

Kiedy dokładność klasyfikacji to za mało: Wyjaśnianie oceny wiarygodności tekstu i pomiar reakcji użytkowników

Piotr Przybyła
piotr.przybyla@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

29 listopada 2021



Praca we współpracy:

- IPI PAN – projekt HOMADOS,
- Universidad Nacional del Sur (Argentyna) – Axel Soto.

Szczegóły w:

P. Przybyła, A. J. Soto, "When classification accuracy is not enough: Explaining news credibility assessment," *Information Processing & Management*, vol. 58.5, 2021.

Motywacja



Dużo badań nad automatyczną oceną wiarygodności tekstu metodami NLP/ML:

- *fake news* / *real news*,
- wykrywanie manipulacji propagandowych [Dimitrov et al., 2021],
- wykrywanie botów [Rangel and Rosso, 2019],
- *fact-checking* [Thorne et al., 2018].

... ale jak wykorzystać rezultaty?

- automatyczne filtrowanie wypełnia kryteria cenzury w rozumieniu praw człowieka [Llansó, 2020],
- ostrzeżenia w UI sieci społecznościowych mają ograniczony wpływ na postrzeganie treści [Pennycook et al., 2020, Clayton et al., 2019] i przekazywanie dalej [Mena, 2020],
- kwestionowanie informacji kojarzonych politycznie może je wzmacniać [Nyhan and Reifler, 2010].

Credibilator



Szanse:

- Zmiana poglądu możliwa pod wpływem alternatywnego wytłumaczenia [Nyhan and Reifler, 2015] lub prezentacji graficznej [Nyhan and Reifler, 2019],
- Graficzne tłumaczenie zwiększa zaufanie do automatycznych decyzji [Ribeiro et al., 2016, Yang et al., 2020].

Czy model NLP/ML może zmienić postrzeganie wiarygodności tekstu?

Credibilator: rozszerzenie do przeglądarki implementujące model oceny wiarygodności w celu analizy metod NLP/ML pod kątem kontaktu z użytkownikiem.

Crediblator



Trump Just Bombarde
MAJOR Surprise That

By Prissy Holly

The text on this page is...
1.88% credible (stylometric)
Why?

Style Neural

Images: Donald Trump, a man, a woman with glasses, and another woman.



Klasyfikacja

Schemat ML



Klasyfikatory opisane szczegółowo wcześniej [Przybyła, 2020].

- Zbiór treningowy $\sim 100,000$ tekstów oznaczonych wiarygodnością według źródła:
 - Wiarygodne: zaufane dla większości populacji USA w sondażu *Pew Research Center* [Mitchell et al., 2014],
 - Niewiarygodne: kategoryzowane jako *Fake News* lub *Impostor* przez *Politifact* [Gillin, 2017],
- Klasyfikatory zaprojektowane, by polegać na *stylu* tekstu wyróżniającym fake news [Bakir and McStay, 2017],
- Ewaluacja uwzględniająca scenariusze użycia:
 - *document CV*: dokumenty testowe wybrane losowo,
 - *topic CV*: dokumenty testowe dot. nowego tematu,
 - *source CV*: dokumenty testowe z nowego źródła.

Klasyfikator stylometryczny



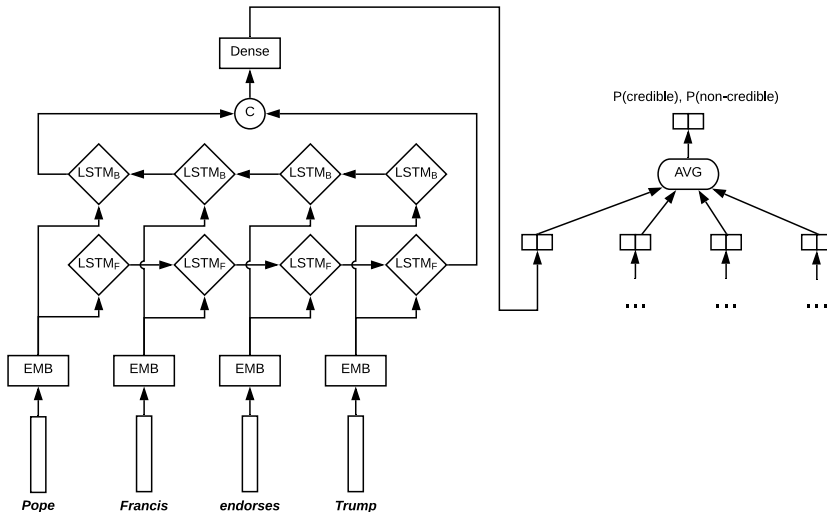
Cechy używane w stylometrii:

- podstawowe statystyki (długość tekstu, zdań i słów),
- liczba słów wg. wielkości liter (*słowa*, *Słowa*, *SŁOWA* i *sŁoWa*),
- częstości unigramów, bigramów i trigramów POS (obecnych w co najmniej 5 dokumentach),
- częstości słów z 182 klas słownika GI (*General Inquirer*), poszerzonego pięciokrotnie w przestrzeni osadzeń.

Budowa modelu liniowego:

1. Sprawdzenie aktywności j -tej cechy $b_{i,j} = \mathbf{1}[x_{i,j} \neq 0]$,
2. Usunięcie j -tej cechy jeśli $|\text{cor}(\mathbf{b}_j, \mathbf{y})| \leq 0.05$
3. Budowa modelu regresji logistycznej z regularyzacją L1 (*glmnet*).

Klasyfikator neuronowy





Dostosowanie

Przetwarzania wstępne



Niezbędne dostosowania:

1. Zastąpienie *CoreNLP* bibliotekami NLP dostępnymi w JS:
 - `compromise` do tokenizacji (wyrażenia regularne),
 - `javascript-lemmatizer` do lematyzacji (słownik),
2. Zastąpienie heurystyki do wydobywania tekstu z HTML przez bibliotekę `unfluff`,

Unfluffing: przed



BREAKING: Malia Obama Sex Tape Surfaces And It's GROSS - **American Flavor**

ShareTweet

Admit it: you knew it was going to happen the moment you first saw video of her twerking and doing drugs in public. Malia Obama's sex tape has surfaced.

(...)

The Obamas love to pretend that they're perfect. Maybe it will help King Barack and King Michelle remember what it's like to be human now that their daughter is a wrecking ball of depravity.

Meanwhile, President Trump's daughters are all successful, hard-working assets to America rather than drug-using embarrassments to humanity.

Share on Facebook Share

Share on Twitter Tweet

Share on Google Plus Share

Share on Pinterest Share

Share on LinkedIn Share

Share on Digg Share

<http://americanflavor.news/2017/09/03/breaking-malia-obama-sex-tape-surfaces-gross/>

Unfluffing: po



Malia Obama Sex Tape Surfaces And It's GROSS

Admit it: you knew it was going to happen the moment you first saw video of her twerking and doing drugs in public. Malia Obama's sex tape has surfaced.

(...)

The Obamas love to pretend that they're perfect. Maybe it will help King Barack and King Michelle remember what it's like to be human now that their daughter is a wrecking ball of depravity.

Meanwhile, President Trump's daughters are all successful, hard-working assets to America rather than drug-using embarrassments to humanity.

<http://americanflavor.news/2017/09/03/breaking-malia-obama-sex-tape-surfaces-gross/>

Kompresja modelu neuronowego



Problem:

- Model neuronowy ma 85M parametrów,
- Rozszerzenie przeglądarki wykonywane na kliencie o nieznanych zasobach, w JS,
- Najwięcej zajmuje tablica 300-wymiarowych osadzeń 929.019 słów z *word2vec*.

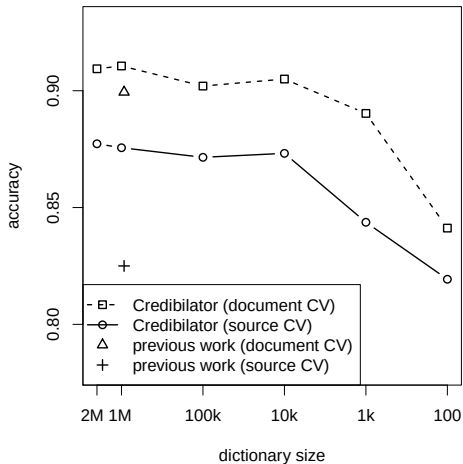
Kompresja słownika:

- Wykorzystanie osadzeń z *fastText* uszeregowanych według częstości,
- Uwzględnienie najczęstszych k słów,
- Zastąpienie reszty tokenem [UNK] z osadzeniem **uczonym wraz z siecią**.

Wpływ kompresji



Jak dużego słownika potrzebujemy?



Dokładność klasyfikacji



Przetwarzanie	Klasyfikator	Dokładność		Rozmiar
		doc. CV	source CV	
oryg.	Stylometryczny (oryg.)	0.9274	0.8097	939
	Neuronowy (oryg.)	0.8994	0.8250	85.18M
	BERT	0.9976	0.7960	110M
adap.	Stylometryczny (adap.)	0.8680	0.8081	617
	Neuronowy (adap.)	0.9050	0.8732	3.26M
	BERT	0.9629	0.9292	110M



Interakcja

Interpretowalność



Interpretowalność: własność systemu opisująca zdolność użytkowników do rozumienia jego sposobu działania [Lipton, 2016].
Stopniowalna [Doran et al., 2017]:

- system **nieprzejrzysty** (*opaque*): przekształcenie wejść na wyjścia jest ukryte przed użytkownikiem $\Rightarrow$ wyjaśnienia zewnętrzne, np. typu LIME [Ribeiro et al., 2016],
- system **zrozumiały** (*comprehensible*): nie ujawnia sposobu działania, ale emituje dodatkowe symbole ułatwiające użytkownikowi zrozumienie $\Rightarrow$ klasyfikator neuronowy
- system **interpretowalny** (*interpretable*): użytkownik może poznać i zrozumieć przekształcenie wejść na wyjścia $\Rightarrow$ klasyfikator stylometryczny,

Interpretowalność: klasyfikator stylometryczny 1



Regresja logistyczna dla pomiaru nie-wiarygodności (s) przykładu $\mathbf{x}_i$:

$$s(\mathbf{x}_i) = \frac{1}{1 + e^{-t(\mathbf{x}_i)}}, \quad t(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{i,j}.$$

Jak zdefiniować **wkład** cechy $c(i, j)$?

- $c(i, j) = \beta_j$
- $c(i, j) = \beta_j x_{i,j}$
- $c(i, j) = \beta_j (x_{i,j} - \bar{x}_j)$

Interpretowalność: klasyfikator stylometryczny 2

Gdy wybierzemy $c(i, j) = \beta_j(x_{i,j} - \bar{x}_j)$ mamy:

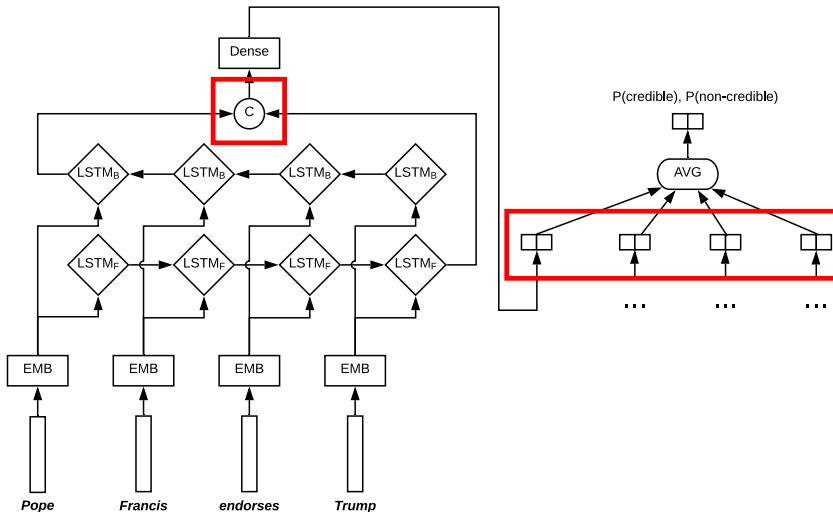
$$s(\mathbf{x}_i) = \frac{1}{1 + e^{-t(\mathbf{x}_i)}}, \quad t(\mathbf{x}_i) = t_0 + \sum_{j=1}^p c(i, j),$$

gdzie $t_0 = \beta_0 + \sum_{j=1}^p \beta_j \bar{x}_j$ to wyraz wolny odpowiadający $t(\mathbf{x})$ dla $\mathbf{x}$ o *średnich* cechach.

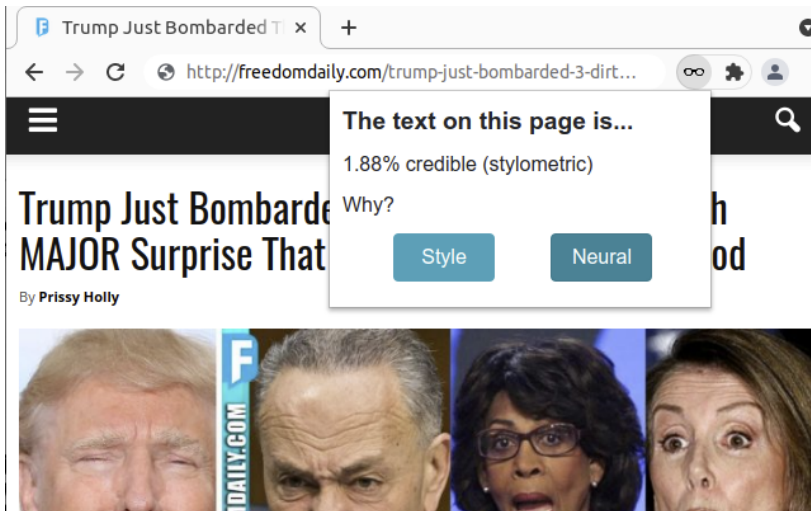
Np. dla udziału słów pisanych małymi literami mamy $\bar{x}_j = 0.83$ i $\beta_j = -2.98$. Teraz c jest dodatnie (zwiększona nie-wiarygodność) dla dokumentów o mniej niż 83% słów małymi literami i *vice versa*.

Równoważne *SHAP values* dla modelu liniowego [Lundberg and Lee, 2017].

Interpretowalność: klasyfikator neuronowy



UI



Ul: tryb neuronowy 1

Credibilator

Stylometric ☒ Neural

Credibility  100%

Human view Machine view

Anthony Bourdain's Last Interview Shockingly Details
Bill Clinton's 'Disgusting' Behavior - Conservative Daily
Post

[illegible]

Similar sentences

Click on a sentence in the main panel to find other similar sentences

Similar sentences to 'Frightening' 'Boundless Blasted' 'Disgusting' 'Biff' 'Destroying' 'Hillary Celebrity Chef Anthony

1- 'Possessed' eBay Doll Allegedly Wreaks Havoc on Married Men A 'haunted' doll flogged on eBay after it left its owners terrified and covered in scratches has struck again.

2- "The Mysterious Sounds Are In the Below Feature Video-When you get to them, they are best heard through stereo headphones" Note: I want to make it clear that I am not prophesying this event, as I am merely a watchman and not a prophet.

3- 'Crabzilla': 50 ft Crab Spotted Near Wharf In England This photo of a giant crab was featured on the website 'Weird Whistable' and has people of this England community startled. dailybuzzle.com

4 - "Obesity Is Becoming A National Security Issue" (LVR) - Soon, America will be too fat to fight. dailybuzzlive.com

was walking the aisle at the Grammy Awards Sunday night when she noticed a young woman with a "Make America Great Again" shirt.

6 - Info Media 179 SHARES ShareTweet Julian Assange is saying that all terror plots are planned by themselves.

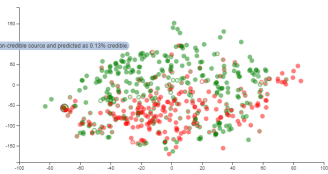
anything but normal. www.socialsavethings.com

9- "His First Four Sentences Are Amazing." www.npr.org politicalmayhem.news

10-11:30: Attorney Rick Warren has commented a federal investigation and will have the full support of the Department of Justice. www.cbsnews.com

Sentence map

Drag and drop to zoom in, double-click to zoom-out



Serwer *mongoDB* przechowujący 2.845.002 zdań z korpusu:

- *locality-sensitive hashing* [Indyk et al., 1997] z 10-bitowymi skrótami do wydobywania 0.1% kandydatów,
- ranking przez podobieństwo cosinusowe kandydatów.

UI: tryb neuronowy 2



Human view

Machine view

Anthony Bourdain's Last Interview Shockingly Details Bill Clinton's 'Disgusting' Behavior - Conservative Daily Post

'Frightening' Bourdain Blasted 'Disgusting' Bill, 'Destroying' Hillary Celebrity Chef Anthony Bourdain shocked the culinary world with his suicide last month. Even more surprisingly, published yesterday was a particularly frank and scathing "bombshell" assessment of Bill Clinton's "rapey, gropey and disgusting" character, from what would turn out to be the final interview before his death. For decades, "conspiracy theorists" have been tracking suspicious suicides and other unusual deaths of people connected with Bill and Hillary Clinton. Suitable to his award-winning career as a food reporter, Bourdain's name sadly gets added to the list at number 0, according to some. Snopes calls the list "false" only because the allegations can not be proven. They admit they all can not be ruled out either. Bourdain was legendary for the way he scrounged the offbeat corners of the world to document the dishes ordinary people enjoy every day. His show "Parts Unknown" received several awards and was about to be nominated for 0 more Emmy's this year. He was dating Italian actress Asia Argento. Everything was going his way. Another thing big in Bourdain's life was the #MeToo movement. He was known as one of the most vocal male voices in the discussion, ever since Argento came forward as one of Harvey Weinstein is very first identified victims. Weinstein has been accused by 0 women so far, including actresses Ashley Judd, Rose McGowan, Angelina Jolie and Gwyneth Paltrow. Allegations range from harassment to rape, go back for decades, and span the globe. Weinstein was "a generous donor to Hillary and Bill Clinton" for a very long time. According to Mercury News, "he donated \$10,000 to Bill

Human view

Machine view

Anthony Bourdain's Last Interview Shockingly Details Bill Clinton's 'Disgusting' Behavior - Conservative Daily Post

'Bill', 'Hillary' Chef Anthony Bourdain shocked the world with his suicide last month. Even more surprisingly, published yesterday was a particularly and "assessment of Bill Clinton's " and " character, from what would turn out to be the final interview before his death. For decades, "conspiracy theorists" have been tracking suspicious and other unusual deaths of people connected with Bill and Hillary Clinton. Suitable to his award-winning career as a food reporter, name sadly gets added to the list at number 0, according to some. Snopes calls the list "false" only because the allegations can not be proven. They admit they all can not be ruled out either. Bourdain was legendary for the way he the corners of the world to document the dishes ordinary people enjoy every day. His show "Parts Unknown" received several awards and was about to be nominated for 0 more this year. He was dating Italian actress Asia Argento. Everything was going his way. Another thing big in Bourdain's life was the #MeToo movement. He was known as one of the most vocal male voices in the discussion, ever since Argento came forward as one of Harvey Weinstein is very first identified victims. Weinstein has been accused by 0 women so far, including Ashley Judd, Rose McGowan, Angelina Jolie and Gwyneth Paltrow. Allegations range from harassment to rape, go back for decades, and span the globe. Weinstein was "a generous donor to Hillary and Bill Clinton" for a very long time. According to Mercury News, "he donated \$10,000 to Bill

UI: tryb neuronowy 3



Similar sentences

Click on a sentence in the main panel to find other similar sentences

Similar sentences to 'Frightening 'Bourdain Blasted 'Disgusting 'Bill , 'Destroying 'Hillary Celebrity Chef Anthony Bourdain shocked the culinary world with his suicide last month .:

1- 'Possessed' eBay Doll Allegedly Wreaks Havoc on Married Men A 'haunted' doll flogged on eBay after it left its owners terrified and covered in scratches has struck again. worldtruth.tv

2- *The Mysterious Sounds Are In the Below Feature Video-When you get to them, they are best heard through stereo headphones* Note: I want to make it clear that I am not prophesying this event, as I am merely a watchman and not a prophet. worldtruth.tv

3- 'Crabzilla' 50 ft Crab Spotted Near Wharf in England This photo of a giant crab was featured on the website 'Weird Whistable' and has people of this England community startled. dailybuzzlive.com

4- "Obesity Is Becoming A National Security Issue" LVRJ – Soon, America will be too fat to fight. dailybuzzlive.com

5- - News Conservative Madonna, pretentious liberal icon who thought blowing up the White House would be fun, was walking the aisle at the Grammy Awards Sunday night when she noticed a young woman with a "Make America Great Again" shirt. newsconservative.com

6- - Info Media 179 SHARES ShareTweet Julian Assange is saying that all terror plots are planned by themself. freeinfomedia.com

7- - Social Every Things Divorces nowadays are no longer unusual, but the divorce this man is anything but normal. www.social.everythings.com

8- - Political Mayhem Saturday, June 3, 2017 Sign in / Join Home Contact Us Privacy Policy Sign in Welcome!
9- "His First Four Sentences Are Amazing" Published on a credible source and predicted as 94.10% credible www.npi.org calmayhem.news

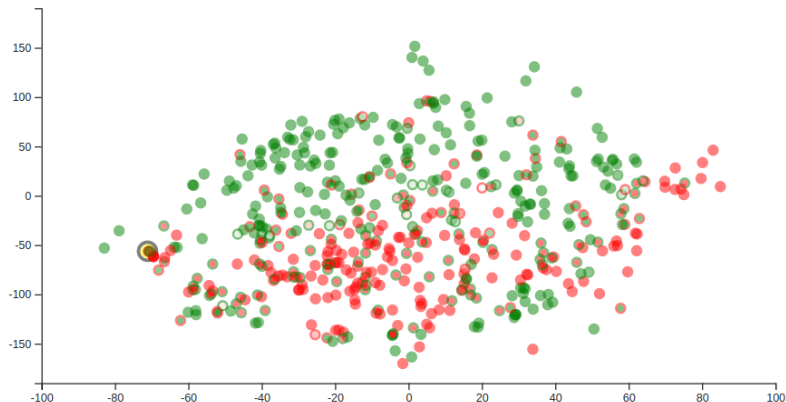
10- "U. S. Attorney Rick Mountcastle has commenced a federal investigation and will have the full support of the Department of Justice. www.cbsnews.com

UI: tryb neuronowy 4



Sentence map

Drag and drop to zoom in, double-click to zoom-out



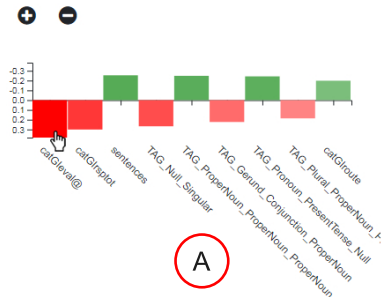
UI: tryb stylometryczny 1



Feature contribution

Click +/- to add or remove a feature.

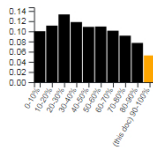
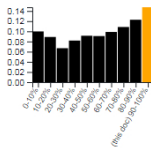
Click on a bar to inspect feature distribution.



Feature distribution

Non-credible news

Credible news



catGieval@
words which imply judgment and evaluation (and similar)



UI: tryb stylometryczny 2



Anthony Bourdain's Last Interview Shockingly Details Bill Clinton's 'Disgusting' Behavior - Conservative Daily Post

Frightening Bourdain Blasted **Disgusting** Bill, 'Destroying' Hillary Celebrity Chef Anthony Bourdain shocked the culinary world with his suicide last month. Even more surprisingly, published yesterday was a particularly frank and scathing "bombshell" assessment of Bill Clinton's "rapey, gropey and **disgusting**" character, from what would turn out to be the final interview before his death. For decades, "conspiracy theorists" have been tracking **suspicious** suicides and other **unusual** deaths of people connected with Bill and Hillary Clinton. **Suitable** to his award-winning career as a food reporter, Bourdain's name sadly gets added to the list at number 57, according to some. Snopes calls the list "false" only because the allegations cannot be proven. They admit they all can't be ruled out either. Bourdain was legendary for the way he scrounged the offbeat corners of the world to document the dishes ordinary people enjoy every day. His show "Parts Unknown" received several awards and was about to be nominated for 6 more Emmy's this year. He was dating Italian actress Asia Argento. Everything was going his way. Another thing big in Bourdain's life was the #MeToo movement. He was known as one of the most vocal male voices in the discussion, ever since Argento came forward as one of Harvey Weinstein's very first identified victims. Weinstein has been accused by 83 women so far, including actresses Ashley Judd, Rose McGowan, Angelina Jolie and Gwyneth Paltrow. Allegations range from harassment to rape, go back for decades, and span the globe. Weinstein was "a generous donor to Hillary and Bill Clinton" for a very long time. According to Mercury News, "he donated \$10,000 to Bill Clinton's legal defense fund after he was impeached in the Monica Lewinsky scandal, and he served as a fundraiser and informal adviser during Hillary Clinton's 2000 Senate campaign." He wrote checks to Hillary's presidential campaign too. Hillary Clinton had to say something in front of the cameras, quick. She claimed she was "appalled" at what Weinstein was doing "behind closed doors." "It was something that was just **intolerable** in every way... this was a different side of a person who I and many others had known in the past." That was not what Bourdain was expecting to hear. He took to Twitter in outrage. Her response is **shameful** in its



Evaluacja

Badania z użytkownikami (A)



Czy korzystanie z interfejsu zmienia pogląd użytkowników na wiarygodność tekstu? Czy nowa ocena jest bardziej z zgodna z oceną ekspertów?

Schemat badania:

1. Wyświetlenie artykułu o nieznanej wiarygodności a w formie tekstowej,
2. Ocena wiarygodności $\hat{a} \in \{1, 2, 3, 4, 5\}$ (*unaided*),
3. Analiza treści ze wskazanym modelem Credibilatora,
4. Ocena wiarygodności $\hat{a} \in \{1, 2, 3, 4, 5\}$ (*aided*).

Tematy i dokumenty



Artykuły dot. bieżących wydarzeń w źródłach o znanej wiarygodności:

- A: Nomination of Donald Trump for Nobel peace prize
- B: State of Virginia passing new legislation on policing
 - NC: *enVolve – DANGER: Virginia's Radically Left Democrats Pass SICKENING Bill That Opens Up Police To Assault!*
 - C: *The Washington Post – Virginia Senate passes sweeping police overhaul bill*
- C: Water contamination in Flint, Michigan
- D: NFL opening the season with a minute of silence
- E: Kanye West releasing a new album

Zadania



User	Model					Credibility				
	A	B	C	D	E	A	B	C	D	E
1	S	N	S	N	S	C	C	NC	NC	C
2	S	N	S	N	S	NC	NC	C	C	C
3	N	S	N	S	N	NC	C	C	C	NC
4	S	N	S	N	S	C	C	C	NC	NC
5	N	S	N	S	N	NC	NC	C	C	NC
6	S	N	S	N	S	NC	C	C	NC	NC
7	N	S	N	S	N	C	C	NC	NC	NC
8	S	N	S	N	S	C	NC	NC	NC	C
9	N	S	N	S	N	NC	NC	NC	C	C
10	S	N	S	N	S	C	C	NC	NC	C
11	N	S	N	S	N	C	NC	NC	C	C
12	S	N	S	N	S	NC	NC	C	C	C
13	S	N	S	N	S	C	C	C	NC	NC
14	S	N	S	N	S	NC	C	C	NC	NC

Ewaluacja



Mierzymy:

- **dokładność:** $\frac{|\{i:(a_i=C \wedge \hat{a}_i > 3) \vee (a_i=NC \wedge \hat{a}_i < 3)\}|}{N}$
- **pewność:** $\frac{\sum_i \frac{\hat{a}_i - 3}{2}}{N}$

ze względu na:

- pracę z lub bez Credibilatora,
- używanie modelu stylometrycznego lub neuronowego,
- prawdziwą ocenę wiarygodności.

Badanie B



Obecność czynników wpływających na wynik niezależnie od jakości narzędzia:

- dokładniejsza ocena wiarygodności przy drugim czytaniu artykułu,
- kopiowanie oceny bez próby zrozumienia wizualizacji,

Rozwiązanie: część zadań w grupie kontrolnej (*placebo*) – *uszkodzony* Credibilator, pokazujący losowe wyniki.

Wyniki: badanie A



	Dokładność		Pewność	
	Unaided	Aided	Unaided	Aided
Model stylometryczny	75.68%	89.19%	62.16%	74.32%
Model neuronowy	69.70%	81.82%	63.64%	65.15%
Artykuł wiarygodny	72.22%	88.89%	59.72%	75.00%
Artykuł niewiarygodny	73.53%	82.35%	66.18%	64.71%
Suma	72.86%	85.71%*	62.86%	70.00%

Wyniki: badanie B



	Dokładność		Pewność	
	Unaided	Aided	Unaided	Aided
Uszkodzony Credibilator	61.90%	66.67%	57.14%	57.14%
Prawdziwy Credibilator	66.67%	85.71%	52.38%	66.67%



Konkluzja

Wnioski



1. TAK, można modele klasyfikacyjne dostosować do ograniczonych zasobów bez straty wydajności,
2. TAK, wsparcie narzędzia zwiększa dokładność oceny wiarygodności przez użytkowników,
3. ALE model o większej dokładności *in vitro* może działać lepiej *in vivo*.

Ograniczenia:

- niewielki rozmiar badania na użytkownikach: 23 uczestników $\times$ 5 tematów $\times$ 2 oceny,
- brak oceny indywidualnych elementów interfejsu: wszystko albo nic,
- brak pomiaru stopnia zrozumienia interpretowalnych elementów.

Zasoby



Credibilator dostępny do instalacji z *Chrome Web Store*.

Dodatkowo, w repozytorium

<https://github.com/piotrmp/credibilator>:

- wyczyszczony korpus dokumentów,
- kod źródłowy rozszerzenia,
- film z instrukcją użytkownika,
- kody źródłowe serwera.

Projekt HOMADOS (*Hampering Misinformation by Assessing Credibility of Online Sources*) finansowany z programu *Polskie Powroty* Narodowej Agencji Wymiany Akademickiej (NAWA), grant nr PPN/PPO/2018/1/00006.



Dziękuję!

Literatura (1)



- Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.7. URL <https://aclanthology.org/2021.semeval-1.7>.
- Francisco Rangel and Paolo Rosso. Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*. CEUR-WS.org, 2019.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018.
- Emma J Llansó. No amount of “AI” in content moderation will solve filtering’s prior-restraint problem. *Big Data and Society*, 7(1), jan 2020. ISSN 2053-9517. doi: 10.1177/2053951720920686. URL <http://journals.sagepub.com/doi/10.1177/2053951720920686>.
- Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, pages 1–14, feb 2020. ISSN 0025-1909. doi: 10.1287/mnsc.2019.3478. URL <http://pubsonline.informs.orghttp://www.informs.orgMANAGEMENTSCIENCEhttp://pubsonline.informs.org/journal/mnsc>.

Literatura (2)



- Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, feb 2019. ISSN 0190-9320. doi: 10.1007/s11109-019-09533-0. URL <http://link.springer.com/10.1007/s11109-019-09533-0>.
- Paul Mena. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet*, 12(2):165–183, jun 2020. ISSN 1944-2866. doi: 10.1002/poi3.214. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.214>.
- Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, jun 2010. ISSN 01909320. doi: 10.1007/s11109-010-9112-2.
- Brendan Nyhan and Jason Reifler. Displacing Misinformation about Events: An Experimental Test of Causal Corrections. *Journal of Experimental Political Science*, 2(1):81–93, apr 2015. ISSN 2052-2630. doi: 10.1017/XPS.2014.22. URL https://www.cambridge.org/core/product/identifier/S2052263014000220/type/journal_article.
- Brendan Nyhan and Jason Reifler. The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, 29(2):222–244, apr 2019. ISSN 17457297. doi: 10.1080/17457289.2018.1465061. URL <https://www.tandfonline.com/doi/abs/10.1080/17457289.2018.1465061>.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You? Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016. ISBN 9781450321389. doi: 10.1145/2939672.2939778.

Literatura (3)



- Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20*, pages 189–201, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377480. URL <https://doi.org/10.1145/3377325.3377480>.
- Piotr Przybyła. Capturing the Style of Fake News. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, volume 34, pages 490–497, New York, USA, 2020. AAAI Press. doi: 10.1609/aaai.v34i01.5386. URL <https://aaai.org/ojs/index.php/AAAI/article/view/5386>.
- Amy Mitchell, Jocelyn Kiley, Jeffrey Gottfried, and Katerina Eva Matsa. Political Polarization & Media Habits. Technical report, Pew Research Center, 2014. URL <https://www.journalism.org/2014/10/21/political-polarization-media-habits/>.
- Joshua Gillin. PolitiFact's guide to fake news websites and what they peddle. *PolitiFact*, 2017. URL <http://www.politifact.com/punditfact/article/2017/apr/20/politifact-guide-fake-news-websites-and-what-they/>.
- Vian Bakir and Andrew McStay. Fake News and The Economy of Emotions: Problems, causes, solutions. *Digital Journalism*, 6(2):154–175, 2017. ISSN 2167082X. doi: 10.1080/21670811.2017.1345645.
- Zachary C. Lipton. The Mythos of Model Interpretability. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, USA, jun 2016. URL <http://arxiv.org/abs/1606.03490>.
- Derek Doran, Sarah Schulz, and Tarek R. Besold. What does explainable AI really mean? A new conceptualization of perspectives. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence*, Bari, Italy, 2017.

Literatura (4)



Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017.

Piotr Indyk, Rajeev Motwani, Prabhakar Raghavan, and Santosh Vempala. Locality-preserving hashing in multidimensional spaces. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 618–625, 1997.