The Information Disorder Level Index: A Method for Assessing the Factuality of Machine-Generated Content

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

International Scientific Conference Dubrovnik Media Days
Dubrovnik, September 30, 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

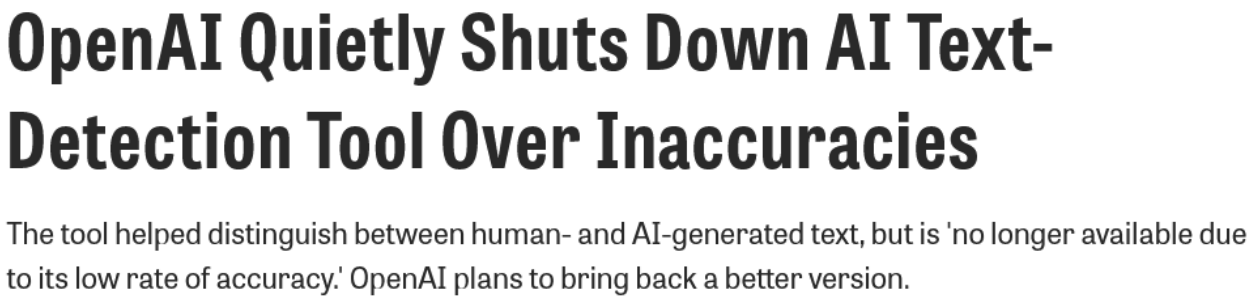
Errors also related to the quality of unsupervised training data
(a lot of issues with ChatGPT, such as UGC and biased data)

Black-box nature of the system explains its malfunctions (Li, 2023)

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)

Black box detection (online tools): create false positives/negatives, dependence to English language (does not work well on non-native written texts)



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.

White box detection (“invisible” watermark): alternative through traceability technologies (i.e., digital watermarking), still need to be improved because several technical aspects still need to be solved (Jiang et al., 2023; Kirchenbauer et al., 2023) + relevance for written content

The challenges of semantic detection

Semantic detection and fact-verification consists of a third approach for developing detection systems, considering the phenomenon of “artificial hallucination”

Human limitations: verification and fact-checking still require a human touch (developing a critical and nuanced approach) + epistemological misunderstanding (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)

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 (criticized 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)

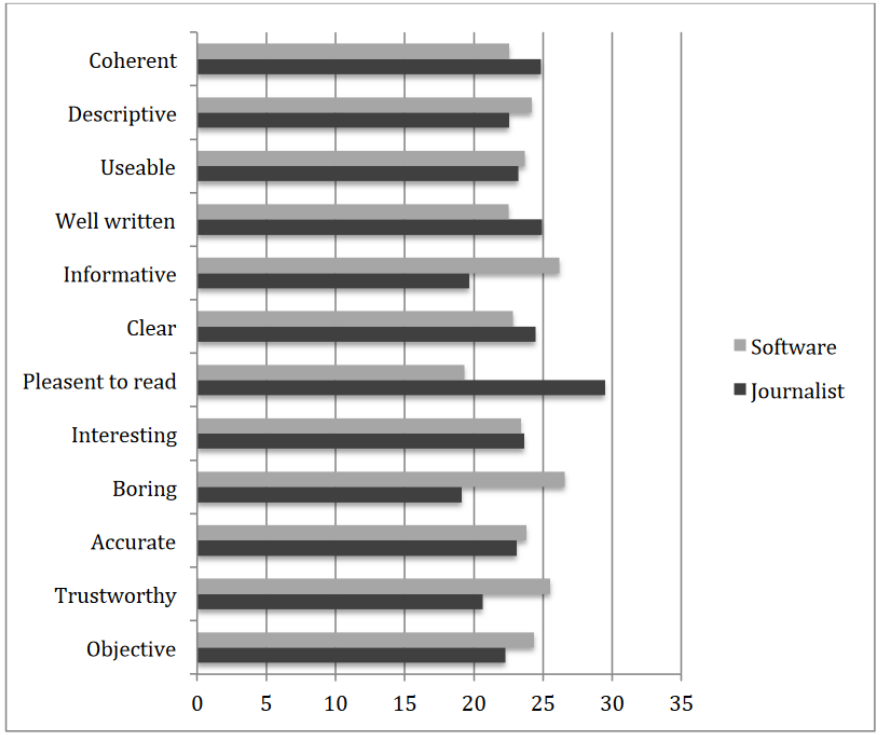


Figure 1 - Mean rank values for each descriptor, clutred by group (journalist or software)

Experimenting text generation with ChatGPT

Building a corpus of 40 texts based on:

- made up events

- real events (after 2021)

- 5 styles: factual (news agency), high quality paper, columnist, sensationalist, pro-Russian

Observations: Lots of additional content added to the prompt requirements / Adding disinformation

Problem: How to quantify or objectify it?

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.

Building a method & a tool

News article = ability to report the facts “as they are” in a structured text that commonly follows the inverted pyramid structure that arranges facts in a descending order of importance to answer the four or five “w-questions” –i.e., who, when, where, what and why–, in simple, concise and informative sentences (Pöttker, 2003; Kulkarni et al., 2022; Walters, 2017).

Considering narratives can take various forms in journalism, from breaking news to opinion, a means to compare 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, such as dots, question marks or exclamation marks, or ellipsis (Jurish & Würzner, 2013; Matusov et al., 2005)

The Information Disorder Level index (IDL index)

The assessment of automatically generated content is most reliable when done humanely. It involves critical thinking about the meaning of the information and its informative or disinformative value, which is related to the accuracy and reliability of the news. The IDL index is a human-based judgement metric. It is language-independent and was developed to evaluate content generated by large language model systems in the context of academic research.

Paste the generated AI text

In a devastating incident that unfolded on a quiet street in Norway, four individuals tragically lost their lives as a result of a collision involving a car and a bus. The incident, which took place on [insert street name], has left the local community in shock and mourning.

According to eyewitness accounts, the victims, whose identities have not yet been released pending notification of their families, were crossing the street when a car struck them. The force of the impact was so severe that it drew the attention of nearby witnesses, who immediately dialed emergency

+ Add prompt (optional)

Reference: add the prompt used to generate the text

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 nothing
It happened somewhere in Norway
Write a factual news article on this

Submit

The Information Disorders Level index (IDL index)

Rate each sentence, according to the following rules

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 a devastating turn of events, four individuals lost their lives in a tragic accident that occurred earlier today on a street in Norway.

True

☑ Check if this sentence is an opinion or a comment

2. 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.

True

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:

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

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

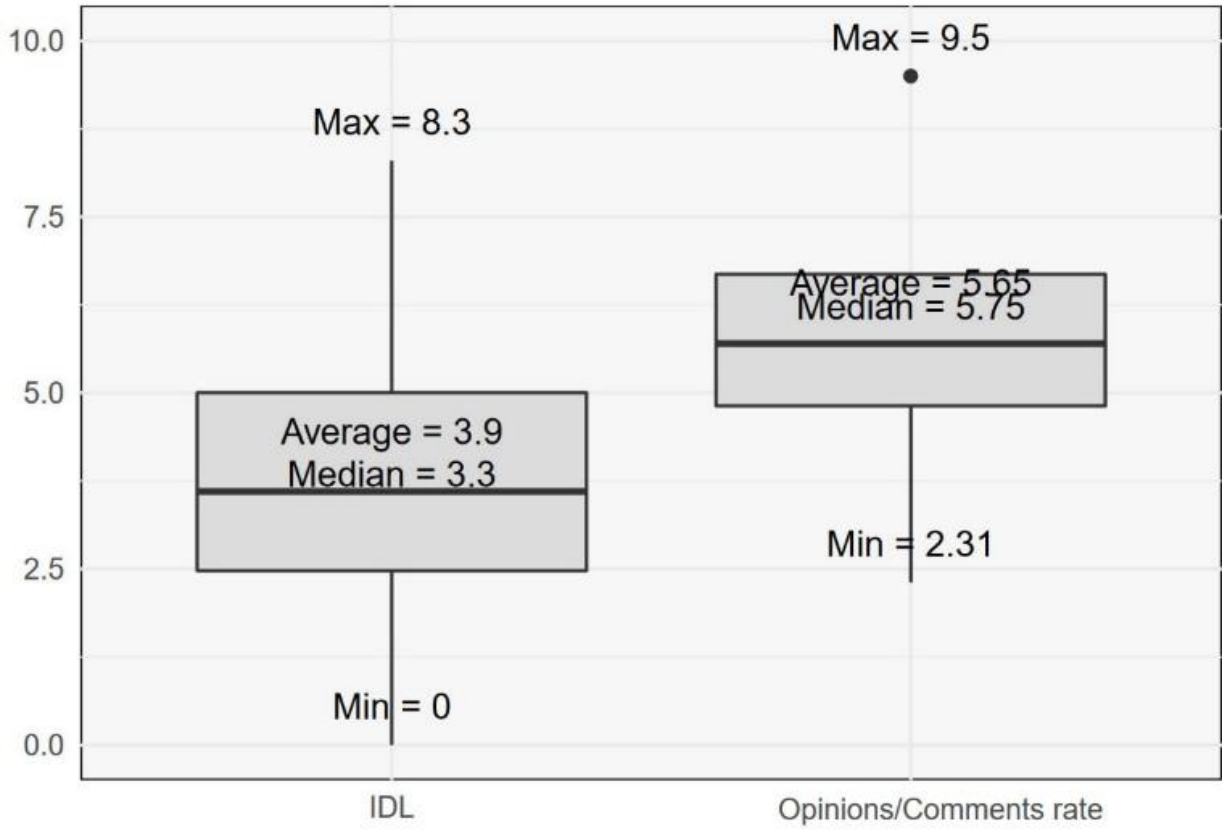
0
TRUE
INFORMATIVE

10
FALSE
DIS-MIS-MAL-INFORMATIVE

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

The tool also include the OCR (Opinion Comment Rate) to quantify it.

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

Limits and way forward

Human-based judgements involved several judges (more than one)

Corpus size, adding real events that occurred before 2021

Building training data based on the tool?

Integrating omissions?

Translating the interface?

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

https://laurence001.github.io/idl/

Lessons learned at this stage

From a research tool to a digital media literacy tool (help to understand to what extend GenAI also generates non-factual content)

The quality of the prompt is also related to the content quality: possible improvements in a journalistic context to mitigate “hallucinations” ?

Hallucinations or fictionalisation (coherence of the narrative)?

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

As ChatGPT 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