Doświadczenie z wykrywania fake news metodami sztucznej inteligencji – projekt H2020 SocialTruth

Dr hab. inż. Rafał Kozik, prof. PBŚ,
Dr hab. inż. Michał Choraś, prof. PBŚ

Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich

Konsorcjum projektu SocialTruth

Start projektu

1 grudnia 2018

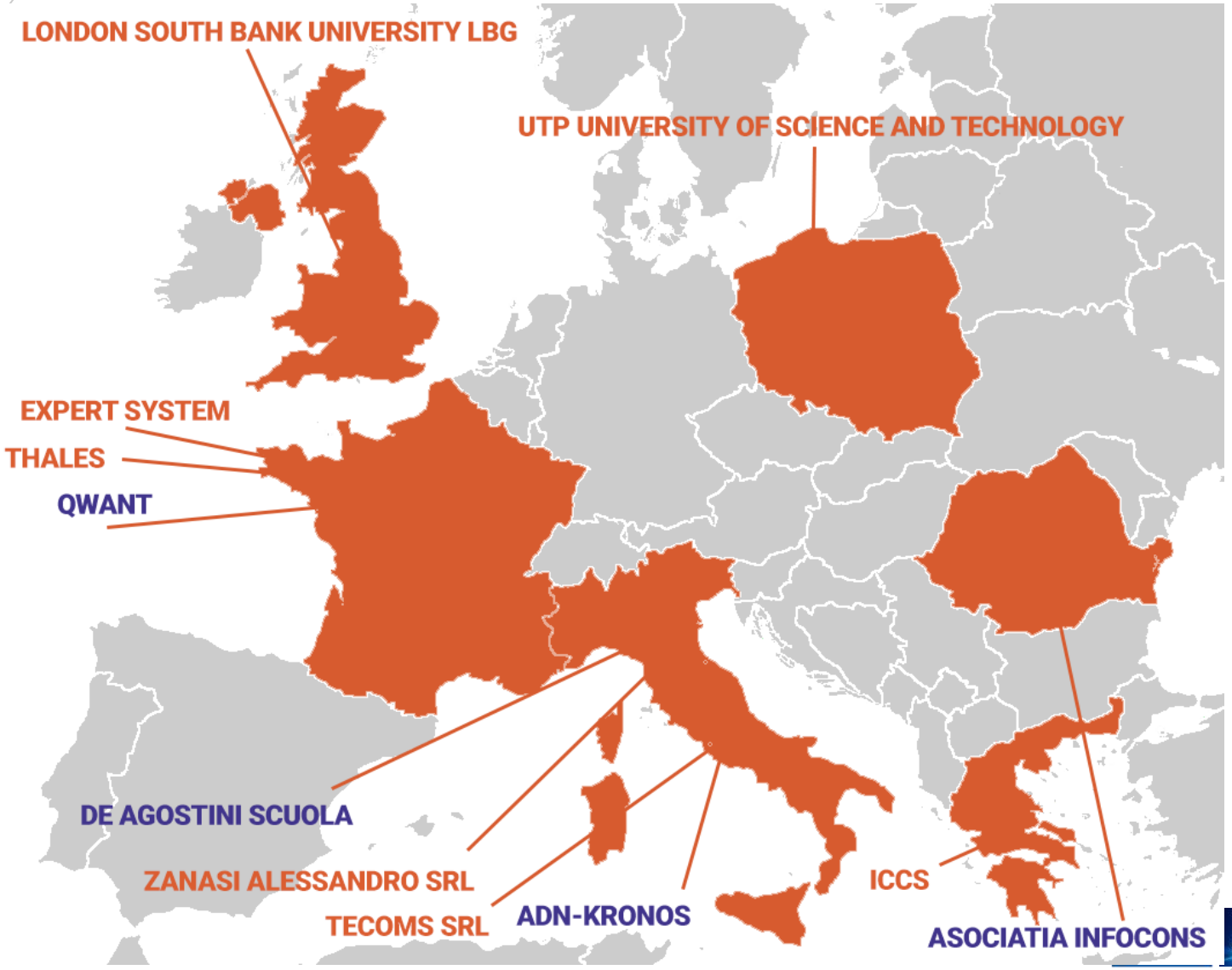
Koniec projektu

30 listopada 2021

Łączny budżet

€ 3 178 110

11 **Partnerów**



Wyzwania

- Błyskawiczny rozwój oraz mediów społecznościowych
- Coraz większa rola mediów cyfrowych w porównaniu do mediów tradycyjnych
- Wpływ fałszywych informacji (fake newsów, dezinformacji) na bezpieczeństwo państw oraz społeczeństw
- Wpływ pandemii COVID-19 na sposób spędzania czasu przez ludzi
- Różni użytkownicy, różne potrzeby i wymagania

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Wyzwania

- Brak ustandaryzowanej definicji fake newsów i dezinformacji online

- Różne aspekty wiarygodności: opinie, interpretacje, dyskusja, fakty nawzajem wykluczające się
- Różne formy ekspresji: satyra, fikcja, teorie spiskowe, propaganda

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Wyzwania – fake news jako przestępstwo

- Brak prawnego umocowania
- Wyzwania dotyczące dochodzenia i dowodów
- Wyjaśnialność

- Problemy z atrybucją
- Nieświadome szerzenie fałszywych informacji (np. poprzez re-tweety)
- Fake newsy generowane przez mechanizmy sztucznej inteligencji

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Strategie

- Analiza treści – różne rodzaje danych
- Indeksowanie informacji w celu odnoszenia się do nich
- Punktowanie reputacji
- Porównywanie podobnych informacji z wielu źródeł
- **Wykorzystanie uczenia maszynowego w celu analizy cech treści**
- Analiza semantyczna z wykorzystaniem ontologii

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Główne założenia detekcji fake newsów

- **Treści cyfrowe** definiowane jako tekst, zdjęcia lub wideo.
- Treść jest **dekomponowana** na poszczególne elementy składowe, a następnie wywoływane są odpowiednie **usługi weryfikacyjne**, których wyniki analizy są ostatecznym rezultatem działania systemu.
- Projekt SocialTruth **integruje usługi weryfikacji treści** za pomocą inteligentnego **algorytmu metaweryfikacji** i zapewnia łatwość użytkowania dzięki elastycznemu narzędziu oferowanemu użytkownikowi końcowemu (wtyczka **Digital Companion**).

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Wizja

- Weryfikacja nie może zostać powierzona jednej instytucji
- Celem nie jest „jeden doskonały algorytm”
- Weryfikacja treści powinna być wygodna
- Musi być też otwarta, demokratyczna, pluralistyczna i rozdystrybuowana

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

SECTIONS SEARCH A NEW YORK YANKEES BASEBALL PODCAST NEW YORK POST LISTEN NOW TIPS SIGN UP

OPINION

Favoring renewable energy over natural gas investment has led to the mess in Texas

By Kevin D. Williamson February 20, 2021 | 12:14pm | Updated

A line forms outside HEB grocery store in Austin, Texas, on Feb. 17 as millions in the state go without power, partly thanks to outdated natural gas pipeline infrastructure.
Bloomberg via Getty Images

I'm writing from Texas, so I'll try to finish this column before the electricity goes out.

MORE ON: TEXAS

COLUMNISTS

Nicole Gelinas

NY should use Biden Bucks to give restaurants and retailers a tax holiday

Maureen Callahan

Meghan Markle's interview was full of bull

Karol Markowicz

Texas shows the way on COVID rationality

SEE ALL COLUMNISTS

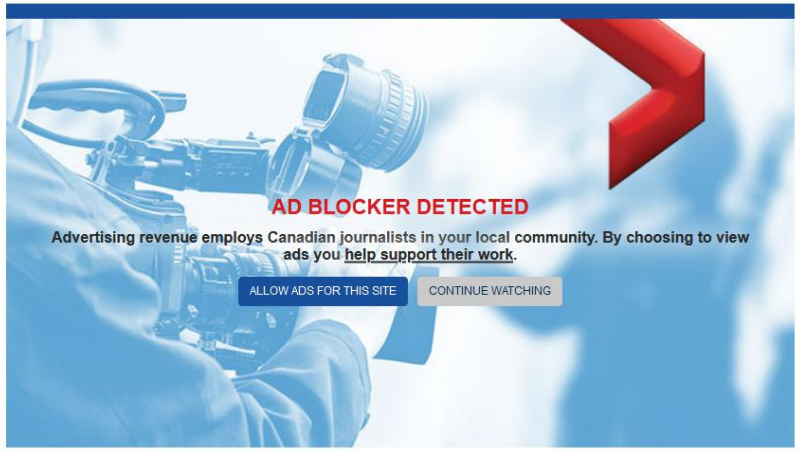
TODAY'S COVER

Global NEWS

WORLD

No, renewable energy is not primarily to blame for Texas power failures

By Ali Swenson and Arijeta Lajka · The Associated Press
Posted February 18, 2021 1:16 am · Updated February 18, 2021 1:17 am



Winter storm keeps millions of Texans in the frigid cold and the dark – Feb 17, 2021

-A A+

With millions of Texas residents still without power amid frigid temperatures, conservative commentators have falsely claimed that wind turbines and solar energy were primarily to blame.

"We should never build another wind turbine in Texas," read a Tuesday Facebook post from Texas Agriculture Commissioner Sid Miller. "The experiment failed big time."

"This is a perfect example of the need for reliable energy sources like natural gas & coal," tweeted U.S. Sen. Steve Daines, a Republican from Montana, on Tuesday.

READ MORE: Texas mayor quits after saying 'only the strong will survive' snowstorm

In reality, failures in natural gas, coal and nuclear energy systems were responsible for nearly twice as many outages as frozen wind turbines and solar panels, the Electric Reliability Council of Texas, which operates the state's power grid, said in a press conference Tuesday.

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

SECTIONS SEARCH A NEW YORK YANKEES BASEBALL PODCAST NEW YORK POST LISTEN NOW TIPS SIGN UP

SocialTruth.eu DIGITAL COMPANION LIGHT

Request a new Evaluation

URL to check

https://nypost.com/2021/02/20/texas-reliance-on-renewable-energy-has-led-to-this-winter-mess/

SEND

Score: ★★★★

A line forms outside HEB grocery store in Austin, Texas, on Feb. 17 as millions in the state go without power, partly thanks to outdated natural gas pipeline infrastructure.

Bloomberg via Getty Images

MORE ON: TEXAS

I'm writing from Texas, so I'll try to finish this column before the electricity goes out.

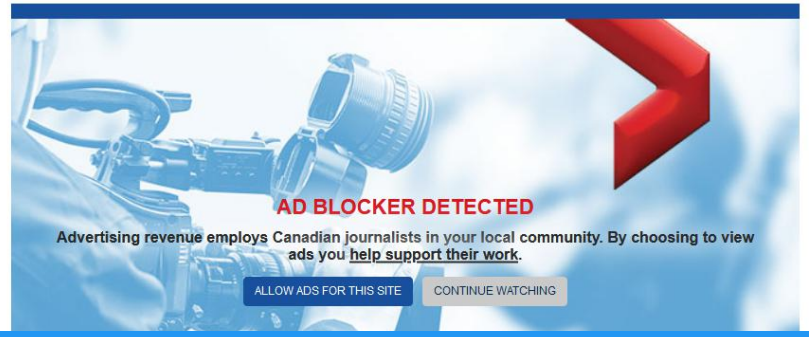
Global NEWS

WORLD

No, renewable energy is not primarily to blame for Texas power failures

By **Ali Swenson and Arijeta Lajka · The Associated Press**

Posted February 18, 2021 1:16 am · Updated February 18, 2021 1:17 am



Request a new Evaluation

URL to check

https://globalnews.ca/news/7647520/texas-power-renewable-energy/

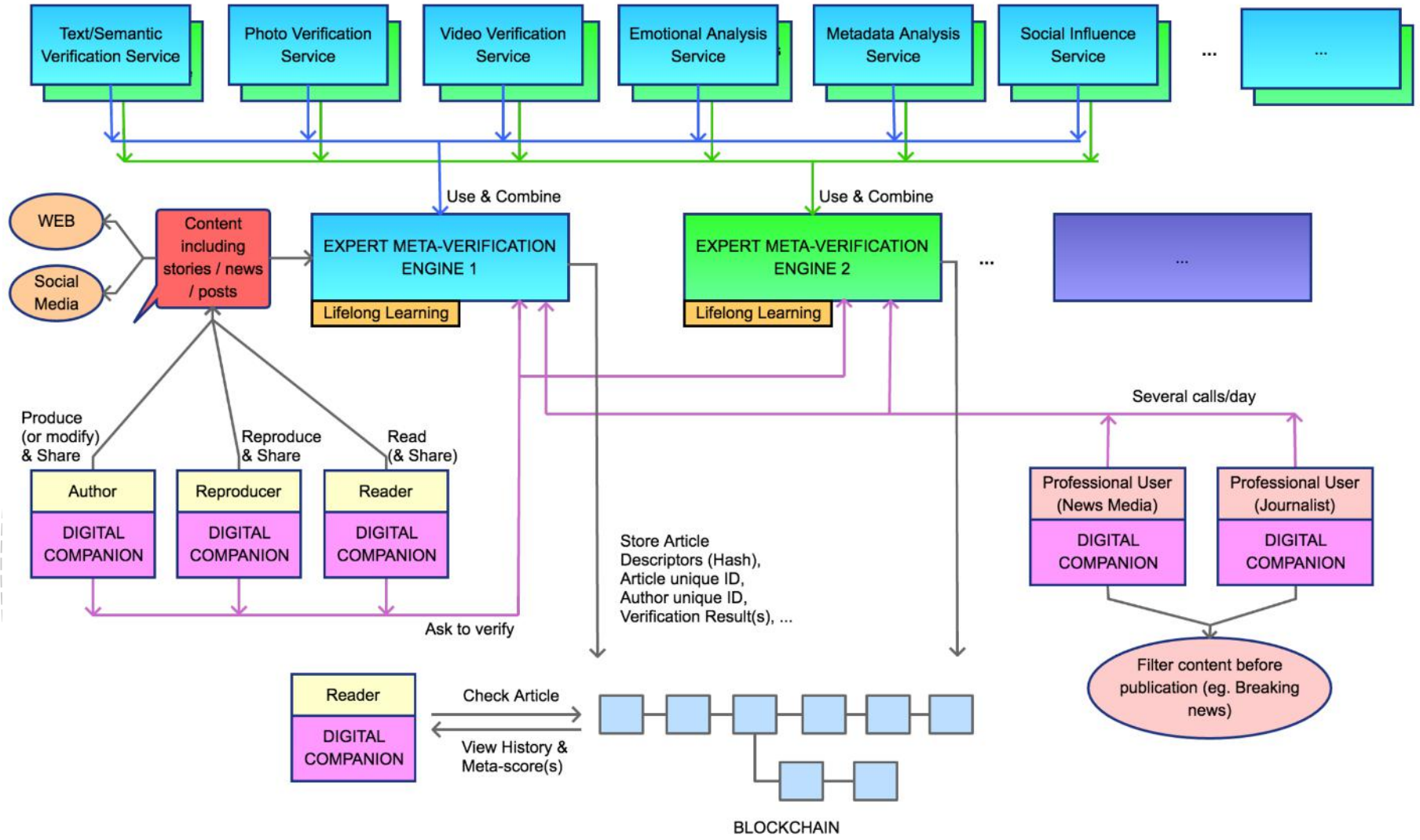
SEND

Score: ★★★★

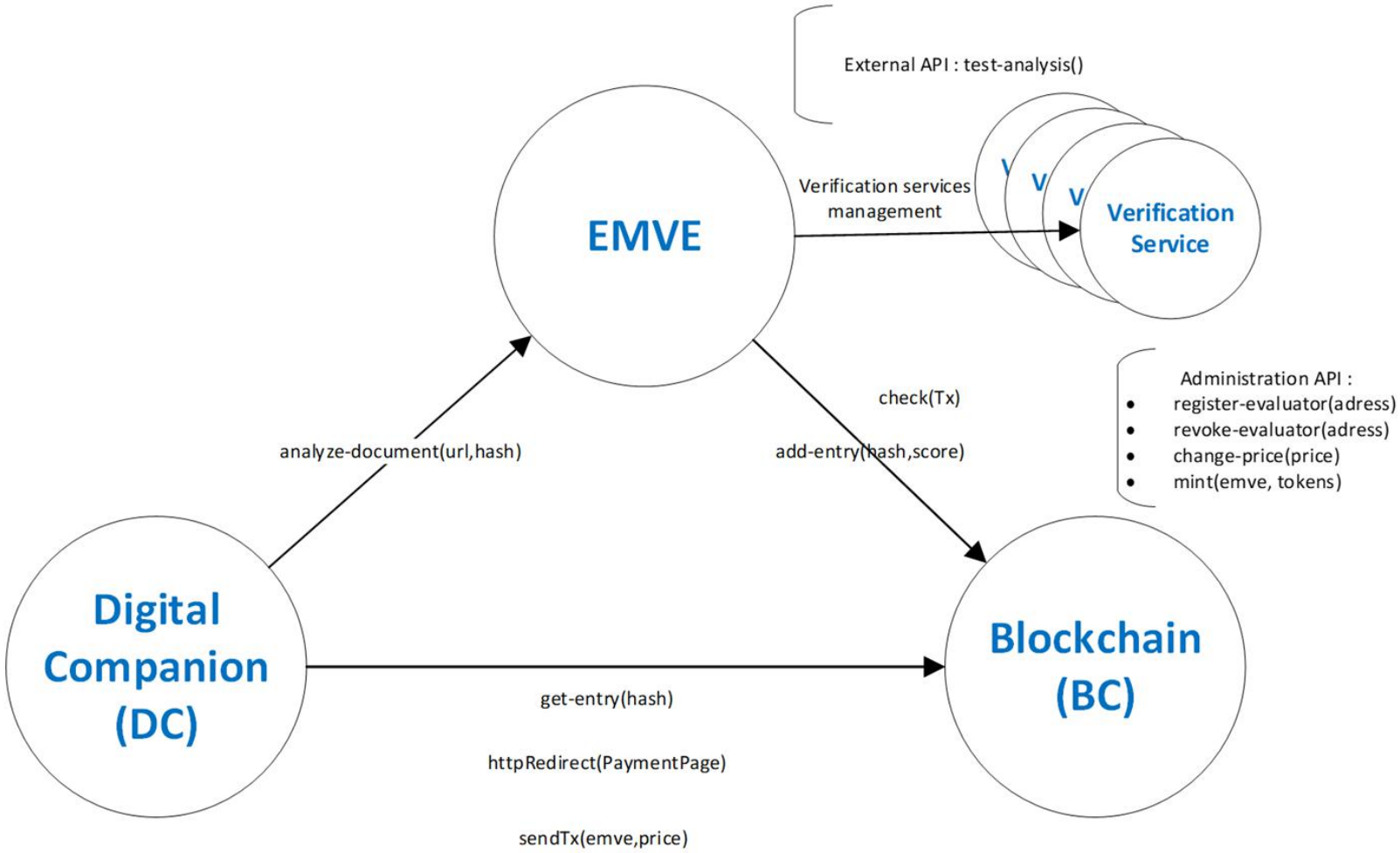
Część techniczna

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Architektura logiczna

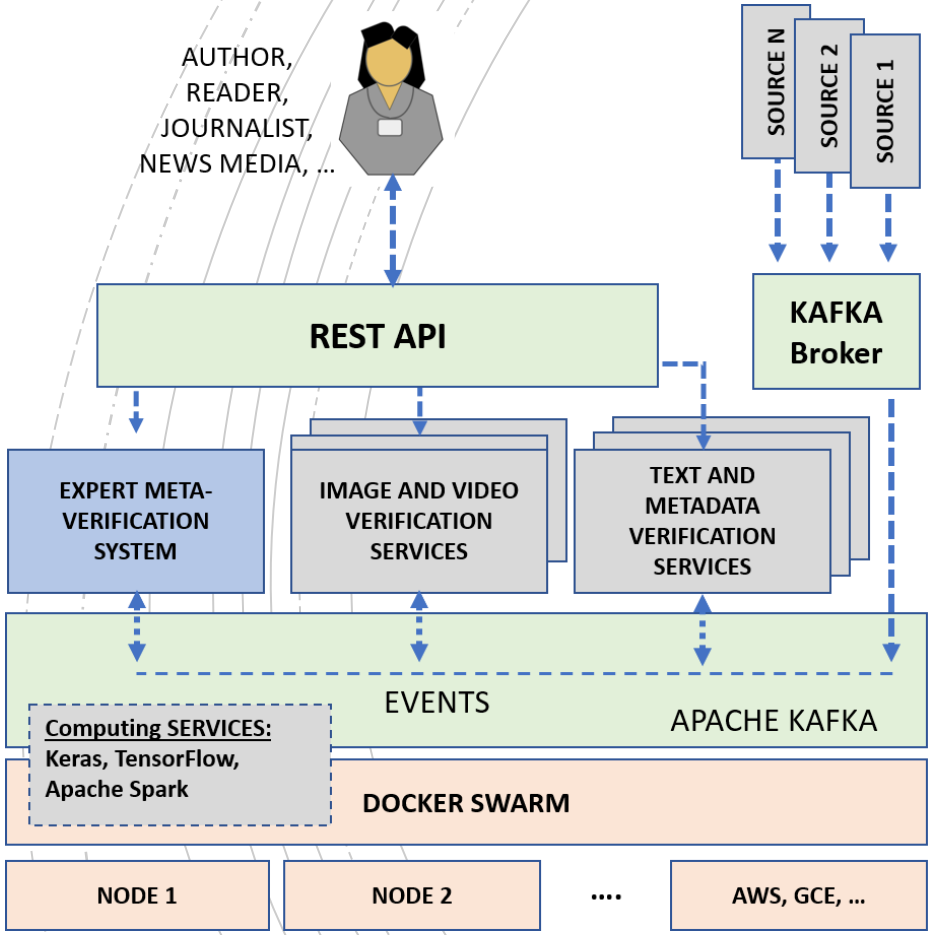


Główne komponenty



SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

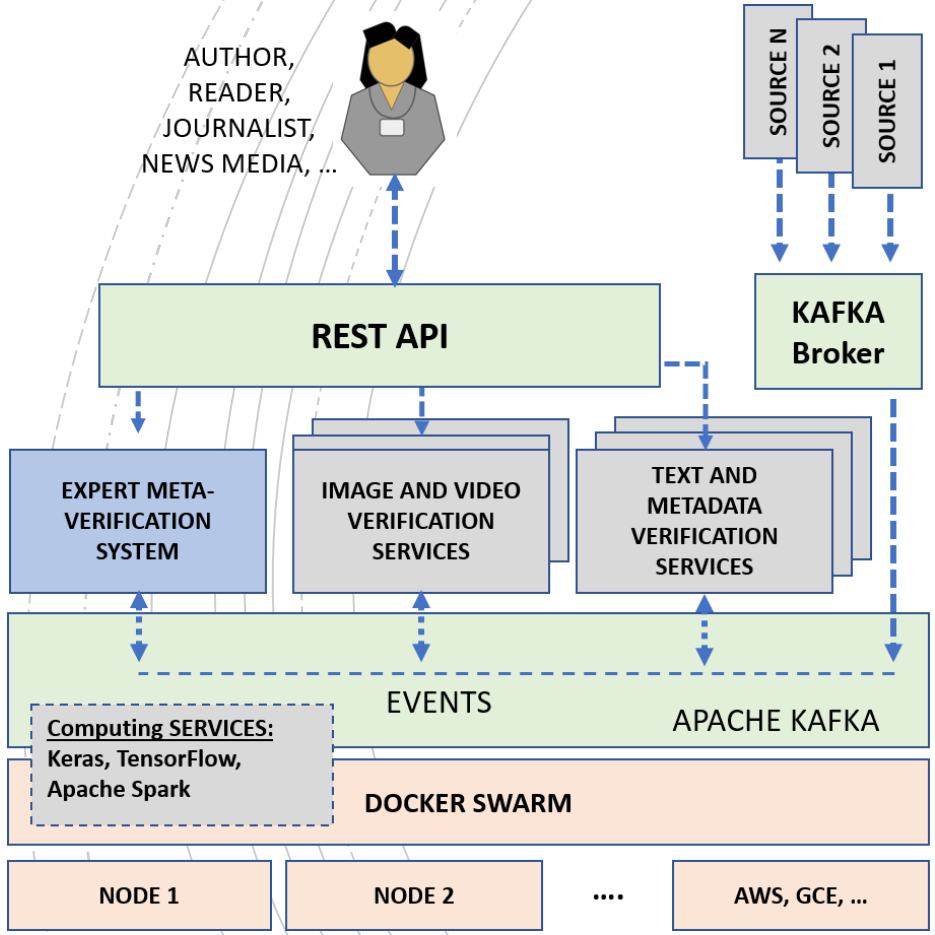
Stos technologiczny (1/2)



- Elementy logiczne: fizyczne węzły i ich orkiestracja, komunikacja i przetwarzanie zdarzeń, usługi weryfikacyjne.
- Architektura wykorzystująca mikro-usługi zaadaptowana razem z narzędziem orkiestracji kontenerów Docker Swarm.
- Usługi weryfikacyjne są najważniejszymi elementami platformy.

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

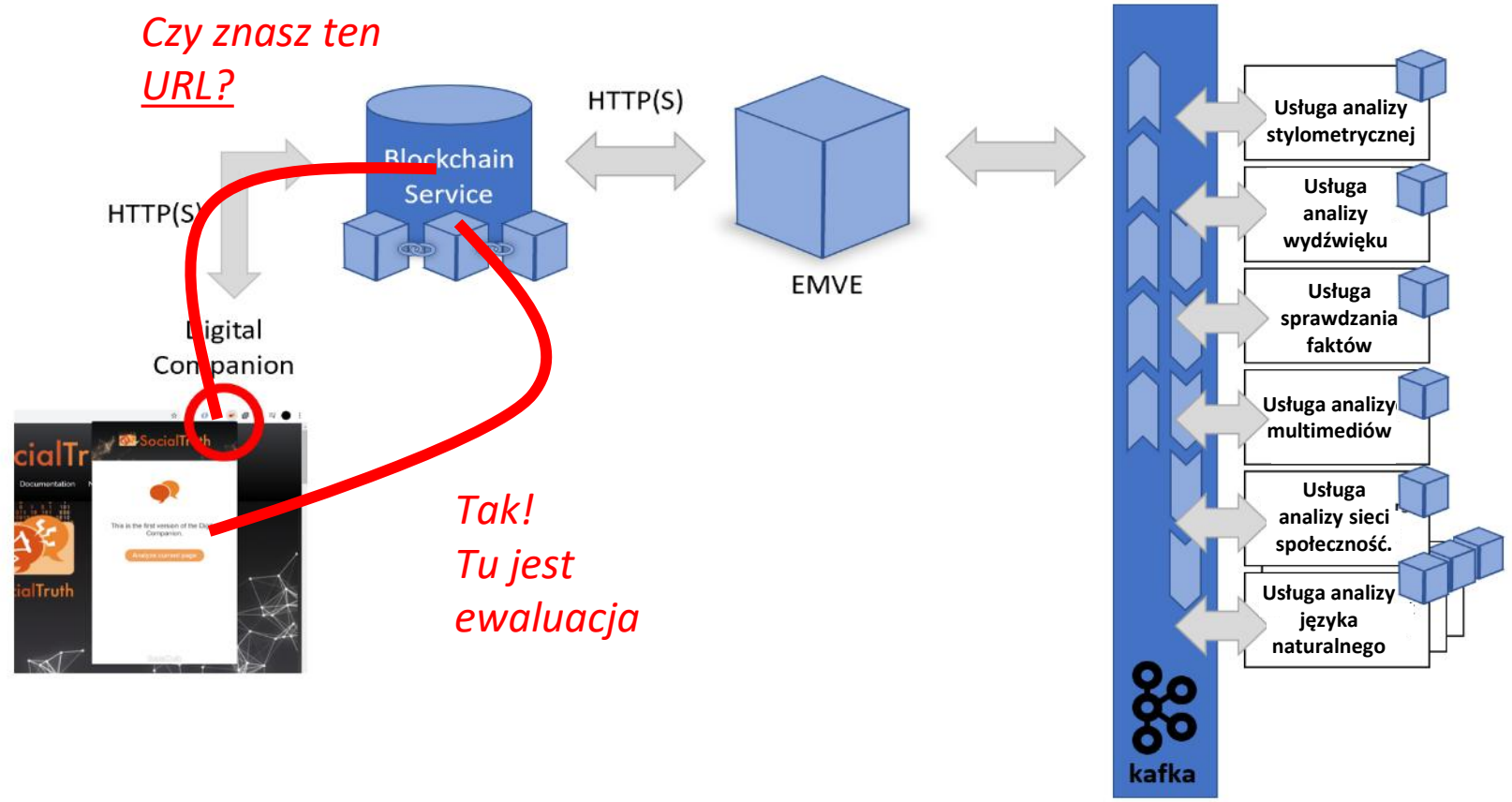
Stos technologiczny (2/2)



- Dane wejściowe są pozyskiwane do platformy za pośrednictwem narzędzia Apache Kafka (przez użytkownika lub dedykowane narzędzia przeszukujące i indeksujące treści w sieci).
- Dane są przekazywane do odpowiednich tzw. „tematów” (poszczególne grupy wiadomości), a następnie konsumowane przez mikro-usługi.
- Ogólny przepływ danych:
  - Żądanie użytkownika przechodzi przez bramę Rest API do komponentu meta-weryfikacji, który przekształca je w zdarzenie Apache Kafka
  - Wydarzenie jest przetwarzane przez usługi weryfikacyjne, a wyniki są przesyłane z powrotem do Kafki
  - Komponent meta-weryfikacji zbiera wyniki z różnych tematów i prezentuje je użytkownikowi

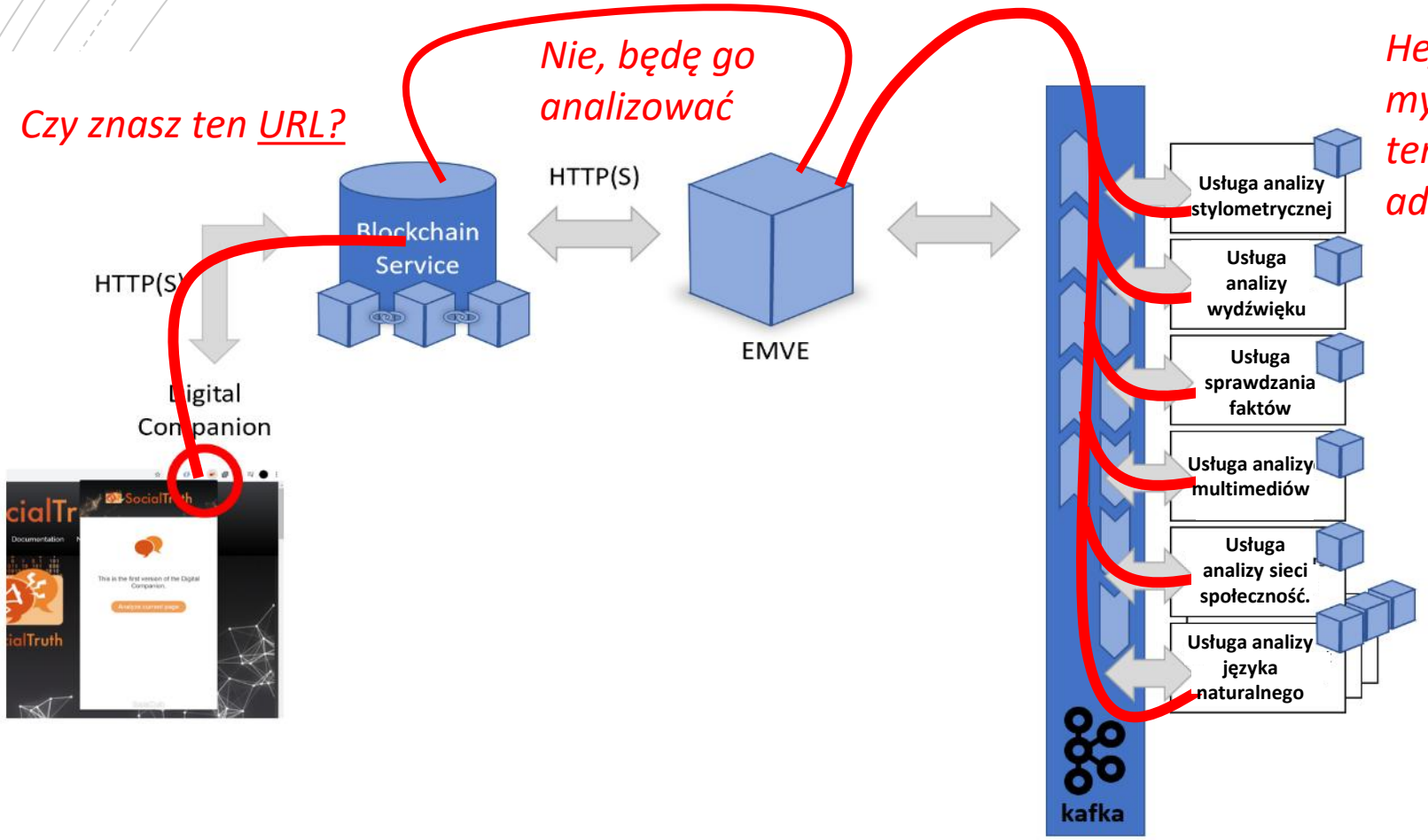
SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Przepływ informacji (1/2)



SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Przepływ informacji (2/2)



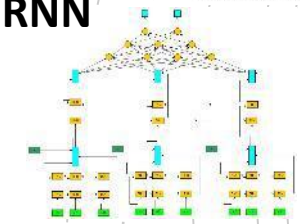
SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Usługi weryfikacyjne

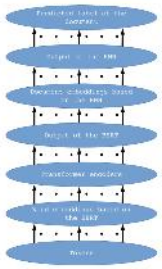
Przetwarzanie języka naturalnego

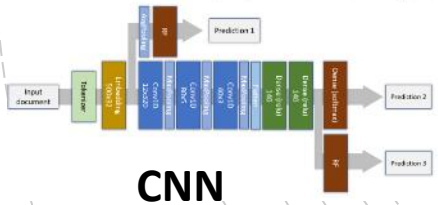
POLITECHNIKA BYDGOSKA im. Jana i Jędrzeja Śniadeckich

RNN



BERT

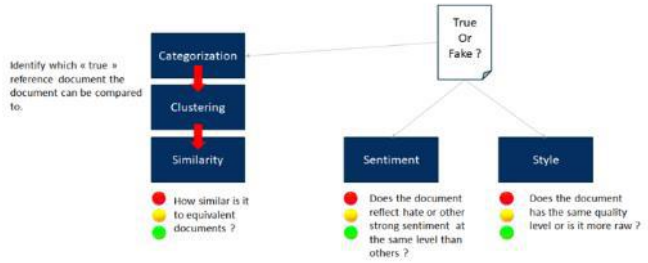


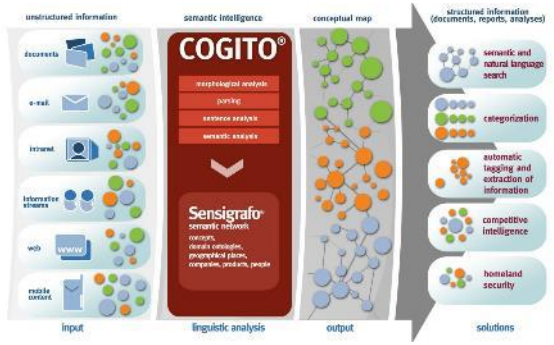


CNN

Analiza semantyczna

EXPERT SYSTEM

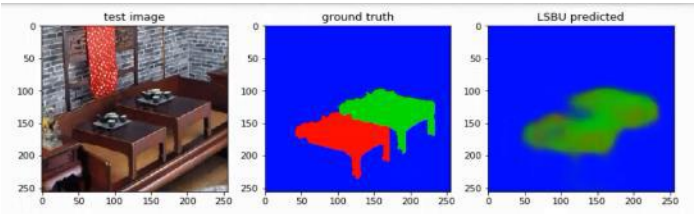




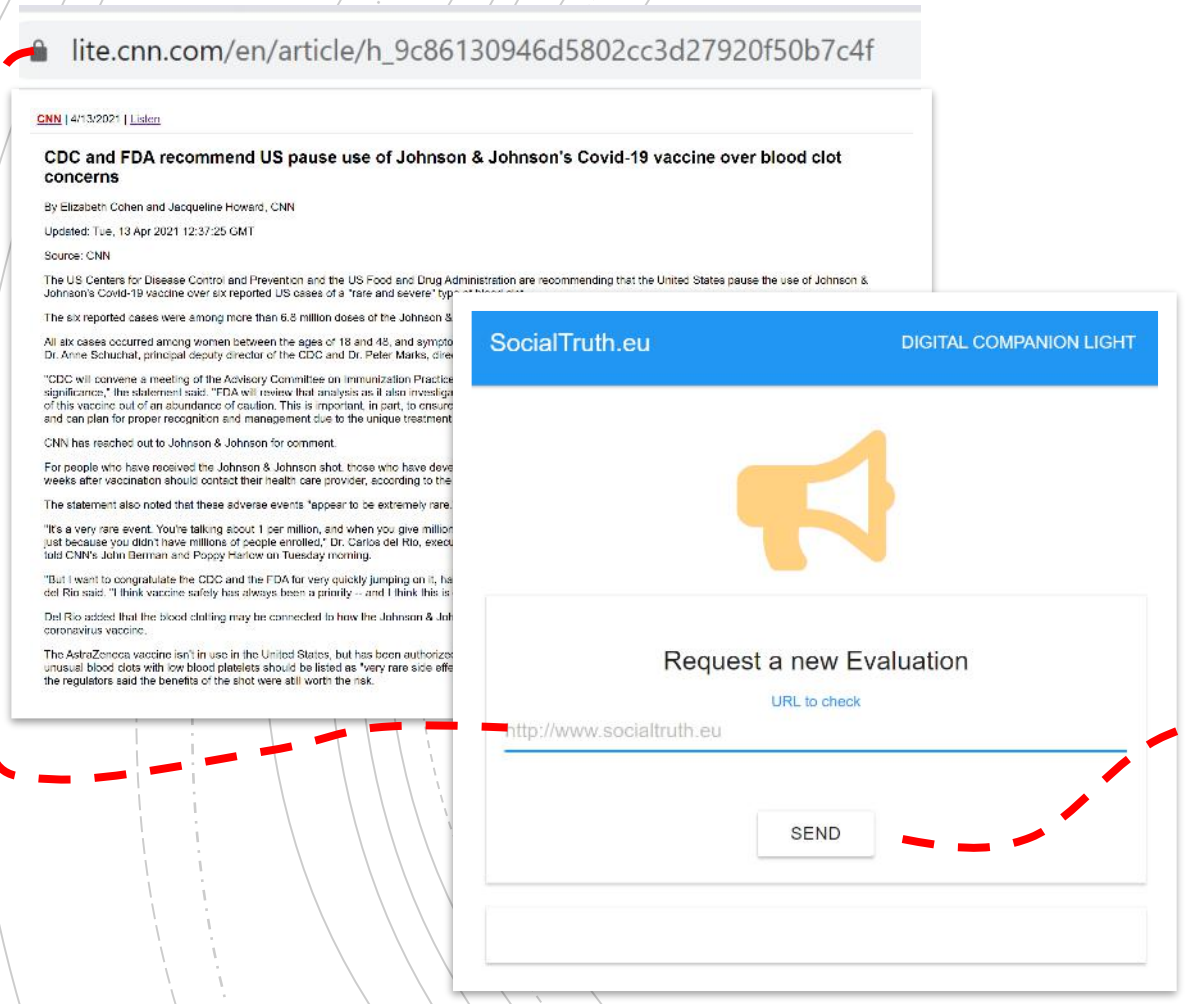
Analiza obrazów

London South Bank University

Rozpoznawanie manipulacji obrazu



Interfejs aplikacji internetowej



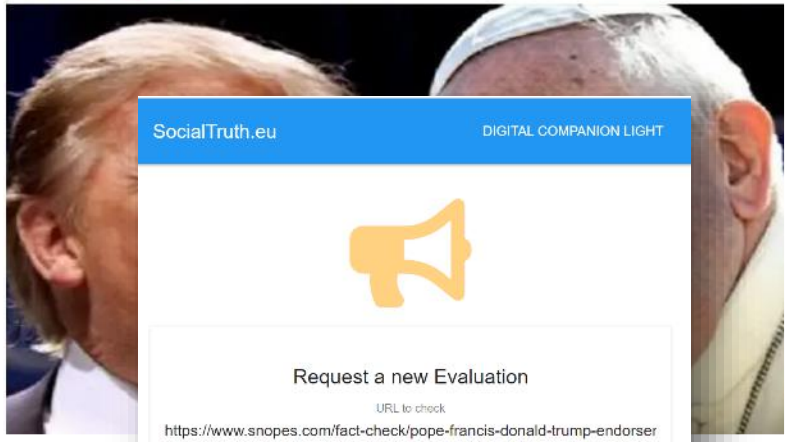
Koncept „4 gwiazdek”

Pope Francis Shocks World, Endorses Donald Trump for President

Reports that His Holiness has endorsed Republican presidential candidate Donald Trump originated with a fake news web site.

By Dan Evon

Published 10 July 2016



SocialTruth.eu DIGITAL COMPANION LIGHT

Request a new Evaluation

URL to check

https://www.snopes.com/fact-check/pope-francis-donald-trump-endorser

SEND

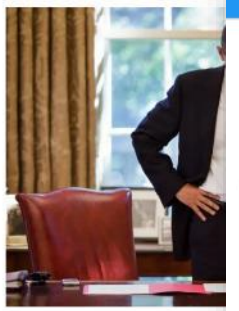
Score: ★☆☆☆

NEWS

Frustrated Obama Sends Nation Rambling 75,000-Word E-Mail

11/26/10 8:00AM

WASHINGTON—Having admittedly "reached the end of [his] rope," President Barack Obama sent a rambling 75,000-word e-mail to the entire nation Wednesday, revealing deep frustrations with America's political culture, his presidency, U



The president's e-mail also CC'd a

SocialTruth.eu DIGITAL COMPANION LIGHT

Request a new Evaluation

URL to check

https://politics.theonion.com/frustrated-obama-sends-nation-rambling-75

SEND

Score: ★☆☆☆

DEMO

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Ocena formalna

CISIS'20

Table 4: Resulted metrics for testing the models based on collection 2 ('text') for the label fake (the comparison between word embeddings techniques, 'bert-base-multilingual-cased', 'bert-base-cased', 'bert-base-uncased', 'bert-base-cased-finetuned-mrpc')

Metric	'multilingual-cased'	'base-cased'	'base-uncased'	'finetuned-mrpc'
true positive (TP)	2246	2271	2260	2213
true negative (TN)	2112	2087	2110	2107
false positive (FP)			33	36
false negative (FN)			17	64
precision			0.9856	0.9840
recall			0.9925	0.9719
f1-score			0.9890	0.9779
accuracy	97.84%	97.34%	97.84%	95.68%

97%

ICCS'20

Table 5. Resulted metrics for testing of models for the label fake (the comparison between word embeddings techniques, 'glove','news', 'twitter', 'crawl')

Metric	'glove'	'news'	'twitter'	'crawl'
true positives (TP)	2275	2080	2269	2217
true negatives (TN)	2139	2121		
false positives (FP)	4	22		
false negatives (FN)	2	197		
precision	0.9982	0.9895		
recall	0.9991	0.9135		
accuracy	99.86%	95.04%	98.03%	97.67%

98%

- Obecnie prototyp systemu przechodzi proces ewaluacji u rzeczywistych użytkowników końcowych (dziennikarze, agencie prasowe).
- Część wyników naukowych i testów technicznych została już opublikowana.
- Wyniki są obiecujące 😊

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

SocialTruth

Projekt SWAROG

SWAROG: System WykrywAnia dezinfoRmacji metOdami sztucznej inteliGencji

- Konkurs NCBIR INFOSTRATEG I - strategiczny program badań naukowych i prac rozwojowych NCBiR
- Konsorcjum: MATIC SA (koordynator), Politechnika Wrocławska, Politechnika Bydgoska
- Start w grudniu, długość projektu: 36 miesięcy
- 3 fazy: prototyp na poziomach TRL4, 6 i 9

matic IT & DEFENCE SYSTEMS

Politechnika Wrocławska

POLITECHNIKA BYDGOSKA im. Jana i Jędrzeja Śniadeckich

HORIZON 2020

Wybrane publikacje

- Kula S., Kozik R., Choraś M., **Implementation of the BERT-derived architectures to tackle disinformation challenges** , Neural Computing and Applications, Springer, July 2021, Open Access, (IF=5,606)
- Choraś M., Demestichas K., Gielczyk A., Herrero A., Ksieniewicz P., Remoundou K., Urda D., Woźniak M., **Advanced Machine Learning Techniques for Fake News (Online Disinformation) Detection: A Systematic Mapping Study**, Applied Soft Computing Elsevier, vol. 101, March 2021 (IF=6,725).
- Kula S., Choraś M., Kozik R. and Woźniak M., **Transformer Based Models in Fake News Detection**, in: Paszynski M. et. al (Eds.): Computational Science – ICCS 2021, LNCS 12745, 28038, Springer, May 2021 (CORE A).
- Ksieniewicz P., Zyblewski P., Choraś M., Kozik R., Gielczyk A., Wozniak M. **Fake News Detection from Data Streams** , in Proc. of International Joint Conference on Neural Networks, IJCNN 2020, IEEE, Glasgow, 2020 (CORE A).
- Kula S., Choraś M., Kozik R., Ksieniewicz P., Wozniak M., **Sentiment Analysis for Fake News Detection by Means of Neural Networks** , in Krzhizhanovskaya V.V. et al. (Eds.): Computational Science – ICCS 2020, LNCS vol. 12140, 652-667, Springer, Amsterdam, 2020 (CORE A).

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477

Dziękujemy za uwagę!

Michał Choraś:

chorasm@pbs.edu.pl

Rafał Kozik:

rkozik@pbs.edu.pl

SocialTruth: projekt współfinansowany przez Komisję Europejską, w ramach programu ramowego Horyzont 2020, grant nr 825477