Digital battlegrounds: The role of Wikipedia in armed conflict information warfare

First authors:
Marieth Coetzer
Leopold Augustin

Supervisor:
Fabian Braesemann

General Online Research Conference
Berlin, 1st April 2025

OXFORD INTERNET INSTITUTE
UNIVERSITY OF OXFORD
GOR organized by DGOF

Why is Wikipedia important?

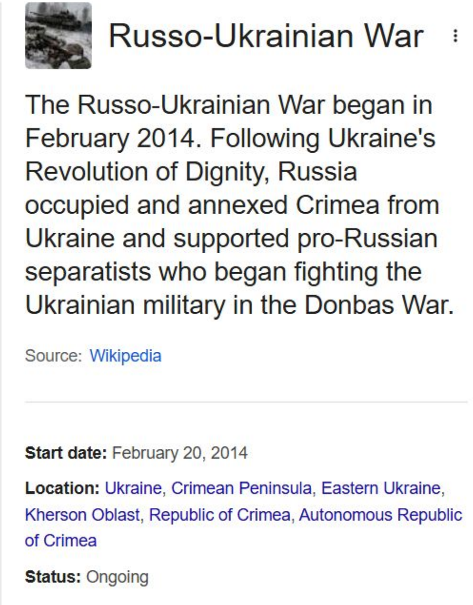
Wikipedia is considered as source of truthful information

Consensus truth

Wikipedia is viewed 18 billion times a month; often appears first in search engines

Wikipedia acts as reality check and as trustworthy source of information (compared to many others in the internet)

Wikipedia is a digital memory place; it stores the “truth” for future generations



Russo-Ukrainian War

The Russo-Ukrainian War began in February 2014. Following Ukraine's Revolution of Dignity, Russia occupied and annexed Crimea from Ukraine and supported pro-Russian separatists who began fighting the Ukrainian military in the Donbas War.

Source: Wikipedia

Start date: February 20, 2014

Location: Ukraine, Crimean Peninsula, Eastern Ukraine, Kherson Oblast, Republic of Crimea, Autonomous Republic of Crimea

Status: Ongoing

» The importance of Wikipedia makes it a possible target of manipulation

Source: New York Times article (Link), Dwivedi et al. 2023 (Link)

“Without Wikipedia, generative A.I. wouldn’t exist.”

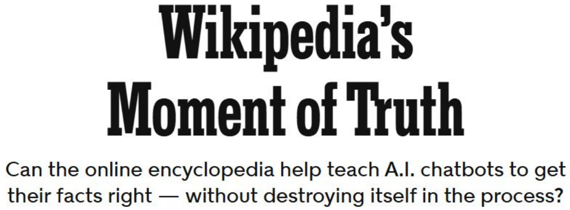
Source: New York Times article (Link), Quote by Nicholas Vincent

Large-language model training data

Wikipedia makes up significant percentage of, e.g., Metas or Google’s training data (2nd after patent data)

Used by virtual assistants to answer questions about products and brands

The plug-in solution of ChatGPT 4 for answers on events later than 2021 relies solely on Wikipedia data



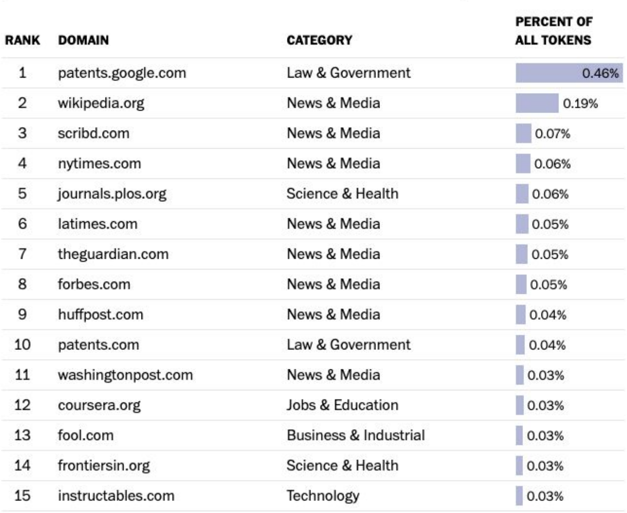
Wikipedia’s Moment of Truth

Can the online encyclopedia help teach A.I. chatbots to get their facts right — without destroying itself in the process?

Source: New York Times article (Link), voicebot.ai article (Link)

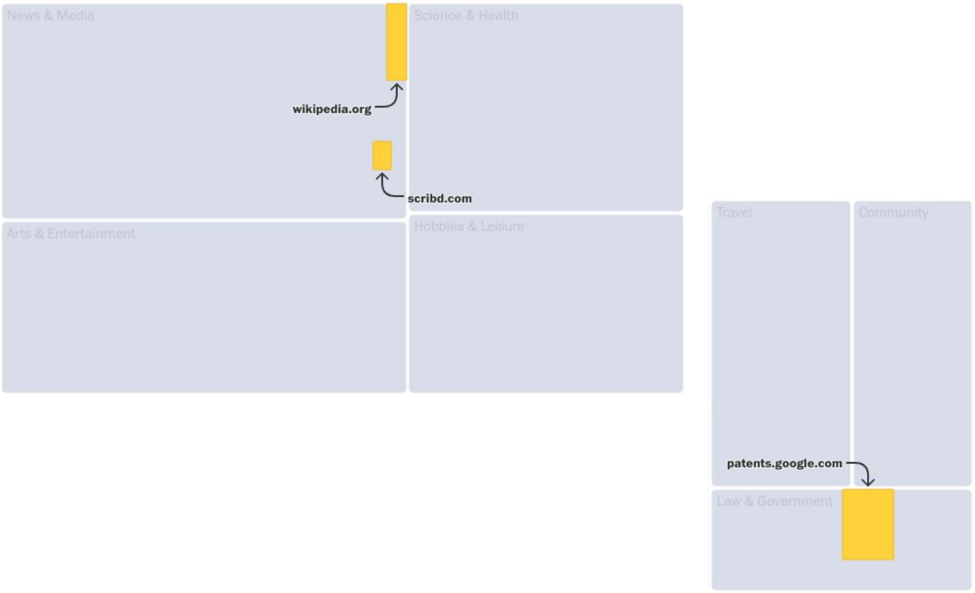
Wikipedia is the second most dominant datasource in Google’s C4 dataset

The top websites in Google’s C4 dataset



RANK	DOMAIN	CATEGORY	PERCENT OF ALL TOKENS
1	patents.google.com	Law & Government	0.46%
2	wikipedia.org	News & Media	0.19%
3	scribd.com	News & Media	0.07%
4	nytimes.com	News & Media	0.06%
5	journals.plos.org	Science & Health	0.06%
6	latimes.com	News & Media	0.05%
7	theguardian.com	News & Media	0.05%
8	forbes.com	News & Media	0.05%
9	huffpost.com	News & Media	0.04%
10	patents.com	Law & Government	0.04%
11	washingtonpost.com	News & Media	0.03%
12	coursera.org	Jobs & Education	0.03%
13	fool.com	Business & Industrial	0.03%
14	frontiersin.org	Science & Health	0.03%
15	instructables.com	Technology	0.03%

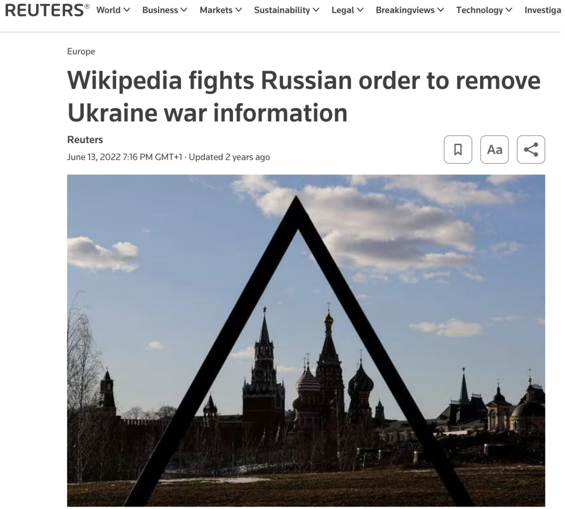
The three most dominant datasources visualized



» Models of Google or Meta have been trained on C4, in other words, on Wikipedia data

Source: Washinton post article (Link)

Consequently, important events such as wars influence Wikipedia

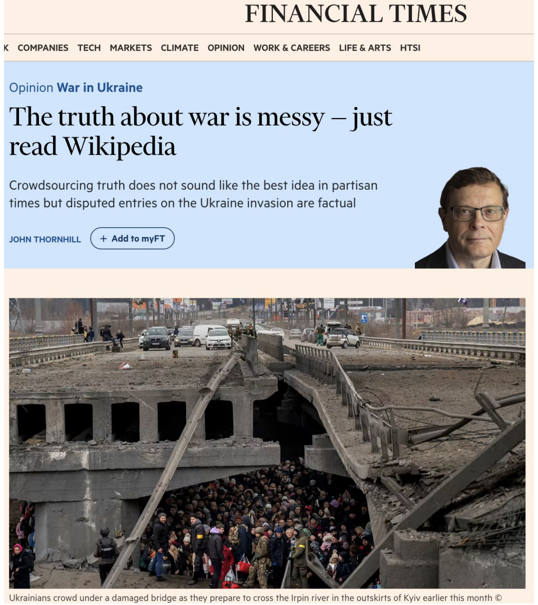


Europe

Wikipedia fights Russian order to remove Ukraine war information

Reuters

June 13, 2022 7:16 PM GMT+1 · Updated 2 years ago



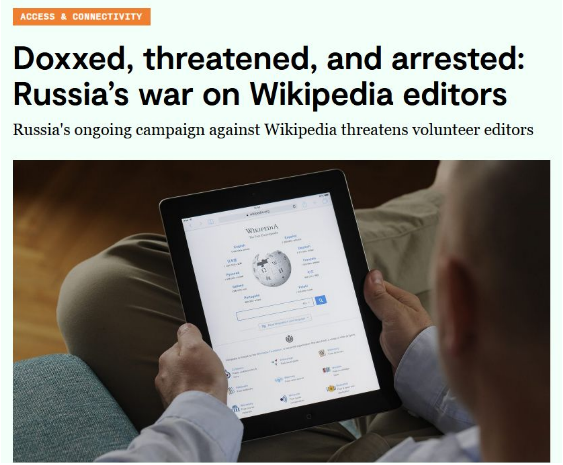
Opinion War in Ukraine

The truth about war is messy – just read Wikipedia

Crowdsourcing truth does not sound like the best idea in partisan times but disputed entries on the Ukraine invasion are factual

JOHN THORNHILL

Ukrainians crowd under a damaged bridge as they prepare to cross the Irpin river in the outskirts of Kyiv earlier this month ©



Doxxed, threatened, and arrested: Russia's war on Wikipedia editors

Russia's ongoing campaign against Wikipedia threatens volunteer editors

» These actions distort the "consensus" truth and LLM training data

Source: Reuters article (Link), Financial times article (Link), Rest of World article (Link)

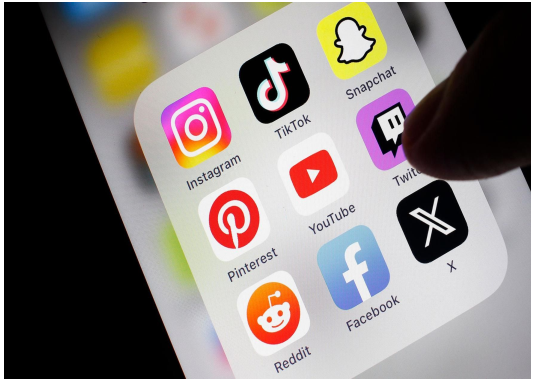
What are the known battlegrounds of information warfare?

Information war takes place on several '*digital battlegrounds*'

Mass media

Newspapers

Social media



- We argue that Wikipedia is another digital battleground of the Russian-Ukrainian information war due to its central role in global information networks

Source: Babacan et al. (2022), Ntanos et al. (2018), Khaldarova et al. (2020), Doroshenko et a. (2021), Treyger et al. (2022)

Our Research:

Is there a relationship between territorial and digital dispute on Wikipedia?

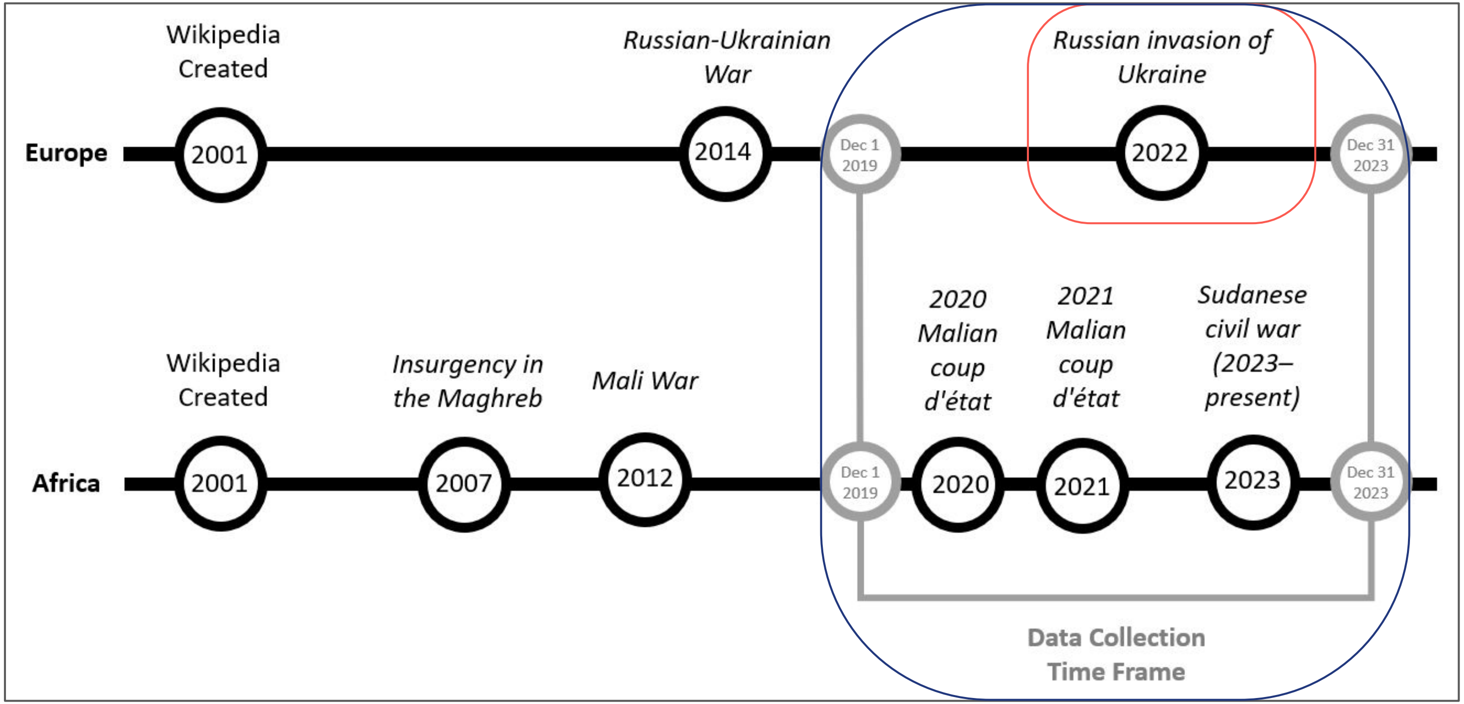
Research question 1

Did the 2022 invasion of Ukraine lead to more attention and disputes on Wikipedia articles about contested Ukrainian regions?

Research question 2

Can we develop an early-warning tool to predict disputes on Wikipedia pages using internal metrics and exogenous sources?

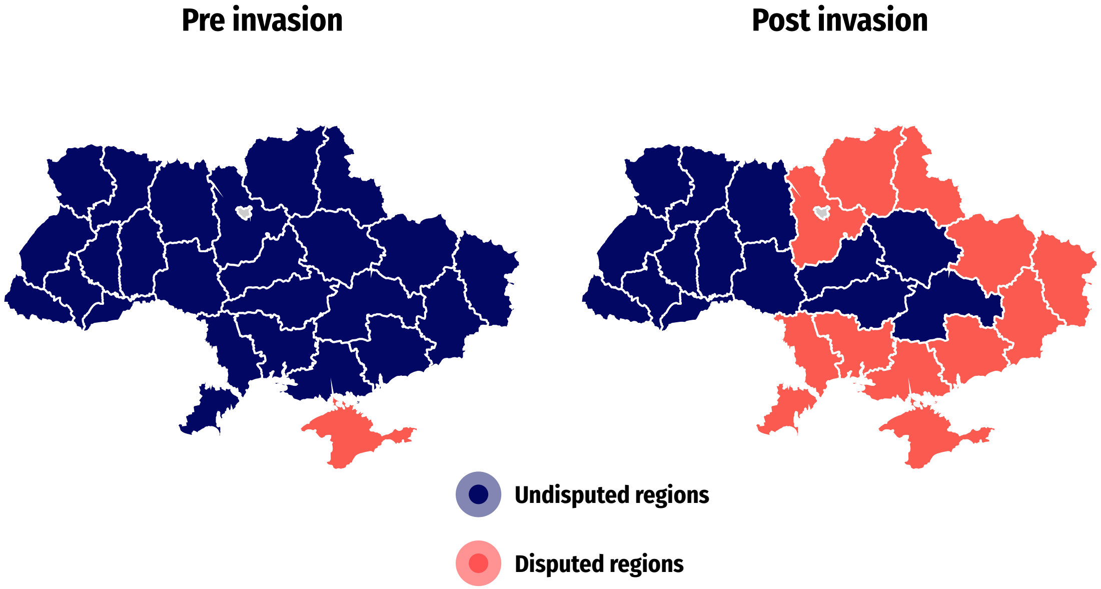
We examined disputes on both a granular and broad perspective



Research question 1

Research question 2

To answer our first research question, we divide Ukrainian regions into disputed and undisputed territories using the ACLED conflict database



Pre invasion

Post invasion

We use three metrics to operationalise digital attention & dispute

Metric 1

Revision

An author releasing a new version of Wikipedia page incl. edits (= *digital attention*)

Metric 2

Revert

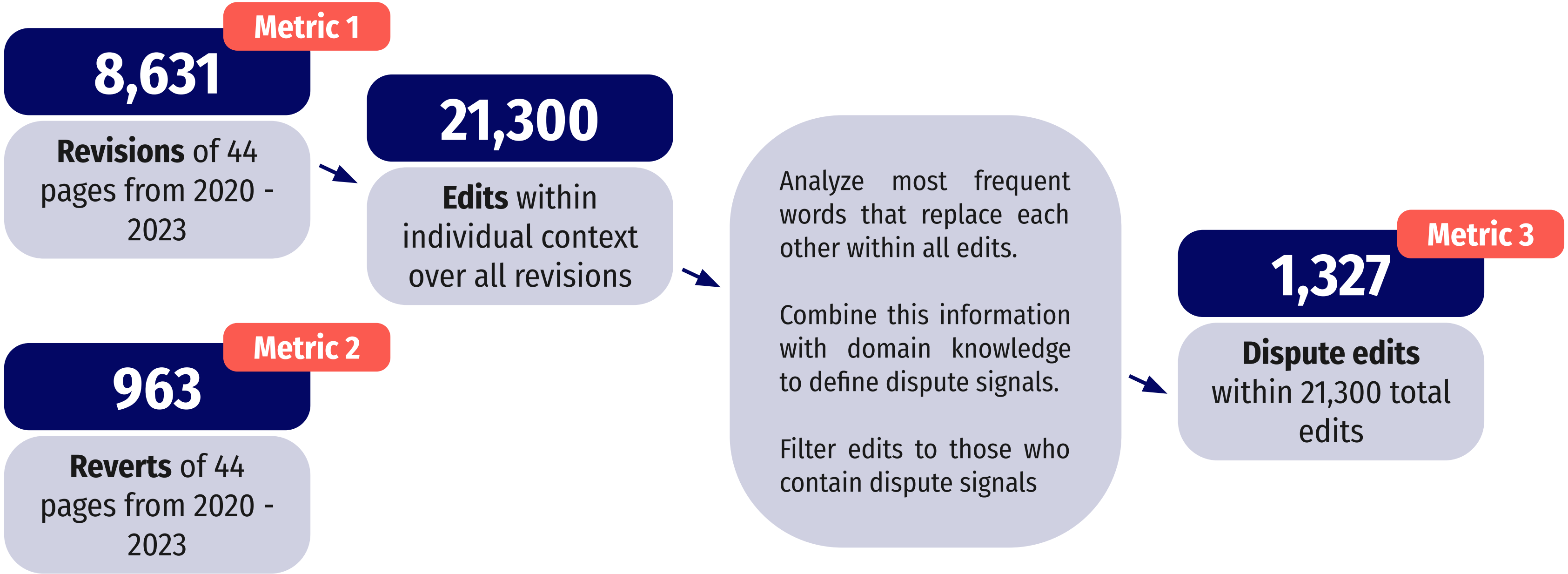
Resetting of a Wikipedia page to a former version (= *non-domain specific dispute*)

Metric 3

Dispute edit

Substitution of domain specific words within an edit (= *domain specific dispute*)

Then, we count all three metrics throughout all the edit histories of relevant Wikipedia pages (44 regional pages in total):



Metric 1

8,631

Revisions of 44 pages from 2020 - 2023

21,300

Edits within individual context over all revisions

Analyze most frequent words that replace each other within all edits.

Combine this information with domain knowledge to define dispute signals.

Filter edits to those who contain dispute signals

Metric 3

1,327

Dispute edits within 21,300 total edits

Metric 2

963

Reverts of 44 pages from 2020 - 2023

Metrics 1, 2, & 3 are the target variables in our empirical model

We formalise our hypothesis in a *difference-in-difference regression*, comparing disputed Ukrainian against undisputed and Polish regions

The Difference-in-difference (DiD) regression measures the effect of the invasion on digital attention and dispute

$$\boxed{E} = \beta_0 + \beta_1 D + \beta_2 P + \beta_3 I + \boxed{\beta_4 ID} + \beta_5 IP + \varepsilon$$

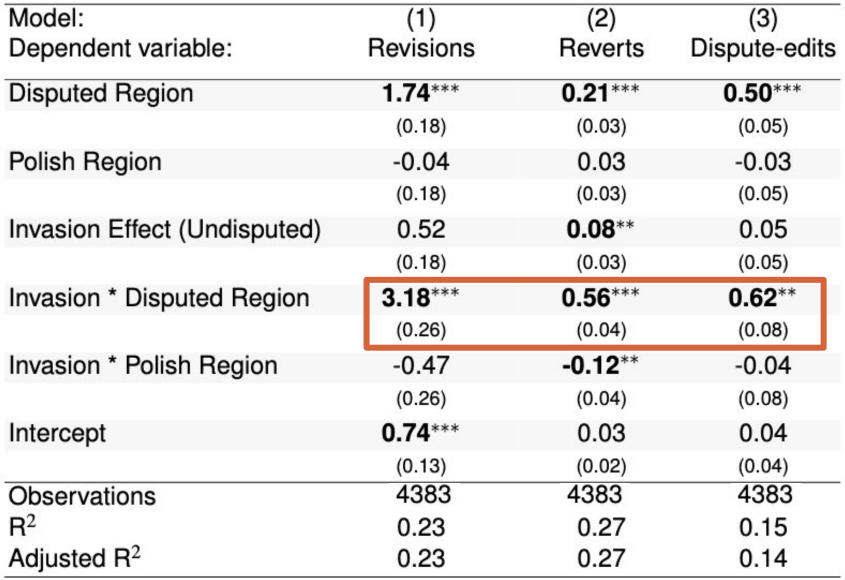
E measures daily sums of **digital attention and dispute across all articles**

Metric 1	Revisions
Metric 2	Reverts
Metric 3	Dispute edits

(I) nvasion * (D) ispute measures the **invasion effect on digital attention and dispute** for **articles about invaded regions** compared to the counterfactual (= undisputed region articles)

We expect a positive and significant effect!

We find that territorially disputed regions see more dispute online



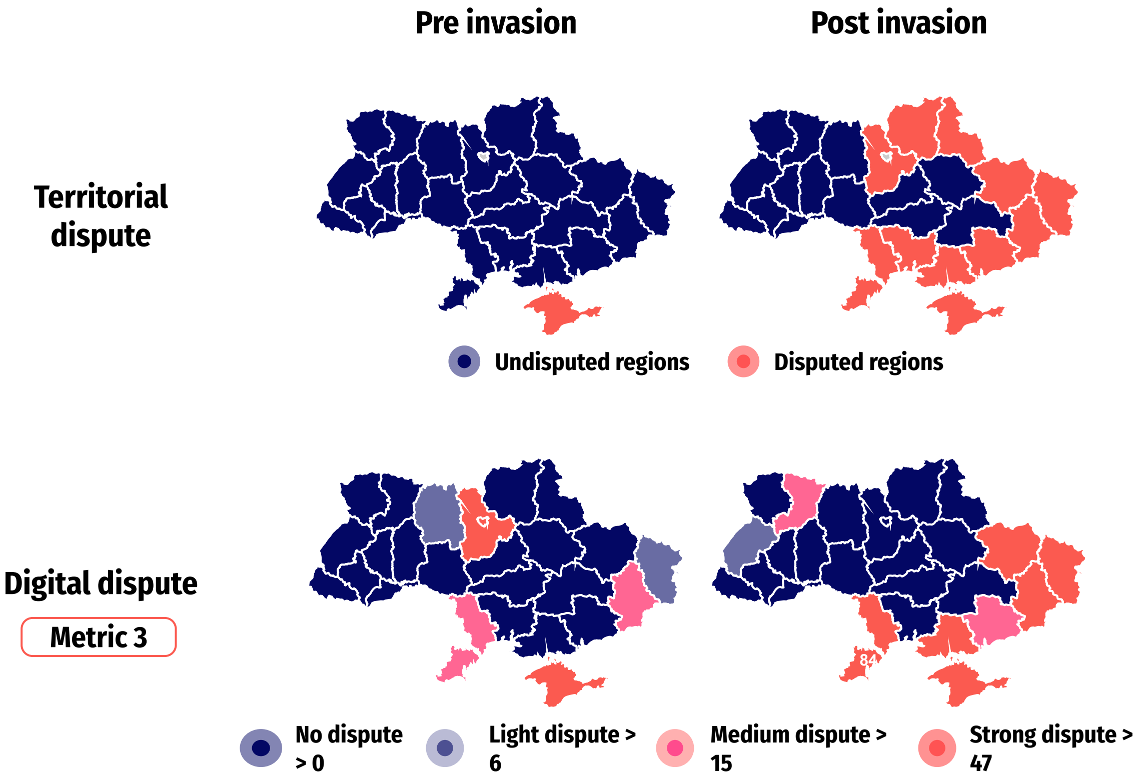
	Metric 1	Metric 2	Metric 3
Model:	(1)	(2)	(3)
Dependent variable:	Revisions	Reverts	Dispute-edits
Disputed Region	**1.74*****	**0.21*****	**0.50*****
	(0.18)	(0.03)	(0.05)
Polish Region	-0.04	0.03	-0.03
	(0.18)	(0.03)	(0.05)
Invasion Effect (Undisputed)	0.52	**0.08****	0.05
	(0.18)	(0.03)	(0.05)
Invasion * Disputed Region	**3.18*****	**0.56*****	**0.62****
	(0.26)	(0.04)	(0.08)
Invasion * Polish Region	-0.47	**-0.12****	-0.04
	(0.26)	(0.04)	(0.08)
Intercept	**0.74*****	0.03	0.04
	(0.13)	(0.02)	(0.04)
Observations	4383	4383	4383
R^2	0.23	0.27	0.15
Adjusted R^2	0.23	0.27	0.14

The relevant coefficient estimates are positive and statistically significant

The invasion's effect on disputed regions is higher than for undisputed regions

In other words: there is an association between *territorial and digital dispute*

We find that territorially disputed regions see more dispute online



Pre invasion

Post invasion

Territorial dispute

● Undisputed regions

● Disputed regions

Digital dispute

Metric 3

● No dispute > 0

● Light dispute > 6

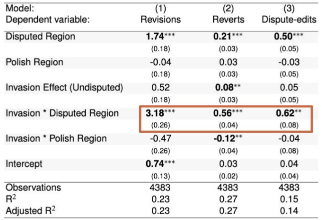
● Medium dispute > 15

● Strong dispute > 47

Coming back to our research question:

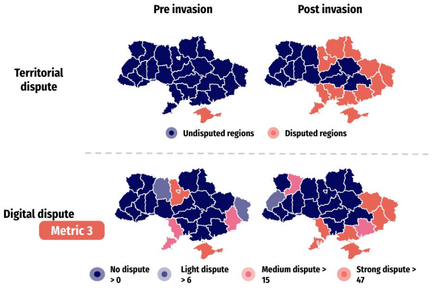
Did the 2022 invasion of Ukraine lead to more attention and disputes on Wikipedia articles about contested Ukrainian regions?

Yes! We find evidence for more attention and dispute on articles about contested Ukrainian regions



Model: Dependent variable:	(1) Revisions	(2) Reverts	(3) Dispute-edits
Disputed Region	**1.74*****	**0.21*****	**0.50*****
	(0.18)	(0.03)	(0.05)
Polish Region	-0.04	0.03	-0.03
	(0.18)	(0.03)	(0.05)
Invasion Effect (Undisputed)	0.52	**0.08****	0.05
	(0.18)	(0.03)	(0.05)
Invasion * Disputed Region	**3.18*****	**0.56*****	**0.62****
	(0.26)	(0.04)	(0.08)
Invasion * Polish Region	-0.47	**-0.12****	-0.04
	(0.26)	(0.04)	(0.08)
Intercept	**0.74*****	0.03	0.04
	(0.13)	(0.02)	(0.04)
Observations	4383	4383	4383
R^2	0.23	0.27	0.15
Adjusted R^2	0.23	0.27	0.14

Tab. 1 Difference-in-differences regression results relating dispute on Wikipedia to to territorial conflict. Territorially undisputed Ukrainian regions are the baseline. We use seven-day left-aligned rolling averages for all three target variables (1) revisions, (2) reverts, and (3) dispute-edits.



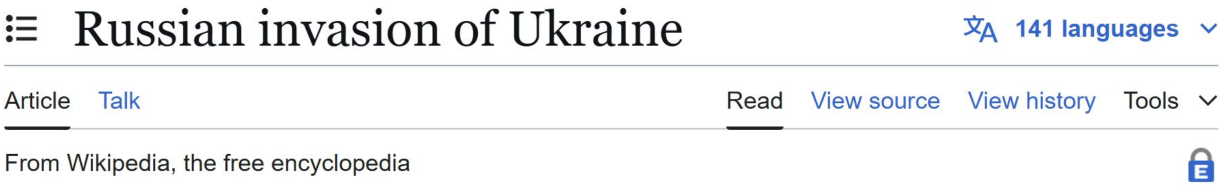
To predict these disputes, we define edit wars

Edit Wars — Occur when information on pages is **heavily contested** or vandalised

Senior Editors — Wikipedians with a history of **high-quality edits**

Metric

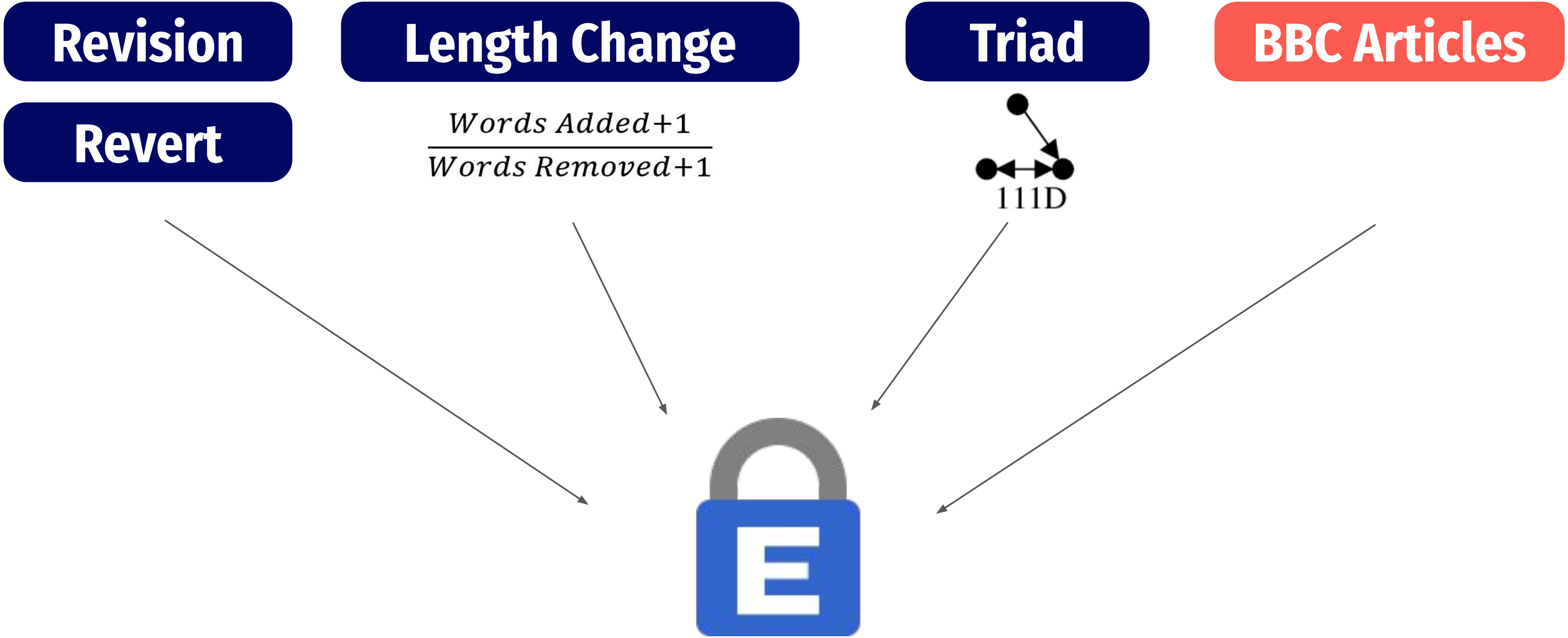
Page Locking — Changing the page setting so **only senior editors can change** or add to the page



Russian invasion of Ukraine

From Wikipedia, the free encyclopedia

We predict edit wars using internal and exogenous metrics



Revision

Revert

Length Change

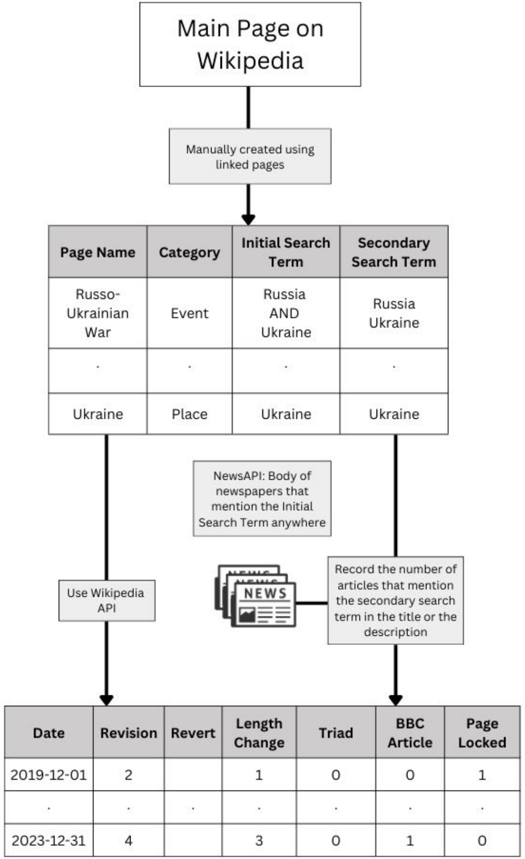
$$\frac{Words\ Added+1}{Words\ Removed+1}$$

Triad

BBC Articles

Source: Sepehri-Rad & Barbosa (2015), Yasseri et al. (2012), Ford et al. (2013)

We compiled a dataset for each Wikipedia page



Main Page on Wikipedia

Manually created using linked pages

Page Name	Category	Initial Search Term	Secondary Search Term
Russo-Ukrainian War	Event	Russia AND Ukraine	Russia Ukraine
.	.	.	.
Ukraine	Place	Ukraine	Ukraine

NewsAPI: Body of newspapers that mention the Initial Search Term anywhere

Use Wikipedia API

Record the number of articles that mention the secondary search term in the title or the description

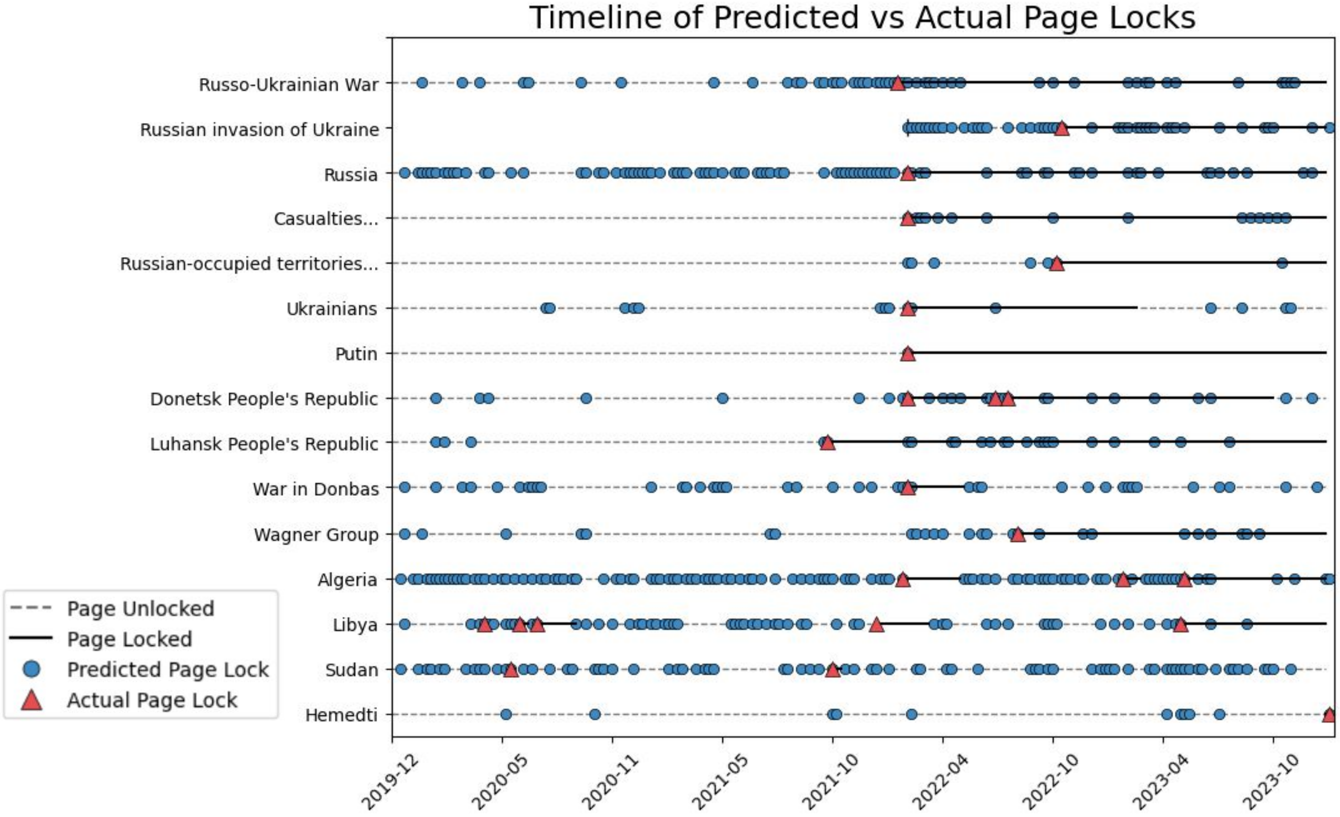
Date	Revision	Revert	Length Change	Triad	BBC Article	Page Locked
2019-12-01	2		1	0	0	1
.	.	.	.	.	.	.
2023-12-31	4		3	0	1	0

Examined 122 related pages of armed conflicts of which 38 had been locked

Used the Wikipedia page names to find related BBC articles on NewsAPI

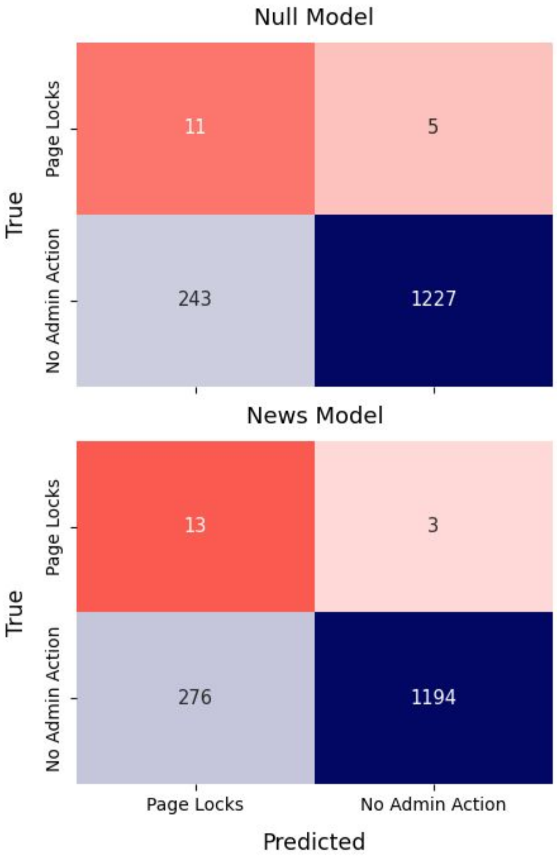
Predicted page locks using a random forest classifier

A granular view of the model’s performance



Given how rarely a pagelock occurs, the data is unbalanced, leading to many **false positives**

The BBC Title metric identified more true positives



Null Model

True \ Predicted	Page Locks	No Admin Action
Page Locks	11	5
No Admin Action	243	1227

News Model

True \ Predicted	Page Locks	No Admin Action
Page Locks	13	3
No Admin Action	276	1194

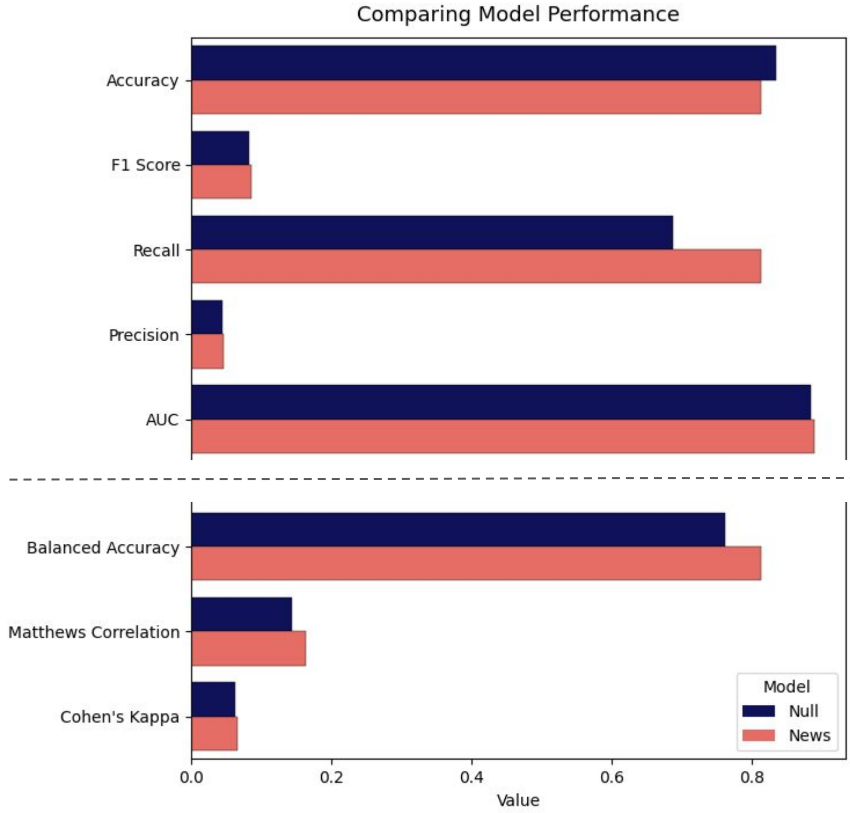
Null Model

Contains only internal Wikipedia Metrics

News Model

Contains BBC Articles as well as internal Wikipedia metrics

Analysing other metrics made for unbalanced data shows the News Model still outperforms the Null Model



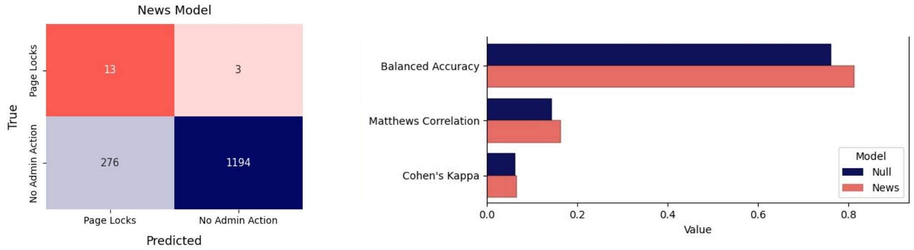
Traditional machine learning evaluation metrics

Metrics to evaluate unbalanced data

Coming back to our research question:

Can we develop an early-warning tool to predict disputes on Wikipedia pages using internal metrics and exogenous sources?

Yes! We predict 81% of the page locks in our test set with our model that uses both internal Wikipedia metrics and BBC titles



Implications

Territorial disputes are pushing into new *digital battlegrounds*

Our approach uncovers Wikipedia as a digital battleground of information warfare

Internal metrics and exogenous sources ca be used as an early-warning tool to predict edit wars

Early Warnings: Analysing and Forecasting Disputes on Wikipedia Armed Conflict Pages

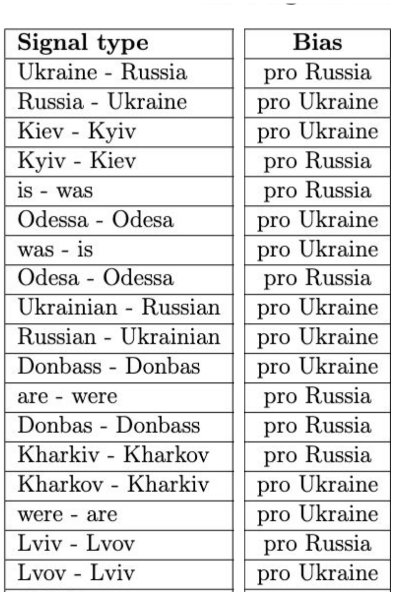
First authors:
Marieth Coetzer
Leopold Augustin

Supervisor:
Fabian Braesemann

General Online Research Conference
Berlin, 1st April 2025

Backup Slides

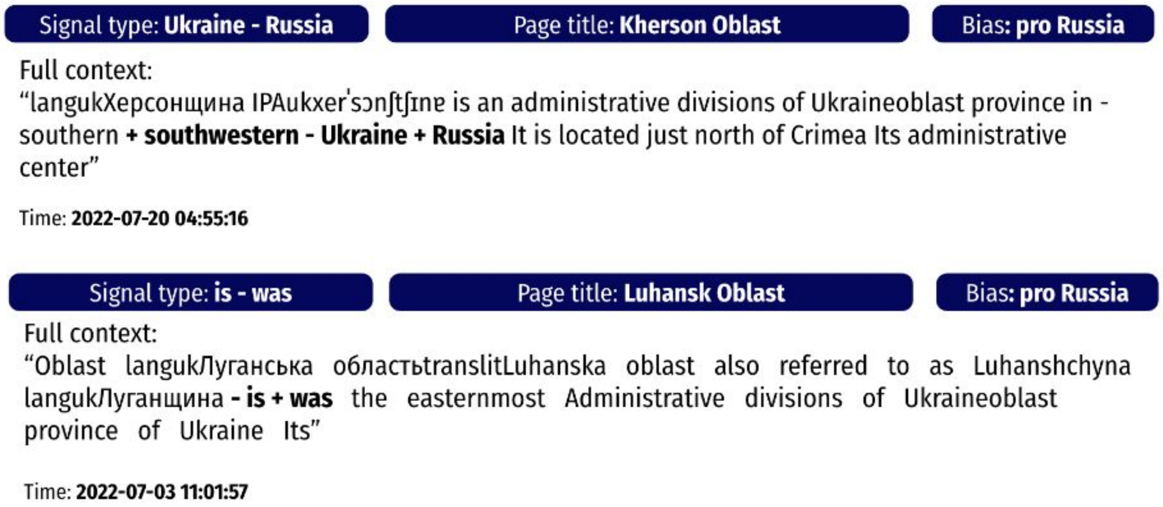
Full list of dispute signals - if these word substitutions were part of an edit, we classify as dispute edit



Signal type	Bias
Ukraine - Russia	pro Russia
Russia - Ukraine	pro Ukraine
Kiev - Kyiv	pro Ukraine
Kyiv - Kiev	pro Russia
is - was	pro Russia
Odessa - Odesa	pro Ukraine
was - is	pro Ukraine
Odesa - Odessa	pro Russia
Ukrainian - Russian	pro Ukraine
Russian - Ukrainian	pro Ukraine
Donbass - Donbas	pro Ukraine
are - were	pro Russia
Donbas - Donbass	pro Russia
Kharkiv - Kharkov	pro Russia
Kharkov - Kharkiv	pro Ukraine
were - are	pro Ukraine
Lviv - Lvov	pro Russia
Lvov - Lviv	pro Ukraine

Examples of dispute edits

Figure 3: **Examples of dispute edits**



Signal type: **Ukraine - Russia** | Page title: **Kherson Oblast** | Bias: **pro Russia**

Full context:
“langukХерсонщина IPAukxerˈsɔnʃtʃɪnɐ is an administrative divisions of Ukraineoblast province in - southern **+ southwestern - Ukraine + Russia** It is located just north of Crimea Its administrative center”

Time: **2022-07-20 04:55:16**

Signal type: **is - was** | Page title: **Luhansk Oblast** | Bias: **pro Russia**

Full context:
“Oblast langukЛуганська областьtranslitLuhanska oblast also referred to as Luhanshchyna langukЛуганщина **- is + was** the easternmost Administrative divisions of Ukraineoblast province of Ukraine Its”

Time: **2022-07-03 11:01:57**

Examples of two dispute edits from the actual data body that discuss the nationhood of Kherson and Luhansk.

Names of Ukrainian Oblasts and Polish Voivodeships

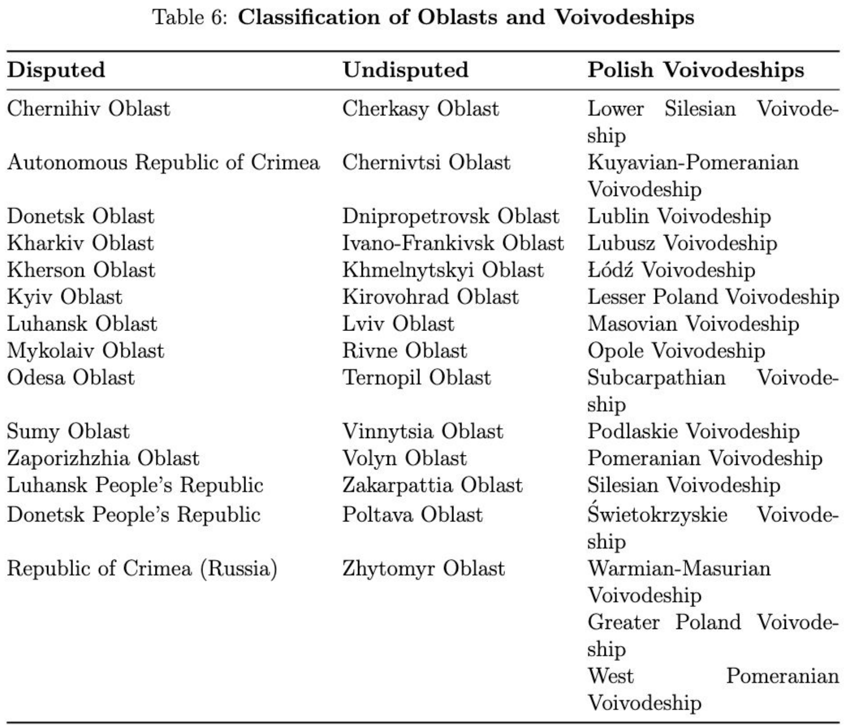
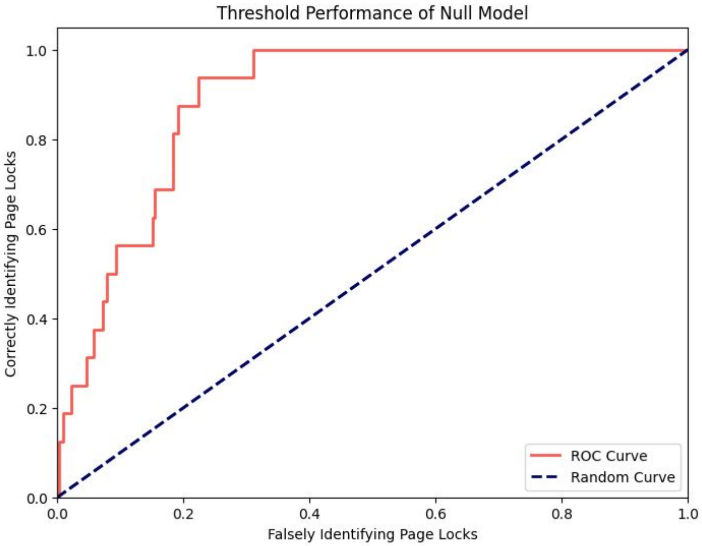
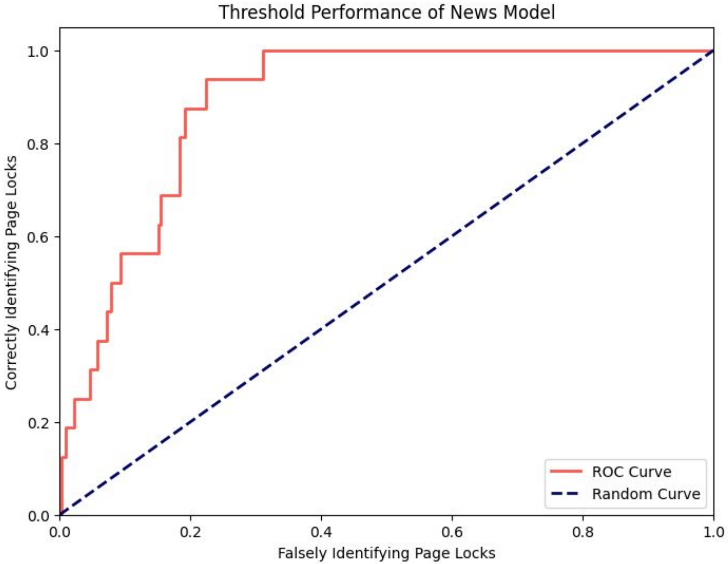


Table 6: **Classification of Oblasts and Voivodeships**

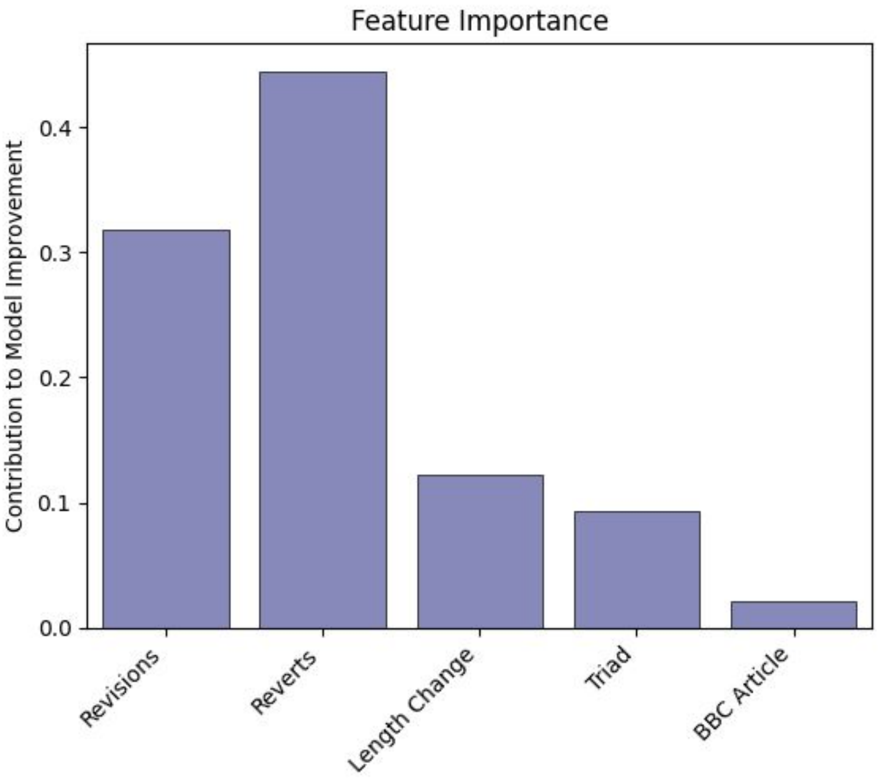
Disputed	**Undisputed**	**Polish Voivodeships**
Chernihiv Oblast	Cherkasy Oblast	Lower Silesian Voivodeship
Autonomous Republic of Crimea	Chernivtsi Oblast	Kuyavian-Pomeranian Voivodeship
Donetsk Oblast	Dnipropetrovsk Oblast	Lublin Voivodeship
Kharkiv Oblast	Ivano-Frankivsk Oblast	Lubusz Voivodeship
Kherson Oblast	Khmelnytskyi Oblast	Łódź Voivodeship
Kyiv Oblast	Kirovohrad Oblast	Lesser Poland Voivodeship
Luhansk Oblast	Lviv Oblast	Masovian Voivodeship
Mykolaiv Oblast	Rivne Oblast	Opole Voivodeship
Odesa Oblast	Ternopil Oblast	Subcarpathian Voivodeship
Sumy Oblast	Vinnytsia Oblast	Podlaskie Voivodeship
Zaporizhzhia Oblast	Volyn Oblast	Pomeranian Voivodeship
Luhansk People's Republic	Zakarpattia Oblast	Silesian Voivodeship
Donetsk People's Republic	Poltava Oblast	Świetokrzyskie Voivodeship
Republic of Crimea (Russia)	Zhytomyr Oblast	Warmian-Masurian Voivodeship
		Greater Poland Voivodeship
		West Pomeranian Voivodeship

Model Thresholds

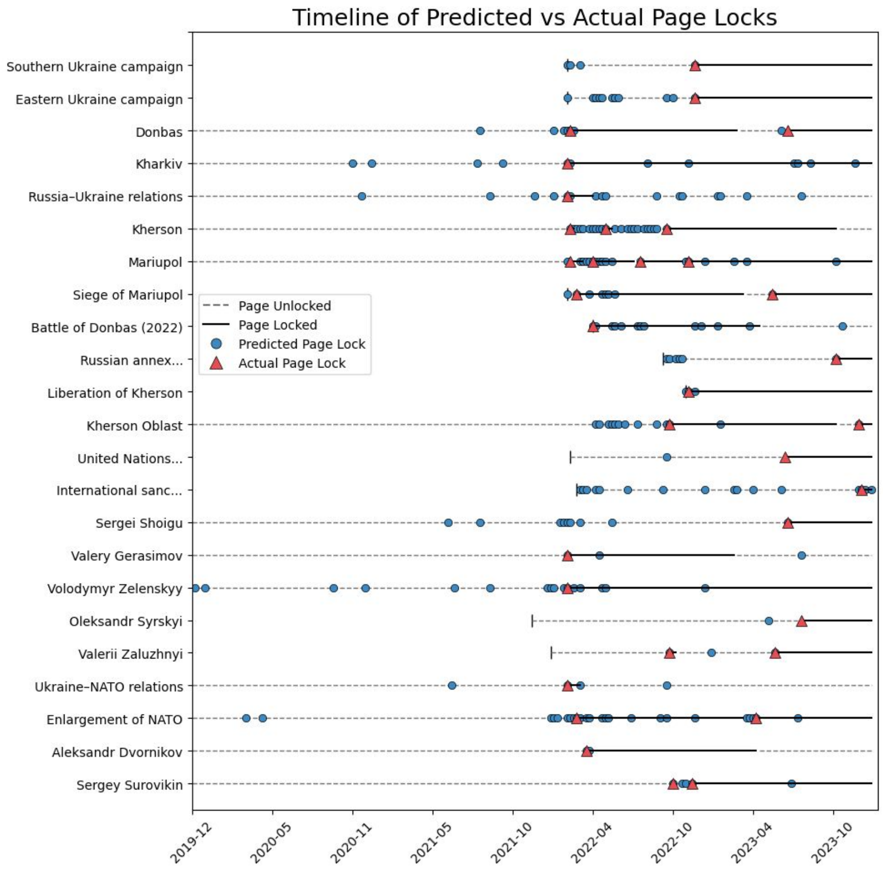




Model Feature Importance



Remaining Pages



Metrics Define

$$\boldsymbol{Balanced\ Accuracy} = \frac{\overbrace{\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}}^{\textbf{Se}} + \overbrace{\frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}}}^{\textbf{Sp}}}{\mathbf{2}}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

P_o - Probability of Agreement Observed

P_e - Probability of Agreement **by Chance**