Ocena wiarygodności dokumentów na podstawie stylu

Piotr Przybyła
piotr.przybyla@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

9. marca 2020

Przykłady z korpusu

NEWS

Caitlyn Jenner Discusses Her Desire To Transition Into A Black Woman

NEWS

3 Year Old Girl Dies After Accidentally Being Tickled To Death By Her Mother

NEWS

18 Year Old Girl Marries Her Father In Arkansas, “We Belong Together”

NEWS

Hispanic Woman Claims, “Donald Trump Paid Me For Sex In Cancun, This Is Our Love Child”

NEWS

4,500 Pound Crocodile Found Roaming The Streets Of Houston, Texas Floods

NEWS

Blac Chyna Signs $2 Million Porn Deal With Brazzers

NEWS

Girl Infects 586 Men with HIV on Purpose, Plans on Infecting 2,000 More Before 2018

NEWS

Dead Body Of Homeless Man Turns Out To Be The Legendary Elvis Presley

Źródło: `https://web.archive.org/web/20171109112741/http://now8news.com/`

Plan prezentacji

Na podstawie: P. Przybyła, "Capturing the Style of Fake News," w *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (AAAI-2020), Nowy Jork, USA, 2020. `https://home.ipipan.waw.pl/p.przybyla/bib/capturing.pdf`

Co to znaczy „fake news”?

Gelfert [2018]

Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design.

Istotne:

- raczej czynność (*deliberate presentation*) niż rezultat (*fake news*),
- ma znaczenie prezentacja jako wiadomości (*as news*) i zamiar twórcy (*by design*),
- nieprawdziwość treści efektem ubocznym, a nie celem (*typically*).

$\longrightarrow$ Konieczność oceny w kontekście.

$\longrightarrow$ Na podstawie samego tekstu można ocenić jedynie **wiarygodność**.

fake⇎false

fake ⇏ false

niedokładna informacja podana przez wiarygodne medium,

- *Kometa 46P/Wirtanen obiega Słońce po stabilnej orbicie eliptycznej.*
- *Wielka Brytania opuści Unię Europejską najpóźniej 29 marca 2019 roku.*
- *W pożarze Greenfell Tower zginęło 79 osób.*

fake ⇎ false

dezinformacji i wprowadzanie w błąd bez fałszywych stwierdzeń,

- wybiórcze przedstawianie informacji pasujących do narracji,
- manipulowanie przekonaniem o powszechności poglądów (np. przez boty),
- *clickbait*.

Inne „fake news”

Niektórzy używają termin *fake news* w odniesieniu do:

- silnie stronniczych mediów (*hyperpartisanship*) [Potthast et al., 2018],
- dokumentów pochodzących z generatorów tekstu (np. GPT-2) [Zellers et al., 2019],
- wiadomości zmyślonych w celach satyrycznych [Burfoot and Baldwin, 2009],
- szkodliwych plotek w mediach społecznościowych [Zhao et al., 2015],
- jakichkolwiek informacji niesprzyjających im politycznie. . .

Zadanie

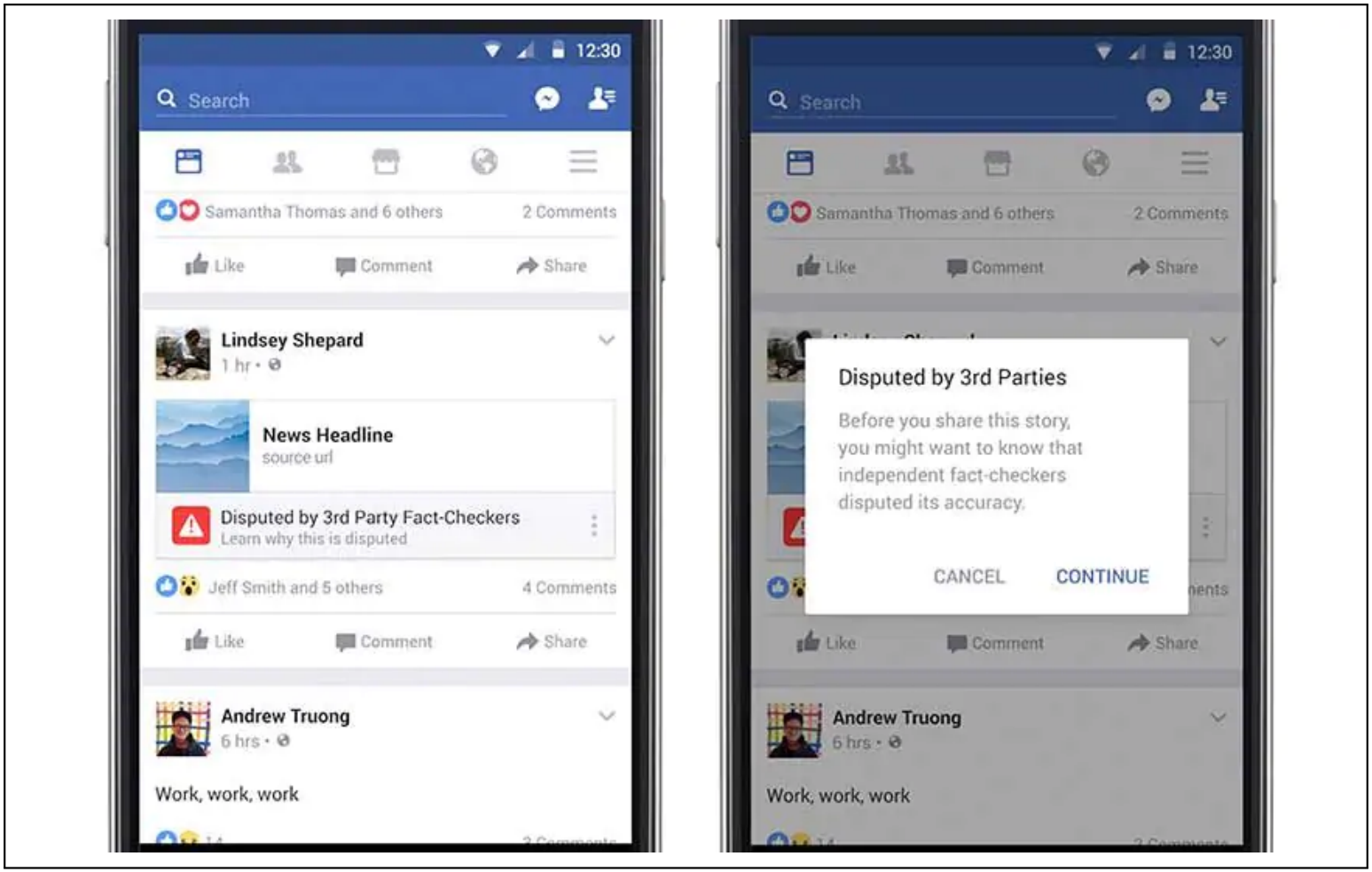
Czemu warto oceniać

Ocena wiarygodności tekstu wiadomości w Internecie na wielu etapach:

- przygotowywania treści przez dziennikarzy – postulowana [Chen et al., 2015],
- rozpowszechniania w sieciach społecznościowych – realizowana:
  - Facebook oznacza treści jako *Disputed* lub *Rated false* oraz ogranicza ich dystrybucję,
  - Google uwzględnia *authoritativeness* w tworzeniu rankingu,
  - Twitter pracuje nad oznaczaniem wiadomości *harmfully misleading*,
- konsumpcji treści przez czytelników – postulowana [Berghel, 2017]

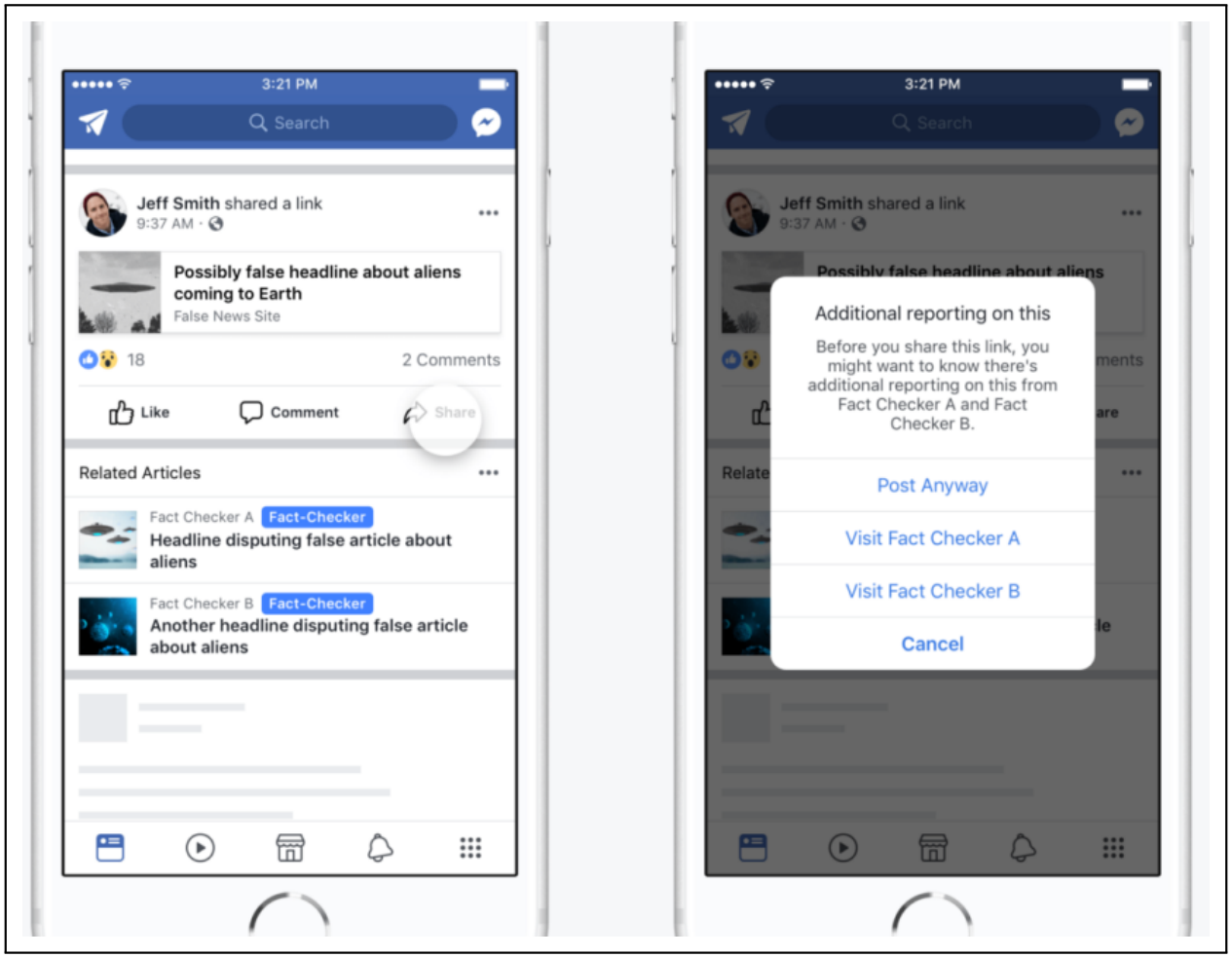
`https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019`
`https://www.nbcnews.com/tech/tech-news/twitter-testing-new-ways-fight-misinformation-including-community-based-points-n1139931`

W praktyce (1)



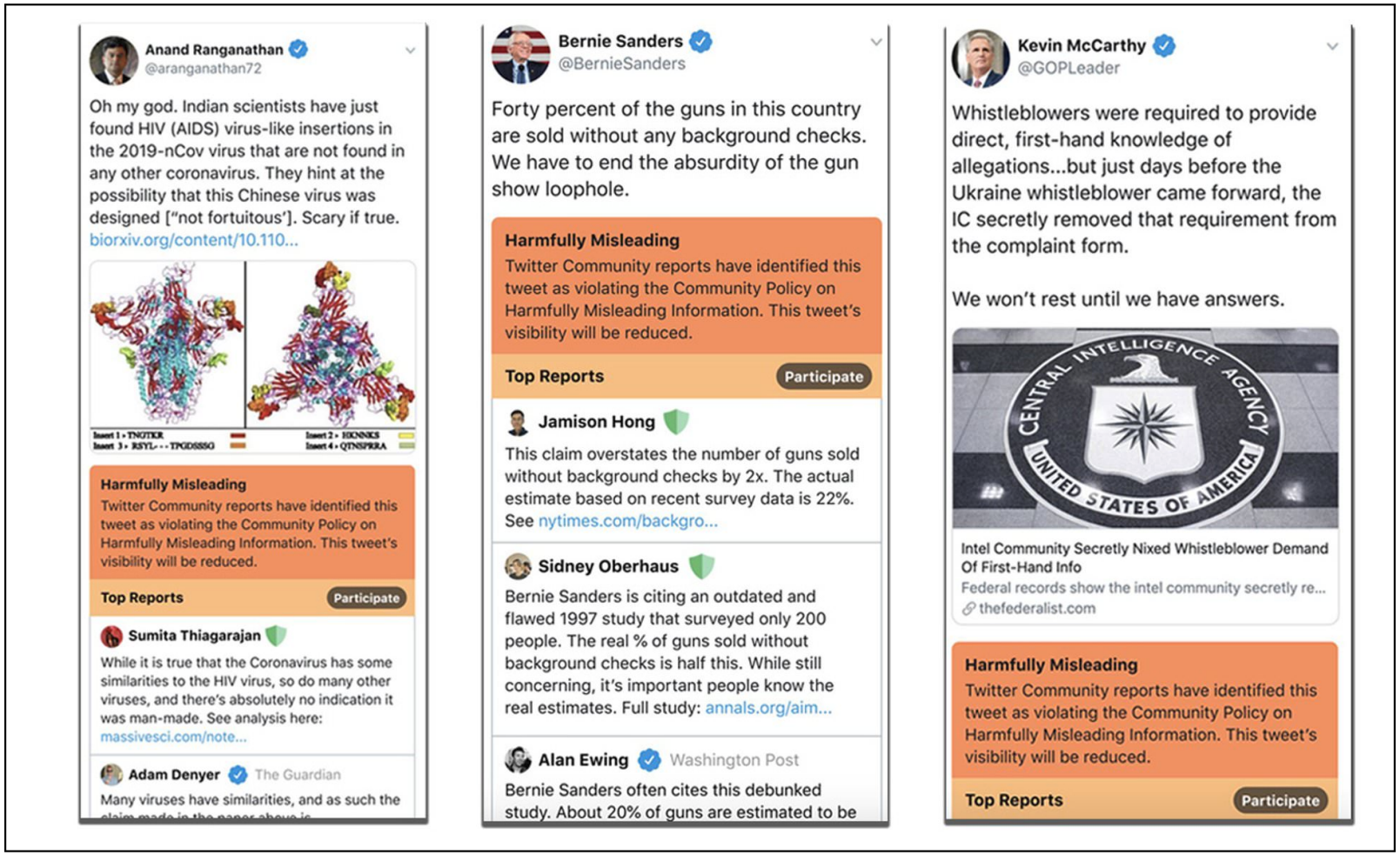
Źródło: https://www.telegraph.co.uk/technology/2016/12/16/facebook-roll-new-tools-tackle-fake-news/

W praktyce (2)



Źródło: `https://techcrunch.com/2017/12/20/facebook-will-ditch-disputed-flags-on-fake-news-and-display-links-to-trustworthy-articles-instead/`

W praktyce (3)



Źródło: `https://www.dailymail.co.uk/sciencetech/article-8026171/Twitter-testing-feature-labels-misinformation-fake-news-Harmfully-Misleading.html`

Czemu nie warto oceniać

Trudności w stosowaniu automatycznej oceny:

- kontekstowość definicji *fake news* utrudnia zadanie (np. materiały satyryczne):
  → możliwa ocena wiarygodności (*credibility*),
  → w wielu zastosowaniach konieczna ręczna ocena,
- binarna etykieta niekoniecznie spełnia cel:
  → eksperymenty na FB wykazały niewielką skuteczność [Clayton et al., 2019],
  → możliwy efekt przeciwny w polaryzujących kwestiach, np.:
  - polityki [Nyhan and Reifler, 2010],
  - globalnego ocieplenia [Hart and Nisbet, 2012],
  - czy szczepionek [Nyhan et al., 2014],

  → może pomóc dodanie alternatywnego wyjaśnienia [Nyhan and Reifler, 2015].

Jak oceniać?

Dwa podejścia:

- Weryfikacja stwierdzeń,
  - weryfikacja prawdziwości stwierdzeń w dokumencie (*fact-checking*),
  - *to nieprawda*
- Klasyfikacja tekstu,
  - uwzględnia całość dokumentu, w tym tematykę, słownictwo i styl,
  - *to brzmi niewiarygodnie*

Źródło: `https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/`

Trudności z weryfikacją stwierdzeń

- weryfikacja stwierdzeń główną metodą dyskredytacji *fake news* stosowaną przez ludzi (np. *PolitiFact*, *FactCheck.org*, *Snopes*),
- próby automatyzacji przez porównywanie z treścią baz wiedzy: tekstowych (*Wikipedia*) lub symbolicznych (*Knowledge Graph*),
- ograniczenia baz wiedzy:
  - częstotliwość aktualizacji (z dnia dzisiejszego),
  - ograniczona wiarygodność (*zm. 5 czerwca 2018 w Watykanie*),
  - ograniczone pokrycie danych nieencyklopedycznych (*Papież Franciszek powiedział, że . . .*),
  - brak informacji negatywnych (*Papież Franciszek NIE powiedział, że . . .*).

Rozwiązania polegają np. na analizie odległości w grafie wiedzy [Ciampaglia et al., 2015] jako przybliżeniu prawdopodobieństwa stwierdzenia.

Klasyfikacja tekstu

Proste zastosowanie ogólnego klasyfikatora tekstu daje mu swobodę wyboru użytecznych cech dokumentu:

- **tematów**, bo *fake news* skupiają się na kilku aktualnie gorących zagadnieniach [Bakir and McStay, 2017],
  $\longrightarrow$ dokładność osłabnie, gdy zmienią się bieżące wydarzenia,
- **źródła**, bo wiarygodność treści zwykle jest podobna w ramach tego samego medium
  $\longrightarrow$ strony *fake news* żyją krótko [Allcott and Gentzkow, 2017],
- **stylu**, gdyż niska wiarygodność wiąże się z nieformalnym, sensacyjnym, emocjonalnym językiem [Bakir and McStay, 2017],
  $\longrightarrow$ nie da się porzucić bez ograniczenia zasięgu.

Tematyka *fake news*: przykład

conservativeview.info

Breaking News

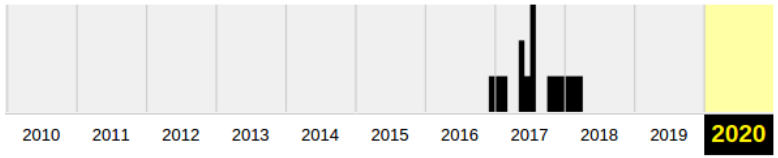
- WATCH: OBAMA'S MUSLIM FAMILY BREAKS THEIR SILENCE – Reveals TRUTH ABOUT OBAMA
- ALERT: Bananas Being Injected With HIV Blood… Here's How You Can Tell
- Right After Germany Attacked Trump, What He Did Next Is Insane!
- Terror Just Erupted In Minnesota At Start Of Ramadan After Governor Welcomed Refugees In Droves
- BREAKING: Trump Just DECIMATED Obama's Cuba Legacy In One UNFORESEEABLE Move
- Millions of People Outraged After Obama Quietly Did This Behind Trump's Back – Look What Hi Is Building!
- Hillary Goes Into Hiding After Republican Presidential Candidate Calls For Her To Be Investigated For Murder
- Anthony Weiner Placed In Protective Custody–Will Turn State's Evidence Against Hillary
- Seconds Ago Trump Stepped To Son Of Fallen Marine And Handed Him The SURPRISE Of A Lifetime
- Trump Just Got On Stage & People Immediately Noticed 1 Thing He Did With His Mouth – We Love It!!

Źródło: `http://conservativeview.info/`

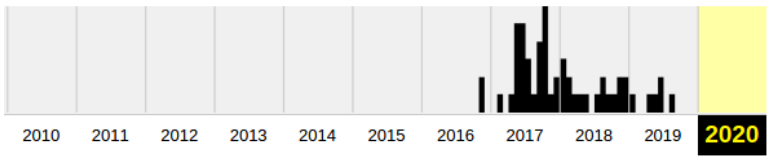
Źródła *fake news*: przykład

Na podstawie liczby kopii w *Web Archive* (web.archive.org)

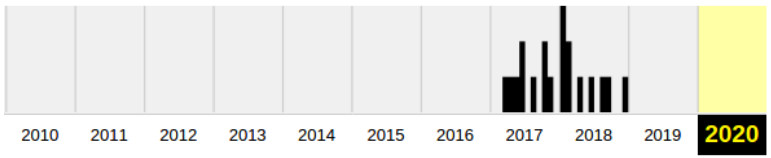
247NewsMedia.com



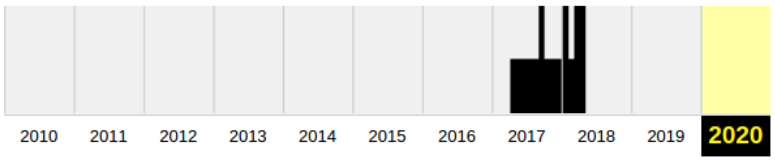
24wpn.com



24x365live.com



AmericanFlavor.news



Literatura: przesłanki klasyfikacji

Poprzednie podejścia do wykrywania *fake news* jako zadania klasyfikacji:

- opierały się o niewielkie dane: 144 [Rubin et al., 2015] lub 110 tekstów zgromadzonych ręcznie [Horne and Adali, 2017],
- wykorzystywały ogólne cechy tekstowe (TF-IDF, n-gramy),
- w rezultacie osiągając zdecydowanie słabsze wyniki w bardziej realistycznych scenariuszach ewaluacji [Pérez-Rosas et al., 2018, Rashkin et al., 2017] oraz polegały na słowach określających temat [Rashkin et al., 2017], np. Syria.

$\longrightarrow$ Analiza wyników wskazuje, że pomimo deklaracji autorów (*style-based*), klasyfikatory te bazują jednak na tematach i źródłach. Jedyny wyjątek: cechy stylometryczne stosowane do pomiaru stronniczości mediów [Potthast et al., 2018].

Stylometria

Styl [Ray, 2015]

the deployment of linguistic resources in written discourse to express and create meaning

Stylometria: pomiar i wykorzystanie różnic w wykorzystaniu zasobów językowych do określenia:

- gatunku tekstu [Argamon et al., 2003]
- tożsamości autora (*authorship attribution*) [Stamatatos, 2009],
- cech autora (*author profiling*):
  - płci [Koppel et al., 2002],
  - poglądów [Diermeier et al., 2011],
  - wieku [Argamon et al., 2007]
  - wykształcenia, przynależności partyjnej [Przybyła and Teisseyre, 2014],
  - (do pewnego stopnia) cech osobowości [Przybyła and Teisseyre, 2015].

Cel pracy

Zbudować klasyfikator wiarygodności dokumentów w j. angielskim w oparciu o styl poprzez:

1. pozyskanie dużego (103,219 documents) korpusu treningowego, by uniknąć niewłaściwego dopasowania,
2. zaprojektowanie modeli uwzględniających przesłanki stylometryczne poprzez cechy lub architekturę,
3. ewaluowanie z użyciem dokumentów z uprzednio nieznanych źródeł lub tematów,
4. zweryfikowanie, czy model istotnie opiera się o przesłanki stylistyczne.

Korpus

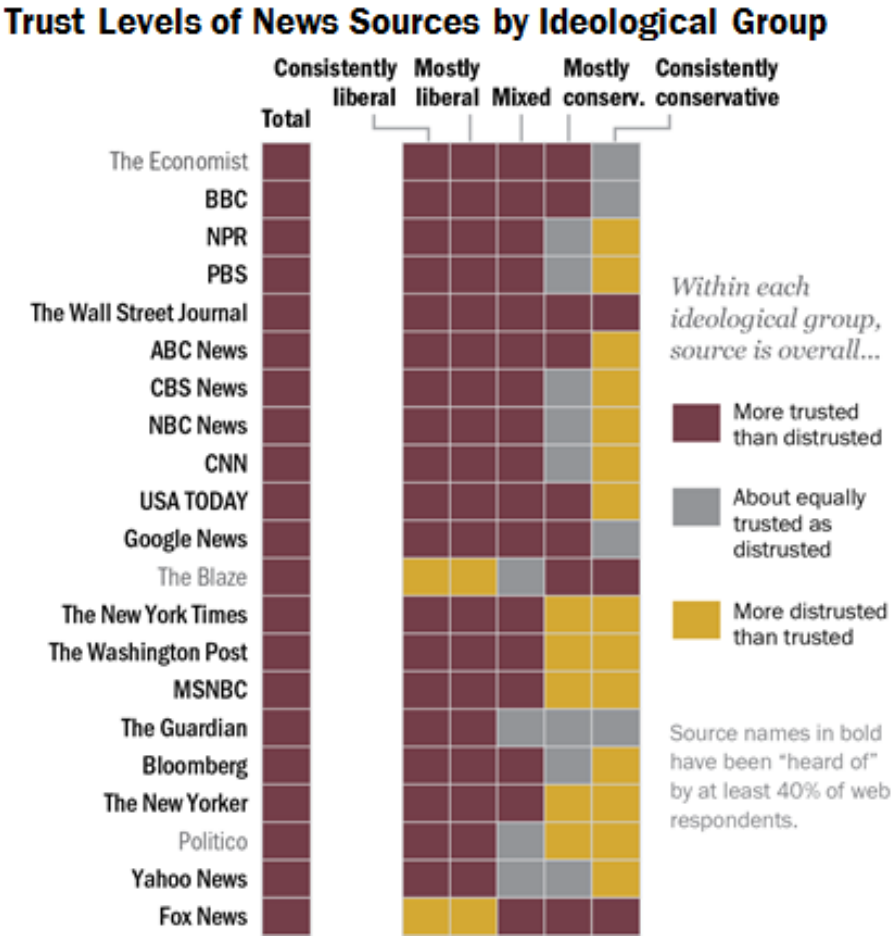
Pozyskanie korpusu: źródła

Wiarygodność dokumentów oceniona na poziomie źródła (domeny) przez ekspertów:

- wiarygodne: 21 mediów, dla którym ufających jest więcej niż nieufających w populacji USA wg sondażu *Pew Research Center*: „*Political Polarization & Media Habits*" [Mitchell et al., 2014],
- niewiarygodne: 241 stron oznaczonych jako *fake news* lub *impostor* by *PolitiFact* w ramach działalności fact-checkingowej [Gillin, 2017] w 2017 roku (1/4 wciąż aktywna w 2019).

Po filtrowaniu zostało 18 i 205, odpowiednio.

Pozyskanie korpusu: źródła



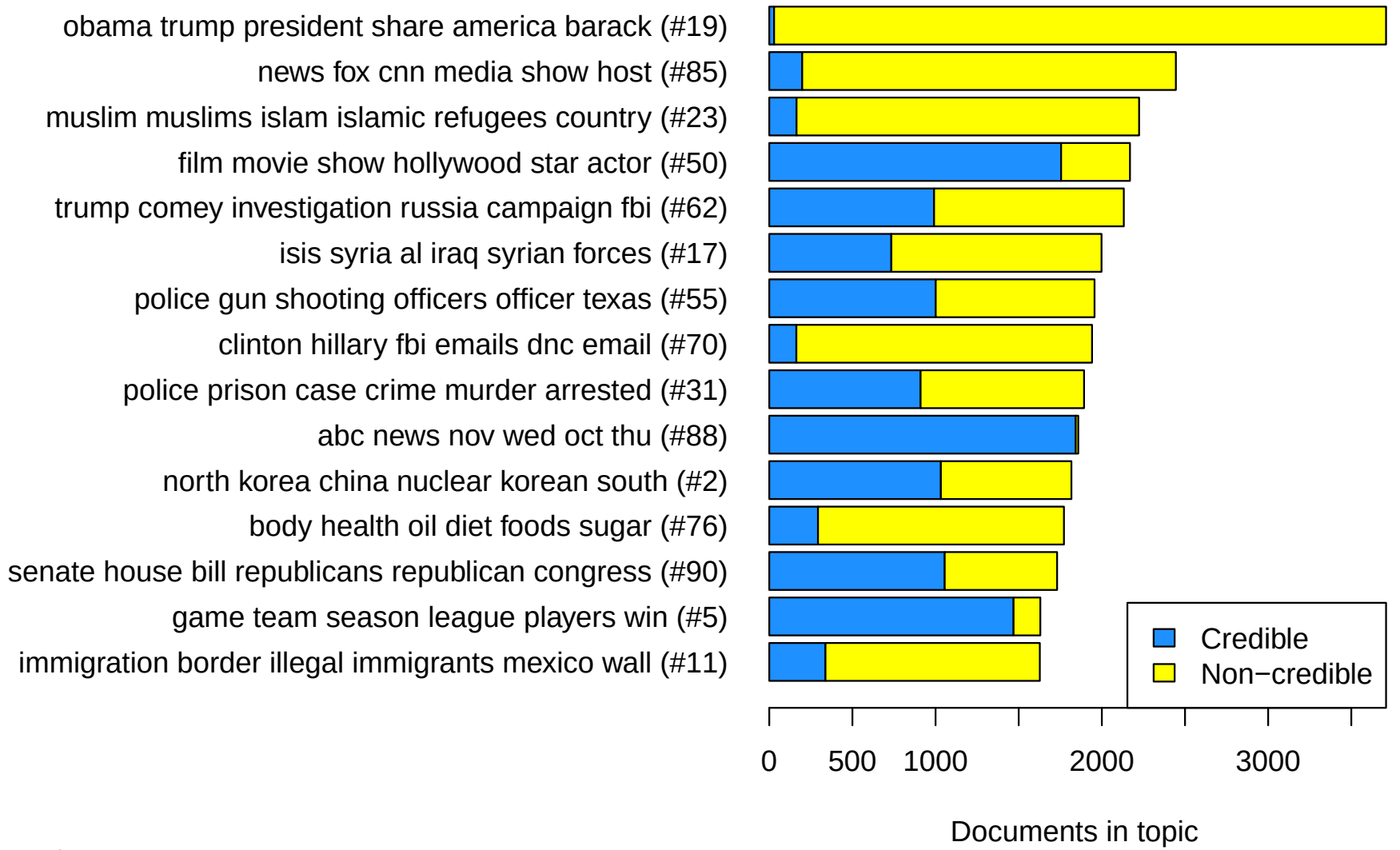
Źródło: `https://www.journalism.org/2014/10/21/political-polarization-media-habits/`

Pozyskanie korpusu: przetwarzanie

1. Pobranie kopii stron z *Web Archive* pomiędzy 01.11.2017 a 09.11.2017 (aktualizacja listy),
2. Pobranie podstron przed śledzenie linków, w ramach domeny i roku 2017, do głębokości 5 linków i 10.000 odwiedzonych na domenę,
3. Wyodrębnienie treści przez prostą heurystykę,
4. Usunięcie tekstów zduplikowanych i o średniej długości linii mniejszej niż 15 słów,
5. Tokenizacja, tagowanie POS przez *Stanford CoreNLP* [Manning et al., 2014] oraz wydobycie tematów LDA przez *Mallet* [McCallum, 2002].

W rezultacie otrzymujemy 103.219 dokumentów (50.429 wiarygodnych i 52.790 niewiarygodnych) z 117M tokenów.

Korpus: tematy

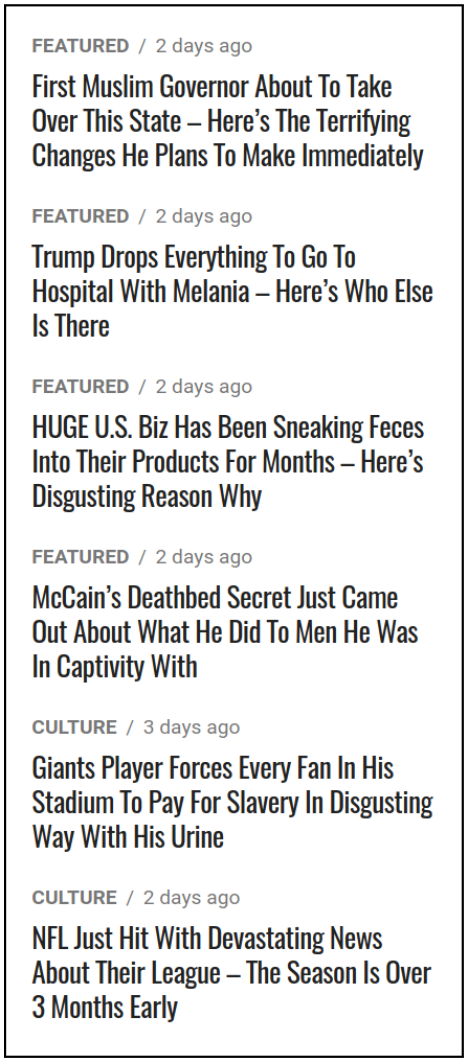


Klasyfikacja

Klasyfikator stylometryczny: cechy

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.



FEATURED / 2 days ago

First Muslim Governor About To Take Over This State – Here's The Terrifying Changes He Plans To Make Immediately

FEATURED / 2 days ago

Trump Drops Everything To Go To Hospital With Melania – Here's Who Else Is There

FEATURED / 2 days ago

HUGE U.S. Biz Has Been Sneaking Feces Into Their Products For Months – Here's Disgusting Reason Why

FEATURED / 2 days ago

McCain's Deathbed Secret Just Came Out About What He Did To Men He Was In Captivity With

CULTURE / 3 days ago

Giants Player Forces Every Fan In His Stadium To Pay For Slavery In Disgusting Way With His Urine

CULTURE / 2 days ago

NFL Just Hit With Devastating News About Their League – The Season Is Over 3 Months Early

Źródło: `https://web.archive.org/web/20171104091151/https://freedomdaily.com/`

Klasyfikator stylometryczny: słownik GI

General Inquirer [Stone et al., 1962] – agregat kilku ręcznie skompilowanych zbiorów słów sklasyfikowanych wg. tematyki, wydźwięku, funkcji itp. 8.640 różnych słów, ale dużo małych kategorii, np. *aquatic*: 20 słów

Procedura rozszerzenia kategorii $\mathcal{C}_i$:

1. Definiowanie zadania klasyfikacji na słowach $w \in V$, gdzie $x_w = \text{word2vec}(w)$ i $y_w = \mathbf{1}[w \in \mathcal{C}]$.
2. Budowa modelu regresji logistycznej M_i estymującego $P(y = 1|x)$,
3. Uporządkowanie słów według malejącej wartości $\hat{y}_w = M_i(x_w)$,
4. Dodanie $4 \times |\mathcal{C}_i|$ nowych słów do $\mathcal{C}_i$.

$\rightarrow$ nowy słownik 34.293 słów.

Klasyfikator stylometryczny: słownik GI

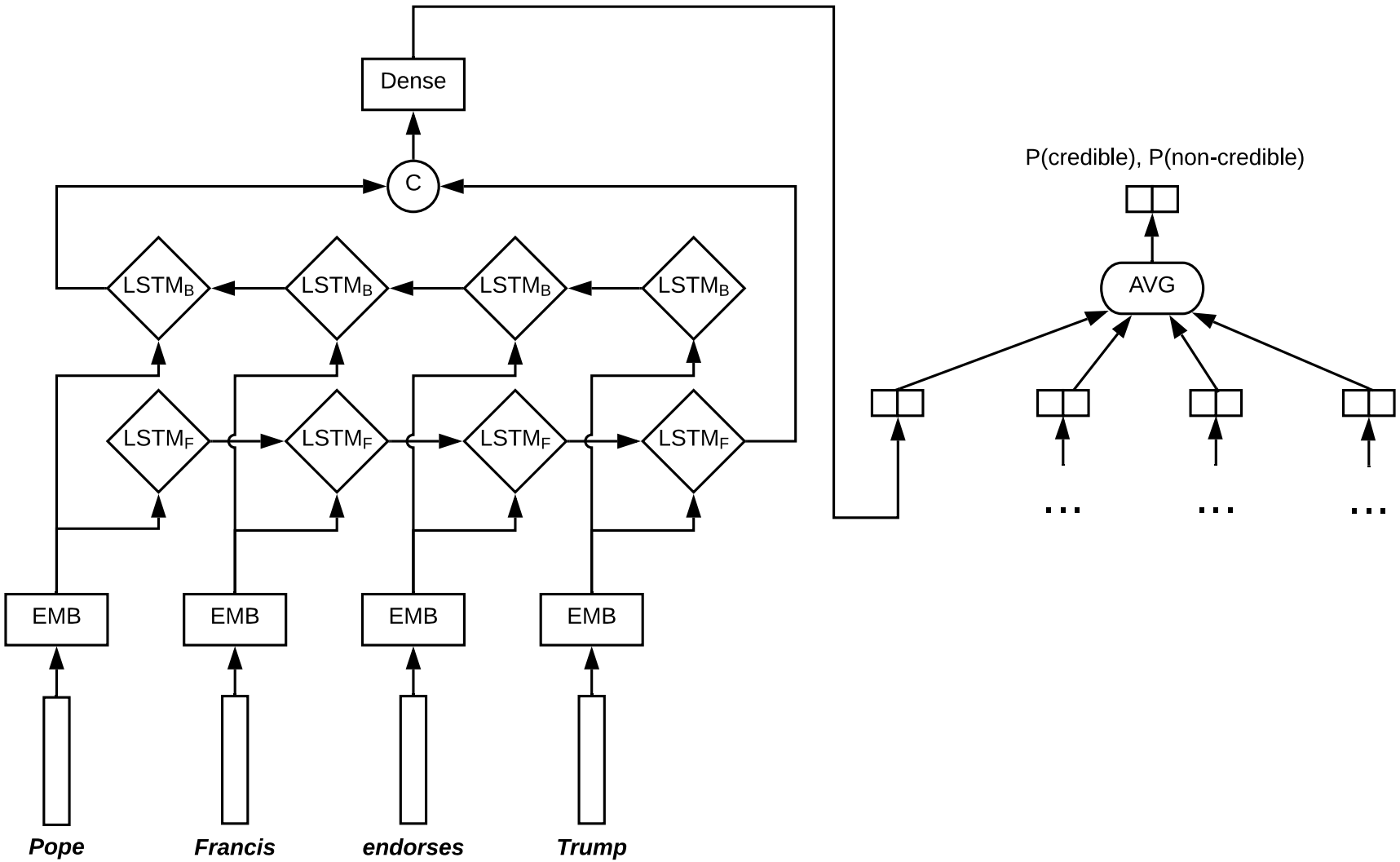
aquatic={***water***, ***channel***, ***pool***, ***sea***, ***river***, ***wave***, ***flood***, ***rapid***, ***lake***, ***stream***, ***ocean***, *bay*, *wash*, *tide*, *harbor*, *creek*, *foam*, *gulf*, *swamp*, *breaker*, *spillway*, *tributary*, *floodwaters*, *rivers*, *waterway*, *floodwater*, *cfs*, *pond*, *lagoon*, *canal*, *weir*, *riverbed*, *waters*, *tides*, *floodway*, *Ganges*, *River*, *riptide*, *dike*, *estuary*, *Danube*, *Yamuna*, *rapids*, *Floodwaters*, *creeks*, *rivulet*, *reservoir*, *seawater*, *tributaries*, *breakwater*, *nullah*, *moat*, *MRGO*, *cusecs*, *whitewater*, *floodgate*, *tailwater*, *Yangtze*, *spillways*, *marsh*, *brook*, *marshes*, *aqueduct*, *watercourse*, *Mommyologist*, *Seine*, *dam*, *Dam*, *basin*, *flume*, *Aquifer*, *sinkhole*, *seas*, *waterfall*, *levee*, *eddies*, *Yalu*, *cesspool*, *Buriganga*, *watercourses*}

Klasyfikator stylometryczny: wybór cech

Dane zawierają 103.219 przypadków opisanych przez 39.235 cech.
Budowa modelu:

1. Sprawdzenie aktywności j-tej cechy $b_{i,j} = \mathbf{1}[x_{i,j} \neq 0]$,
2. Usunięcie j-tej cechy jeśli $|\mathrm{cor}(\mathbf{b}_j, \mathbf{y})| \leqslant 0.05$
   (zamiast $\mathrm{mean}(\mathbf{b}_j) < 0.1$)
3. Budowa modelu regresji logistycznej z regularyzacją L1 (`glmnet`).

Klasyfikator neuronowy: BiLSTMAvg



Klasyfikator neuronowy: trenowanie

Budowa modelu w *TensorFlow* 1.14, przy czym:

- osadzenia 300-wymiarowe dla 282.871 słów (z *word2vec* na *Google News Corpus* [Mikolov et al., 2013]),
- parametry osadzeń ustalone, trenowane tylko warstwy LSTM 100-wymiarowe (2*160.400 parametrów) i gęsta (402 parametry),
- brak regularyzacji,
- maksymalna długość zdania: 120 tokenów, maksymalna długość dokumentu: 50 zdań.
- optymalizator: *Adam*, logistyczna funkcja straty,
- trenowanie przez 10 epok.

Ewaluacja

Baseline

Dla perspektywy dwa klasyfikatory tekstu ogólnego przeznaczenia, podobne do wcześniejszych prac z oceną wiarygodności:

- **Bag of words**: klasyfikacja jak w modelu stylometrycznym, ale cechy to częstości unigramów, bigramów i trigramów form podstawowych,
- **BERT**: model pretrenowany BERT [Devlin et al., 2018] użyty w scenariuszu dostrajania do klasyfikacji (liniowa predykcja wg. wyjścia elementu `[CLS]`).

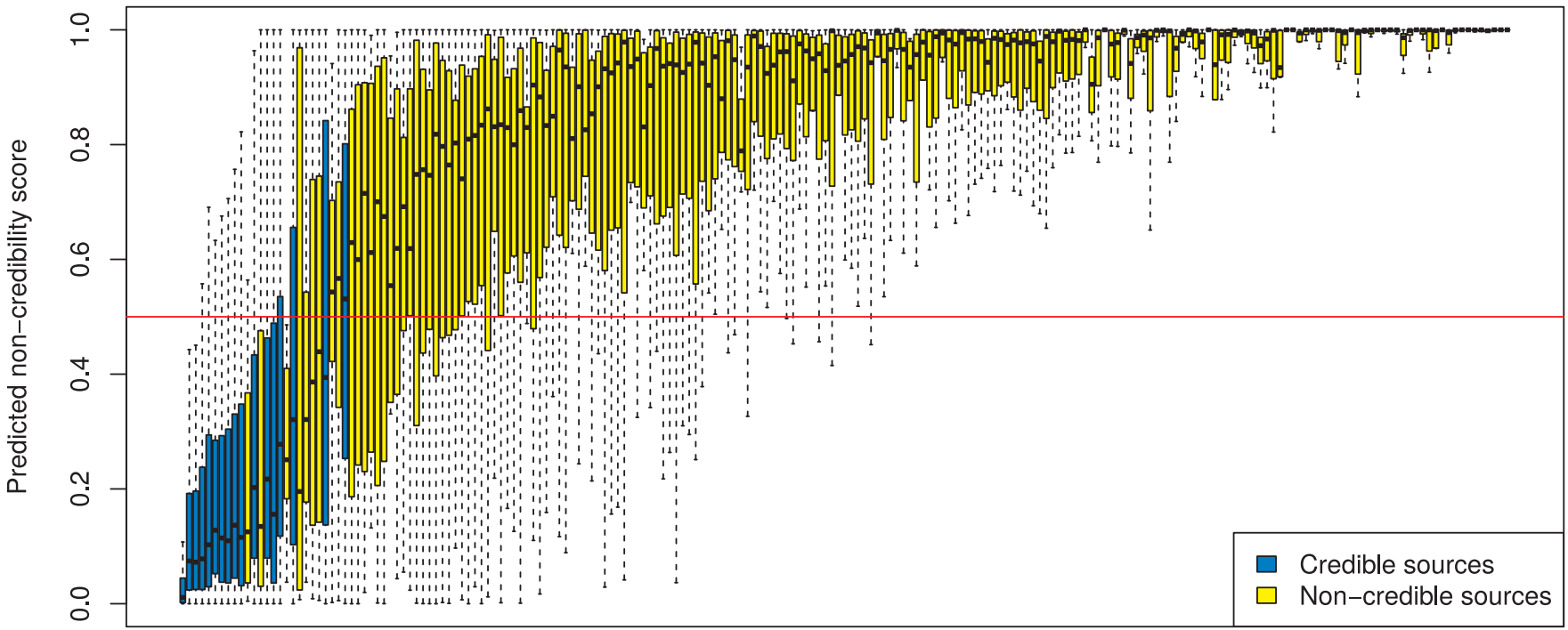
Wyniki: źródła

Scenariusze walidacji krzyżowej:

- document CV – losowy podział na dane treningowe i testowe,
- topic CV – dokumenty testowe z innego tematu (np. nowe wydarzenie),
- source CV – dokumenty testowe z innego źródła (np. nowa strona FN).

Metoda	**document CV**	**topic CV**	**source CV**
Stylometryczna	0.9274	0.9173	0.8097
BiLSTMAvg	0.8994	0.8921	0.8250
Bag of words	0.9913	0.9886	0.7078
BERT	0.9976	0.9965	0.7960

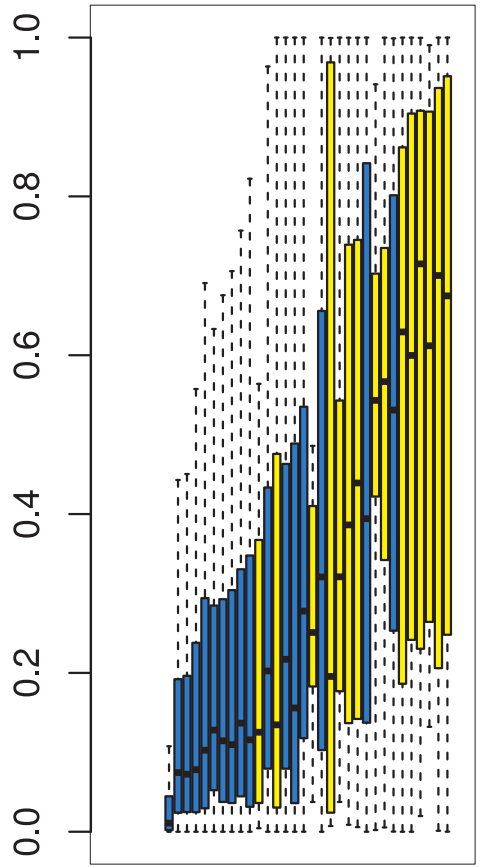
Wyniki: źródła



Źródła graniczne

Źródła, których styl nie pasuje do kategorii:

- wiarygodne:
  - *Fox News* – najmniej zaufane wg. sondażu,
  - *TheBlaze* – najmniej znane wg. sondażu,
- niewiarygodne:
  - *The Times Mexico – imposter site*,
    - *Leaked Audio: Mexican President Agrees to Pay For Wall* (własne)
    - *Snapchat's Physical Footprint Reveals Core Priority of the Brand* (z *Yahoo Finance*)
  - *Before It's News – citizen journalism*,
    - *Worker Says Nanobot Mosquito Killed Woman!*
    - *Mammals Began Daytime Activity After Dinosaur Extinction*



Wyniki: istotne słowa

Trump Just Bombarded These 3 Dirty Dems With MAJOR Surprise That Will Shut Them Up For Good

The liberal snowflake meltdown over the past couple of days following Trump firing FBI Director Comey has been nothing short of hilarious to witness. The very same people who were screaming at the top of their freaking lungs and begging Barack Hussein Obama to fire Comey are now the very same idiots feigning outrage after Trump finally decided to take out the trash. As these morons are now labeling Trump a fascist and even calling for his impeachment over Comey's dismissal, President Trump had finally had enough. Calling out their hypocrisy in a way that only Trump can do, Chuck Schumer, Nancy Pelosi, and Maxine Waters woke up to a nasty surprise this morning that immediately sent the trio of morons flying back into their land of unicorns and rainbows where they belong.

Democrats are the biggest hypocrites on the face of the planet and wasted no time proving that to the world once again following Comey's firing. The usual clowns were out in force spinning the story and pushing their fake and cringeworthy outrage, as Chuck Schumer , Nancy Pelosi, and Maxine Waters lost their minds on all the liberal networks, trying to claim that Trump was trying to get rid of Comey because he has some sort of bombshell evidence about a Russian "collusion."

But Trump immediately decided to set the record straight on Comey and pushed out an epic two-minute video where he tweeted out the truth so Americans could see just how full of crap and hypocrisy liberals truly are, along with the hashtag #DrainTheSwamp.

The Democrats should be ashamed. This is a disgrace! #DrainTheSwamp pic.twitter.com/UfbKEECm2V
— Donald J. Trump (@realDonaldTrump)

Źródło: `http://freedomdaily.com/trump-just-bombarded-3-dirty-dems-major-surprise-will-shut-good/`

Zakończenie

Ograniczenia

- klasyfikacja tekstu w izolacji pozwala oszacować **wiarygodność**, ale nie wykryć **fake news**,
- etykietowanie na poziomie źródeł ignoruje różnice w wiarygodności dokumentów z tego samego źródła (p. źródła graniczne),
- proponowane metody wciąż mają gorszą wydajność dla dokumentów z nieznanych źródeł,
- prosta metoda wydobywania treści z HTML przepuszcza elementy strony zdradzające źródło,
- dobre rezultaty metody neuronowej wskazują, że zasługuje na większą uwagę, np. pod kątem interpretowalności,
- nieznana jest dokładność ręcznej oceny wiarygodności,
- zaobserwowane zależności stylistyczne mogą nie dotyczyć dezinformacji innego pochodzenia (np. motywowanej politycznie, a nie finansowo) lub rodzaju (np. generowanej maszynowo).

Zasoby

- News Style Corpus,
  - lista stron,
  - kod do ściągania kopii z *Web Archive* i konwersji do tekstu,
  - podziały walidacji krzyżowej do powtórzenia ewaluacji,
- Klasyfikator stylometryczny
  - rozszerzony słownik GI,
  - kod do generowania cech,
  - kod w *R* do budowania i ewaluacji modeli,
- Klasyfikator neuronowy BiLSTMAvg
  - implementacja w *TensorFlow*.

udostępnione pod `github.com/piotrmp/fakestyle`.

Projekt

Projekt HOMADOS (*Hampering Misinformation by Assessing Credibility of Online Sources*):

- obejmuje różne podejścia do problemów zarządzania wiarygodnością dokumentów online:
  - wykrywanie niewiarygodnych treści,
  - wykrywanie niewiarygodnych użytkowników (boty),
  - ocena wiarygodności stwierdzeń,
  - promowanie wiarygodnych treści,
- finansowany jest z:
  - programu *Polskie Powroty* Narodowej Agencji Wymiany Akademickiej (NAWA), grant nr PPN/PPO/2018/1/00006,
  - grantów obliczeniowych udzielonych przez *Google Cloud* i *Poznańskie Centrum Superkomputerowo-Sieciowe*.
- poszukuje pracowników! więcej: `piotr.przybyla@ipipan.waw.pl`

Wszystko o projekcie: `homados.ipipan.waw.pl`

Podsumowanie

Czy można ocenić wiarygodność tekstu na podstawie jego własności stylistycznych?
→ Tak, z dokładnością rzędu 80-90%

Czy wystarczy zastosować zwykły klasyfikator tekstu, sprawdzony w innych zadaniach?
→ Nie, gdyż nauczy się on nie tego, co nas interesuje.

Czy klasyfikator stylometryczny rzeczywiście opiera się na słownictwie w swoich decyzjach?
→ Tak, choć bierze też pod uwagę schematy składniowe, trudniejsze do interpretacji.

Czy sieci neuronowe oferują większą dokładność niż ręcznie przygotowane cechy?
→ Tak, choć za cenę niższej interpretowalności.

Dziękuję!

Literatura (1)

Axel Gelfert. Fake News: A Definition. *Informal Logic*, 38(1):84–117, mar 2018. ISSN 0824-2577. doi: 10.22329/il.v38i1.5068. URL https://ojs.uwindsor.ca/index.php/informal{_}logic/article/view/5068.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240. Association for Computational Linguistics, 2018. URL https://www.aclweb.org/anthology/P18-1022.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending Against Neural Fake News. may 2019. URL http://arxiv.org/abs/1905.12616.

Clint Burfoot and Timothy Baldwin. Automatic satire detection: are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164. Association for Computational Linguistics, 2009. URL https://www.aclweb.org/anthology/papers/P/P09/P09-2041/.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. *WWW '15 Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405, 2015. ISSN 01635999. doi: 10.1145/2736277.2741637.

Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology*, 52(1), 2015. ISSN 23739231. doi: 10.1002/pra2.2015.145052010081.

Hal Berghel. Lies, Damn Lies, and Fake News. *Computer*, 50(2):80–85, feb 2017. ISSN 0018-9162. doi: 10.1109/MC.2017.56.

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.

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.

P. Sol Hart and Erik C. Nisbet. Boomerang Effects in Science Communication. *Communication Research*, 39(6): 701–723, dec 2012. ISSN 0093-6502. doi: 10.1177/0093650211416646. URL http://journals.sagepub.com/doi/10.1177/0093650211416646.

Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L. Freed. Effective messages in vaccine promotion: A randomized trial. *Pediatrics*, 133(4):e835–e842, apr 2014. ISSN 10984275. doi: 10.1542/peds.2013-2365.

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.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational Fact Checking from Knowledge Networks. *PLOS ONE*, 10(6):e0128193, jun 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0128193.

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.

Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.211.

Literatura (3)

Victoria L Rubin, Niall J Conroy, and Yimin Chen. Towards news verification: Deception detection methods for news discourse. *Proceedings of the Hawaii International Conference on System Sciences*, 2015. doi: 10.13140/2.1.4822.8166.

Benjamin D. Horne and Sibel Adali. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. In *Proceedings of the 2nd International Workshop on News and Public Opinion at ICWSM*. Association for the Advancement of Artificial Intelligence, 2017. URL http://arxiv.org/abs/1703.09398.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic Detection of Fake News. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, 2018. URL https://aclanthology.info/papers/C18-1287/c18-1287.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1317.

Brian Ray. *Style: An Introduction to History, Theory, Research, and Pedagogy*. Parlor Press, The WAC Clearinghouse, 2015. ISBN 9781602356146.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender , Genre , and Writing Style in Formal Written Texts. *Text*, 23(3):321–346, 2003.

Efstathios Stamatatos. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60:538–556, 2009. ISSN 15322882. doi: 10.1002/asi.21001.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412, nov 2002. ISSN 0268-1145. doi: 10.1093/llc/17.4.401. URL http://llc.oxfordjournals.org/content/17/4/401.abstract.

Literatura (4)

Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. Language and Ideology in Congress. *British Journal of Political Science*, 42(01):31–55, may 2011. ISSN 0007-1234. URL http://journals.cambridge.org/abstract{_}S0007123411000160.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Mining the Blogosphere: Age, gender and the varieties of self-expression. In *First Monday*, volume 12, sep 2007. URL http://ojphi.org/ojs/index.php/fm/article/view/2003/1878.

Piotr Przybyła and Paweł Teisseyre. Analysing Utterances in Polish Parliament to Predict Speaker's Background. *Journal of Quantitative Linguistics*, 21(4):350–376, 2014. ISSN 0929-6174. doi: 10.1080/09296174.2014.944330. URL http://www.tandfonline.com/doi/abs/10.1080/09296174.2014.944330.

Piotr Przybyła and Paweł Teisseyre. What do your look-alikes say about you? Exploiting strong and weak similarities for author profiling - Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Labs and Workshops, Notebook Papers*, Toulouse, France, 2015. CEUR-WS.org. URL https://www.uni-weimar.de/medien/webis/events/pan-15/pan15-papers-final/pan15-author-profiling/przybyla15-notebook.pdf.

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/politifacts-guide-fake-news-websites-and-what-they/.

Literatura (5)

Christopher D. Manning, John Bauer, Jenny Finkel, Steven J. Bethard, Mihai Surdeanu, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2014. ISBN 9781941643006. URL `http://aclweb.org/anthology/P14-5010`.

Andrew Kachites McCallum. MALLET: A machine learning for language toolkit, 2002. URL `http://mallet.cs.umass.edu`.

Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498, 1962. doi: 10.1002/bs.3830070412.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2018. URL `http://arxiv.org/abs/1810.04805`.