Nie ma uniwersalnej definicji **fake news** - nawet w dziennikarstwie. Poniższa tabela porządkuje różne inne powiązane terminy, często współwystępujące z ideą fake news. Wyszczególniono tu trzy kluczowe cechy:

- **autentyczność**: posiadanie jakiegokolwiek stwierdzenia nie opartego na faktach,
- **intencja**: czy celem jest wprowadzanie w błąd czy też rozrywka,
- **wiadomości**: czy informacja to wiadomości (*news*).

Table 1. Comparison Between Concepts Related to Fake News

Concept	Authenticity	Intention	News?
Deceptive news	Non-factual	Mislead	Yes
False news	Non-factual	Undefined	Yes
Satire news	Non-unified[2]	Entertain	Yes
Disinformation	Non-factual	Mislead	Undefined
Misinformation	Non-factual	Undefined	Undefined
Cherry-picking	Commonly factual	Mislead	Undefined
Clickbait	Undefined	Mislead	Undefined
Rumor	Undefined	Undefined	Undefined

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

W literaturze można spotkac różne próby dookreślenia fake news, np.

A news article that is intentionally and verifiably false

A news article or message published and propagated through media, carrying false information regardless of the means and motives behind it

przy czym pokrywają się one częsciowo z opisem `false news` oraz *disinformation*.

Szeroka definicja fake news

Fake news to fałszywe wiadomości (*false news*).

Wąska definicja fake news

Fake news to intencjonalnie fałszywe wiadomości (*false news*) obpulikowane przez portal informacyjny (*news outlet*).

Podstawowe teorie (głownie związane z naukami społecznymi oraz ekonomią) stanowią bezcenne źródło odniesienia w celu opisu oraz modelowania zjawiska fake news. Można wyróżnić tu dwie grupy: jedna odnosi się do **tekstu**, druga do **użytkowników**.

Table 2. Fundamental Theories in Social Sciences (Including Psychology and Philosophy) and Economics

		Theory	Phenomenon
News-Related Theories		Undeutsch hypothesis [Undeutsch 1967]	A statement based on a factual experience differs in content style and quality from that offantasy.
		Reality monitoring [Johnson and Raye 1981]	Actual events are characterized by higher levels of sensory-perceptual information.
		Four-factor theory [Zuckerman et al. 1981]	Lies are expressed differently in terms of arousal, behavior control, emotion, and thinking from truth.
		Information manipulation theory [McCornack et al. 2014]	Extreme information quantity often exists in deception.
...gements and Roles in Fake News Activities)	Social Impacts	Conservatism bias [Basu 1997]	The tendency to revise one's belief insufficiently when presented with new evidence.
		Semmelweis reflex [Bálint and Bálint 2009]	Individuals tend to reject new evidence because it contradicts with established norms and beliefs.
		Echo chamber effect [Jamieson and Cappella 2008]	Beliefs are amplified or reinforced by communication and repetition within a closed system.
		Attentional bias [MacLeod et al. 1986]	An individual's perception is affected by his or her recurring thoughts atthe time.
		Validity effect [Boehm 1994]	Individuals tend to believe information is correct after repeated exposures.
		Bandwagon effect [Leibenstein 1950]	Individuals do something primarily because others are doing it.
		Normative influence theory [Deutsch and Gerard 1955]	The influence of others leading us to conform to be liked and accepted by them.
		Social identity theory [Ashforth and Mael 1989]	An individual's self-concept derives from perceived membership in a relevant social group.
		Availability cascade [Kuran and Sunstein 1999]	Individuals tend to adopt insights expressed by others when such insights are gaining more popularity within their social circles
User-Related Theories (User's Engage...	Self-Impact	Confirmation bias [Nickerson 1998]	Individuals tend to trust information that confirms their preexisting beliefs or hypotheses.
		Selective exposure [Freedman and Sears 1965]	Individuals prefer information that confirms their preexisting attitudes.
		Desirability bias [Fisher 1993]	Individuals are inclined to accept information that pleases them.
		Illusion of asymmetric insight [Pronin et al. 2001]	Individuals perceive their knowledge to surpass that of others.
		Naive realism [Ward et al. 1997]	The senses provide us with direct awareness of objects as they really are.
		Overconfidence effect [Dunning et al. 1990]	A person's subjective confidence in his judgments is reliably greater than the objective ones.
	Benefits	Prospecttheory [Kahneman and Tversky 2013]	People make decisions based on the value oflosses and gains rather than the outcome.
		Contrast effect [Hovland et al. 1957]	The enhancement or diminishment of cognition due to successive or simultaneous exposure to a stimulus oflesser or greater value in the same dimension.
		Valence effect [Frijda 1986]	People tend to overestimate the likelihood of good things happening rather than bad things.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Można też stworzyć pewną prosta typologię , która odnosi się do sposobu (metody) wykrywania fake news

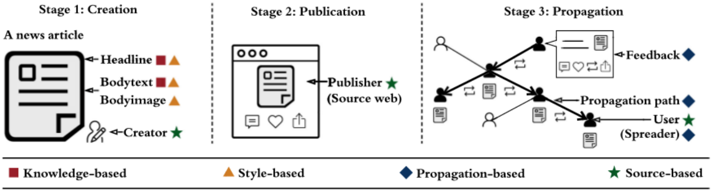


Fig. 1. Fake news life cycle and connections to the four fake news detection perspectives in this survey.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Knowledege-based

Oparte na wiedzy: sprawdzamy, czy zawartość wiadomości jest oparta na faktach

Propagation-based

Oparte na sposobie propagacji: to, jak się rozchodzi, może świadczyć o fake news.

Style-based

Oparte na stylu - np. zakładamy, że fake news mogą zawierać b. silne emocje.

Source-based

Oparte na wiarygodności zródeł, które zapoczątkowały propagację

W przypadku metod opartych na widzy mamy kilka mozliwości: (i) fakty są sprawdzane **manualnie** (ii) dysponujemy **automatyczną** detekcją. W tym pierwszym przypadku najczęście mamy do czynienia albo z podejściem **eksperckim** albo wykorzystaniem **crowd-sourcingu**.

(a) (Expert-based) PolitiFact: the PolitiFact scorecard

(b) (Crowd-sourced) Fiskkit: the tag distribution

Fig. 2. Illustrations of manual fact-checking websites.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

W ciągu ostatnich lat pojawiło się wiele serwisów (stron WWW), sprawdzających wiarygodność wiadomości. Poniższa tabela opisuje najważniejsze z nich

Table 3. Comparison Among Expert-Based Fact-Checking Websites

Website	Topics Covered	Content Analyzed	Assessment Labels
PolitiFact[3]	American politics	Statements	True; Mostly true; Half true; Mostly false; False; Pants on fire
The Washington Post Fact Checker[4]	American politics	Statements and claims	One pinocchio; Two pinocchio; Three pinocchio; Four pinocchio; The Geppetto checkmark; An upside-down Pinocchio; Verdict pending
FactCheck[5]	American politics	TV ads, debates, speeches, interviews, and news	True; No evidence; False
Snopes[6]	Politics and other social and topical issues	News articles and videos	True; Mostly true; Mixture; Mostly false; False; Unproven; Outdated; Miscaptioned; Correct attribution; Misattributed; Scam; Legend
TruthOrFiction[7]	Politics, religion, nature, aviation, food, medical, etc.	Email rumors	Truth; Fiction; etc.
FullFact[8]	Economy, health, education, crime, immigration, law	Articles	Ambiguity (no clear labels)
HoaxSlayer[9]	Ambiguity	Articles and messages	Hoaxes, scams, malware, bogus warning, fake news, misleading, true, humor, spams, etc.
GossipCop[10]	Hollywood and celebrities	Articles	0–10 scale, where 0 indicates completely fake news and 10 indicates completely true news

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Oczywiście, metody manualne w żaden sposób nie skalują się z liczbą wiadomości (szczególnie w przypadku mediów społecznościowych). Aby obejść ten problem stosuje się automatyczną detekcję fake news, korzystająca z metod IR (*information retrieval*), NLP oraz ML (*machine learning*)

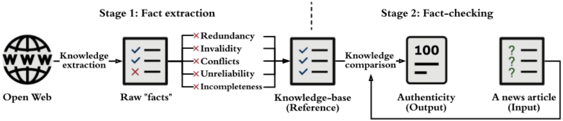


Fig. 3. Automatic news fact-checking process.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Wiedza

To zbiór trójelementowych krotek `(Podmiot, Orzeczenie, Dopełnienie)` uzyskanych z danej informacji.

Np. zdanie *Donald Trump jest prezydentem USA* może zostać zapisane jako `(Donald Trump, Zawód, Prezydent)`

Fakt

To wiedza (tzn. krotka `POD`) zweryfikowana jako prawda.

W odróżnieniu od metod opartych na wiedzy, które bazując na ocenie faktów przedstawianych w wiadomości, metody oparte na stylu są w stanie ocenić **intencję wiadomości** (tzn. czy ma ona za zadanie wprowadzić kogoś w błąd czy nie). Ogólna intuicja sprowadza się do założenia, że osoby, które tworzą fake news, tworzą je w **specjalnym stylu**, który ma w zamyśle zachęcać do ich przeczytania i uznania zawartości za prawdę.

W kwestii metodologii, rozpatrujemy tu dwie dobrze znane nam grupy cech:

- ogólne cechy (*features*) tekstowe, takie jak częstotliwość wyrazów (uzyskiwana np. za pomoca BOW) czy POS; głebsze struktury zdaniowe mogą być określane przy pomocy PCFG (*probabilistic context-free grammars*); struktury retoryczne daje nam RST (*rethorical structure theory*) a część semantyczną ocenimy korzystając z LIWC (*linguistic inquiry and word count*)
- cechy latentne (ukryte) - tu opieramy się na omówionych na poprzednich wykładach metodach zanurzeniowych

Poniżej tabela z różnymi wykorzystywanymi cechami:

Table 4. Semantic-Level Features in News Content

Attribute Type	Feature	[Zhou et al. 2004b]	[Fuller et al. 2009]	[Afroz et al. 2012]	[Siering et al. 2016]	[Zhang et al. 2016]	[Bond et al. 2017]	[Potthast et al. 2017]	[Perez-Rosas et al. 2017]	[Zhou et al. 2019a]
Quantity	# Characters			✓						✓
	# Words	✓	✓	✓	✓	✓				✓
	# Noun phrases	✓								
	# Sentences	✓	✓	✓	✓					✓
	# Paragraphs							✓		✓
Complexity	Average # characters per word	✓	✓	✓	✓					✓
	Average # words per sentence	✓	✓	✓	✓	✓				✓
	Average # clauses per sentence	✓			✓					
	Average # punctuations per sentence	✓	✓	✓	✓					
Uncertainty	#/% Modal verbs (e.g., "shall")	✓	✓	✓	✓					
	#/% Certainty terms (e.g., "never" and "always")	✓	✓	✓	✓		✓			✓
	#/% Generalizing terms (e.g., "generally" and "all")		✓							
	#/% Tentative terms (e.g., "probably")		✓	✓			✓			✓
	#/% Numbers and quantifiers			✓						
	#/% Question marks			✓						✓
Subjectivity	#/% Biased lexicons (e.g., "attack")									✓
	#/% Subjective verbs (e.g., "feel" and "believe")	✓				✓				
	#/% Report verbs (e.g., "announce")									✓
	#/% Factive verbs (e.g., "observe")									✓
Non-immediacy	#/% Passive voice	✓	✓							
	#/% Self reference: 1st person singular pronouns	✓	✓	✓	✓	✓				
	#/% Group reference: 1st person plural pronouns	✓	✓	✓	✓	✓				
	#/% Other reference: 2nd and 3rd person pronouns	✓	✓	✓	✓	✓				
	#/% Quotations			✓				✓		
Sentiment	#/% Positive words	✓	✓	✓	✓	✓	✓			✓
	#/% Negative words	✓	✓	✓	✓	✓	✓			✓
	#/% Anxiety/angry/sadness words						✓			✓
	#/% Exclamation marks			✓						✓
	Content sentiment polarity									✓
Diversity	Lexical diversity: #/% unique words or terms	✓	✓	✓	✓	✓				✓
	Content word diversity: #/% unique content words	✓	✓			✓				✓
	Redundancy: #/% unique function words	✓	✓	✓		✓				✓
	#/% Unique nouns/verbs/adjectives/adverbs									✓
Informality	#/% Typos (misspelled words)	✓			✓	✓				
	#/% Swear words/netspeak/assent/nonfluencies/fillers									✓
Specificity	Temporal/spatial ratio	✓	✓				✓			
	Sensory ratio	✓	✓		✓		✓			✓
	Causation terms		✓				✓			✓
	Exclusive terms		✓							
Readablity (e.g., Flesch-Kincaid and Gunning-Fog index)				✓				✓	✓	✓

The studies labeled with gray background color investigate news articles.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Jeśli chodzi o metody związane ze stylem, to możemy korzystać tu za całej gamy podejść ML od nienadzorowanych, półnadzorowanych, nadzorowanych, jak również ogólnie DL (*deep learning*).

Table 6. Feature Performance (Accuracy (Acc.) and F_1-score) in Fake News Detection Using Traditional ML (RF and XGBoost Classifiers) [Zhou et al. 2019a]

Feature Group			PolitiFact data [Shu et al. 2018]				BuzzFeed data [Shu et al. 2018]			
			XGBoost		RF		XGBoost		RF	
			Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Non-Latent Features	Lexicon	BOWs (f_s)	0.856	0.858	0.837	0.836	0.823	0.823	0.815	0.815
		Unigram+bigram (f_r)	0.755	0.756	0.754	0.755	0.721	0.711	0.735	0.723
	Syntax	POS tags (f_s)	0.755	0.755	0.776	0.776	0.745	0.745	0.732	0.732
		Rewrite rules ($r*$, f_s)	**0.877**	**0.877**	0.836	0.836	0.778	0.778	0.845	0.845
		Rewrite rules ($r*$, f_r)	0.749	0.753	0.743	0.748	0.735	0.738	0.732	0.735
	Semantic	LIWC	0.645	0.649	0.645	0.647	0.655	0.655	0.663	0.659
		Theory-driven [Zhou et al. 2019a]	0.745	0.748	0.737	0.737	0.722	0.750	0.789	0.789
	Discourse	Rhetorical relationships	0.621	0.621	0.633	0.633	0.658	0.658	0.665	0.665
	Combination	[Zhou et al. 2019a]	0.865	0.865	**0.845**	**0.845**	**0.855**	**0.856**	**0.854**	**0.854**
Latent Features		Word2Vec [Mikolov et al. 2013]	0.688	0.671	0.663	0.667	0.703	0.714	0.722	0.718
		Doc2Vec [Le and Mikolov 2014]	0.698	0.684	0.712	0.698	0.615	0.610	0.620	0.615

Results show that (i) non-latent features can outperform latent ones, (ii) combining features across levels can outperform using single-level features, and (iii) the (standardized) frequencies of lexicons and rewrite rules better represent fake news content style and perform better (while being more time consuming to compute) than other feature groups.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Z drugiej strony, czasem istotniejsze są wzorce (*patterns*), które mogą świadczyć o tym, że fake news cechują się określonymi wyróznikami

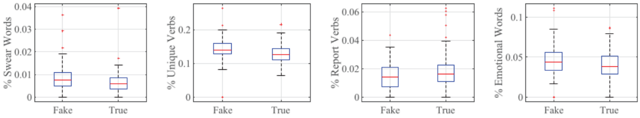


Fig. 7. Fake news textual patterns [Zhou et al. 2019a] (PolitiFact, data is from FakeNewsNet [Shu et al. 2018]). Compared to true news text, fake news text has (i) higher informality (% swear words), (ii) diversity (% unique verbs), and (iii) subjectivity (% report verbs), and is (iv) more emotional (% emotional words).

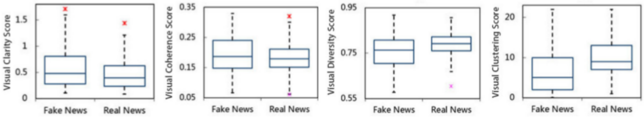


Fig. 8. Fake news visual patterns (Twitter+Weibo, directly from Jin et al. [2017]). Compared to true news images, fake news images often have higher clarity and coherence but lower diversity and clustering score (see Table 5 for a description of these features).

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

W tym przypadku podstawowym pojęciem jest kaskada wiadomości (*news cascade*), na podstawie której możemy wyróżnić kilka(naście) odrębnych cech, stanowiących następnie wejście do metod ML.

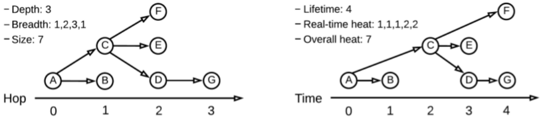


Fig. 9. Illustrations of news cascades.

Table 7. News Cascade Features

Feature Type	Feature Description	H	T	P
Cascade size	Overall number of nodes in a cascade [Castillo et al. 2011; Vosoughi et al. 2018]	✓	✓	✓
Cascade breadth	Maximum (or average) breadth of a news cascade [Vosoughi et al. 2018]	✓		✓
Cascade depth	Depth of a news cascade [Castillo et al. 2011; Vosoughi et al. 2018]	✓		✓
Structural virality	Average distance among all pairs of nodes in a cascade [Vosoughi et al. 2018]	✓		✓
Node degree	Degree of the root node of a news cascade [Castillo et al. 2011]	✓		
	Maximum (or average) degree of non-root nodes in a news cascade [Castillo et al. 2011]	✓		
Spread speed	Time taken for a cascade to reach a certain depth (or size) [Vosoughi et al. 2018]	✓		✓
	Time interval between the root node and its child nodes [Wu et al. 2015]		✓	
Cascade similarity	Similarity scores between a cascade and other cascades in the corpus [Wu et al. 2015]	✓		

H, hop-based news cascades; T, time-based news cascades; P, pattern-driven features.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Tak jak w przypadku stylu, tu również możemy się opierać na wzorcach

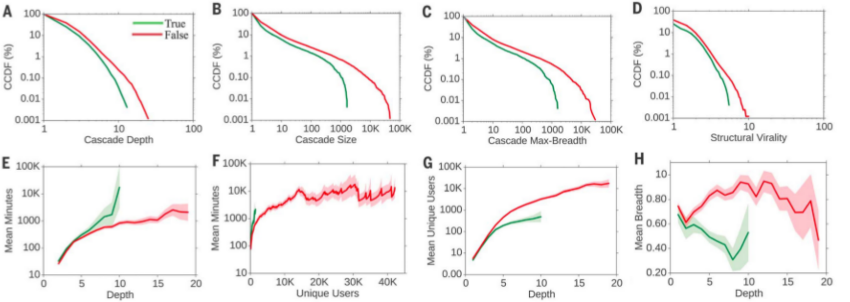


Fig. 10. Fake news cascade-based propagation patterns (Twitter data, directly from Vosoughi et al. [2018]). (A–D) Complementary cumulative distribution function (CCDF) distributions of cascade depth, size, max-breadth, and structural virality of fake news are always above that of true news. (E, F) The average time taken for fake news cascades to reach a certain depth and a certain number of unique users are both less than that for true news cascades. (G, H) For fake news cascades, their average number of unique users and breadth at a certain depth are always greater than that of true news cascades.

[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

Ostatni w typologii jest problem źródła: tu śledzimy skąd wychodzą połaczenia (tzn jak

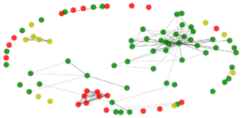


Fig. 16. Coauthorship network (directly from Sitaula et al. [2020]). Nodes are news authors associated with fake news (red), true news (green), or both (yellow), and edges indicate that two authors collaborate only once (dashed) or at least twice (solid).

Fig. 17. Content sharing network (directly from Horne et al. [2019]). Nodes are news publishers, and edges are the flows of news articles among publishers. Orange: Russian/conspiracy community; yellow: right-wing/conspiracy community; green: U.S. mainstream community; magenta: left-wing blog community; and cyan: UK mainstream community.

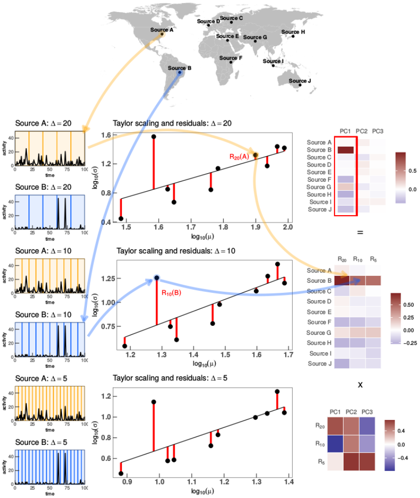
[Źródło: X. Zhou & R. Zafarani, ACM Computing Surveys 53, 109 (2020)]

- przykład nie do końca dotyczy fake news, a raczej korelacji pomiędzy sposobem pisania w niektórych news outlets a pogladami politycznymi,
- rozpatrujemy tu omawiane wcześniej prawo Taylora, mówiące o skalowaniu się odchylernia standardowego w funkcji wartości oczekiwanej $\sigma \sim \mu^{\alpha}$
- samo prawo Taylora jest dość powszechnie osberwoane, natomiast ponieważ jest ono prawem statystycznym, występują od niego odchyłki - pytanie jak te odchyłki się grupują i czy widac wśród nich jakieś prawidłowości,
- wreszcie prawo Taylora można obserwować dla różnych skal czasowych i wygodnie jest skorzystac z miary, która agregowałaby te wkłady,
- stąd też wyznaczono *R* (rezydua) zdefiniowane jako

  $$R_{s,\Delta} = \log_{10} \frac{\sigma_{s,\Delta}}{B_{\Delta}\mu_{s,\Delta}^{\alpha_{\Delta}}}$$

  czyli po prostu róznicę logarytmów zmierzonych odchyłek i tych oczekiwanych, tzn wynikających z prawa Taylora
- następnie na takiej macierzy wykonano PCA i pierwszą skłądową główną uznano za *RA* - reaktywność danego tematu

Poniższa grafika powinna rzucać trochę światła na podejście



Poniżej rysunek dla konkretnych danych. Pokazano jedynie 4 najistotniejsze składowe głowne. Widać, że pierwsza z nich tłumaczy ponad połowę zmienności i odpowiada praktycznie za średnią wartość wkładów od poszczególnych rezyduów.

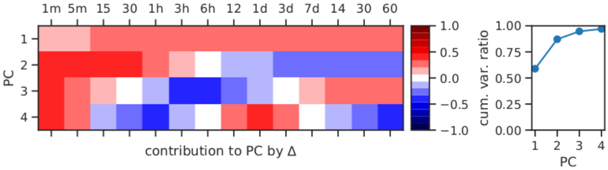


Figure 4. Principal Component Analysis of residuals by time window size for the keyword *European Union* (clean data). (left) Contributions of residuals $R_{s,\Delta}(k)$ for the Δ analyzed to the first four PCs (first four rows of matrix $\hat{\mathbf{G}}$). The first principal component is roughly the arithmetic mean of the residuals over different timescales, the second PC has opposite loadings for long and short timescales. (right) A cumulative explained variance ratio for first four PCs. The first four PCs explain typically around 97% of variance.

Reaktywność daje możliwość oceny, czy dany temat jest poruszany częściej czy rzadziej niż wynikałoby to z czystego skalowania.

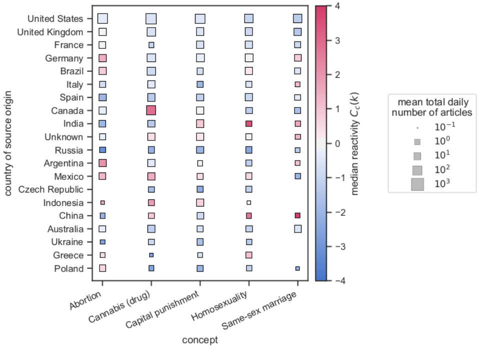
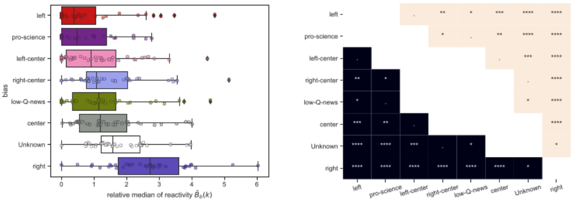


Figure 6. The median reactivity $C_c(k)$ of the reactivity for all sources from a given country c for keywords k related to "polarizing" concepts. Square size is proportional to the mean daily number of articles published by sources from a given country c on a given topic k; color represents the median reactivity $C_c(k) = \langle RA_s(k) \rangle_{s \in \mathscr{S}_c}$ of sources from the country c. Missing squares indicate that there were no publishers from the country that published at least 36 articles on the topic in our dataset. Red symbols correspond to topics that were reactively discussed in the country in 2018.

Co więcej, można zauważyć korelacje pomiedzy względną medianą *RA* a nacechowaniem politycznym (na podstawie wspomnianego serwisu Fact Check / Media Bias)



(a) (b)

Figure 7. Comparison of relative median reactivity $\tilde{B}_b(k)$ between political bias groups for all keywords. *Left*-oriented sources are generally less reactive than the *right*-oriented sources. (a): values of relative median reactivity $\tilde{B}_b(k)$ (see Eq. 4) for all keywords by political bias. (b): results of pairwise Dunn median tests with a two-step Benjamini-Krieger-Yekutieli FDR adjustment[30]; the adjusted p-values describe the likelihood of the observed difference in medians of two samples assuming there is no difference between the medians of their populations. Thus, the lower the p-value, the more statistically significant the difference between the two group medians. $**** - p \in [0, 0.001)$, $*** - p \in [0.001, 0.005)$, $** - p \in [0.005, 0.01)$, $* - p \in [0.01, .05)$, $. - p \in [0.05, 0.1)$, otherwise $p \in [0.1, 1]$. The color indicates which group had the higher median – black if the median of the group from the corresponding row was higher than the median of the group from the corresponding column; white – the opposite case.