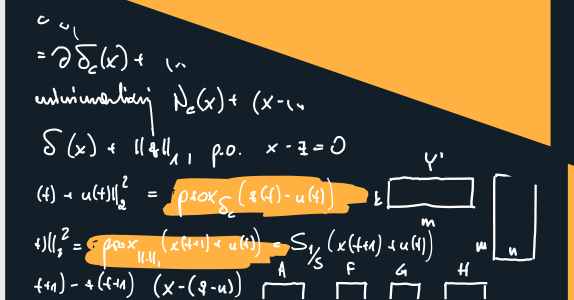
DETEKCJA DEEPFAKES ORAZ FAŁSZYWYCH REKLAM POPRZEZ ANALIZĘ TEKSTU I TECHNIK MANIPULACYJNYCH

mgr inż. Alicja Martinek

dr inż. Ewelina Bartuzi - Trokielewicz



17. lutego 2025

SCIENCE NASK

Czym jest DEEPFAKE

technika manipulacji materiału audiowizualnego

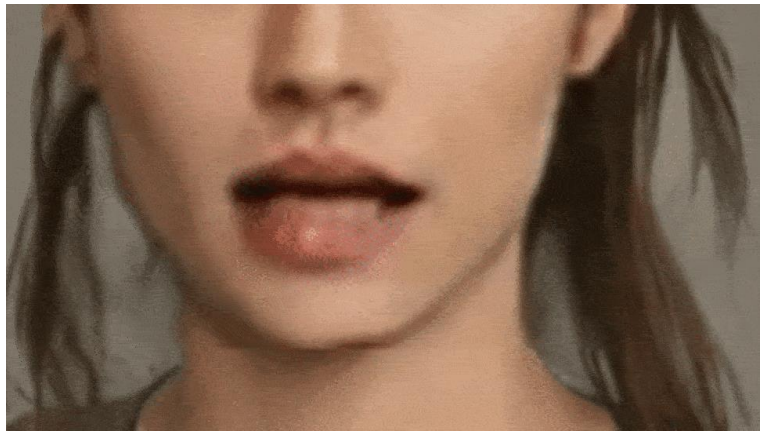
przy użyciu sztucznej inteligencji

w których twarz lub głos osoby są zastąpione inną treścią

CZYM JEST DEEPFAKE

- AUDIO

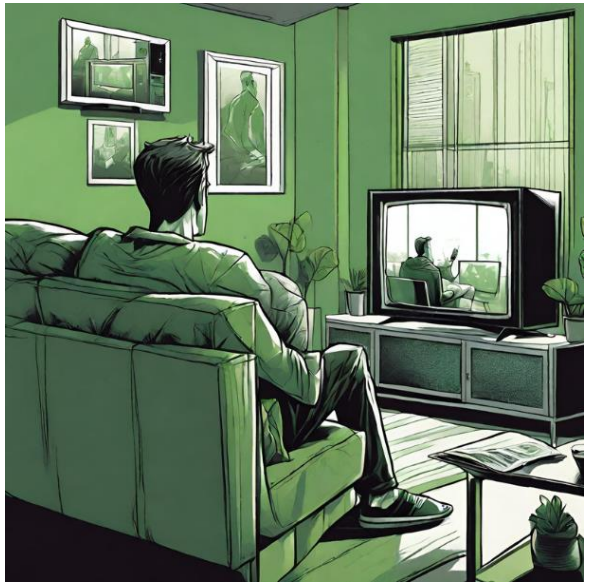
- OBRAZ



- WIDEO

Dlaczego **DEEPFAKE** są tak skuteczne i będą stanowiły coraz większy problem?

większa wiara **w to co widzimy i słyszymy,** ograniczone zaufanie dotyczy raczej tekstu



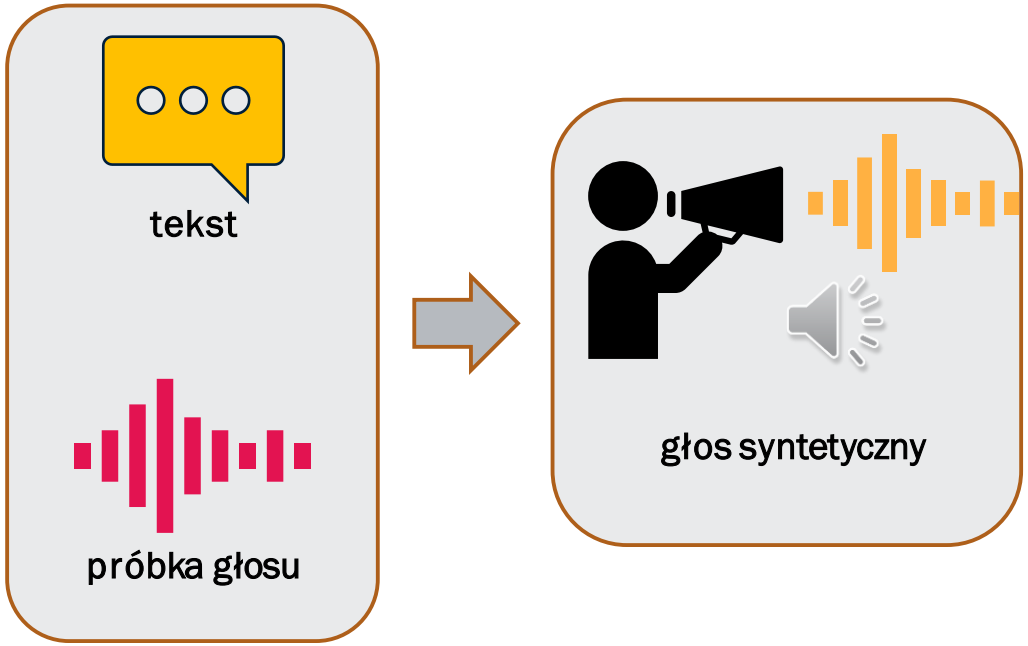
przyzwyczailiśmy się do przekazu **wiadomości** w formie **audiowizualnej**

rozwija się coraz więcej łatwo dostępnych **metod generowania deepfake**

internet staje **się najlepszym medium** informacyjnym → **więcej frontów ataku**

RODZAJE KLONOWANIA GŁOSU

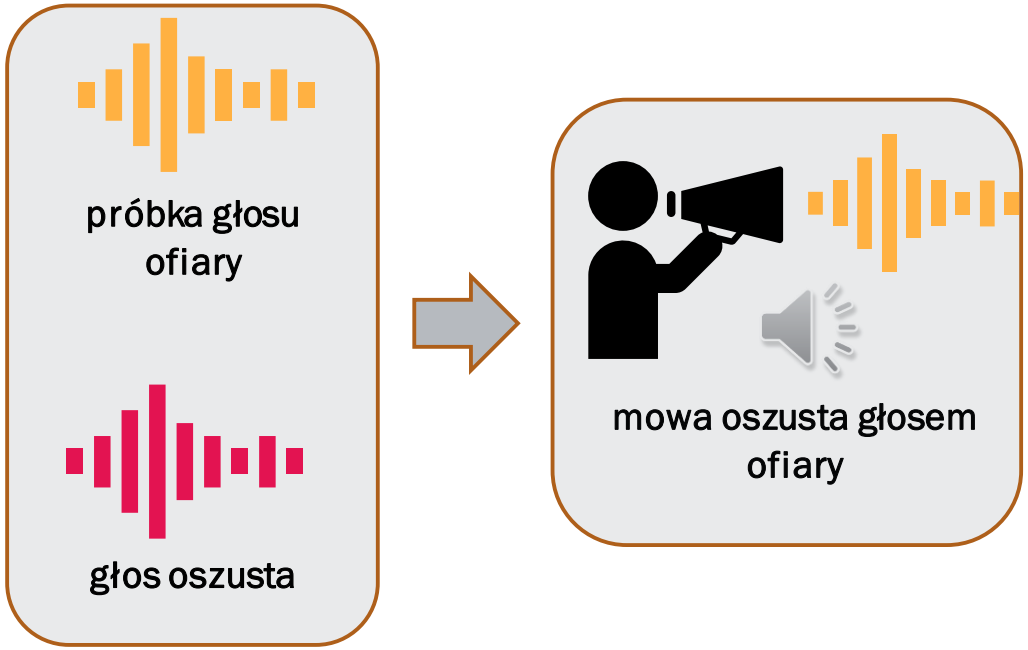
TEXT TO SPEECH - TTS



Cechy syntetycznej mowy:

- mowa maszynowa
- nacechowanie stylem mowy z próbki głosowej

SPEECH TO SPEECH - STS



Cechy syntetycznej mowy:

- mowa nacechowana emocjami
- trudno odwzorować styl mówienia ofiary
- jakość zwiększa się z podobieństwem barwy głosu ofiary do głosu oszusta

POTRZEBNE DANE DO AUDIO DEEPFAKE:

TTS

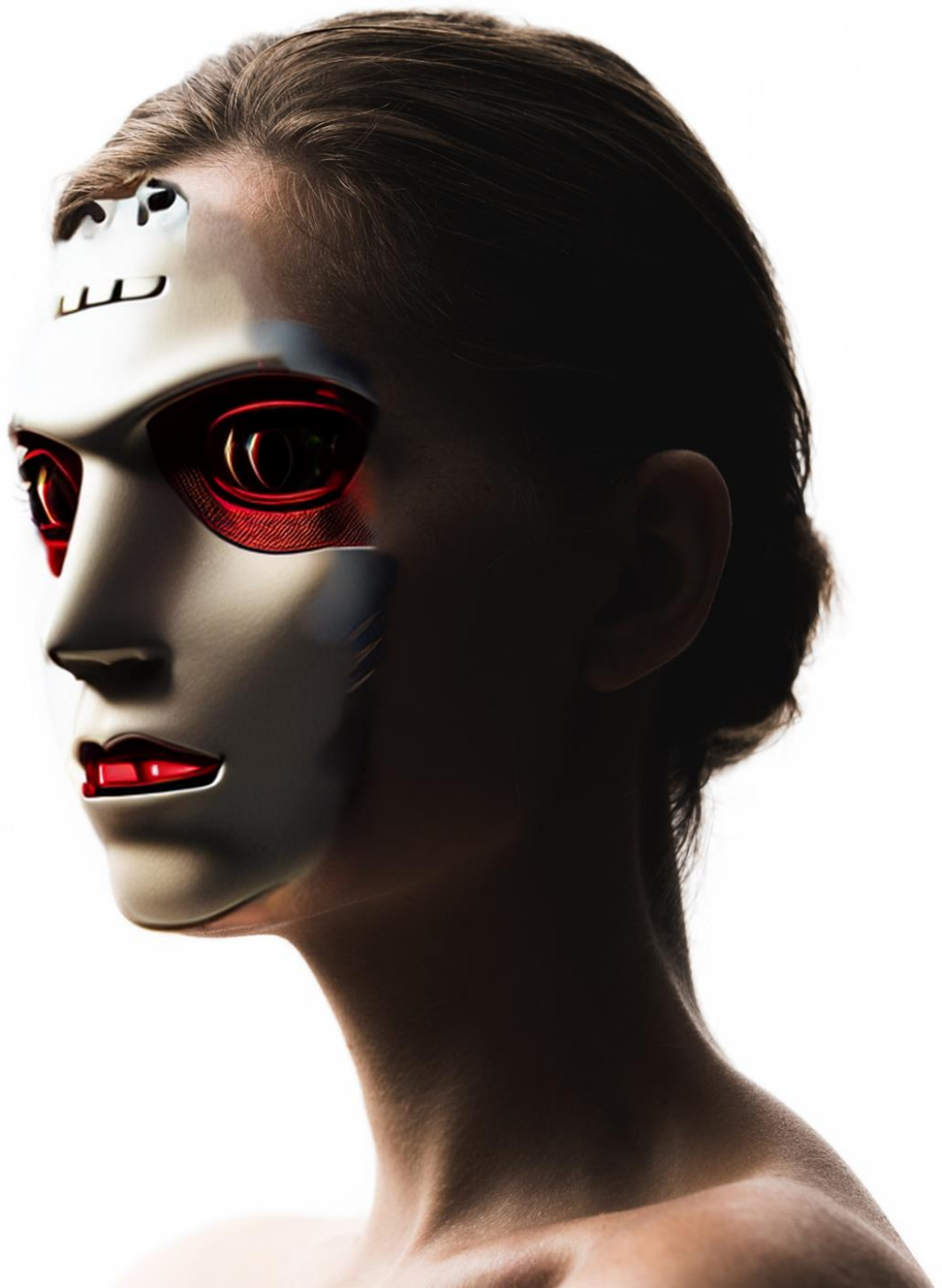
<1 min
nagrania audio

STS

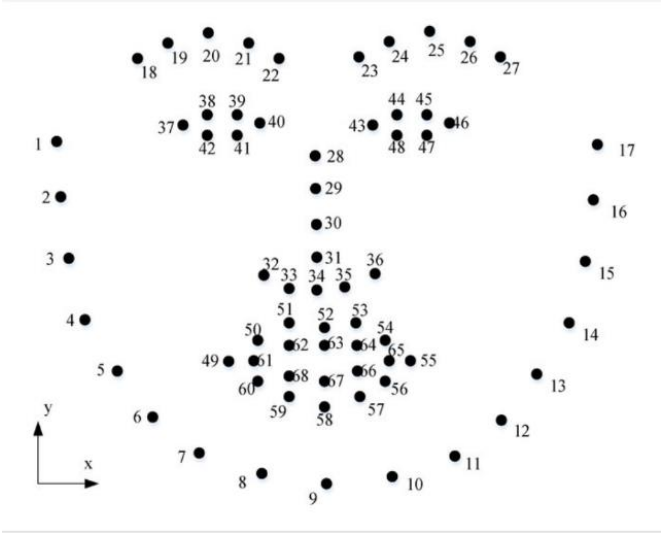
~ 5 min
nagrania audio

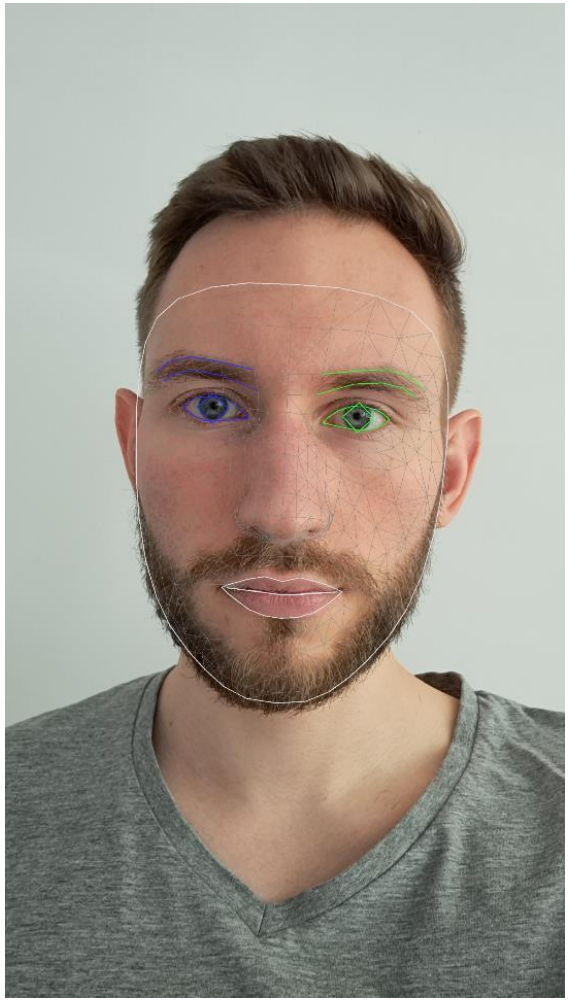
RODZAJE DEEPFAKE’ów TWARZY

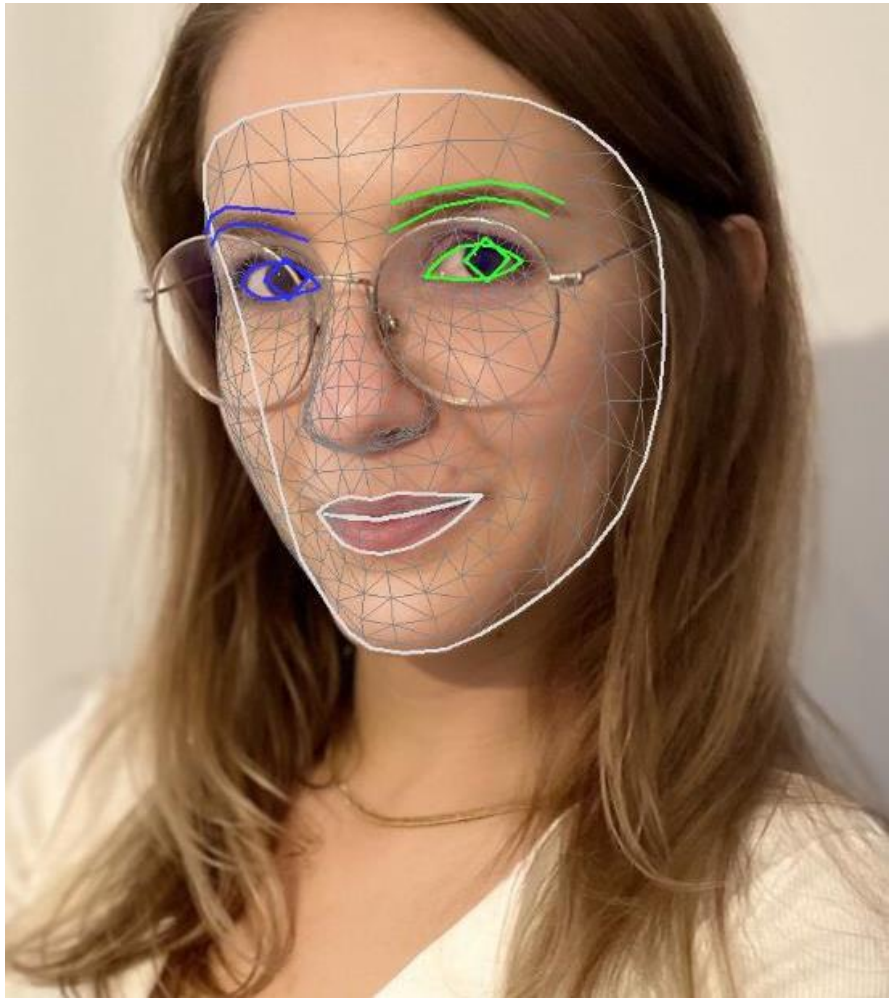
Jak maszyny widzą twarz?



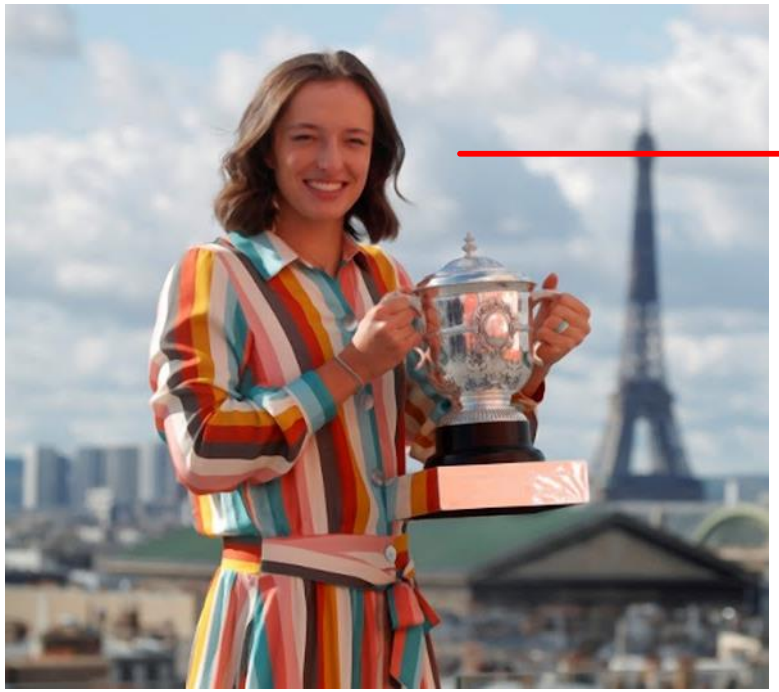
JAK MASZYNY WIDZĄ TWARZ?



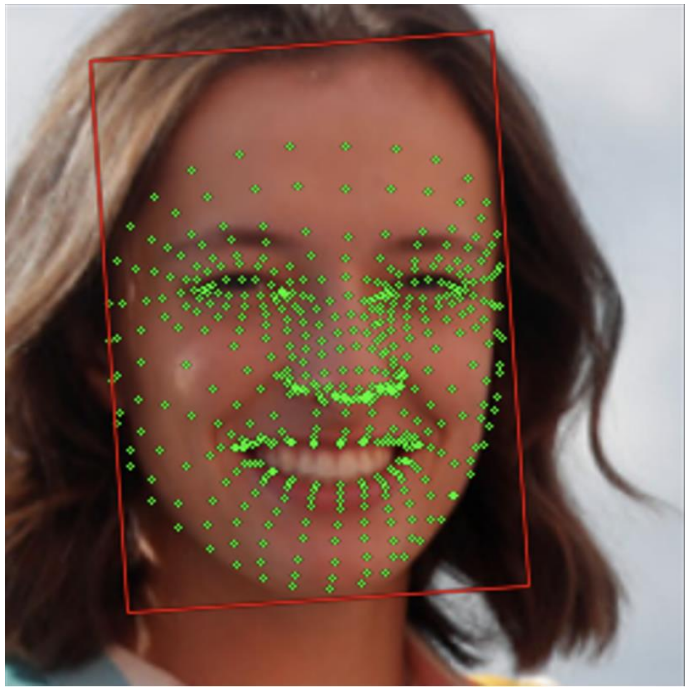




PODSTAWY TECHNOLOGICZNE PODMIANY TWARZY

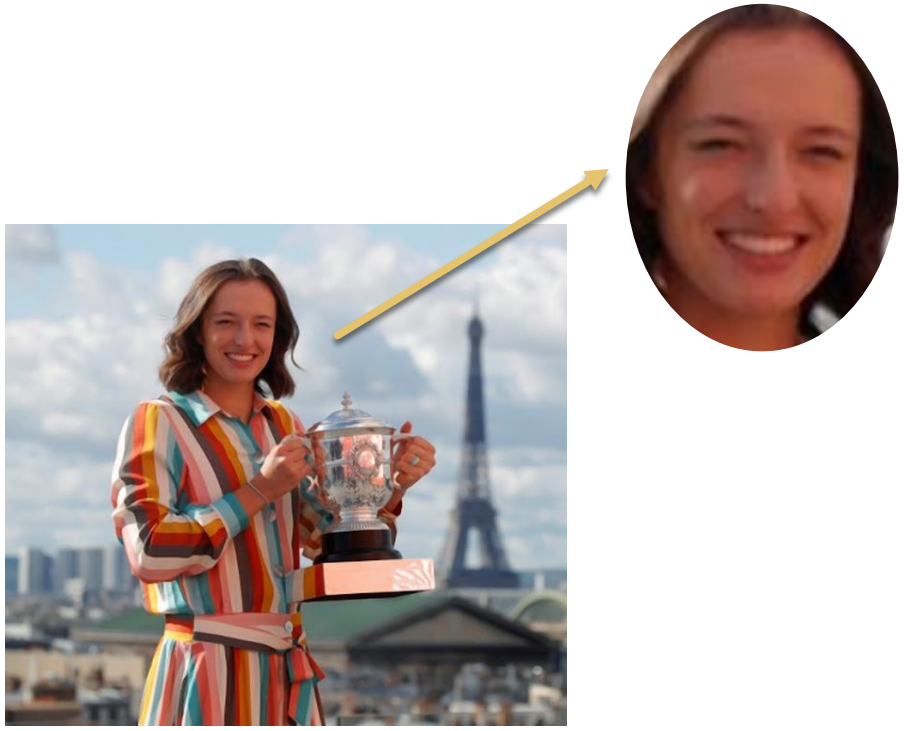


ekstrakcja twarzy



tak maszyny widzą twarze

PODSTAWY TECHNOLOGICZNE PODMIANY TWARZY

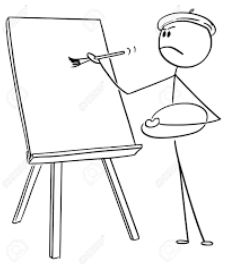


ekstrakcja twarzy

PODSTAWY TECHNOLOGICZNE PODMIANY TWARZY

ekstrakcja twarzy

enkoder



PODSTAWY TECHNOLOGICZNE PODMIANY TWARZY

ekstrakcja twarzy

enkoder

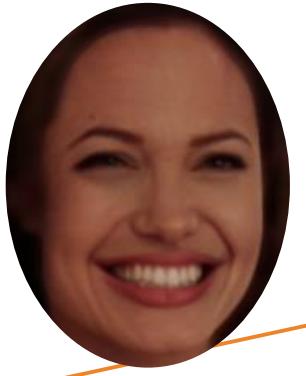
przestrzeń ukryta

PODSTAWY TECHNOLOGICZNE PODMIANY TWARZY

ekstrakcja twarzy

enkoder

przestrzeń ukryta



dekoder

PODSTAWY TECHNOLOGICZNE PODMIANY TWARZY

ekstrakcja twarzy

enkoder

przestrzeń ukryta

dekoder

podmiana twarzy

Podmiana twarzy - *face swap*

Łączenie twarzy - *face morph*

Modyfikacja atrybutów - *atribute modification*

Animowanie twarzy - *face animation*

Syntetyczne twarza - *synthetic face*

Rozwój DEEPFAKE

https://aifinance.medium.com/generative-everything-the-technology-that-brought-you-deepfakes-soon-brings-you-3777a8bccee7

Synchronizacja ust ze ścieżką audio

- *lipsync*

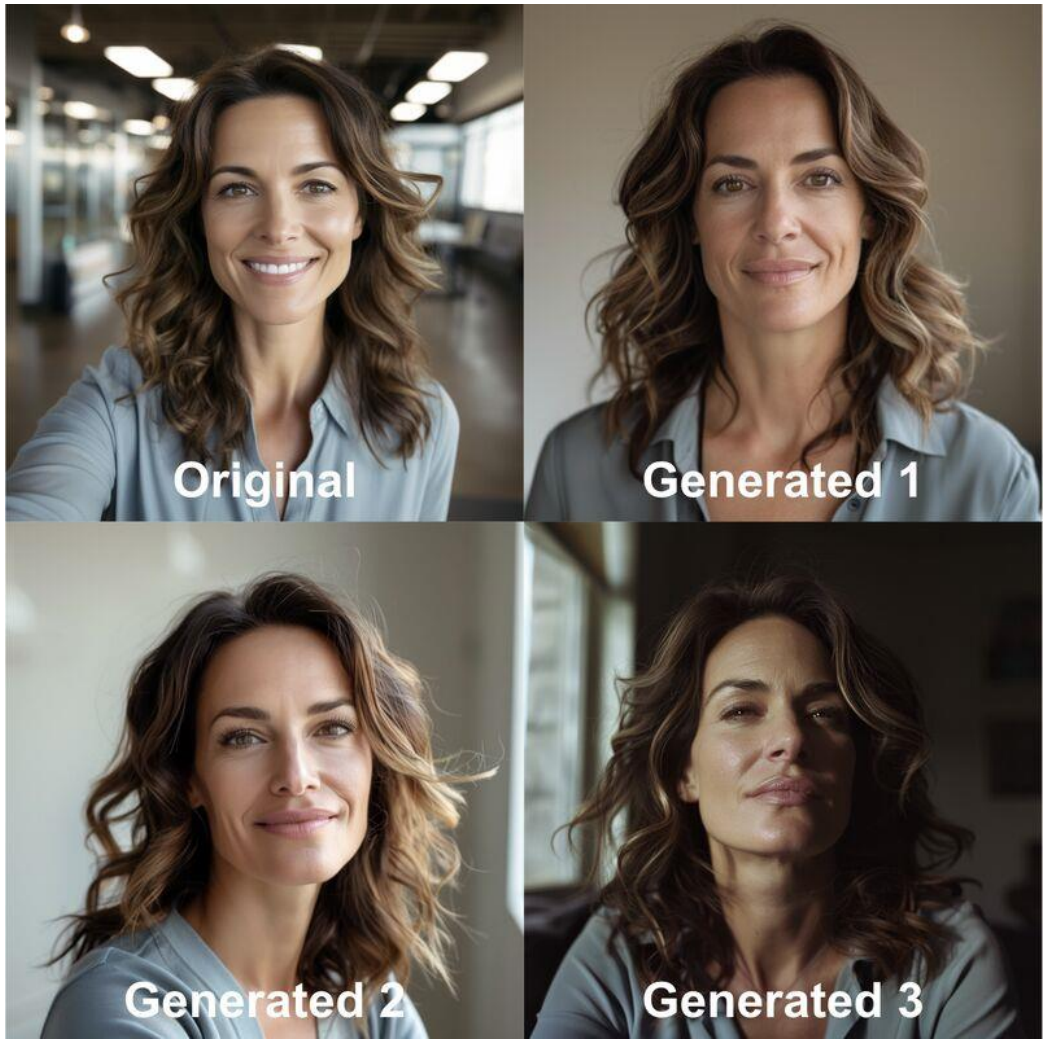
Synchronizacja ust ze ścieżką audio

- *lipsync*

- **wideo** referencyjne + nowa **ścieżka głosowa**

Prompt-base deepfake

Prompt-base deepfake



Linkedin/midjourney

- nowa funkcja Midjourney, Character Reference, generuje biometrycznie dokładne twarze

EMO – EMOTE PORTRAIT ALIVE

- **jedno** zdjęcie + **ścieżka głosowa**

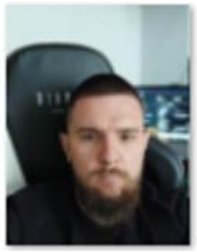
https://humanaigc.github.io/emote-portrait-alive/

ILE DANYCH POTRZEBUJEMY ABY WYTWORZYĆ DEEPFAKE TWARZY?

POTRZEBNE DANE DO WIZUALNYCH DEEPFAKE:

dziesiątki tysięcy w przypadku tworzenia modelu twarzy

POTRZEBNE DANE DO WIZUALNYCH DEEPFAKE:



... ale czasem wystarczy też jedno, dobre zdjęcie

Dlaczego o tym mówimy?

TTS

- **generowane głosy** często zawierają błędy w wymowie słów, intonacji, czy mają nieprawidłowy akcent

ASR

– występują **błędy w konwersji mowy na tekst**

- „zaokrąglanie” słów
- halucynacje,
- ucinania i powtórzenia

Maskowanie informacji

- coraz częściej fałszywe materiały są maskowane – zarówno w obszarze wizualnym, jak i audio, jednak przekaz pozostaje podobny

Wszystkie te czynniki wpływają na analizę treści!

Zagrożenia związane z deepfake’ami

NISZCZENIE REPUTACJI

The Twitch Community Has Been Rocked By A Deepfake Porn Scandal. “It Was Quite Horrifying,” One Victim Said.

“I was wishing for eye bleach,” streamer Sweet Anita told BuzzFeed News.

Steven Asarch
BuzzFeed Contributor

Posted on February 6, 2023 at 4:45 pm

Tworzą erotyczne zdjęcia koleżanek ze szkoły: "Tak się bawimy". Coraz więcej ofiar oszustwa

CYBERPRZESTĘPCZOŚĆ

DEZINFORMACJA

MANIPULACJA OPINIĄ PUBLICZNĄ

Kolejny deepfake zebrał żniwa. Tym razem celem był Joe Biden

publikacja
2024-02-26 22:24

Steve Kramer, konsultant sztabu Deana Philipsa, rywala prezydenta USA Joe Bidena w prawyborach Demokratów, przyznał się do skonstruowania i użycia deepfake'a, który udawał głos Bidena i zniechęcał do wzięcia udziału w styczniowych prawyborach w stanie New Hampshire.

oszustwa finansowe

WPŁYW NA LUDZI

- kradzież tożsamości
- oszustwa i wyłudzenia
- fałszywe reklamy inwestycji

oszustwa finansowe

WPŁYW NA LUDZI

- kradzież tożsamości
- oszustwa i wyłudzenia
- fałszywe reklamy inwestycji

oszustwa finansowe

WPŁYW NA LUDZI

- kradzież tożs
- oszustwa i wy
- fałszywe rekl

https://www.knf.gov.pl/dla_rynku/CSIRT_KNF / facebook.com

oszustwa finansowe

WPŁYW NA LUDZI

- kradzież tożsamości
- oszustwa i wyłudzenia
- fałszywe reklamy inwestycji
- fałszywe rekomendacje

Znany lekarz padł ofiarą oszustów. W oparciu o deepfake jego wizerunek wykorzystano do reklamy pseudoleków

Autorzy: PAP; oprac. JW • Źródło: PAP • Opublikowano: 21 listopada 2023 15:45 • Zaktualizowano: 21 listopada 2023 18:37

Dr Michał Sutkowski padł ofiarą kradzieży wizerunku i jego twarz pojawiła się w reklamach pseudoleków kardiologicznych i rzekomych preparatów na pasożyty - poinformowała reprezentująca go kancelaria prawna. Podkreśliła, że w sprawie złożone zostało już zawiadomienie o podejrzeniu popełnieniu przestępstwa.

Dr Michał Sutkowski, lekarz rodzinny, padł ofiarą złodziei wizerunku. Fot. PAP/Tytus Żmijewski

oszustwa finansowe

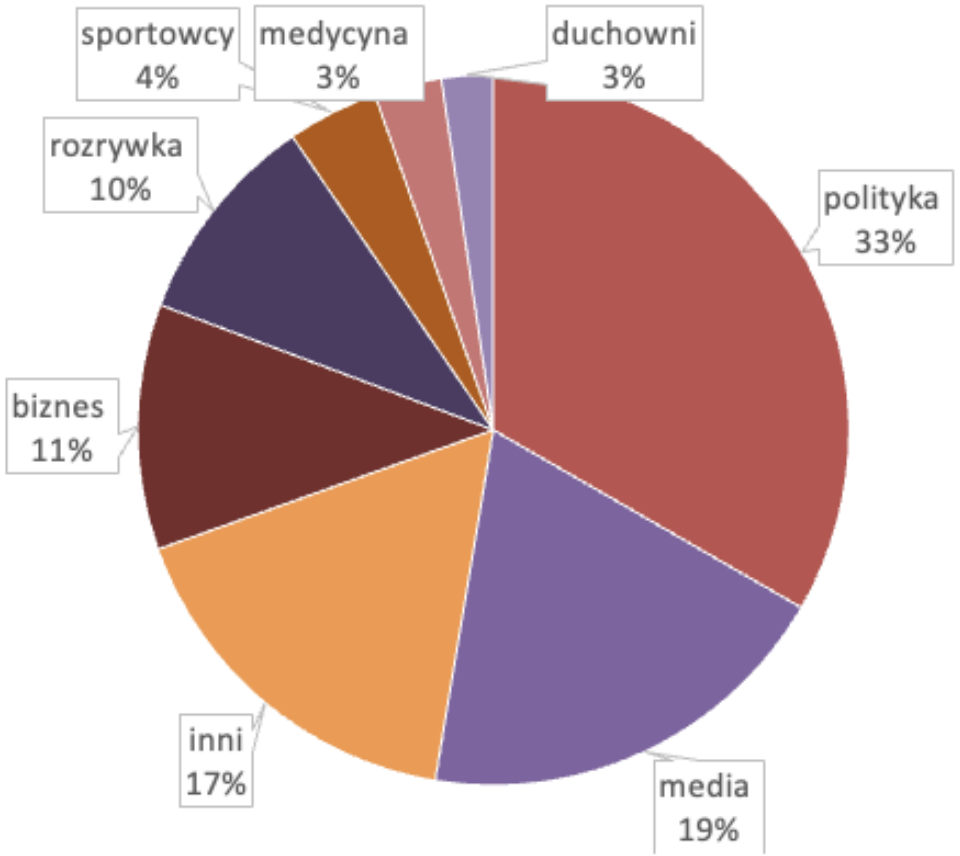
WPŁYW NA LUDZI

- Fałszywa pomoc

www.facebook.com

www.facebook.com

Wykorzystywane wizerunki w fałszywych reklamach - pierwsza połowa 2024

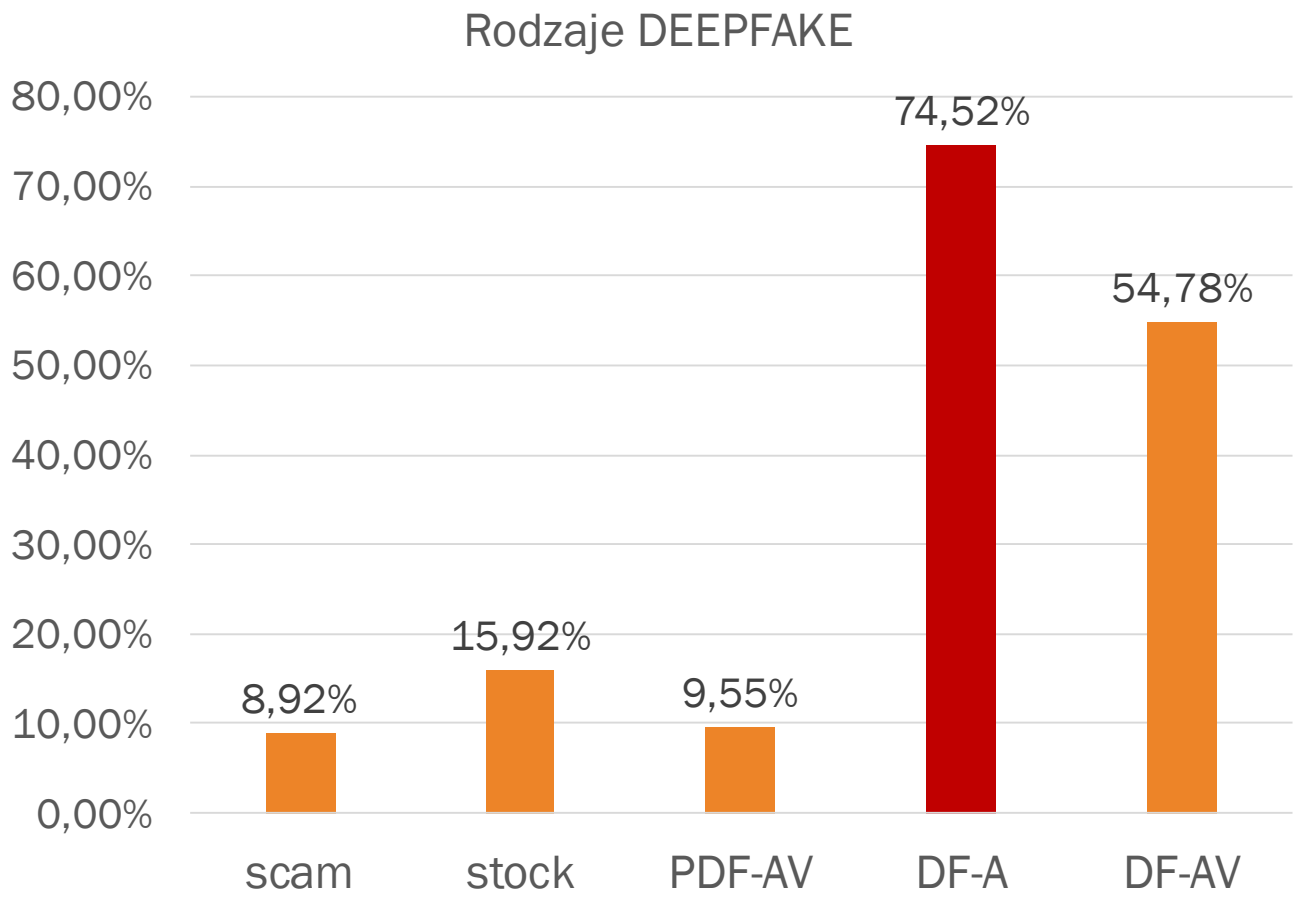


ponad **7600** fałszywych reklam ze zmanipulowaną treścią audiowizualną

wykorzystano ponad **170** wizerunków znany osób

Analiza na podstawie danych zbieranych przez NASK i KNF

STATYSTYKI Z ZEBRANYCH DANYCH

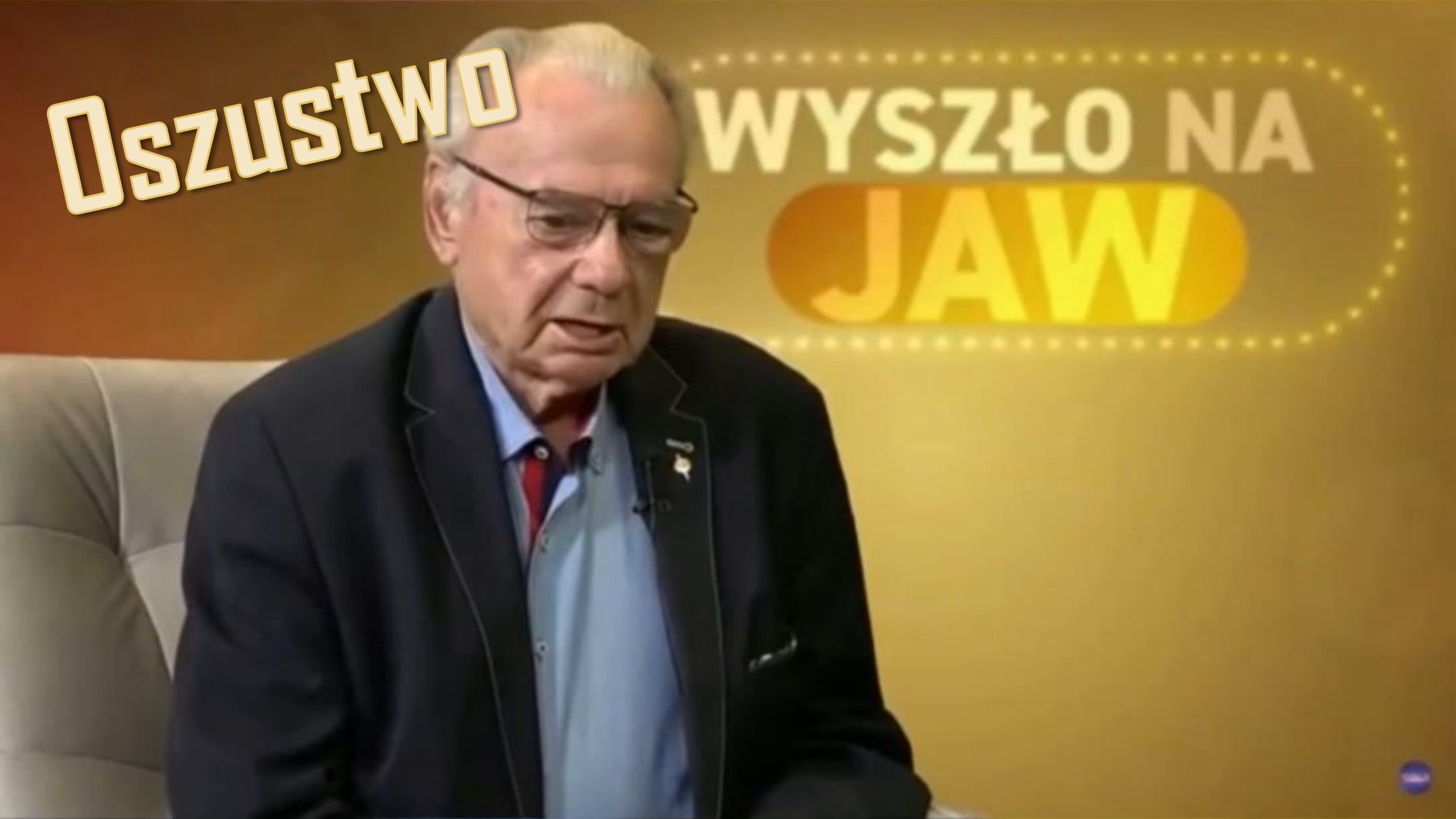


Wnioski:

- największą popularnością cieszą się materiały DF-A, czyli **audio** DF, gdzie cała ścieżka dźwiękowa jest zmanipulowana,
- niecałe 55% danych stanowią **jednolite** DF-AV
- wiele materiałów zawiera dodatkowo materiały **prawdziwe** i wizualizacje **stockowe** (PDF-AV)

legenda: **DF** – deepfake; **PDF** – partial-DF; **A** - audio; **V** - video

WyszłoNaZdrowie. "Kradzież wizerunku znanego lekarza. Prof. Krzysztof Bielecki padł ofiarą oszustów." *YouTube*, https://www.youtube.com/watch?v=3_Tog86RO5Y. Dostęp: 5.12.2024


Oszustwo
WYSZŁO NA
JAW

Jak wykryć DEEPFAKE?

- Niezgodność ruchu warg z treścią

www.facebook.com

Jak wykryć **DEEP**FAKE?

- Niezgodność mowy ciała z przekazem

www.facebook.com

Jak wykryć DEEPFAKE?

- Artefakty w obrębie twarzy

www.facebook.com

Jak wykryć DEEPFAKE?

- Błędy w renderowaniu okolicy ust

www.facebook.com

Jak wykryć **DEEP**FAKE?

- Błędy w odtwarzaniu uzębienia

Jak wykryć **DEEPFAKE**?

- Nietypowe intonacje

www.facebook.com

Jak wykryć **DEEP**FAKE?

- Zmiana akcentu

www.facebook.com

Jak wykryć DEEPFAKE?

- Zmiana głośności

www.facebook.com

Jak wykryć DEEPFAKE?

- Nieprawidłowa odmiana słów, głównie liczb

www.facebook.com

Jak wykryć DEEPFAKE?

- Błędy logiczne w treści

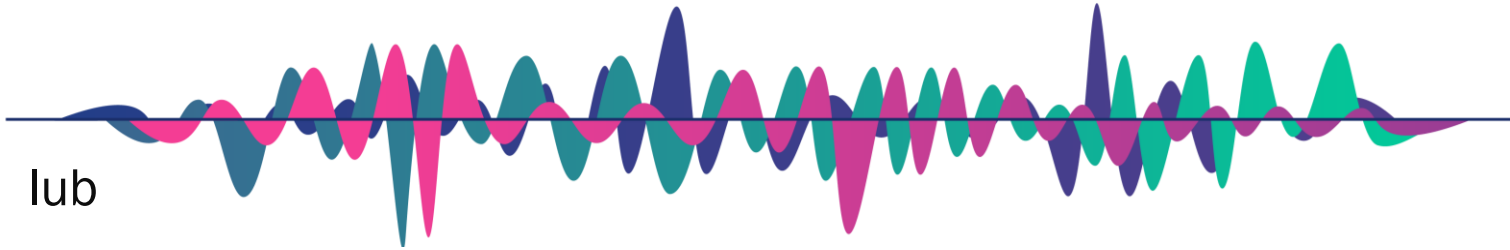
www.facebook.com

Jak wykryć DEEPFAKE?

- Niezgodność barwy głosu

www.facebook.com

Jak wykryć DEEPFAKE?

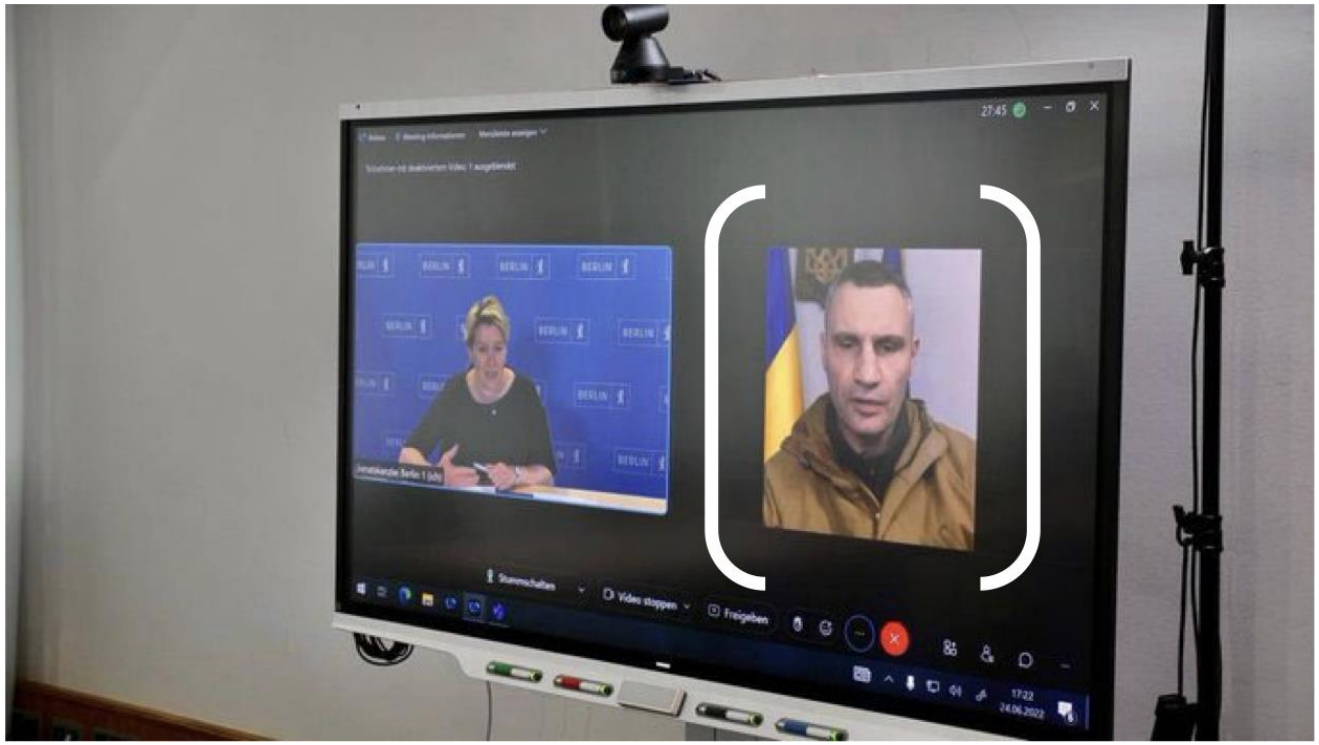


- brak naturalnych odgłosów tła lub niespójne hałas w tle
- „cyfrowe chrząszcze”

www.facebook.com

DLACZEGO POTRZEBUJEMY DETEKTORÓW DEEPFAKE

- > 700 tys. h filmów dziennie pojawia się na YouTube
- aplikacje do tworzenia treści zmanipulowanych są łatwo dostępne i proste w użyciu
- fałszywe reklamy deepfake wprowadzają w błąd, a ich liczba szybko rośnie
- nie wszyscy są w stanie rozpoznać manipulację



Narzędzia do weryfikacji

Jak działają?

- pozwalają na automatyczne **wykrywanie niespójności** zauważalnych percepcyjnie
  - artefakty, błędy generacji
  - niezgodność w teksturze, oświetleniu, cieniach
  - analiza ekspresji, ruchu twarzy, mrugania, ruchu oczu
  - niespójności w głosie
- analiza niezgodności i różnic w sekwencjach czasowych
- wykrywanie cech ukrytych
- analiza spektrogramów
- cechy biometryczne

Dostępne narzędzia

- AI – image-detector
- Duck Duck Goose
- Sensity
- Illuminarty

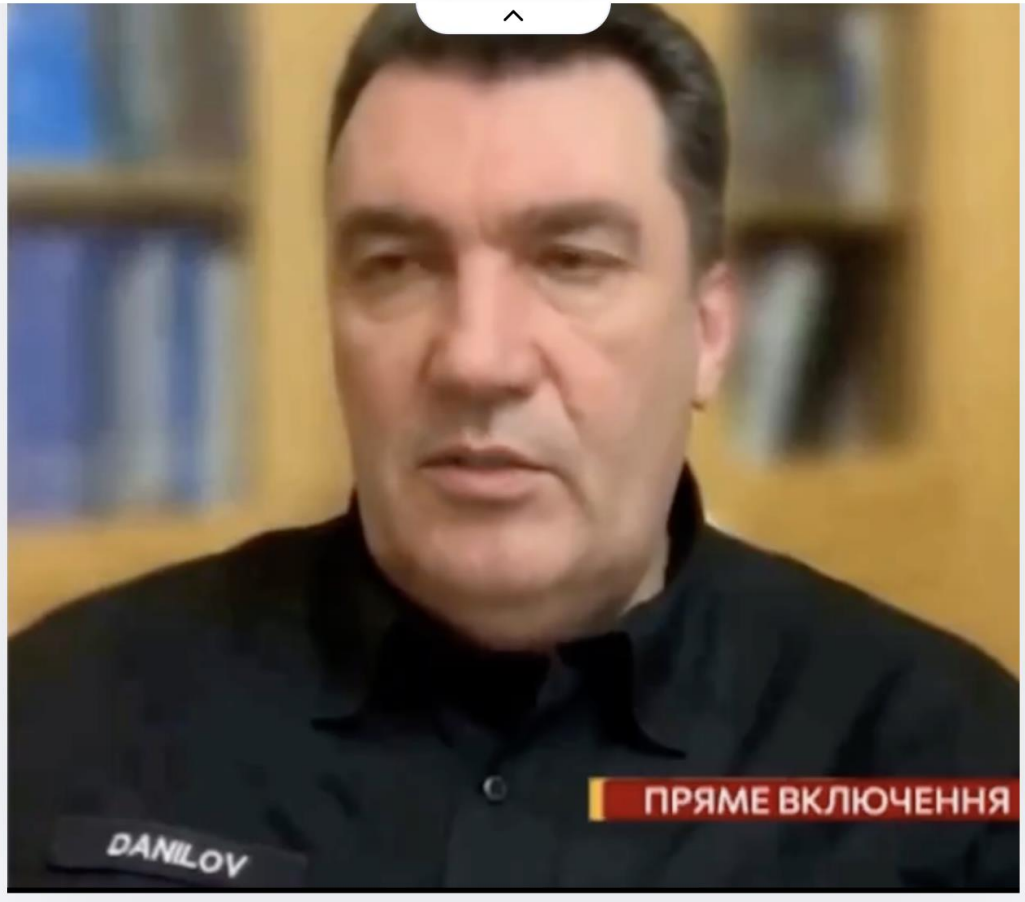
- HIVE
- AIorNOT

- deepfake-total

DETEKTOR 1

AI-Generated Image Detection

Detect AI generated images from popular tools like DALL-E, Midjourney, Stable Diffusion, and others.



Results

Simple | </> JSON

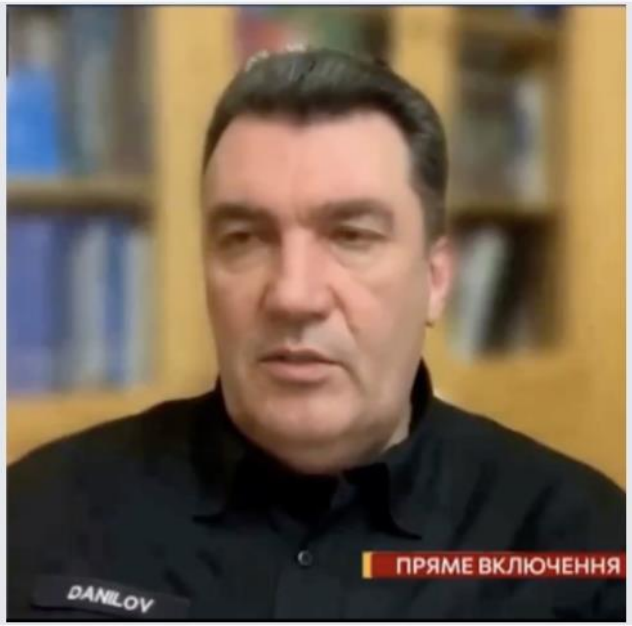
Class	Confidence Score
not_ai_generated	1.00
none	1.00
ai_generated	0.00
inconclusive	0.00
midjourney	0.00
stablediffusion	0.00
dalle	0.00
hive	0.00
gan	0.00
stablediffusionxl	0.00
kandinsky	0.00
adobefirefly	0.00
bingimagecreator	0.00

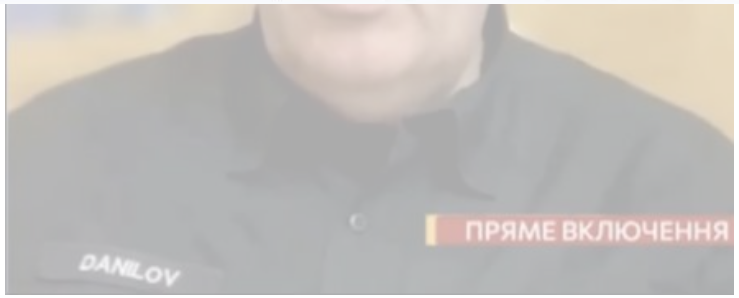
DETEKTOR 1

DETEKTOR 2

REAL

image





output

artificial

artificial	91%
human	9%

DETEKTOR 1

DETEKTOR 2

DETEKTOR 3

REAL

FAKE

DETEKTOR 1

DETEKTOR 2

REAL

FAKE

DeepfakeTotal

Analysing seconds 0 to 30

Fake-O-Meter: 51.6%

DETEKTOR 1

DETEKTOR 2

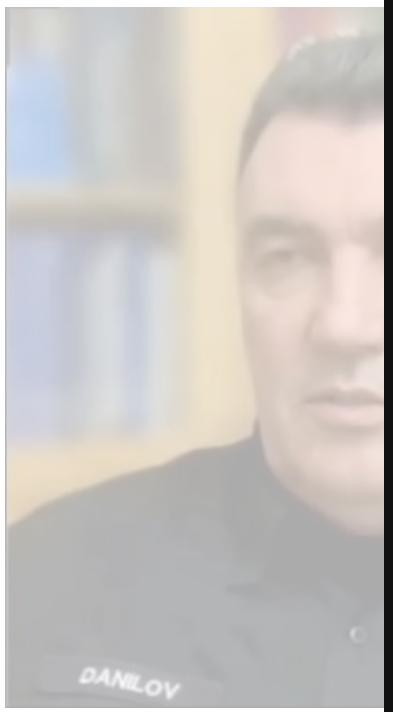
DETEKTOR 3

DETEKTOR 4

REAL

FAKE

DON'T KNOW



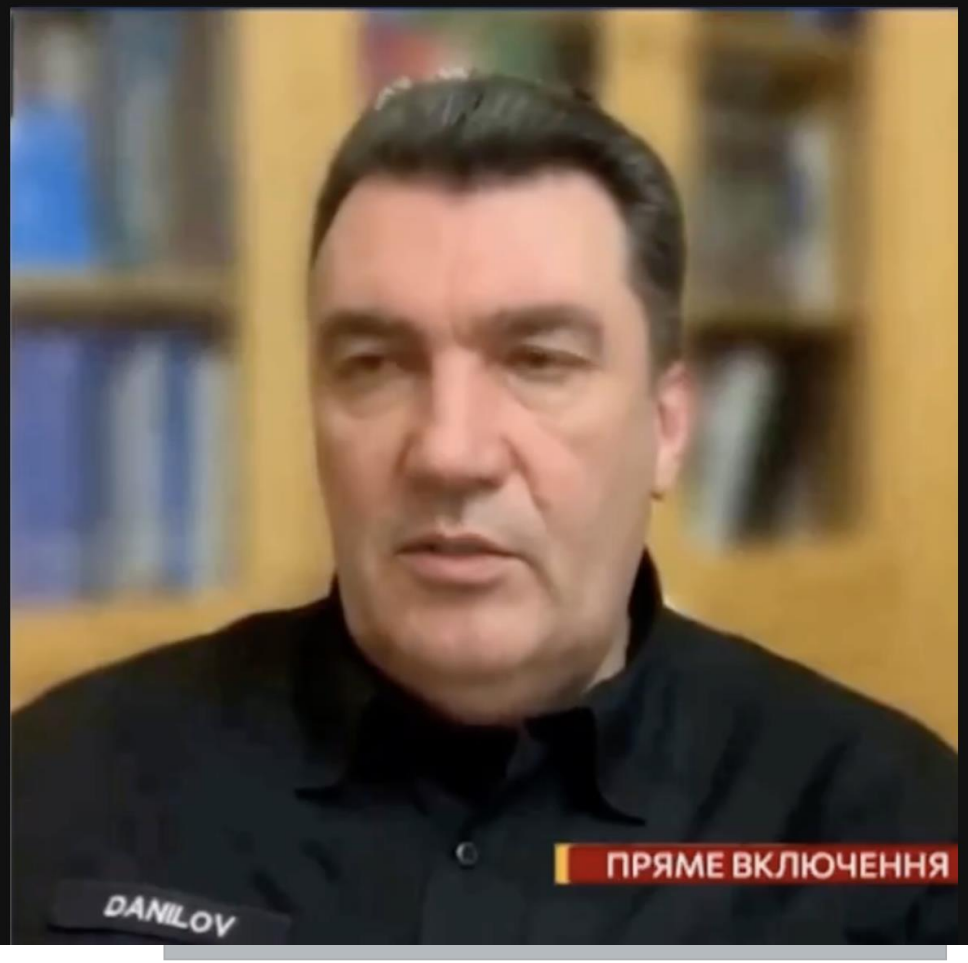
ILLUMINARTY

Analysis API Docs Plans

Sign in

Image

Text



Show Localized

AI Probability: 0.2%

Illuminarty
Image Analysis
Free Plan
Version 1.4

T KNOW

Jeden rabin powie REAL, drugi rabin powie FAKE

DETEKTOR 1 → REAL

DETEKTOR 2 → FAKE

DETEKTOR 3 → DON'T KNOW

DETEKTOR 4 → REAL

DETECTING DEEPFAKES AND FALSE ADS THROUGH ANALYSIS OF TEXT AND SOCIAL ENGINEERING TECHNIQUES

Detecting deepfakes and false ads through analysis of text and social engineering techniques

Alicja Martinek
NASK National Research Institute
ul. Kolska 12, 01-045 Warszawa
AGH University of Kraków
al. Mickiewicza 30, 30-059 Kraków
alicja.martinek@nask.pl

Ewelina Bartuzi-Trokielewicz
NASK National Research Institute
ul. Kolska 12, 01-045 Warszawa
Warsaw University of Technology
Plac Politechniki 1, 00-661 Warszawa
ewelina.bartuzi@nask.pl

Abstract

Existing deepfake detection algorithm frequently fail to successfully identify fabricated materials. These algorithms primarily focus on technical analysis of video and audio, often neglecting the meaning of content itself. In this paper, we introduce a novel approach that emphasizes the analysis of text-based transcripts, particularly those from AI-generated deepfake advertisements, placing the text content at the center of attention. Our method combines linguistic features, evaluation of grammatical mistakes, and the identification of social engineer-

sulted in an estimated 500,000 deepfakes shared on social media platforms in 2023 alone.

Deepfake technology is used in a variety of ways, particularly in financial scams. These schemes often target vulnerable groups, including young internet users and seniors, who may be lured by unrealistically attractive offers. Especially for the latter group, the desire for financial independence from relatives can lead to risky decisions, making them prime targets for scams.

The scripts used in deepfake scams tend to follow a specific and well-structured pattern. These scams typically start with convincing deepfake

TECHNIKI MANIPULACYJNE

- superlatywy
- wywoływanie emocji
- dopasowanie do potrzeb odbiorcy
- obietnica
- wezwanie do działania
- presja czasowa
- bezpośredni zwrot do odbiorcy
- oczernianie
- ...
- konspiracja
- transparentność

Wystarczy jedna filiżanka dziennie, aby przyśpieszyć metabolizm, schudnąć 5 razy szybciej i widocznie się zmienić.

Złóż oficjalny wniosek o niezależność finansową i zacznij zarabiać już dziś.

Ocal siebie i swoich bliskich nim będzie za późno.

Od tego momentu Twoje nowe, szczęśliwe życie stanie się rzeczywistością.

Jeżeli tak jak ja jesteś mamą i chciałabyś dorobić do domowego budżetu siedząc w domu, to zapraszam do współpracy.

Jestem zwykłym chłopakiem, takim jak Ty.

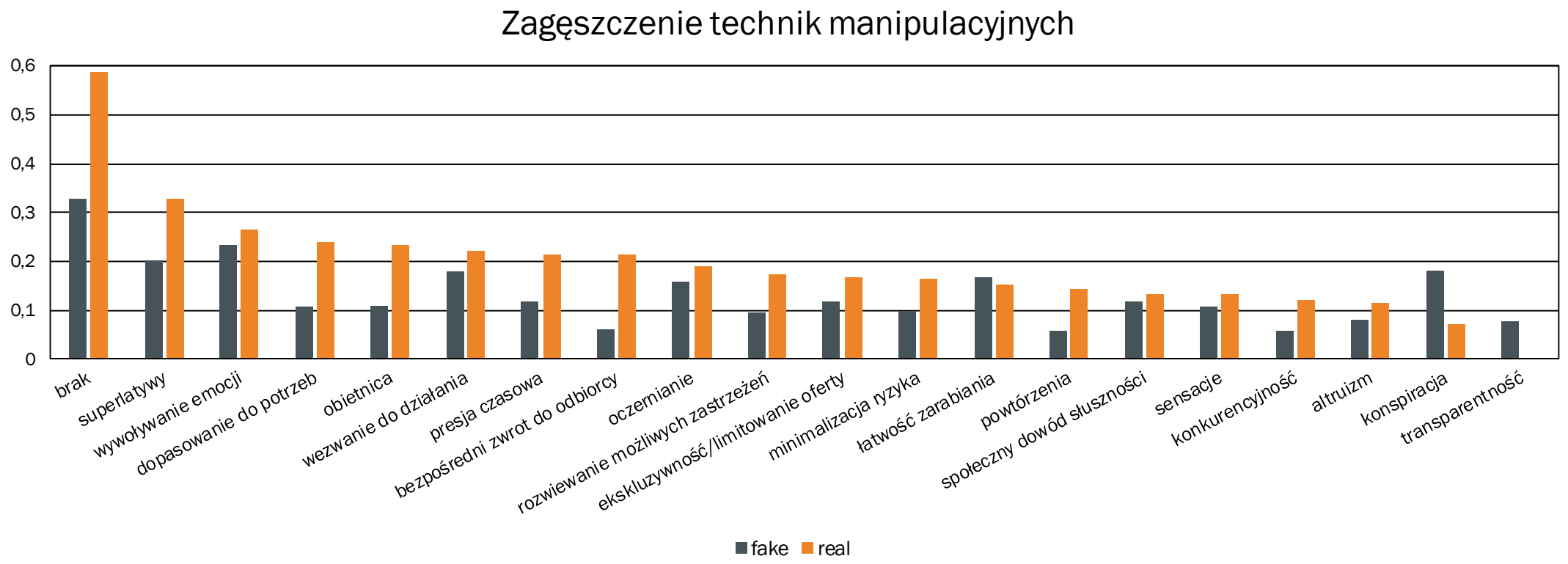
Jeśli będziesz zwlekał, prawdopodobnie skończysz jak 99% innych ludzi.

Większość lekarzy nie dba o to.

Nie jestem pewien jak długo ten film będzie dostępny, ponieważ presja ze strony mafii farmaceutycznej rośnie i prawdopodobnie wkrótce zostanie usunięty.

To nie jest kolejna kampania reklamowa o kimś kto próbuje wyłudzić od Ciebie pieniądze.

PERSWAZJA I TECHNIKI MANIPULACYJNE

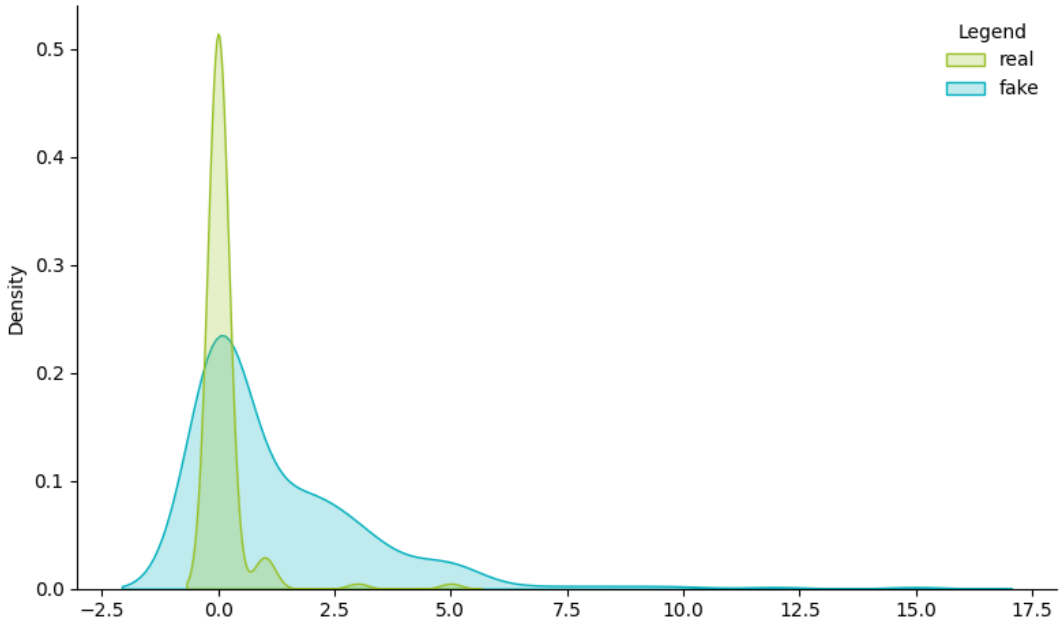


ZBIÓR DANYCH

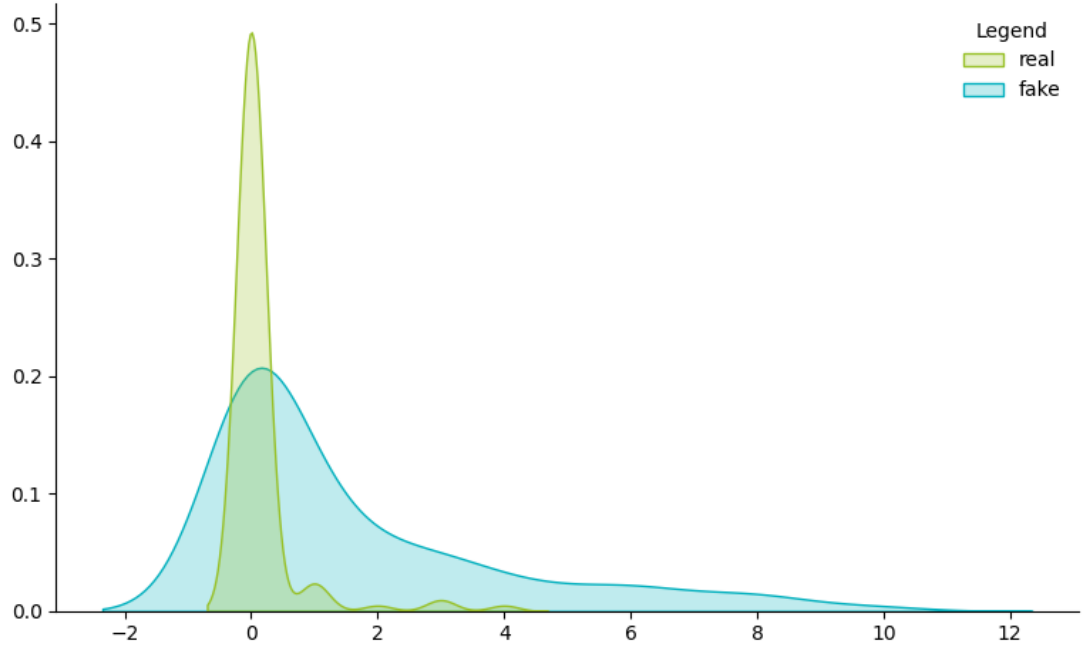
temat	suma	fake	real
inwestowanie	152	135	17
medycyna	92	72	20
inne	36	7	29
oferta pomocy	30	15	15
hazard	10	10	0
	320	239	81

- pochodzą z mediów społecznościowych
- wykorzystują wizerunki osób publicznych
- techniki maskujące mają wpływ na sygnał audio i wideo
- zawierają błędy gramatyczne i logiczne

ZBIÓR DANYCH

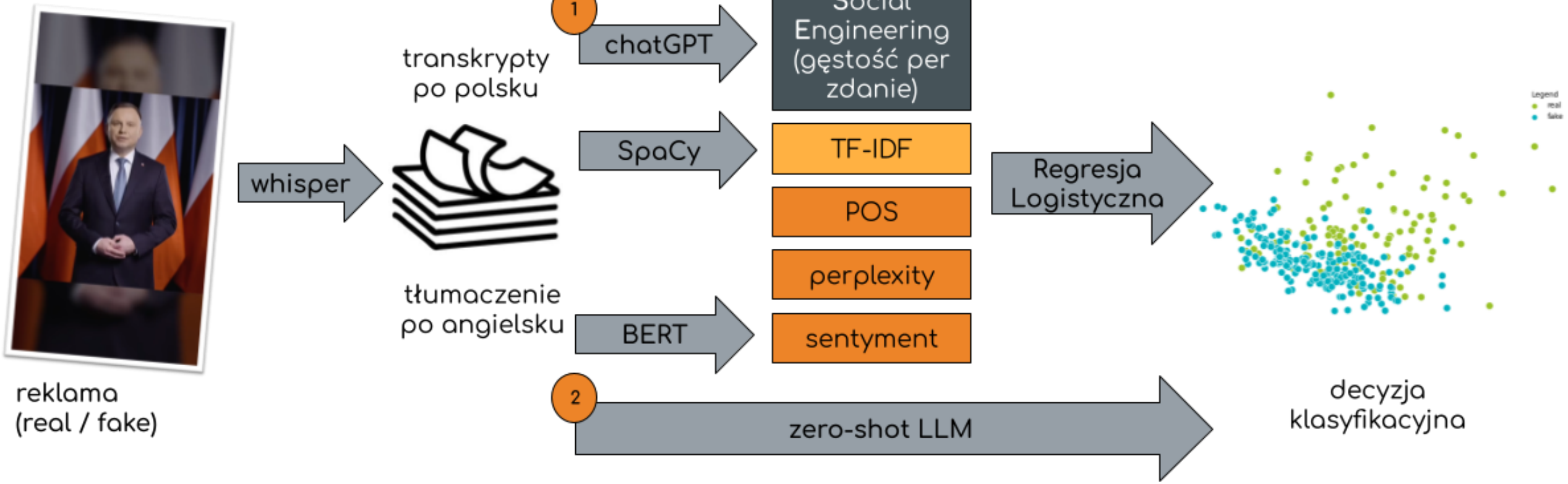


Dystrybucja błędów gramatycznych.



Dystrybucja odwołań do ***„polskości”***.

METODOLOGIA



WYNIKI

model	Polski		Angielski	
	acc	std	acc	std
SE	80.91	5.54	79.36	2.76
SE + TF-IDF	89.91	3.21	89.36	2.76
SE + TF-IDF + Linguistic	90.55	2.88	90.82	1.93
GPT-4o-mini	94.86	6.89	95.40	5.68
Ministral-8B-Instruct	80.54	9.57	84.59	19.78
BERT + Regresja Logistyczna	80.11	7.50	87.08	4.80

WNIOSKI

- multimodalne podejście do detekcji deepfakes powinno stać się złotym standardem
- analiza transkryptów oferuje dodatkową warstwę weryfikacji materiałów
- LLMy mogą być użyte by wzbogacić klasyfikacyjny framework
- gdy wyjaśnialność jest kluczowa...

  techniki manipulacyjne, TF-IDF i lingwistyczna charakterystyka tekstu przychodzi z pomocą
- 90.55% z użyciem wyjaśnialnych metod oraz 94.86% z użyciem LLMów
- międzyjęzykowa generalizacja metody jest możliwa, ale część treści może zgubić się w tłumaczeniu

OGRANICZENIA BADAŃ

- więcej danych!
- więcej języków!
- więcej zróżnicowania tematycznego!

Zapraszamy do kontaktu!

deepfake@nask.pl

SCIENCE
NASK