The Information Disorder Level (IDL) Index

An Experiment to Assess Machine-Generated Content's Factuality

Laurence Dierickx, Carl-Gustav Lindén, Andreas L. Opdahl
University of Bergen

MASSHINE, GENERATIVE METHODS - AI AS COLLABORATOR AND COMPANION IN THE SOCIAL SCIENCES AND HUMANITIES
Copenhagen, December 06-08, 2023

The paradoxical uses of generative AI

To produce misleading content and disinform at large scale

NewsGuard has so far identified 487 AI-generated news and information sites operating with little to no human oversight, and is tracking false narratives produced by artificial intelligence tools

To produce news content and support journalistic processes

New survey finds half of newsrooms use Generative AI tools; only 20% have guidelines in place

2023-05-25. A new WAN-IFRA survey, conducted in collaboration with SCHICKLER Consulting, sets a barometer of where news publishers stand so far on using Generative AI.

Dealing with biases and artificial hallucinations

Generating misleading content is not always intentional,

ChatGPT known for producing erroneous or inaccurate content
Adding information or generating realistic experiences that do not correspond to any real-world input (semantic noise: omissions, contradictions, invented content)

Well-known phenomenon in LLMs related to the use of vast amount of data

Errors also related to the quality of unsupervised training data
(a lot of issues with ChatGPT, such as UGC and biased data)

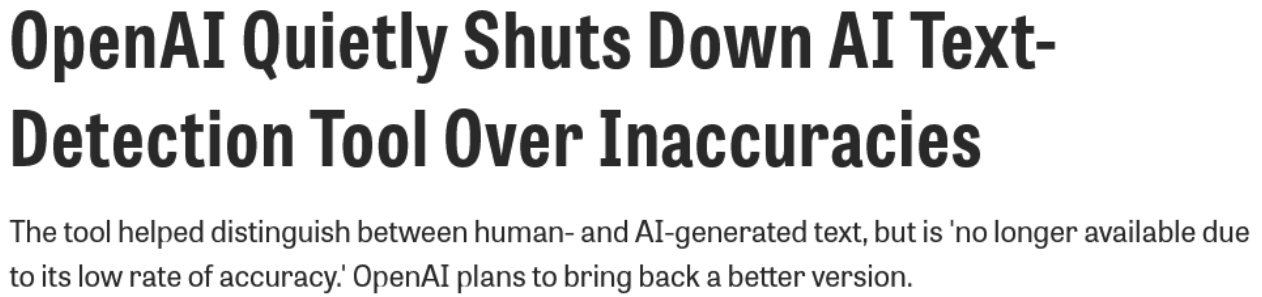
Threat for generating harmful content

Detecting machine-generated content

Distinguishing truthful text from misinformation has become particularly challenging as they present similar writing styles to machine-generated texts with true content (Schuster et al., 2023)

Detectors create false positives/ negatives, dependence on English language (does not work well on non-native written texts).

Watermarking techniques are less relevant for written content.



OpenAI Quietly Shuts Down AI Text-Detection Tool Over Inaccuracies

The tool helped distinguish between human- and AI-generated text, but is 'no longer available due to its low rate of accuracy.' OpenAI plans to bring back a better version.

The challenges of semantic detection

Semantic detection and fact-verification is another approach for developing detection systems, considering “artificial hallucination”

Human limitations: verification and fact-checking still require a human touch (developing a critical and nuanced approach) + difficulty to define complex concepts

Technical limitations: (training) data quality issues mostly related to a lack of expertise (using Wikipedia or non-experts), timeliness (lack of maintenance over time), and language (English-dependency)

Added value of human expertise to evaluate and mitigate artificial hallucinations

Dierickx, L., Lindén, C. G., & Opdahl, A. L. (2023). Automated Fact-Checking to Support Professional Practices: Systematic Literature Review and Meta-Analysis. *International Journal of Communication*, *17*, 21.

Assessing accuracy

Exploring NLP to mitigate intentional or unintentional manipulation in machine-generated content

Long tradition of evaluation in NLP: automated metrics (criticised because of their limitations but independent of language, fast and inexpensive) and human-based judgments (content quality and accuracy). Rating scales to assess intrinsic qualities.

In journalism: Coherence, descriptive value, usability, writing quality, informativeness, clarity, pleasantness, interest, boredom, preciseness, trustworthiness and objectivity (Clerwall, 2014); intelligence, education, reliability, bias, accuracy, completeness, factuality, quality and honesty (van der Kaa & Krahmer, 2014)

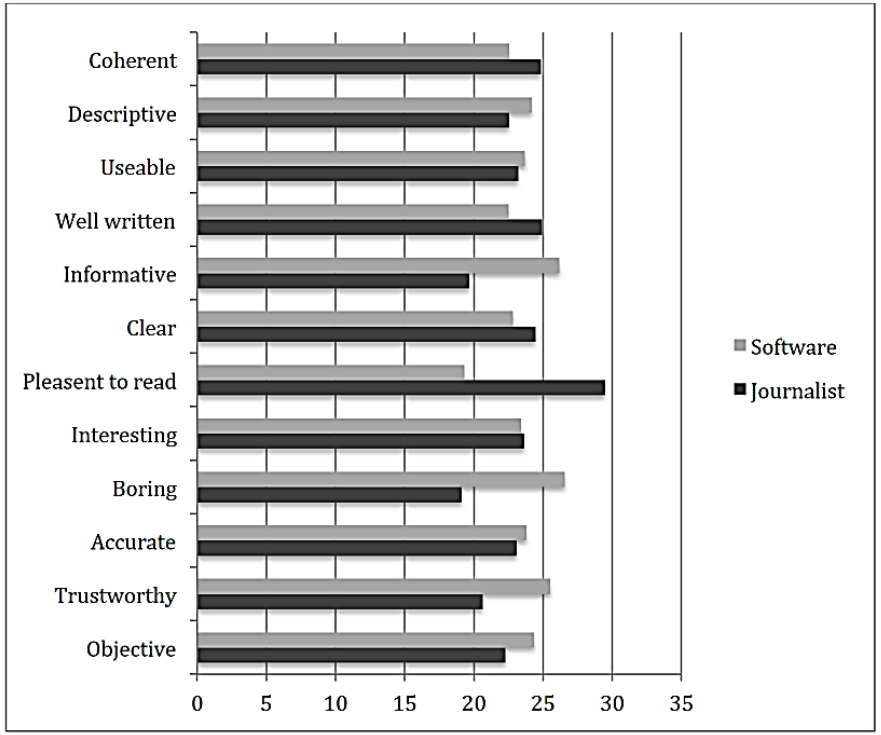


Figure 1 - Mean rank values for each descriptor, clutred by group (journalist or software)

Experimenting text generation with ChatGPT

Building a corpus of 40 texts based on:

- made up events

- real events (after 2021 but connected to events that occur before)

- 5 styles: factual news, high quality, columnist, sensationalist, pro-Russian

Observations

- Lots of additional content added to the prompt requirements

- Adding disinformation (hallucinations)

SAMPLE:
A CAR ACCIDENT
IN NORWAY

PROMPT (FACTS)

4 peoples died
They were crossing the street
A car hurt them
Then, a bus collided the car
911 was called
Rescuers cannot do anything
It happened somewhere in Norway

Write a factual news article on these facts

(…)

The victims, who were attempting to cross the road, were struck by a car and subsequently involved in a collision with a bus, resulting in their untimely demise. The incident has left the local community in a state of shock and mourning.
Authorities have initiated a thorough investigation into the circumstances surrounding the accident. Preliminary findings indicate that the four individuals, whose identities have not been released pending notification of their families, were pedestrians crossing the street when a car collided with them.
Eyewitnesses present at the scene reported a scene of chaos and confusion, as passersby and concerned citizens rushed to aid the victims before the arrival of emergency services.
The community in Norway is coming together to support one another during this difficult time, with grief counseling services being made available to those affected by the loss.
Our thoughts and prayers go out to the families and loved ones of the victims during this challenging period of grief and sorrow.

SAMPLE:
THE WAGNER GROUP AND THE RUSSIAN MILITARY

PROMPT (FACTS)

- Tensions have escalated between the Wagner Group and the Russian military.
- Russia claimed that Ukraine suffered over 3,700 casualties during a Ukrainian offensive in the Donetsk region.
- Ukrainian military had no information about a major attack in the region and called the claims "fake."
- Kyiv officials admitted to offensive actions commencing around the city of Bakhmut.
- Prigozhin's forces withdrew from Bakhmut, transferring control to the Russian army.

Write a neutral factual news article based on these 5 facts

In recent developments, escalating tensions have been observed between the Russian military and the controversial private military company known as the Wagner Group. As both sides engage in a war of words, conflicting reports have emerged regarding casualties during a Ukrainian offensive in the Donetsk region. (…)
In a significant turn of events, it has been reported that forces affiliated with Yevgeny Prigozhin, the Russian businessman linked to the Wagner Group, have withdrawn from Bakhmut. Control of the city has been transferred to the Russian army. The reasons behind this decision and the implications it may have on the overall situation remain unclear at this time. (…)
As the situation continues to unfold, it is crucial for accurate and verified information to be disseminated to ensure an objective understanding of the events and their implications. Both sides should prioritize open communication and a commitment to resolving their differences through peaceful means, in line with international norms and agreements.
The international community remains watchful of the evolving situation in the Donetsk region and hopes for a peaceful resolution that safeguards the well-being of all parties involved and minimizes the impact on the civilian population.

Characteristics of a news article

- Ability to report facts "as they are"
- Structured text following the inverted pyramid structure
- Arranges facts in descending order of importance
- Aims to answer the "w-questions“: who, when, where, what, and why
- Typically presented in simple, concise, and informative sentences
- Narratives in journalism can take various forms: From breaking news to opinion

Building a method & a tool

Hypothesis: a human assessment of machine-generated content is generally most reliable as it involves a critical thinking about the meaning of the information and its informative value, which are related to the accuracy and reliability of the news.

Aims to assessing the accuracy of event reporting, the extent of the system's generation of "artificial hallucinations“, grounded in fact detection and verification.

Comparing two different versions consists of ensuring that they contain the same amount of information sentence by sentence (Emde et al., 2016).

Comparing the prompt with the machine-generated text
Based on sentence segmentation of the text, which is based on the sentences’ boundaries (dots, question marks or exclamation marks, or ellipsis (Jurish & Würzner, 2013; Matusov et al., 2005)

Rating: 0 (True), 1 (Mostly True), 2 (Mostly False), 3 (False)

Considering the total number of assessed sentences (the 'Cannot say' answer is not included in the formula, based on the assumption that, as a joker, it does not provide meaningful input to the evaluation process), the IDL index consists of the sum of the cumulative scores for 'Mostly true' (1 point attributed to each sentence), 'Mostly false' (2 points attributed to each sentence), and 'False' (3 points attributed to each sentence), divided by the total number of sentences assessed multiplied by 3 (the maximum possible score). The index is then normalised on a scale ranging from 0 to 10.

The formula for the IDL index can be expressed as:

$$\text{IDL index} = \left(\frac{(\text{MT} \times 1) + (\text{MF} \times 2) + (\text{F} \times 3)}{(\text{MT} + \text{MF} + \text{F}) \times 3} \right) \times 10$$

where:

$$\text{MT} = \text{number of sentences classified as 'Mostly True'}$$
$$\text{MF} = \text{number of sentences classified as 'Mostly False'}$$
$$\text{F} = \text{number of sentences classified as 'False'}$$

NORDIS

HOME REASSESSMENT ABOUT DEMO

NORDIS

The Information Disorder Level index (IDL index)

Reassessing machine-generated content

This assessment involves the analysis of a text that has already undergone evaluation. Human-based judgments are inherently subjective and often involve the input of multiple judges to capture diverse perspectives. Evaluating previously evaluated auto-generated content allows for the creation of a robust training dataset. The texts in this corpus can be based on facts that really happened and on others that were completely invented. They also depend on the requirements and creativity of previous users of the system.

Machine-generated content to reassess

In recent developments, escalating tensions have been observed between the Russian military and the controversial private military company known as the Wagner Group. As both sides engage in a war of words, conflicting reports have emerged regarding casualties during a Ukrainian offensive in the Donetsk region.According to Russia, Ukraine suffered more than 3700 casualties as a result of the Ukrainian military's offensive in the Donetsk region. However, Ukrainian military officials have

Prompt used

FACTS
1) Tensions have escalated between the Wagner Group and the Russian military.
2) Russia claimed that Ukraine suffered over 3700 casualties during a Ukrainian offensive in

Start the assessment

NORDIS

HOME REASSESSMENT ABOUT DEMO

NORDIS

The Information Disorder Level index (IDL index)

Rate each sentence, according to the following rules

Make your choice:

True: Completely true or accurate and reliable (informative)

Mostly True: Predominantly true with some elements of falsehood

Cannot Say: Difficult to determine accuracy

Mostly False: Predominantly false with some elements of truth

False: Completely false or incorrect (disinformative)

You can also mark the sentences you consider as an opinion or a comment to estimate their proportion in the assessed text (independent from the IDL index).

1:

In recent developments, escalating tensions have been observed between the Russian military and the controversial private military company known as the Wagner Group.

Prompt used

FACTS
1) Tensions have escalated between the Wagner Group and the Russian military.
2) Russia claimed that Ukraine suffered over 3700 casualties during a Ukrainian offensive in the Donetsk region.
3) Ukrainian military had no information about a major attack in the region and called the claims fake.
4) Kyiv officials admitted to offensive actions commencing around the city of Bakhmut.
5) Prigozhin\'s forces withdrew from Bakhmut, transferring control to the Russian army.
Write a neutral factual news article based on these 5 facts.

There are 3 sentences marked as 'True', 3 marked as 'Mostly true', 3 marked as 'Mostly false', and 11 marked as 'False' out of a total of 20 sentences assessed.
You marked 3 sentences as opinion or comment (15% of the total amount of sentences).

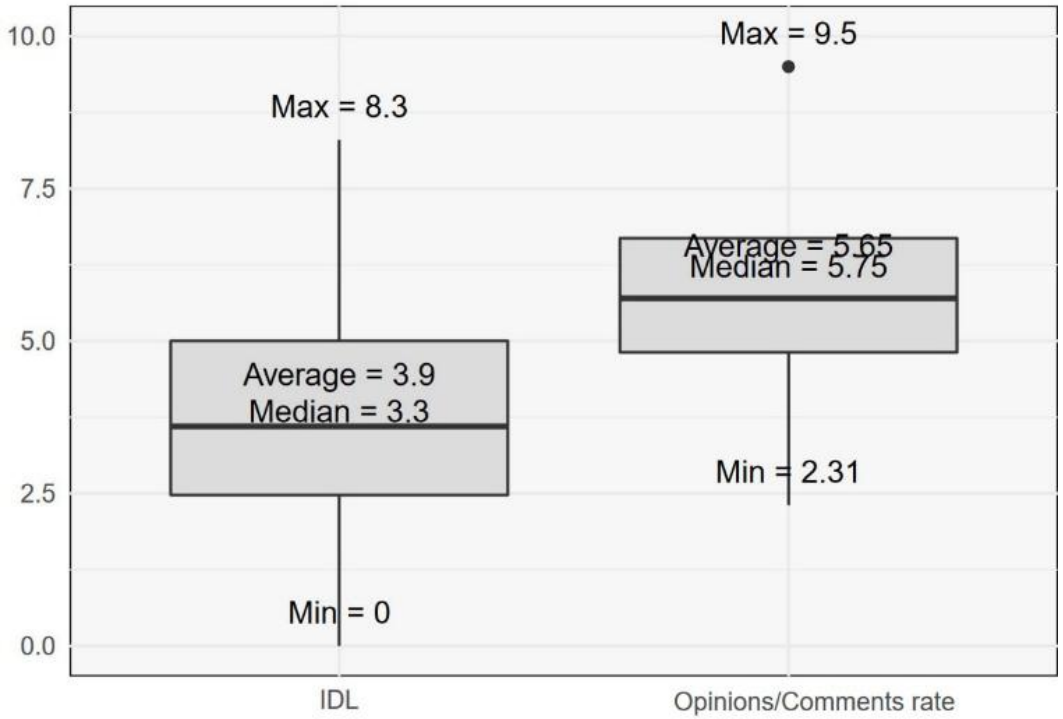
The IDL index for this news is 7

We also observed that ChatGPT tended to add opinions and comments.

**The tool also include the OCR (Opinion/Comments Rate) to quantify it.
It corresponds to the percentage of sentences marked as such.
It is considered a complementary indicator of the informational quality of the machine-generated content.**

Testing the corpus

To mitigate biases in the results, we excluded sensationalist, pro-Russian, and columnist writing styles to examine the OC rate for factual and high-quality newspaper styles. The 14 pieces of text retained for this analysis show an average rate of 3.72, with a minimum of 2.31 and a maximum of 5,45. These scores illustrate that even when asked to write factual pieces, ChatGPT is likely to add comments or opinions.



Correlation analysis: lack of meaningfulness positive correlation between these two variables, which can be due to the difficulty in assessing the factuality or truthfulness of a comment or an opinion.

Linear regression: no strong evidence to suggest that the IDL index has a significant influence or relationship with the OC rate.

Limits and way forward

Human-based judgements involved several judges
(more than one, reassessing already assessed content)

Building training data based on the tool: http://idlindex.net

Corpus size, adding real events that occurred before 2021

Integrating omissions (corresponding to the absence of one or more attributes from the input data)?

Translating the interface?

Improving prompts to mitigate “hallucinations” ?

Benchmarking?

Lessons learned at this stage

Involves critical thought about the meaning of the information and its informative value, which is related to the accuracy and reliability of the news. Thinking on the coherence of the narratives: hallucinations, fictionalisation, simulation, extrapolation (attributes absent from the input data).

From a research tool to a digital media literacy tool (help to understand to what extend GenAI also generates non-factual content and the limits of generative AI; to encourage critical thinking about what makes a report event factual.

Results suggest that using machine-generated content in the news should be humanly supervised.

As ChatGPT regularly added opinions or comments to all the samples of the corpus, it is possible to hypothesise that this mixture of genres is a clue to determining that it consists of a machine-generated piece.

Thank you!

@ohmyshambles @Gusse @andreaslo

Dierickx, L., Lindén, CG., Opdahl, A.L. (2023). The Information Disorder Level (IDL) Index: A Human-Based Metric to Assess the Factuality of Machine-Generated Content. In: Ceolin, D., Caselli, T., Tulin, M. (eds) Disinformation in Open Online Media. MISDOOM 2023. Lecture Notes in Computer Science, vol 14397. Springer, Cham.

https://doi.org/10.1007/978-3-031-47896-3_5

Prototype
https://github.com/laurence001/idl/tree/main

Development Tool
http://idlindex.net