Challenges on fighting Disinformation in Wikipedia

Diego Sáez Trumper
diego@wikimedia.org

Agenda

- Our context
- What we have done so far
- Current & future work

Photo by Elijah O'Donnell on Unsplash

Wikipedia is used as ground truth in (Automatic) Fact Checking tasks

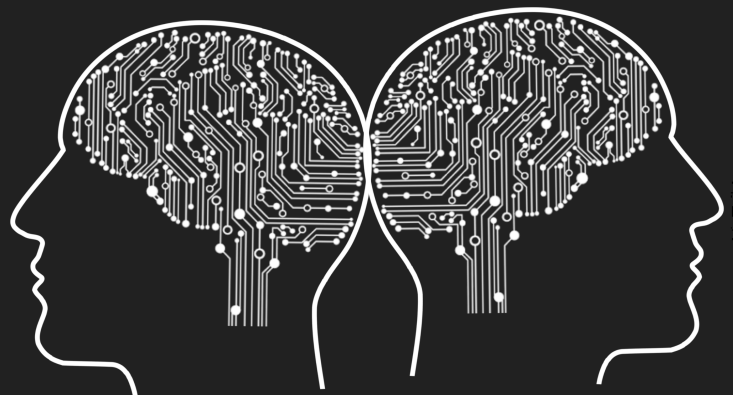


Image by Mohamed Hassan on Pixibay

Wikimedia Foundation, CC BY-SA 3.0 via Wikimedia Commons

Wikipedia is different from Social Networks

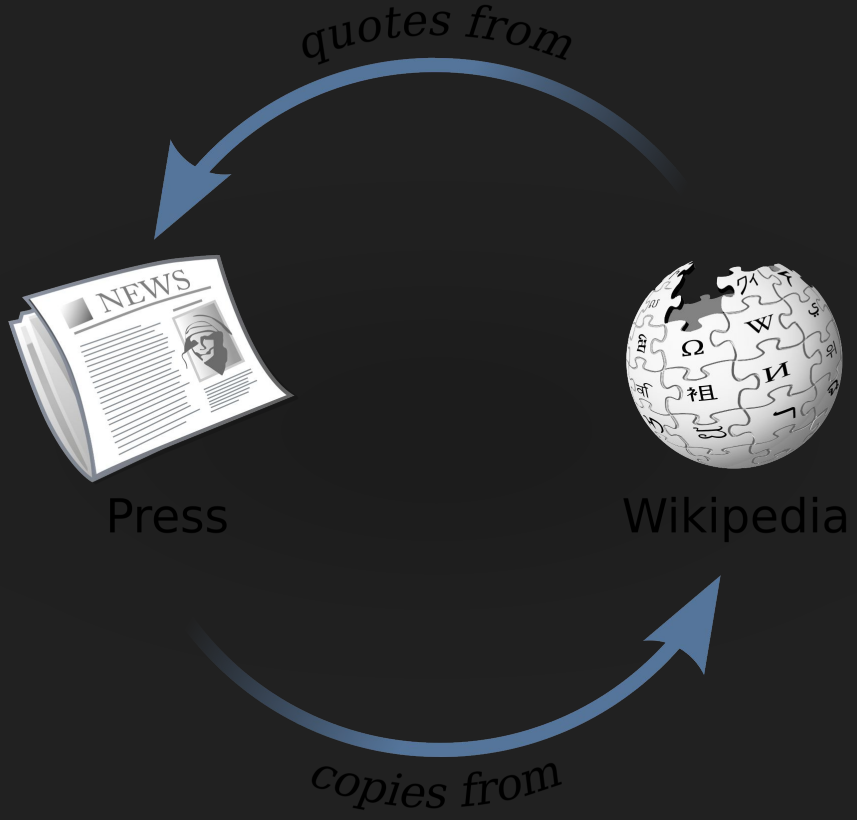
- Social networks are about expressing opinions.
  - Opinions and facts are mixed
  - Popularity is very important

- Wikipedia is about sharing knowledge
  - Content Integrity

Wikimedia Foundation, CC BY-SA 3.0 via Wikimedia Commons

Our challenges

- No ground truth
  - Or no single ground-truth
- Circular reporting
- Subtle attacks
- Imbalances across projects
- Cultural differences
  - Ex. {{USgovtPOV}}



Our approach

Understand	Prevent	Support Workflows
● Create Conceptual Models ● Provide Insights	● Early warnings ● Identify threads	● Machines to support editors in simple but time consuming tasks ● ML to identify potential content policy violations

ONLINE DISINFORMATION AND THE ROLE OF WIKIPEDIA

A PREPRINT

Diego Saez-Trumper
Wikimedia Foundation
diego@wikimedia.org

October 29, 2019

ABSTRACT

The aim of this study is to find key areas of research that can be useful to fight against disinformation on Wikipedia. To address this problem we perform a literature review trying to answer three main questions: *(i)* What is disinformation? *(ii)*What are the most popular mechanisms to spread online disinformation? and *(iii)* Which are the mechanisms that are currently being used to fight against disinformation?.

A taxonomy of (dis)information

	Authenticity	Intention
Disinformation	False	Bad
Misinformation	False	Unknown
Mal-Information	True	Bad
Fake News	False	Bad
Satire News	False	Not Bad
Imposter Content	False	Unknown
Fabricated Content	False	Bad
Manipulated Content	Unknown	Bad
Rumor	Unknown	Unknown

Wikipedia's Vulnerability

Mechanism	Description	Type	Wikipedia's Vulnerability
Bots	Software used to automatize the spread of messages, generating the idea that of a lot people is given an specific opinion or interest about a topic	Technical	Low
Sock-puppets	Multiple Online identities used for purposes of deception.	Social	Medium
Web Brigades	A set of users coordinated to introduce fake content by exploiting the weakness of communities and systems.	Social	High
Click farms	Where a large group of low-paid workers are hired to perform some micro-tasks to deceive online systems.	Social	Medium
Deepfake	AI a technique for human image synthesis that can be used to create fake videos of celebrities or notable people.	Technical	Medium
Data Voids	Exploiting missing data to manipulate search results	Social	Medium
Circular reporting	A situation where a piece of information appears to come from multiple independent sources, but in reality comes from only one source.	Social	High

Social media traffic report pilot

Contact

Jonathan Morgan
Wikimedia Foundation

Collaborators

Isaac Johnson
Wikimedia Foundation

Duration: 2020-February – 2020-May

Open source
via GitHub

[[meta:Research:Social_media_traffic_report_pilot]]

Top articles by social media traffic on 2020-12-14 [edit]

Last updated on 14:00, 15 December 2020 (UTC)

Rank	Platform	Article	Platform traffic 12-14	Platform traffic 12-13	All traffic 12-14	Watchers	Visiting watchers
1	Youtube	BBC World Service	15029	17145	21650	214	< 30
2	Facebook	Kalpana Chawla	11467	< 500	19501	197	< 30
3	Reddit	Chicken Cup (Chenghua)	10644	1571	47837	< 30	< 30
4	Facebook	STS-107	9188	< 500	10325	85	< 30
5	Twitter	SB19	8792	2179	26173	160	108
6	Youtube	PBS	7345	6179	11570	445	56
7	Reddit	Trictena atripalpis	6214	< 500	22105	< 30	< 30
8	Youtube	Deutsche Welle	5563	5583	7701	153	< 30
9	Twitter	Treasure (band)	5260	5120	13998	103	98
10	Reddit	Beechcraft Bonanza	4909	567	23411	95	< 30

Content propagation within Projects

Contact

Diego Saez

Wikimedia Foundation

Collaborators

Giovanni Comarela

Federal University of Espírito Santo

Souneil Park

Teléfonica Research

Rodolfo Valentim

Federal University of Espírito Santo

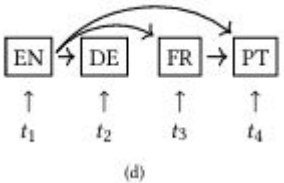


Table 2 - Creation of pages related to the size of Wikipedia Projects.

Number of Languages	1/308	2/307	9/300
Ratio of Items	10/90	20/80	50/50
Large→Large	0%	2.36%	23.18%
Small → Small	79.1%	71.12%	35.63%
Large → Small	14.08%	17.22%	23.82%
Small → Large	6.81%	9.28%	17.36%

[[meta:Research:Exploration_on_content_propagation_across_Wikimedia_projects]]

Aligning Wikipedia and Wikidata

Contact

Diego Saez
Wikimedia Foundation

Collaborators

Meeyoung Cha
KAIST

Ma Jing
CUKH

Cheng-Te Li
NCKU

Yi-Ju
NCKU

Table1: Examples of the consistent/inconsistent data

Label	Examples
Consistent:0	Sentence: ' Biography: He was born in Pavlovsky Posad near Moscow.'
	Wikidata claim: 'place of birth Pavlovsky Posad'
Inconsistent:1	Sentence: 'At the age of thirteen, he entered Dulwich College.
	Wikidata claim: 'educated at Shippensburg University of Pennsylvania'

[[meta:Research:Discovering_content_inconsistencies_between_Wikidata_and_Wikipedia]]

{{Templates}}

Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability

Miriam Redi
Wikimedia Foundation
London, UK

Besnik Fetahu
L3S Research Center
Leibniz University of Hannover

Jonathan Morgan
Wikimedia Foundation
Seattle, WA

Dario Taraborelli
Wikimedia Foundation
San Francisco, CA

ABSTRACT

Wikipedia is playing an increasingly central role on the web, and the policies its contributors follow when sourcing and fact-checking content affect million of readers. Among these core guiding principles, *verifiability* policies have a particularly important role. Verifiability requires that information included in a Wikipedia article be corroborated against *reliable secondary sources*. Because of the manual labor needed to curate and fact-check Wikipedia at scale, however, its contents do not always evenly comply with these policies. Citations (i.e. reference to external sources) may not conform to verifiability requirements or may be missing altogether, poten-

1 INTRODUCTION

Wikipedia is playing an increasingly important role as a "neutral" arbiter of the factual accuracy of information published in the web. Search engines like Google systematically pull content from Wikipedia and display it alongside search results [38], while large social platforms have started experimenting with links to Wikipedia articles, in an effort to tackle the spread of disinformation [37].

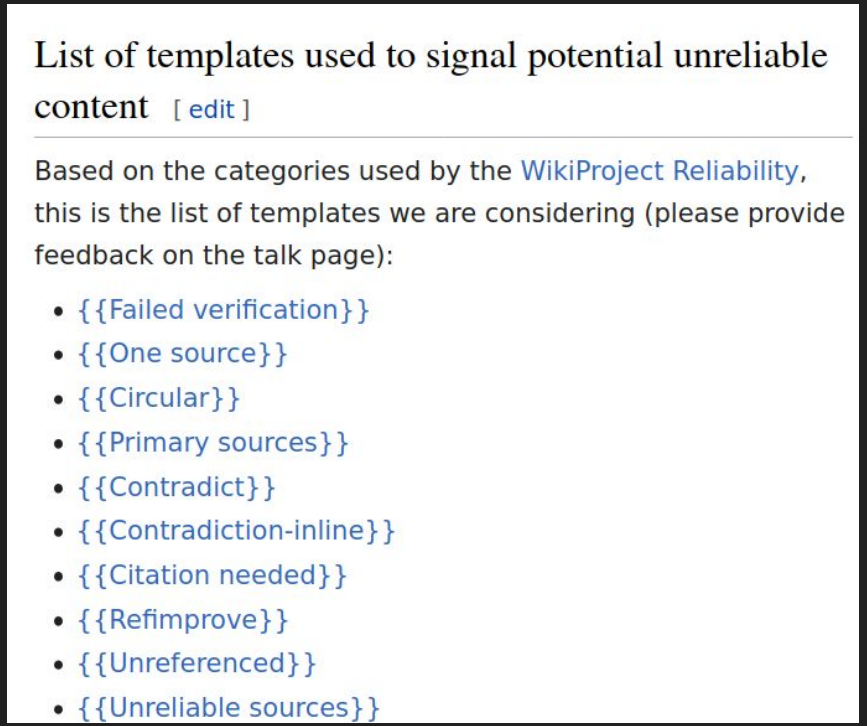
Research on the accuracy of information available on Wikipedia suggests that despite its radical openness—anyone can edit most articles, often without having an account —the confidence that other platforms place in the factual accuracy of Wikipedia is largely

Content reliability related templates

- 41 templates related with content reliability
- Positive and negative examples

Please provide Feedback

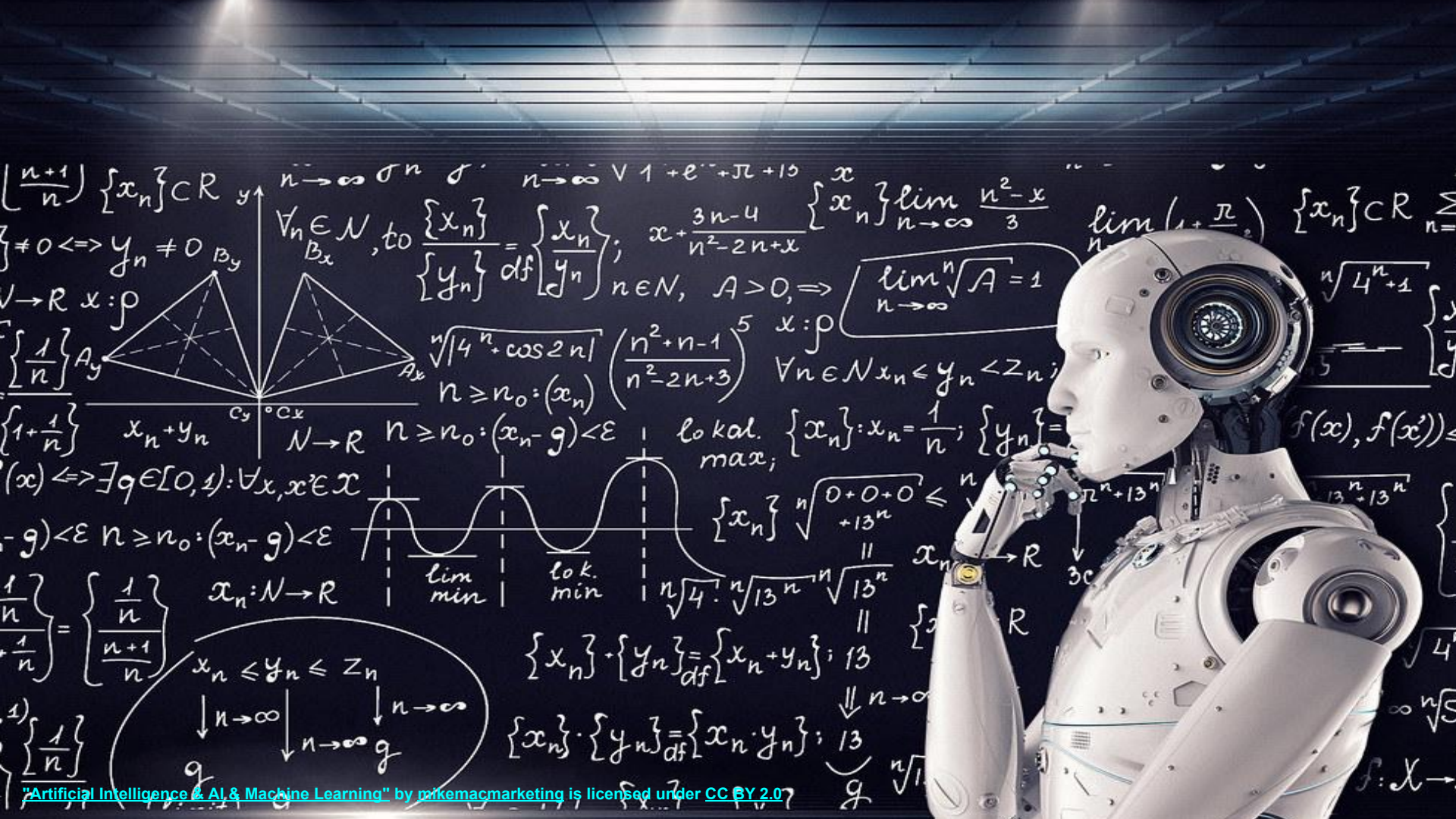
[[meta:User:Diego_(WMF)/templatesReliability]]



List of templates used to signal potential unreliable content [edit]

Based on the categories used by the WikiProject Reliability, this is the list of templates we are considering (please provide feedback on the talk page):

- {{Failed verification}}
- {{One source}}
- {{Circular}}
- {{Primary sources}}
- {{Contradict}}
- {{Contradiction-inline}}
- {{Citation needed}}
- {{Refimprove}}
- {{Unreferenced}}
- {{Unreliable sources}}



"Artificial Intelligence & AI & Machine Learning" by mikemacmarketing is licensed under CC BY 2.0

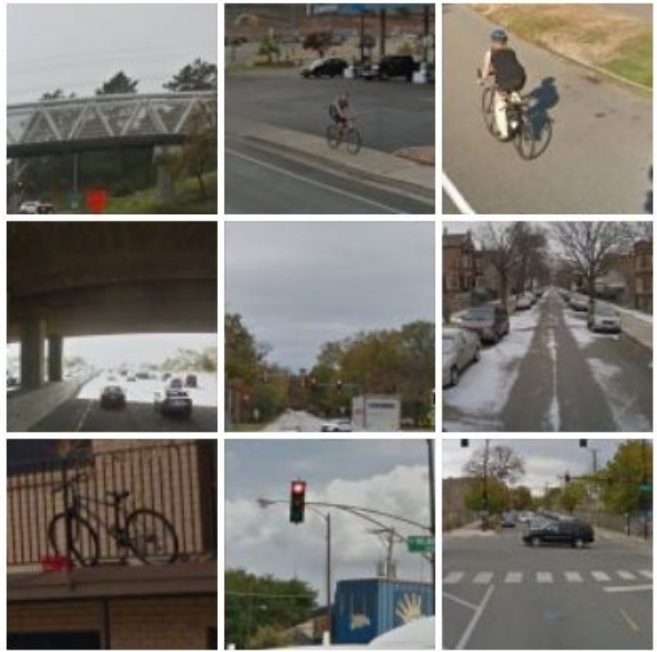
Your Nam

Your

oom is prot

Select all images with

bicycles



VERIFY

Work in Progress & Next steps

- Release a set of new datasets for content reliability in Wikipedia (In collaboration with Miriam Redi and Kay Wong)

- World-wide Internship program
  - Students
    - Yi-Ju Lu (NCKU, Taiwan)
    - Rodolfo Valentim (UFES, Brazil)
    - Kay Wong (Malaysia) [outreachy.org]
  - Profiles currently looking for:
    - NLP
    - Front-end interfaces

diego@wikimedia.org

[[meta:User:Diego_(WMF)/templatesReliability]]