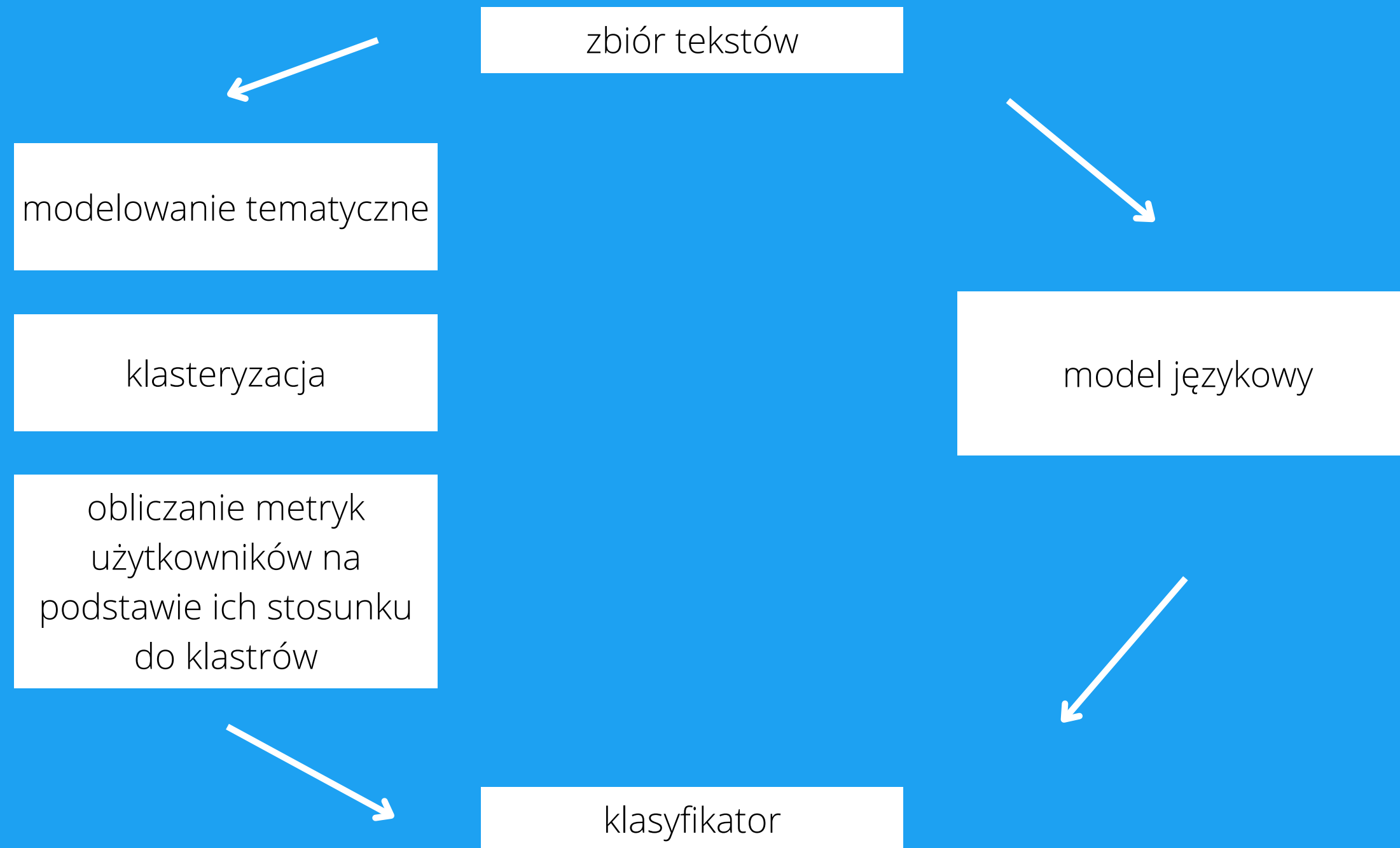
- Metoda opiera swoje działanie o modelowanie tematyczne
- Każdy użytkownik zostaje scharakteryzowany za pomocą metryk, opisujących jego stosunku do różnych grup tematów





# SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

10



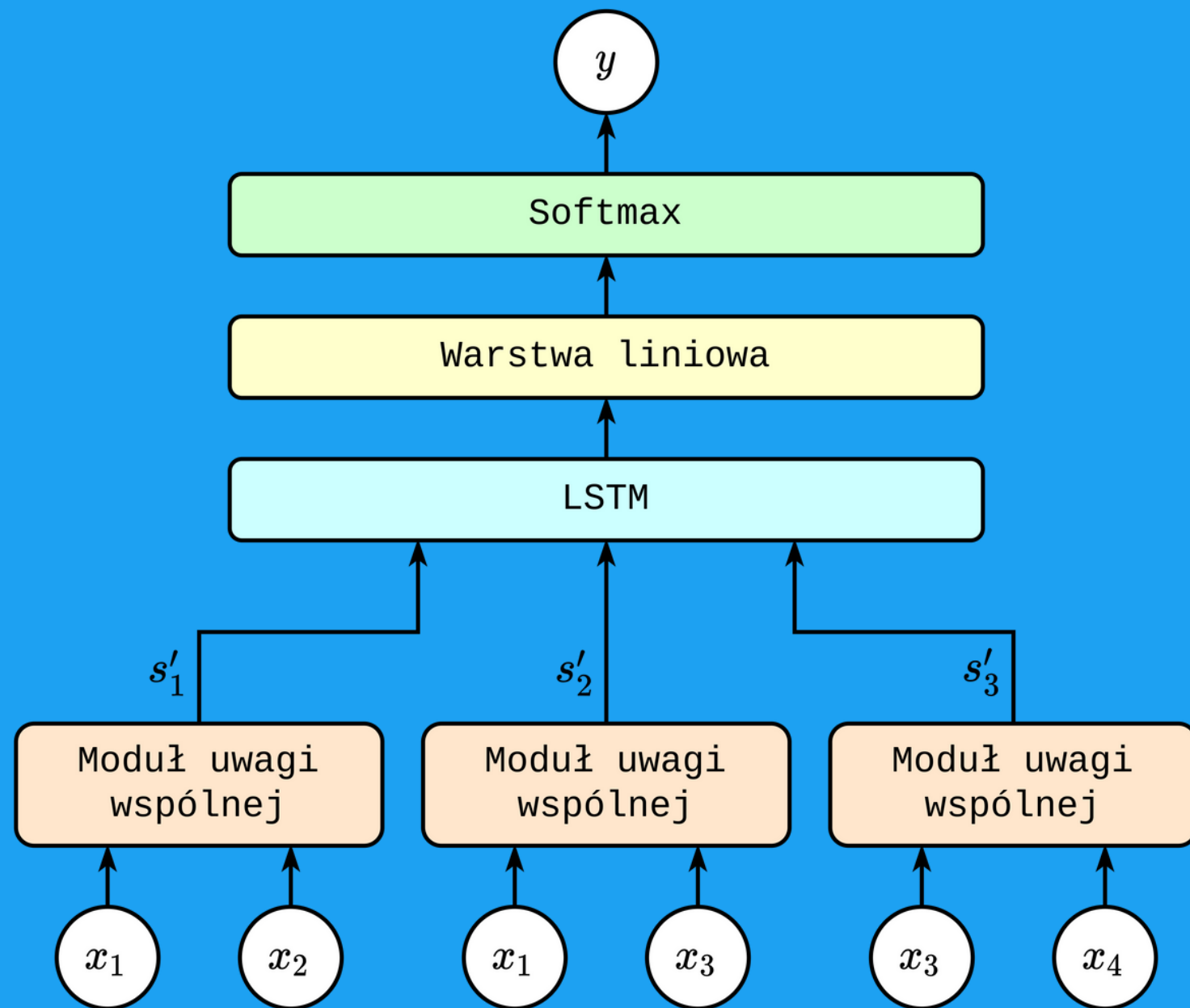
# MODEL WYKRYWANIA FAŁSZYWYCH TREŚCI

11

- Model wywoływany jest na życzenie użytkownika
- Predykcja dokonywana jest na podstawie analizy przebiegu dyskusji związanej z danym postem
- Predykowana jest jedna z trzech klas (prawda, fałsz i niezweryfikowany)

# MODEL WYKRYWANIA FAŁSZYWYCH TREŚCI

12



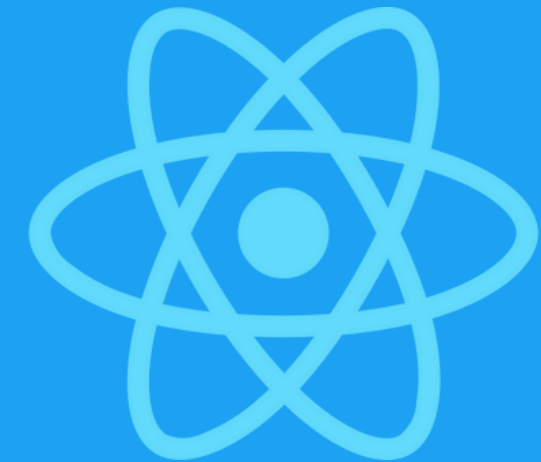
- Zastosowano modelowanie lokalnej interakcji pomiędzy postami, a ich bezpośrednimi odpowiedziami za pomocą modułu uwagi wspólnej (ang. Co-Attention module)
- Modelowane są także zmiany interakcji w czasie

# STACK TECHNOLOGICZNY

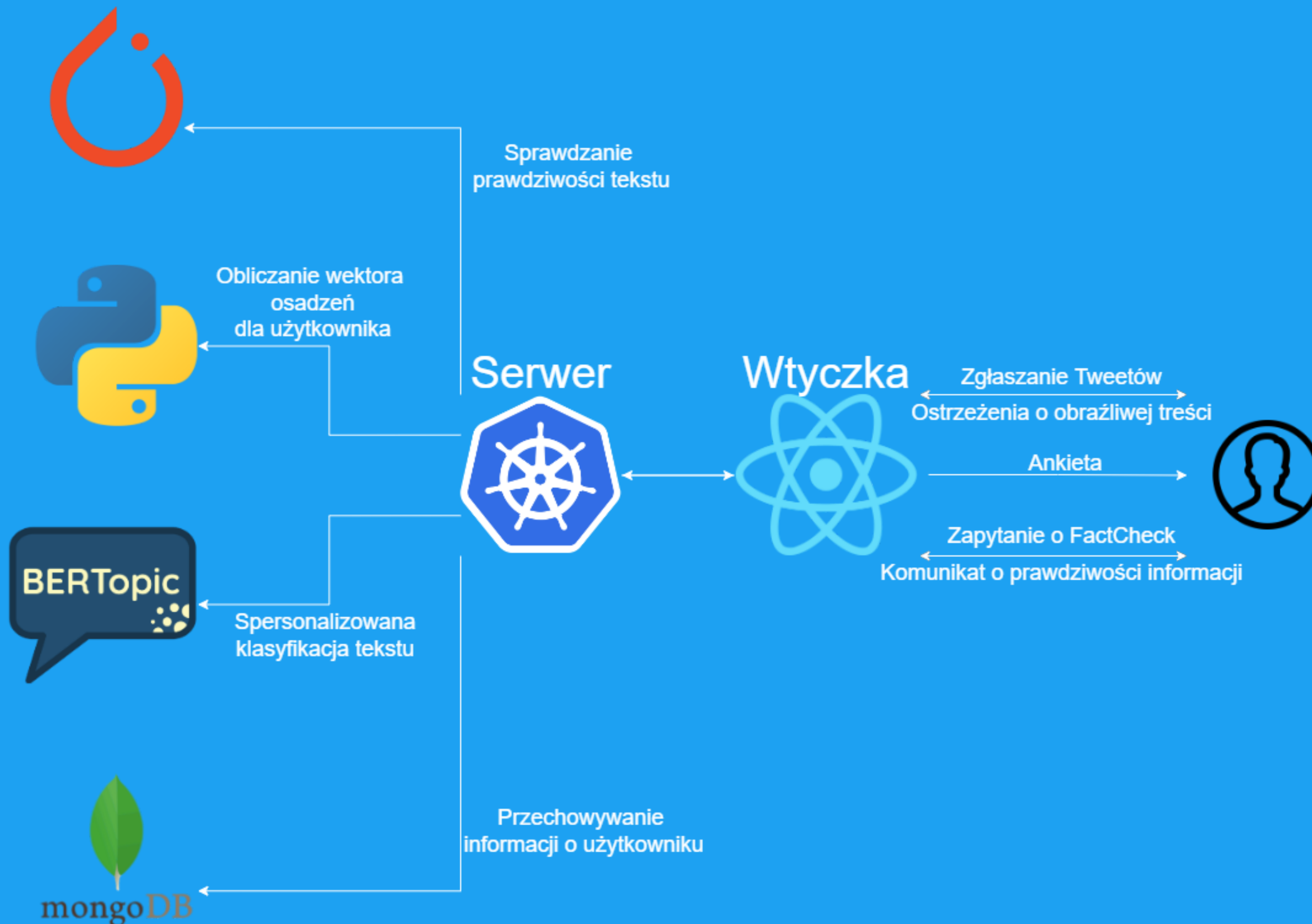
13



 FastAPI



# ARCHITEKTURA SYSTEMU



# WTYCZKA - WSTĘPNA ANKIETA

15

TwitterBuddy

Dashboard

Reports

Dashboard

Mark posts you think are offensive

Some comments may be highly inappropriate

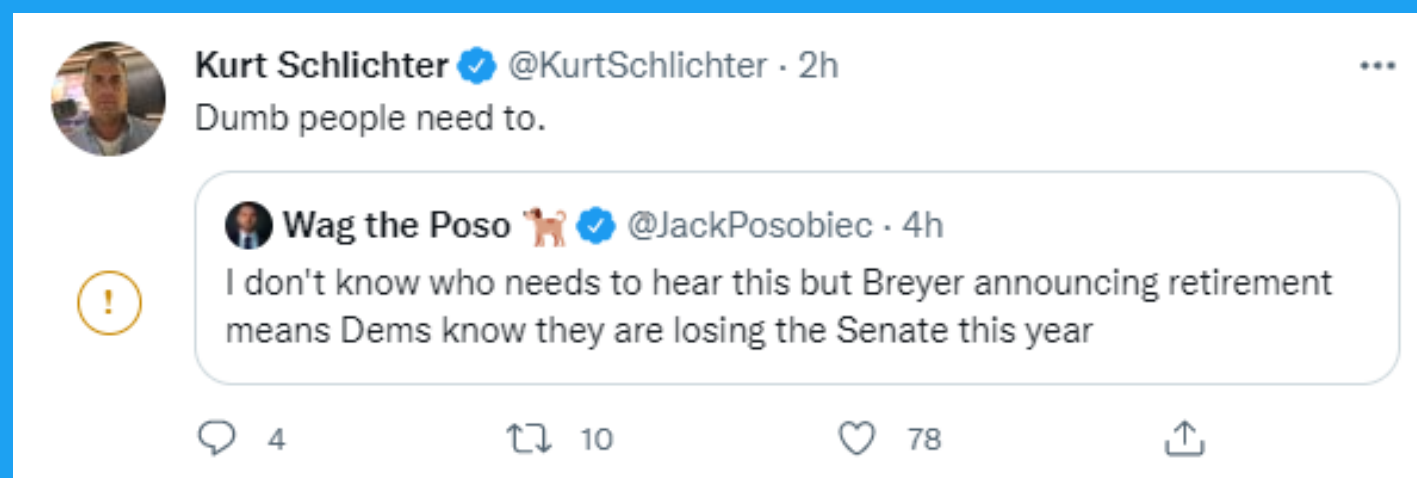
POST	OFFENSIVE
Young one, you can block me all you want. The point is you do not have a life. When you block me in less than 10 minutes at any time of day or night.	<input type="checkbox"/>
Only or last warning? Make up your mind please!!  This was your first warning verbatim: ``This is the only warning you will receive for your disruptive edits. The next time you delete or blank page contents or templates from Wikipedia, as you did to Talk:Glock pistol, you will be blocked from editing. Dave1185 (talk) ``  I did not delete content nor did I blank anything, I deleted a comment that I made while using someone else s account. That was an accident. I did not realize that my brother was signed in and after I though about it I realized that I was kind of a dick to the other editors, I deleted it. You responded by threatening with a ban for deleting what was mine to delete.  After I responded in a condescending way to your threat, you apparently thought that i merited a second warning: ``This is the last warning you will receive for your disruptive comments. If you continue to make personal attacks on other people as you did at User talk:Dave1185, you will be blocked for disruption. Comment on content, not on other contributors or people. Dave1185 (talk) ``  Then there's this: ``Last warning! Do this again and I'll see to it that you get nominated at WP:ANI or WP:AIV faster than you can spell your name out. BTW, I'm not an Admin. ``  Well feel free to ban me or report me for being a dick and deleting something I shouldn't have posted to begin with. Errors while editing happen from time to time. Feel free to nominate me for anything but prom queen. I could really care less if you are offended. Whats more I find that your lack of good faith and threats to be childish. Have a blast, dick.	<input type="checkbox"/>
CONSUMER WATCH!  Never do business with this bank. It's not the real Wachovia bank anymore. First Union, which has always been a bad bank, bought out Wachovia so that it could change its name. First Union had a terrible reputation for incompetence and fraud, whereas Wachovia had a spotless reputation.	



# WTYCZKA - DZIAŁANIE

## 4 rodzaje podejmowanej akcji:

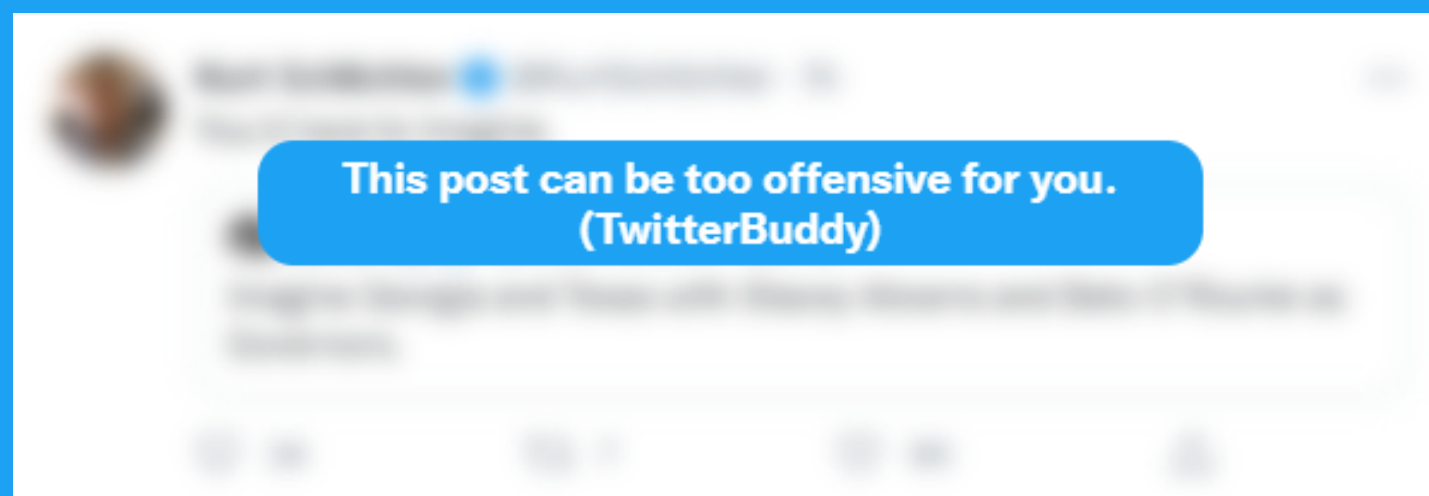
1



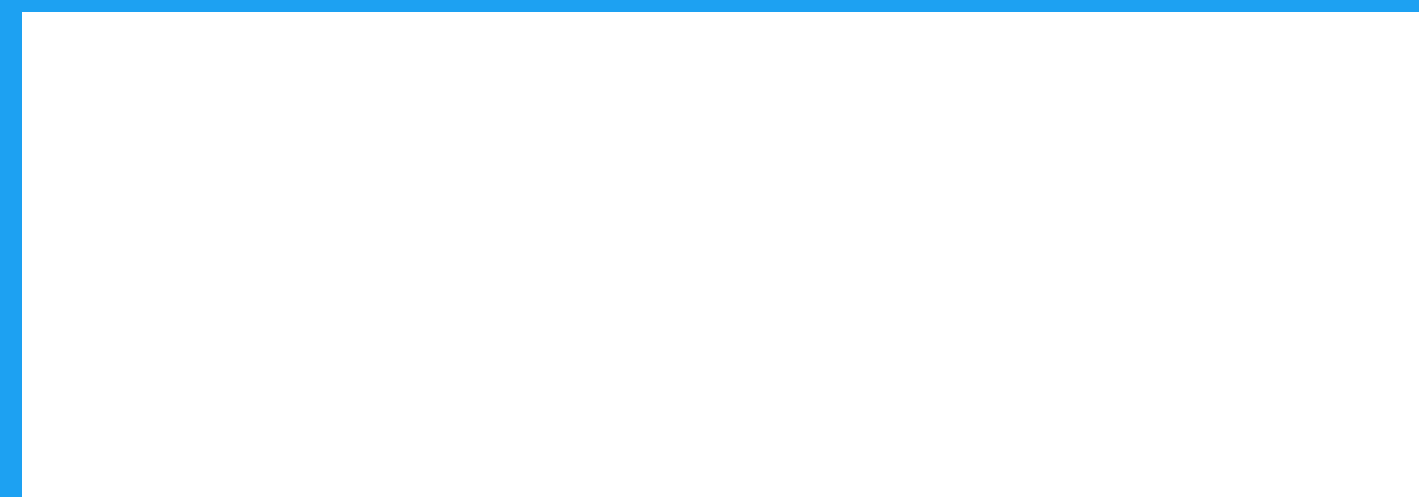
2



3



4



# WTYCZKA - DZIAŁANIE

## Fact check & feedback



**The New York Ti...**  @nytim... · 28m ...

Rodolfo Hernández, empresario, estrella de TikTok, candidato disruptivo y ¿próximo presidente de Colombia?




nytimes.com  
Rodolfo Hernández, empresario, estrella de TikTok, candidato disruptivo y ...




**Prediction feedback**


**Fact-check**

13 10 44




**The New York Ti...**  @nytim... · 26m ...

A New York Times analysis of more than 1,000 photos found that Russia has used hundreds of weapons in Ukraine that are widely banned by international treaties. Russia's attacks made widespread use of weapons that kill, maim and destroy indiscriminately.



nytimes.com  
What Hundreds of Photos of Weapons Reveal About Russia's Brutal War ...



**Prediction feedback**

**Fact-check**

**TRUE**

54 218 356

# #TwitterBuddy



# AGENDA

1. Story (Adam) 1.5min
2. Problem biznesowy (Adam) 1min
3. Propozycja rozwiązania(Adam) 1min
4. Model do hate speech (Jan) 2min
5. Model do fake news(Przemo) 2min
6. Stack technologiczny(Mati) 1min
7. Demo – screeny(Jan) 2min

# WTY CZKA - WSTĘPNA ANKIETA

14

**TwitterBuddy** Dashboard Reports

Dashboard

**We need some basic info about you**

Your gender

Male

Are you a native english speaker?

No

Your education

High school

Send Answers

# WTYCZKA - WSTĘPNA ANKIETA

15

