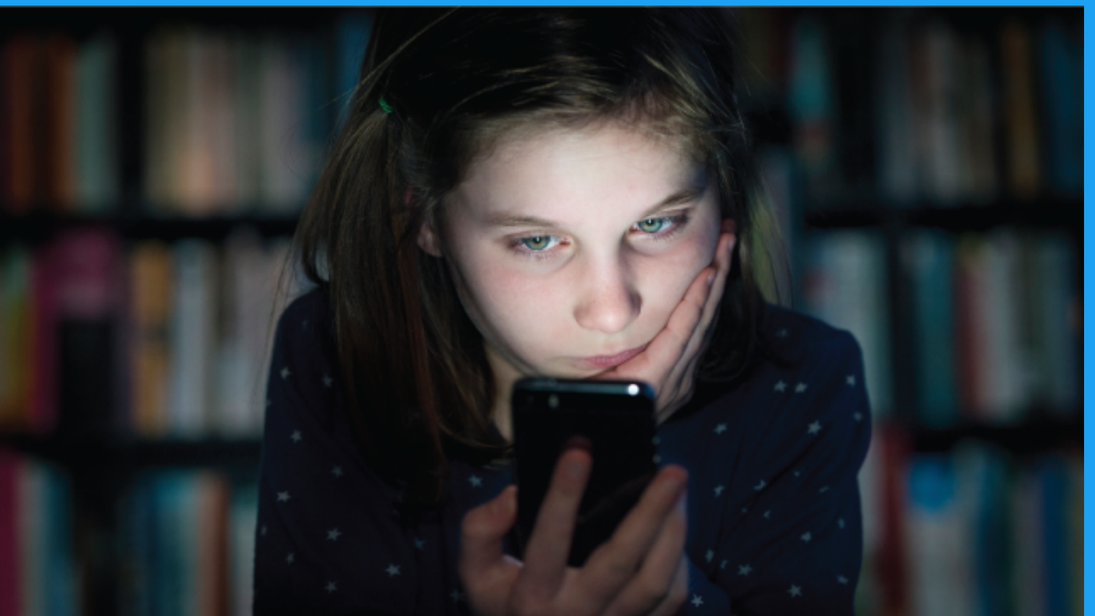
#TwitterBuddy

@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"

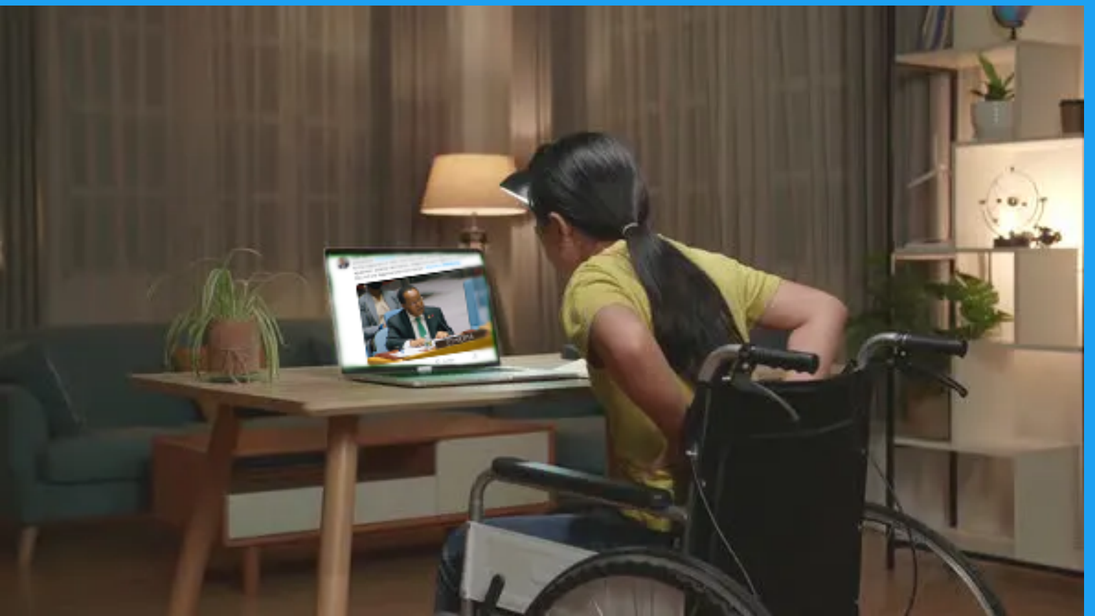
Pew Research Center [1]



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

"73% osób z niepełnosprawnością doświadczyło cyberprzemocy"

Toth et al. [2]



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

"Osoby LGBTQIA+ są niemal dwukrtonie 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

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

"Ostatnie badania podkreślają związek między rozprzestrzenianiem fałszywych informacji a nękaniem"

Maftei et al. [4]

[4] Alexandra Maftei, Andrei-Corneliu Holman, and Ioan-Alex Merlici. 2022. Using fake news as means of cyber-bullying: The link with compulsive internet use and online moral disengagement. Comput. Hum. Behav. 127, (Feb 2022)

PROBLEM BIZNESOWY

Moderacja nie bierze pod uwagę potrzeb indywidualnych użytkowników

Brak możliwości uwzględnienia osobistej wrażliwości każdego człowieka

Całkowite pominięcie procesu weryfikacji prawdziwości stwierdzeń

NASZE ROZWIĄZANIE

Interaktywna wtyczka przeglądarkowa

WARTOŚĆ BIZNESOWA

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

SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

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

Metryki

ankieta wstępna

wektor reprezentujący użytkownika

łatwość douczania

SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

zbiór tekstów

SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

zbiór tekstów

modelowanie tematyczne

SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

zbiór tekstów

modelowanie tematyczne

klasteryzacja

SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

zbiór tekstów

modelowanie tematyczne

klasteryzacja

obliczanie metryk użytkowników na podstawie ich stosunku do klastrów

SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI

zbiór tekstów

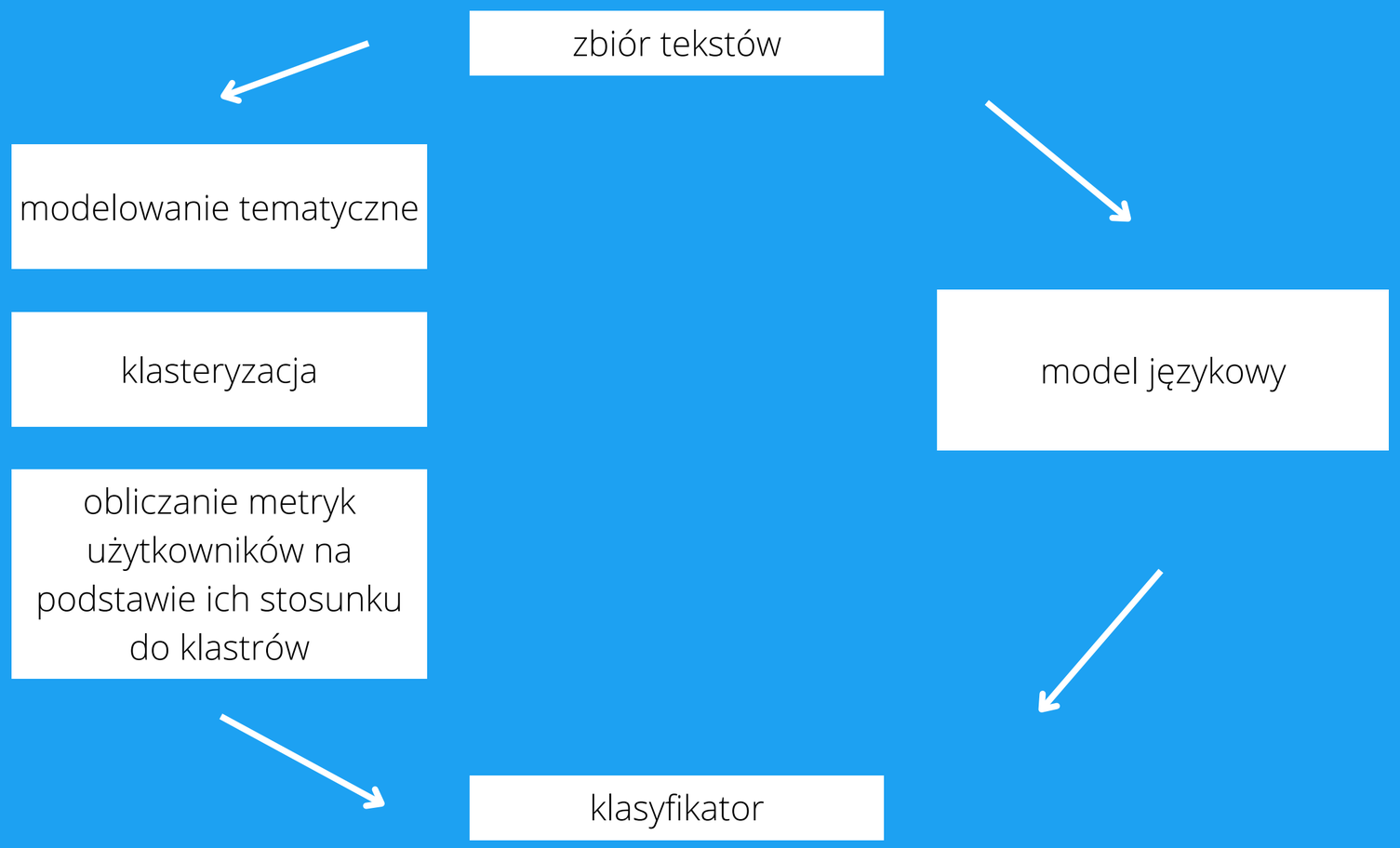
modelowanie tematyczne

klasteryzacja

obliczanie metryk użytkowników na podstawie ich stosunku do klastrów

model językowy

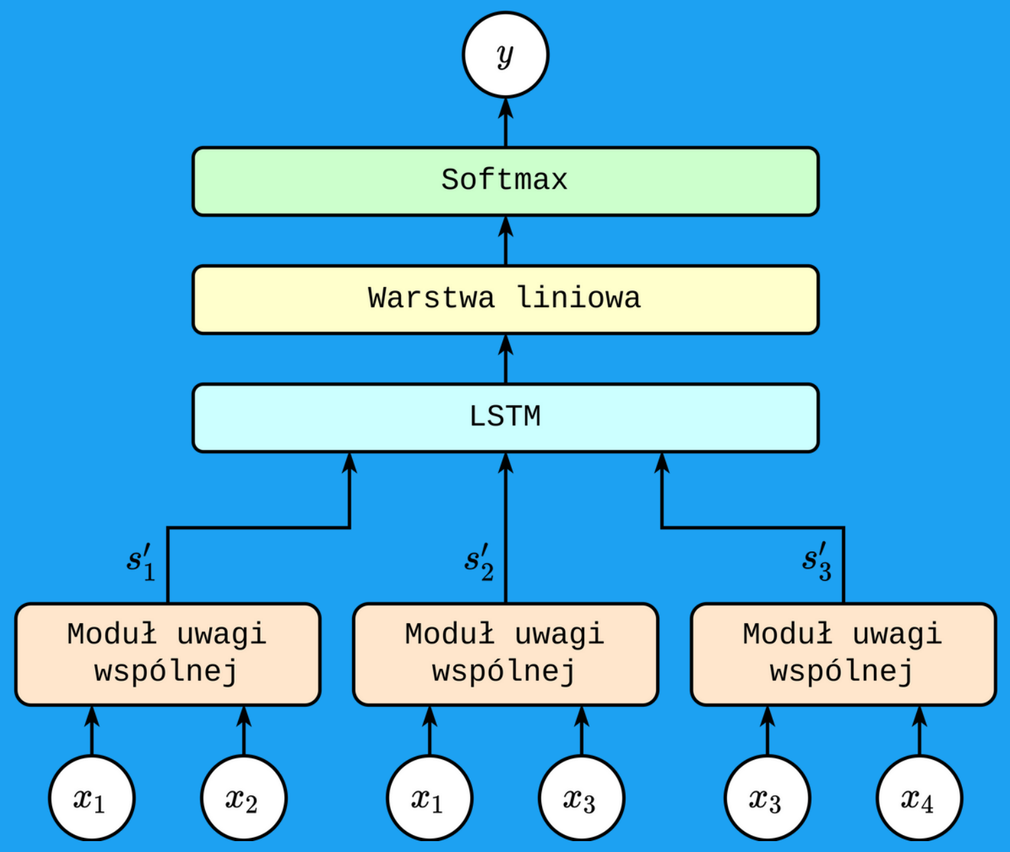
SPERSONALIZOWANE WYKRYWANIE MOWY NIENAWIŚCI



MODEL WYKRYWANIA FAŁSZYWYCH TREŚCI

- 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



- 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

FastAPI

scikit learn

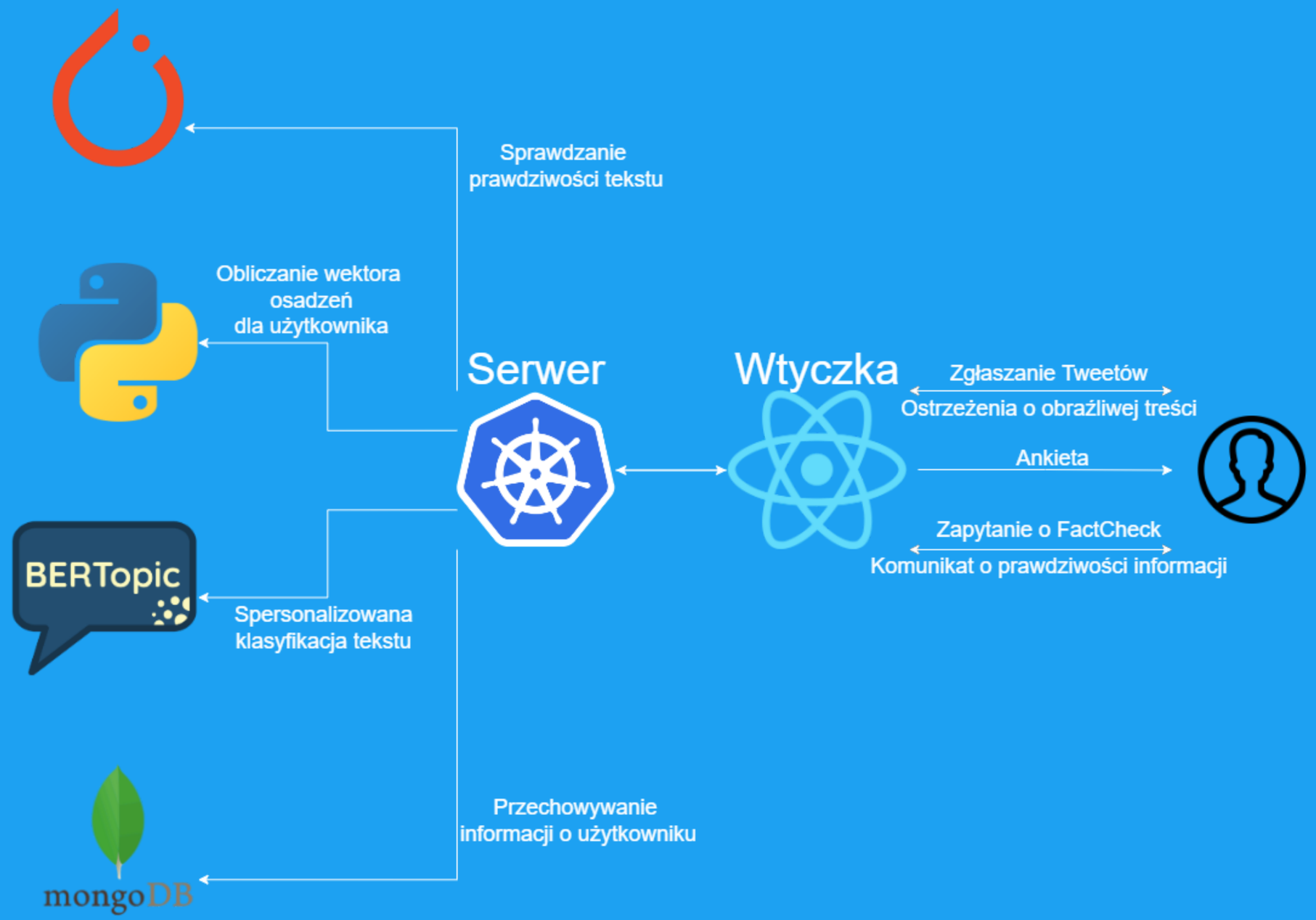
mongoDB

DVC

BERTopic

docker

ARCHITEKTURA SYSTEMU



WTYCZKA - WSTĘPNA ANKIETA

TwitterBuddy | Dashboard | Reports

Dashboard

Mark posts you think are offensive

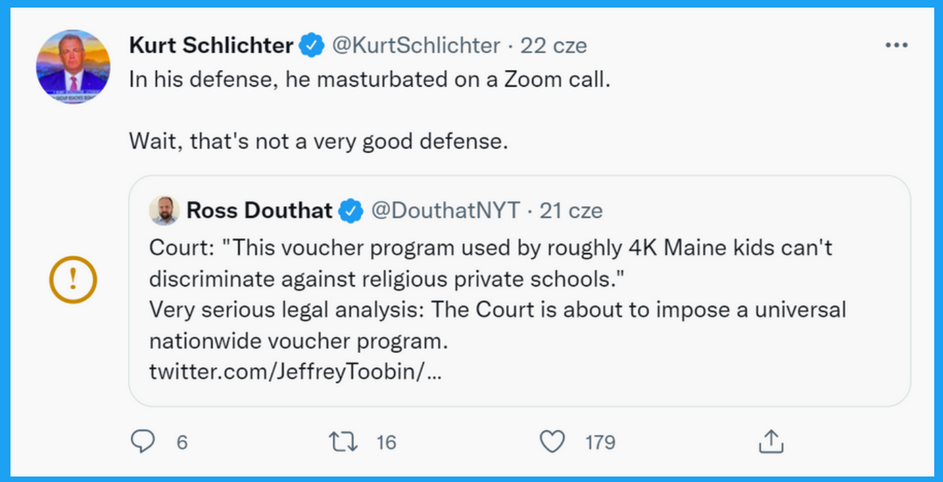
Some comments may be highly inappropriate

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.	☐
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.	☐
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.	

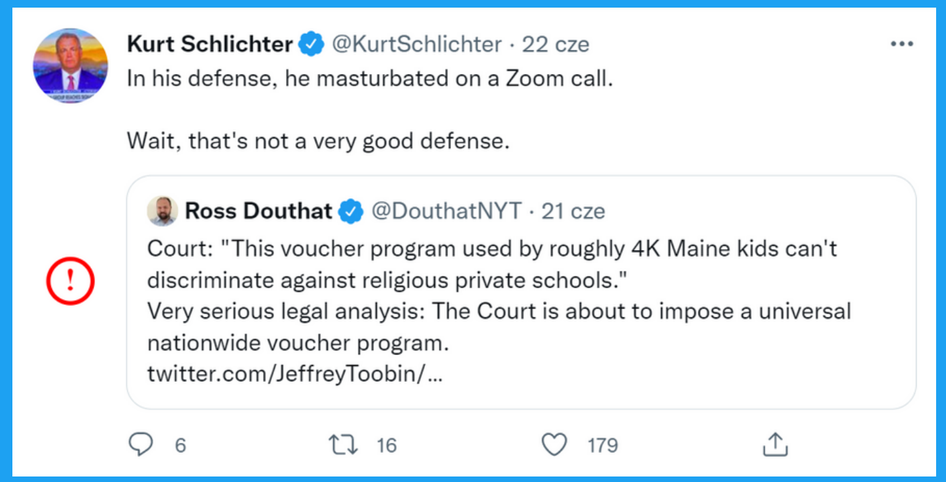
WTYCZKA - DZIAŁANIE

4 rodzaje podejmowanej akcji:

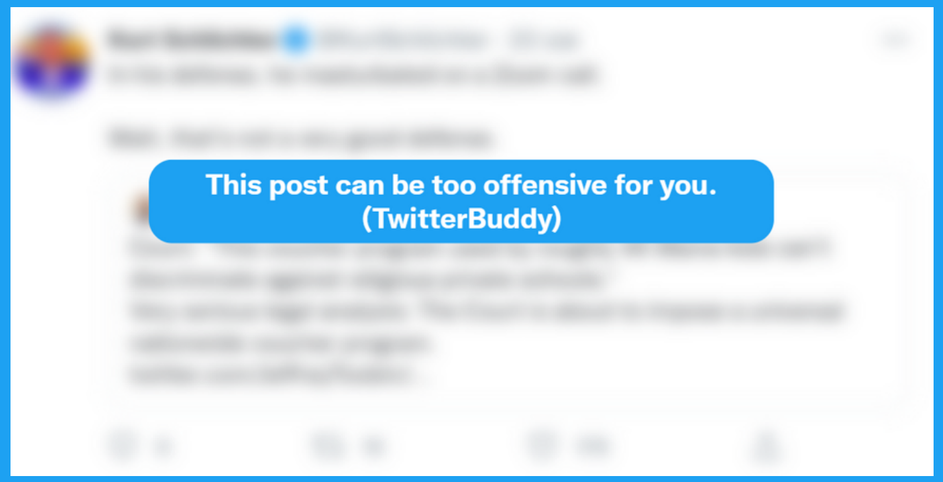
1



2



3



4

WTYCZKA - DZIAŁANIE

Fact check & feedback

TRUE

#TwitterBuddy