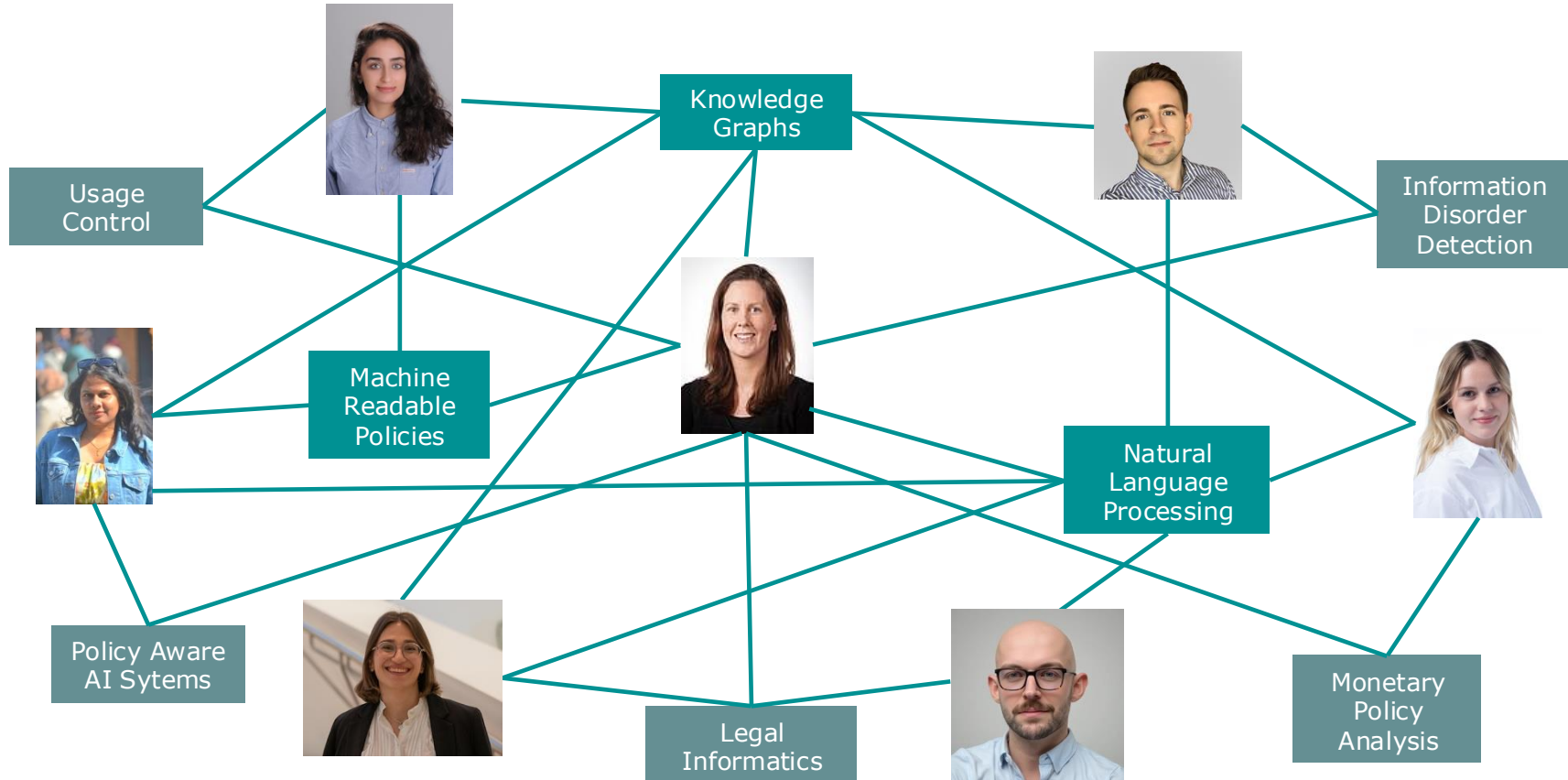


A Comparison and Critical Reflection of Information Disorder Detection Techniques



Sabrina Kirrane, 15th of September 2026
Rensselaer Polytechnic Institute (RPI) CS Colloquium





Institute for Data, Process and Knowledge Management

Axel Polleres: Data Management Group



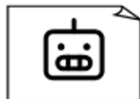
Efficient
Representations of



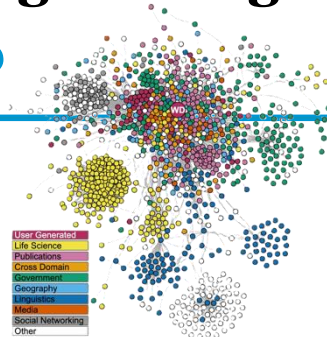
Query Languages and
Benchmarking



Scalable RDF
archiving systems



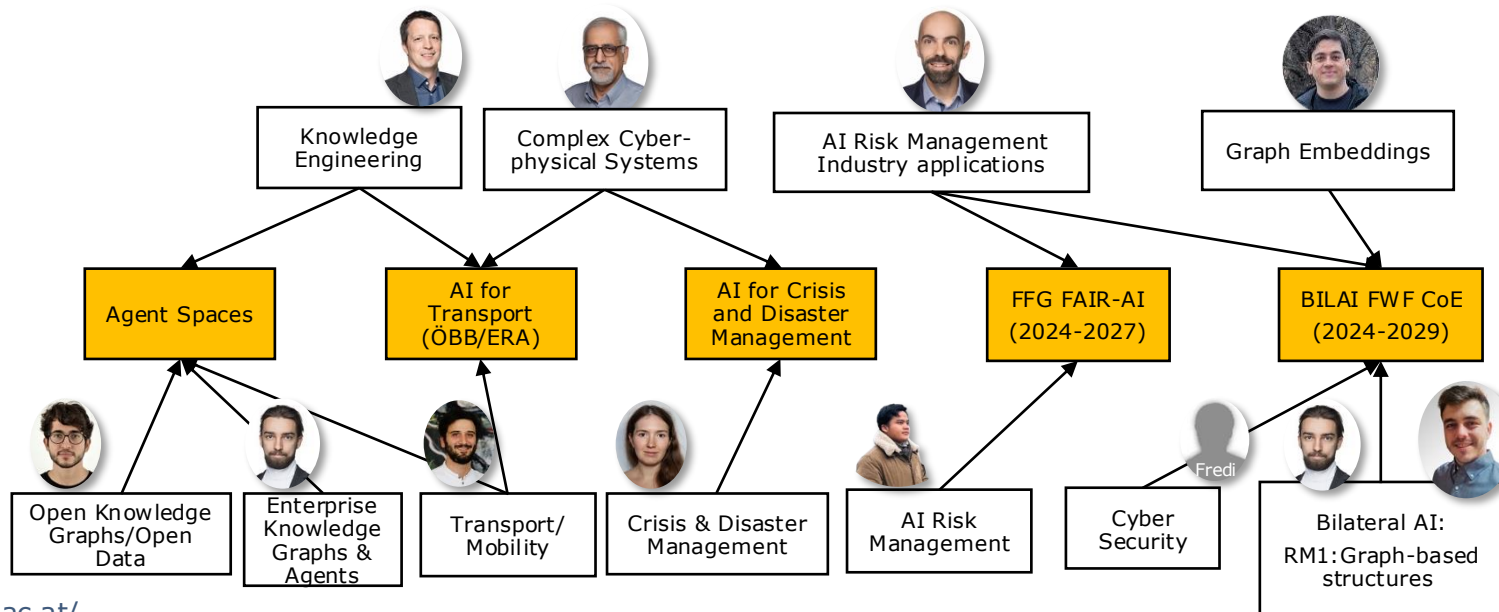
Agentic
LLMs



Research
Topics

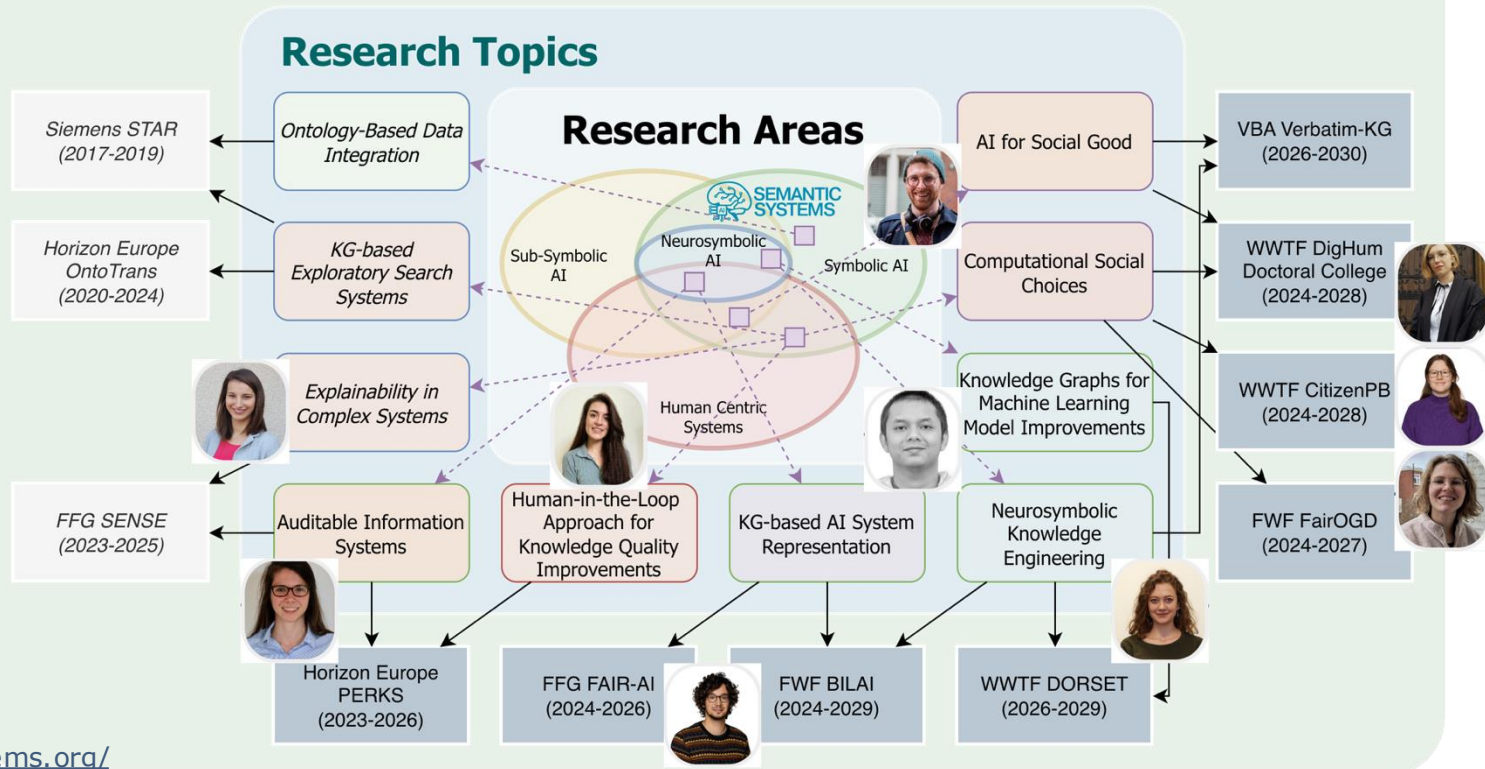
Research
Projects

Application
Domains





Research Projects



Information Disorders

Misinformation, Disinformation, Malinformation

FALSENESS

INTENT TO HARM

Misinformation

Unintentional mistakes such as inaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.

Disinformation

Fabricated and deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.

Malinformation

Deliberate publication of private information for personal or corporate interest, such as revenge porn. Deliberate change of Context, date or time of Genuine content.



<https://theonion.com/>



<https://waterfordwhispersnews.com/>



Information Disorders

Misinformation, Disinformation, Malinformation



... classifies misinformation and disinformation as:
the #1 global risk for the next two years
& the #5 global risk for the next ten years



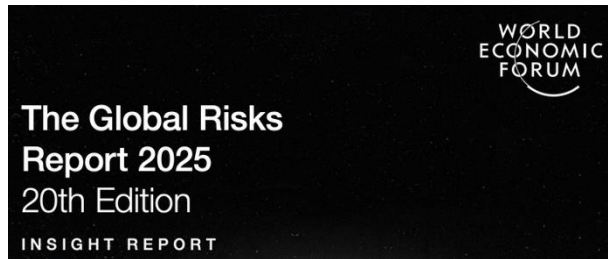
... places it again as #1 for the next two years & #5 for the next 10 years

Information Disorders

Misinformation, Disinformation, Malinformation



.... classifies misinformation and disinformation as:
the #1 global risk for the next two years
& the #5 global risk for the next ten years



.... places it again as #1 for the next two years & #5 for the next 10 years.



.... places it as #2 for the next two years & #4 for the next 10 years.

Short term (2 years)

1. Geoeconomic confrontation
2. Misinformation and disinformation
3. Societal polarization
4. Extreme weather events
5. State-based armed conflict
6. Cyber insecurity
7. Inequality
8. Erosion of human rights and/or of civic freedoms
9. Pollution
10. Involuntary migration or displacement

Long term (10 years)

1. Extreme weather events
2. Biodiversity loss and ecosystem collapse
3. Critical change to Earth systems
4. Misinformation and disinformation
5. Adverse outcomes of AI technologies
6. Natural resource shortages
7. Inequality
8. Cyber insecurity
9. Societal polarization
10. Pollution

Information Disorders

Cross-Data and Cross-Model Evaluation



Contents lists available at [ScienceDirect](#)

Information Fusion

journal homepage: www.elsevier.com/locate/inffus



A Comparison and Critical Reflection of Information Disorder Detection Techniques: Performing a Cross-Data and Cross-Model Evaluation

Mark Nicolas Gruensteidl*, Sabrina Kirrane

Institute for Complex Networks, Vienna University of Economics and Business, Welthandelsplatz 1, Vienna, 1020, Austria

<https://www.sciencedirect.com/science/article/pii/S1566253525008681?via%3Dihub>



Information Disorders

Datasets and Sources

Dataset Name	News Articles	Text	Theme or Domain	Size	Year
ISOT [49,50]	Real + Fake	Title + Text	Political/world news	44,898	2016-17
BuzzFeed [51]	Real + Fake	Title + Text	U.S. Pres. Elections	1,627	2016
LIAR [52]	Real + Fake	Statements	Political/world news	12,836	2007-16
NELA-GT-2018 [53]	Real + Misinfo	Title + Text	Political/world news	713,000	2018
NELA-GT-2022 [54]	Real + Misinfo	Title + Text	Political/world news	1,778,361	2022
WELFake [29]	Real + Fake	Title + Text	Political/world news	72,134	2016-17
COVID-19 [55]	Real + Fake	Statements	COVID-19 news	10,700	2020
EUvsDisinfo [56,57]	Fake	Title + Text	Political/EU news	9,965	2016-23
ENR [58]	Real	Title + Text	Political/EU news	510	2022-24
EMNAD [59,60]	Real	Title + Text	Political/EU news	18,000,000	2017-21

Information Disorders

Pre-trained and Fine-tuned Models

- Bidirectional encoder representations from transformers (**BERT**) is a language model introduced in October 2018 by **Google**.
- BERT is bidirectional, which means that it looks at context from both sides of a word.
- Masked Language Modelling (MLM): It works by masking words and letting the model predict them.
- Next Sentence Prediction (NSP): It predicts if one sentence follows another.

RoBERTa (2019, Facebook AI):

It removed (Next Sentence Prediction) NSP and was trained longer with more data for better performance.

DistilBERT (2019, Hugging Face):

It is a compressed version of BERT (40% smaller, 60% faster), with nearly same accuracy and great deployment.

ALBERT (2019, Google):

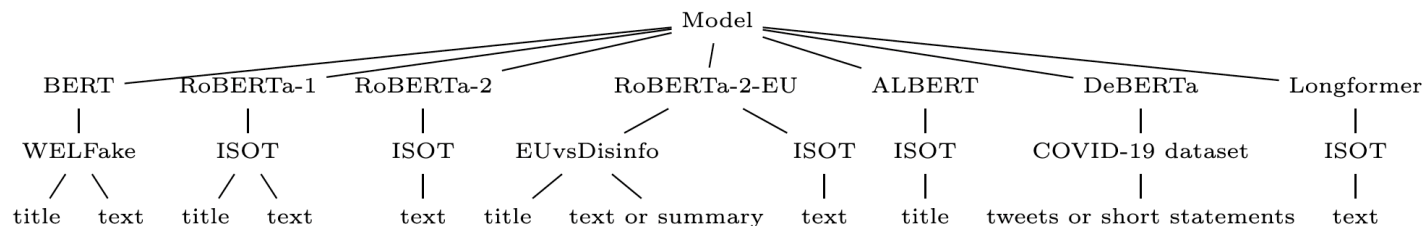
It is the "Lite" version of BERT with fewer parameters, shared weights, and factorised embeddings.

DeBERTa (2020, Microsoft):

Introduced disentangled attention and an enhanced mask decoder.

Existing Model Performance

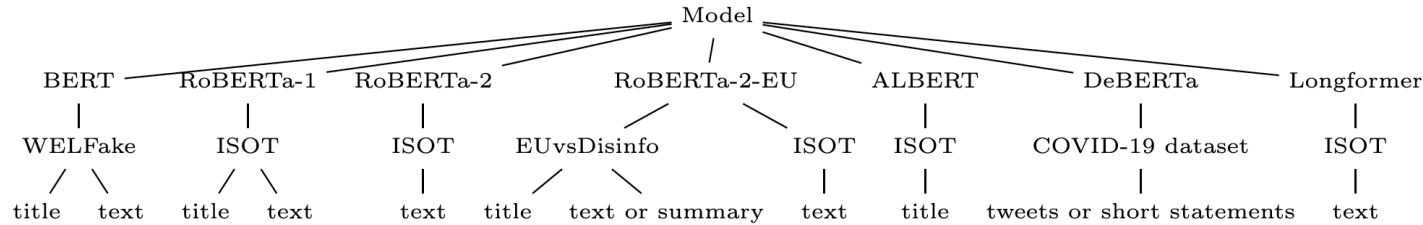
Pre-trained and Fine-tuned Models



Model Name	Architecture/Base-Model	Fine-tuning Data
BERT dima806/ fake-news-classifier [61]	BERT [62] bert-base-cased [63]	WELFake (title and text) [29]
RoBERTa-1 hamzab/ roberta-fake-news-classification [64]	RoBERTa [65] roberta-base [66]	CLMENTBISAILLON - fake-and-real-news-dataset; ISOT (title and text) [49,50,67]
RoBERTa-2 jy46604790/ Fake-News-Bert-Detect [68]		ISOT (potentially only text, without title) [49,50]
RoBERTa-2-EU winterForestStump/ Roberta-fake-news-detector (based on [68] and fine-tuned on EU news) [69]		EUvsDisinfo dataset (title and text or summary) [56,57] and ISOT (potentially only text, without title) [49,50]
ALBERT XSY/ albert-base-v2- fakenews-discriminator [70]	ALBERT [71] albert-base-v2 [72]	ISOT (only title, without text) [49,50]
DeBERTa Denyol/ FakeNews-deberta-base [73]	DeBERTa [74] deberta-base [75]	COVID-19 dataset (tweets or short statements) [55, 76]
Longformer AlexanderHolmes0/ fake-news-detector-long [77]	Longformer [39] longformer-base-4096 [38]	AlexanderHolmes0 - fake dataset (ISOT) (without title, only text) [49,50,78]

Existing Model Performance (Experiment 1)

Pre-trained and Fine-tuned Models



		Predicted	
Actual	0	TN	FP
	1	FN	TP

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Metrics:		Accuracy	Recall	Precision	F1 score
Models:	Sample size:				
BERT [61]	500	0.866 (0.84-0.898)	0.702 (0.642-0.757)	1.0 (-)	0.825 (0.77-0.859)
RoBERTa-1 [64]	500	1.0 (-)	1.0 (-)	1.0 (-)	1.0 (-)
RoBERTa-2 [68]	500	1.0 (-)	1.0 (-)	1.0 (-)	1.0 (-)
RoBERTa-2-EU [69]	500	1.0 (-)	0.0 (-)	0.0 (-)	0.0 (-)
ALBERT [70]	500	0.978 (0.962-0.99)	0.985 (0.962-0.996)	0.973 (0.944-0.992)	0.979 (0.964-0.988)
DeBERTa [73]	500	0.974 (0.957-0.986)	0.988 (0.965-0.996)	0.960 (0.932-0.981)	0.974 (0.957-0.986)
Longformer [77]	500	1.0 (-)	1.0 (-)	1.0 (-)	1.0 (-)

Confidence Intervals are in brackets

Existing Model Performance (Experiment 2)

BERT (WELfake title + text)

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
BERT [61]					
Sample size:	4000	500	500	500	500
Accuracy:	0.215 (0.204-0.227)	0.152 (0.122-0.18)	0.434 (0.394-0.472)	0.614 (0.57-0.648)	0.558 (0.518-0.596)
Recall:	0.0 (-)	0.087 (0.059-0.115)	0.173 (0.137-0.217)	0.106 (0.068-0.15)	0.115 (0.065-0.179)
Precision:	0.0 (-)	0.696 (0.561-0.804)	0.475 (0.382-0.57)	0.216 (0.133-0.311)	0.124 (0.075-0.196)
F1 score:	0.0 (-)	0.155 (0.108-0.199)	0.253 (0.192-0.313)	0.142 (0.092-0.214)	0.119 (0.074-0.177)

- The BERT model is trained on WELfake data and not a good match for the ISOT data
- This holds even when the sample count is raised up to the 4,000
- It also has poor performance over the other datasets

Existing Model Performance (Experiment 2)

RoBERTa-1 (ISOT title + text)

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
RoBERTa-1 [64]					
Sample size:	500	500	500	500	500
Accuracy:	1.0 (–)	0.78 (0.739-0.82)	0.522 (0.474-0.556)	0.564 (0.518-0.604)	0.488 (0.437-0.534)
Recall:	1.0 (–)	0.814 (0.776-0.851)	0.579 (0.523-0.633)	0.682 (0.617-0.763)	0.592 (0.507-0.685)
Precision:	1.0 (–)	0.931 (0.9-0.952)	0.569 (0.493-0.624)	0.377 (0.321-0.434)	0.275 (0.227-0.321)
F1 score:	1.0 (–)	0.868 (0.844-0.893)	0.574 (0.527-0.624)	0.486 (0.431-0.549)	0.376 (0.324-0.441)

- The RoBERTa-1 model obviously performs well on the data it is fine-tuned on
- Additionally it performs quite well on Buzzfeed, which contains similar articles
- It is less suitable for the LIAR and NELA datasets

Existing Model Performance (Experiment 2)

RoBERTa-2 (ISOT text)

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
RoBERTa-2 [68]					
Sample size:	500	500	500	500	500
Accuracy:	1.0 (-)	0.396 (0.35-0.445)	0.456 (0.413-0.498)	0.626 (0.588-0.668)	0.646 (0.608-0.682)
Recall:	1.0 (-)	0.336 (0.294-0.383)	0.076 (0.048-0.114)	0.172 (0.118-0.239)	0.115 (0.069-0.177)
Precision:	1.0 (-)	0.962 (0.927-0.987)	0.583 (0.429-0.76)	0.296 (0.194-0.393)	0.195 (0.119-0.297)
F1 score:	1.0 (-)	0.498 (0.457-0.553)	0.134 (0.087-0.191)	0.218 (0.151-0.302)	0.145 (0.089-0.215)

- The RoBERTa-2 model obviously performs well on the data it is fine-tuned on
- Note that it does not perform as well on the Buzzfeed data without the help of the titles

Existing Model Performance (Experiment 2)

RoBERTa-2-EU (ISOT text + EUvsDisinfo title + text)

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
RoBERTa-2-EU [69]					
Sample size:	500	500	4000	500	500
Accuracy:	0.858 (0.82-0.884)	0.552 (0.511-0.593)	0.443 (0.429-0.456)	0.554 (0.514-0.592)	0.52 (0.486-0.564)
Recall:	1.0 (-)	0.522 (0.477-0.569)	0.001 (0.0-0.003)	0.431 (0.359-0.524)	0.439 (0.367-0.516)
Precision:	0.775 (0.722-0.822)	0.955 (0.929-0.979)	0.5 (0.0-1.0)	0.322 (0.247-0.378)	0.255 (0.192-0.317)
F1 score:	0.873 (0.839-0.898)	0.675 (0.628-0.709)	0.002 (0.0-0.007)	0.368 (0.302-0.431)	0.322 (0.271-0.39)

- Using additional data from another source, creates noise and weakens the ability of the model to detect its initial fine-tuning data
- It has better performance on the Buzzfeed data but is generally worse for the other datasets

Existing Model Performance (Experiment 2)

ALBERT (ISOT title)

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
ALBERT [70]					
Sample size:	500	500	500	500	500
Accuracy:	0.644 (0.608-0.686)	0.474 (0.433-0.514)	0.472 (0.418-0.508)	0.544 (0.497-0.587)	0.562 (0.519-0.608)
Recall:	0.291 (0.241-0.347)	0.422 (0.377-0.469)	0.331 (0.281-0.391)	0.331 (0.262-0.417)	0.339 (0.27-0.426)
Precision:	0.934 (0.816-0.974)	0.974 (0.943-0.995)	0.541 (0.47-0.62)	0.283 (0.216-0.352)	0.249 (0.191-0.315)
F1 score:	0.444 (0.383-0.515)	0.588 (0.547-0.627)	0.411 (0.357-0.459)	0.305 (0.244-0.362)	0.287 (0.226-0.345)

The ALBERT model is fine-tuned on the titles only and does not perform as well when the titles and text are combined.

Existing Model Performance (Experiment 3)

RoBERTa-2-EU (ISOT text + EUvsDisinfo title + text)

Name (Abbreviation)	ENR ^a	EMNAD ^b	EUvsDisinfo ^c
ENR-EMN-EU Mix 1 (M1)	510	9,455	9,965
ENR-EMN-EU Mix 2 (M2)	510	510	1,020
EMN-EU Mix 3 (M3)	–	9,965	9,965
ENR-EU Mix 4 (M4)	510	–	510
ENR-BE Mix 5 (M5)	510	9,455 (BE)	9,965
ENR-BE Mix 6 (M6)	510	510 (BE)	1,020
Country Mix ([code]) ^d	–	9,965 ([code])	9,965

Existing Model Performance (Experiment 3)

RoBERTa-2-EU (ISOT text + EUvsDisinfo title + text)

Name (Abbreviation)	ENR ^a	EMNAD ^b	EUvsDisinfo ^c
ENR-EMN-EU Mix 1 (M1)	510	9,455	9,965
ENR-EMN-EU Mix 2 (M2)	510	510	1,020
EMN-EU Mix 3 (M3)	–	9,965	9,965
ENR-EU Mix 4 (M4)	510	–	510
ENR-BE Mix 5 (M5)	510	9,455 (BE)	9,965
ENR-BE Mix 6 (M6)	510	510 (BE)	1,020
Country Mix ([code]) ^d	–	9,965 ([code])	9,965

- Good performance when ENR and **European Multilingual News Articles Dataset (EMNAD)** balanced or **European Newsroom (ENR)** only.
- Performance deteriorates when EMNAD dominates.

Datasets:	M1	M2	M3	M4	M5	M6	BE	CY	IE	MT	UK
Sample size:	500	500	500	500	500	500	500	500	500	500	500
Seed: 42											
Accuracy:	0.746 (0.708-0.776)	0.89 (0.858-0.916)	0.728 (0.688-0.764)	0.99 (0.98-0.998)	0.714 (0.678-0.75)	0.866 (0.836-0.89)	0.74 (0.7-0.772)	0.788 (0.748-0.825)	0.672 (0.634-0.712)	0.734 (0.698-0.77)	0.472 (0.423-0.519)
Recall:	0.524 (0.474-0.588)	0.774 (0.728-0.815)	0.491 (0.436-0.558)	0.979 (0.955-0.996)	0.464 (0.401-0.519)	0.724 (0.672-0.779)	0.513 (0.451-0.572)	0.603 (0.547-0.655)	0.386 (0.331-0.442)	0.502 (0.438-0.554)	0.472 (0.428-0.511)
Precision:	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)	1.0 (–)
F1 score:	0.688 (0.638-0.735)	0.872 (0.834-0.907)	0.658 (0.607-0.712)	0.989 (0.981-0.998)	0.634 (0.585-0.7)	0.840 (0.806-0.883)	0.678 (0.628-0.724)	0.752 (0.703-0.793)	0.557 (0.492-0.608)	0.668 (0.602-0.709)	0.641 (0.596-0.677)

Existing Model Performance

Summary of our Findings

- The RoBERTa-1 & Longformer models performed better than the other models, however RoBERTa-1 performs better over the Buzzfeed data
- Longformer takes much longer than RoBERTa-1 due to the extended sequence length input limit of 4,096 tokens
- The RoBERTa-1, fine-tuned on EUvsDisinfo data, shows good performance on unseen ENR data, however it was not as strong on the EMNAD data
- ISOT is a good starting dataset for fine-tuning but needs further contextual data

Information Disorders

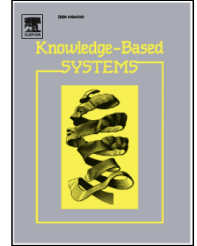
Explainability Evaluation



Contents lists available at [ScienceDirect](#)

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



An Evaluation of XAI Methods applied to Information Disorder Detection Models

Mark Nicolas Gruensteidl *, Sabrina Kirrane 

Institute for Complex Networks, Vienna University of Economics and Business, Welthandelsplatz 1, Vienna, 1020, Austria

<https://www.sciencedirect.com/science/article/pii/S0950705126000493?via%3Dihub>



Explainability Evaluation

Datasets and Data Sources

Dataset Name	Articles	Text	Theme or Domain	Size	Year
ISOT [104,105]	Real + Fake	Title + Text	Political/world news	44,898	2016-17
EUvsDISINFO [107,108]	Fake	Title + Text	Political/EU news	9,965	2016-23
ENR-EUvsDISINFO-Mix [32,109]	Real + Fake	Title + Text	Political/EU news	1,020	2022-24
EMNAD-EUvsDISINFO-Mix [32,110,111]	Real + Fake	Title + Text	Political/EU news	19,930	2017-21

- ENR-EUvsDISINFO-Mix = M4 is a balanced mix of 510 real **European Newsroom (ENR)** articles and 510 fake news articles
- EMNAD-EUvsDISINFO-Mix = M3 is a balanced mix of 9,965 real **European Multilingual News Articles Dataset (EMNAD)** articles and 9,965 fake news articles

Explainability Evaluation

Models and XAI

1. The jy46604790 RoBERTa model fine-tuned on the ISOT dataset

AKA RoBERTa-2 (ISOT text)

2. The winterForestStump RoBERTa model fine-tuned on ISOT and EUvsDISINFO data

AKA RoBERTa-2-EU (ISOT text + EUvsDisinfo title + text)

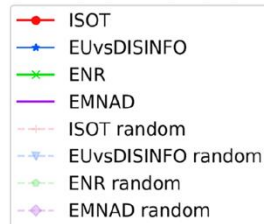
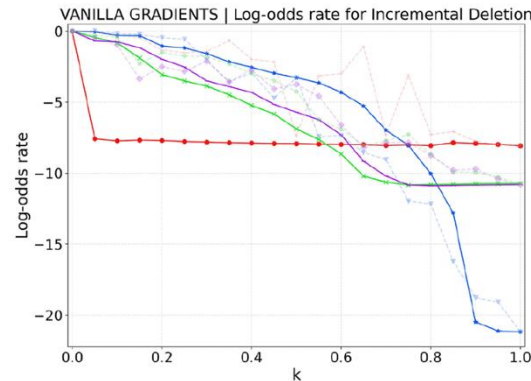
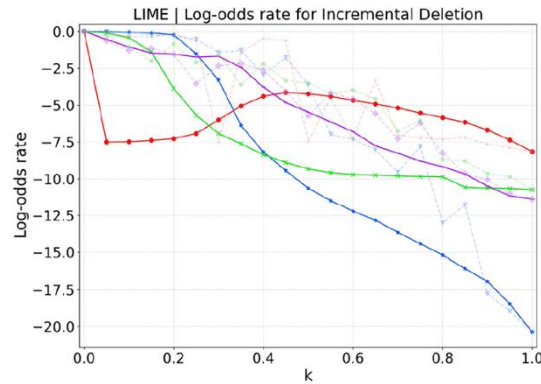
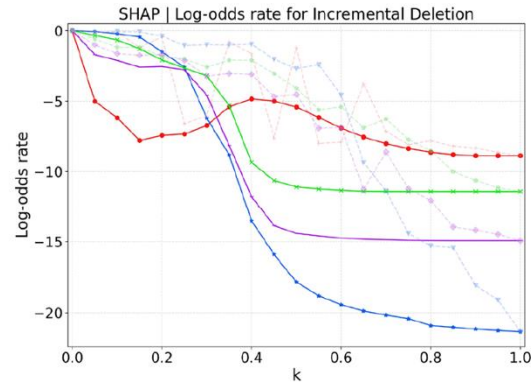
Method	Reference	Scope	Data Type	Foundation	Year
SHAP	[25]	Local/Global	Tabular, text, image	Shapely values	2017
LIME	[23]	Local	Tabular, text, image	Local linear approximation Computes the gradient of the class score with respect to the input	2016
Vanilla Gradients	[26,113,114]	Local	Tabular, text, image		2014/ 2017

1 Correctness

Incremental Deletion Log-odds rate

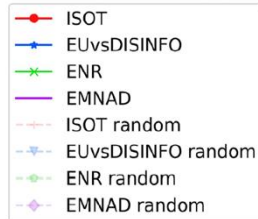
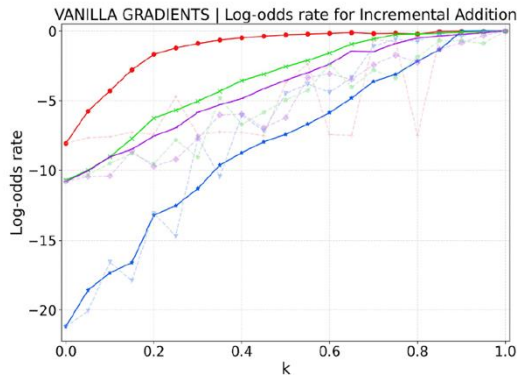
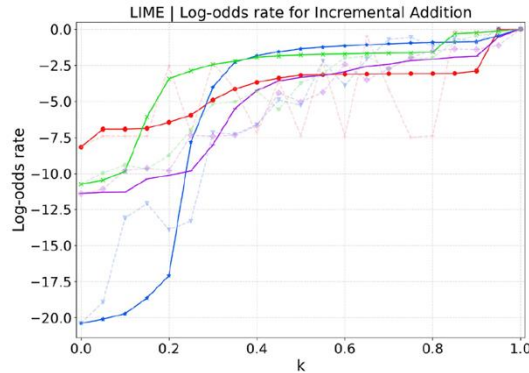
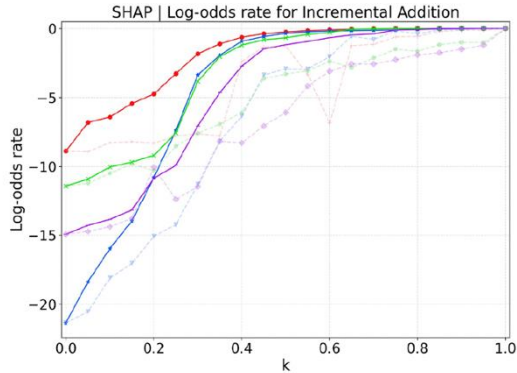
The trustfulness or faithfulness of an explanation.

Incremental Deletion deletes features iteratively based on the feature importance list.



1 Correctness

Incremental Addition Log-odds rate

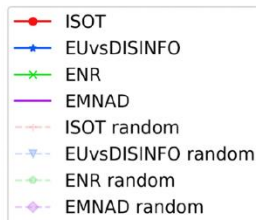
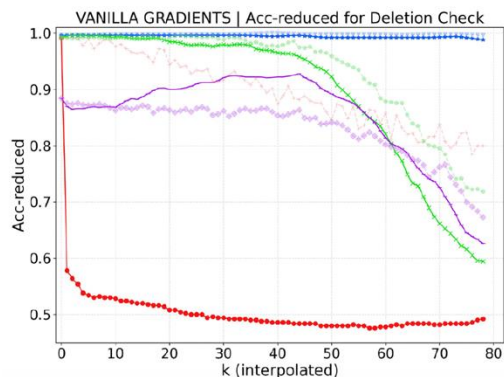
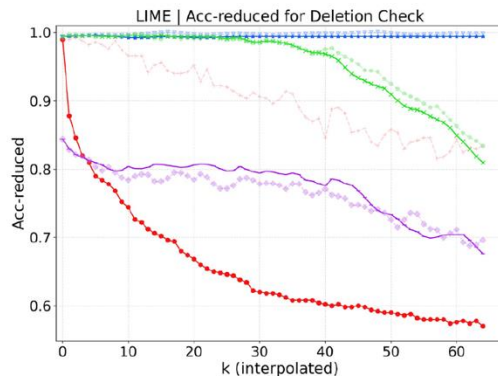
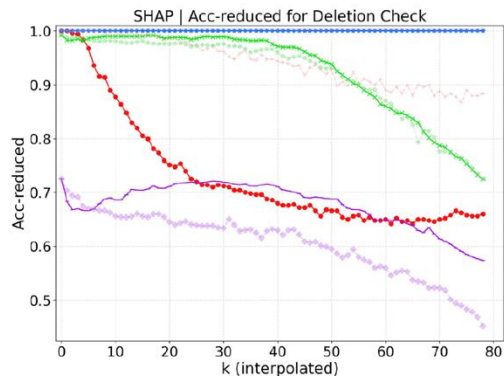


The trustfulness or faithfulness of an explanation.

Incremental Addition adds features iteratively based on the feature importance list.

2 Completeness

Deletion Check Acc-reduced

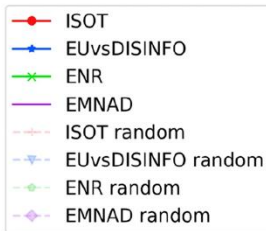
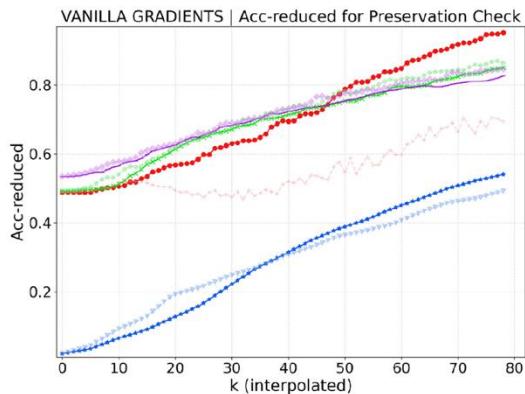
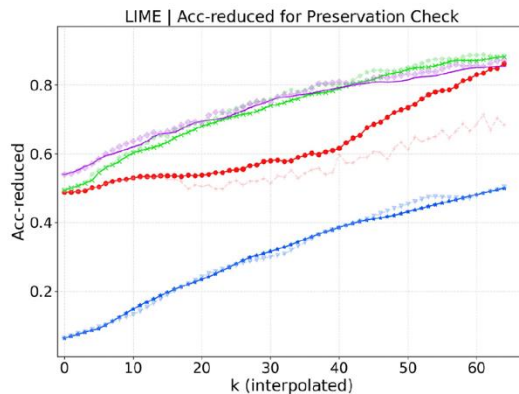
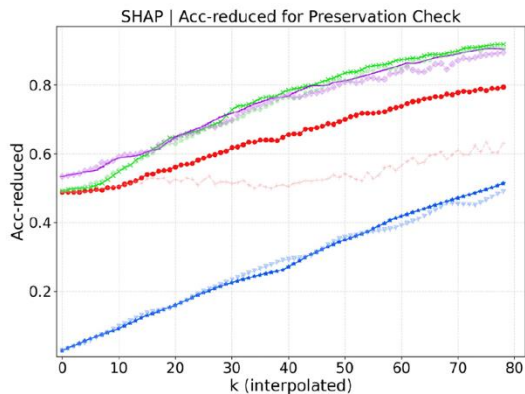


Determining to what degree aspects of the black box behaviour are shown in the explanation output

The Deletion Check is done by removing the top features and using the remaining ones as a model's input based on the local importance of features. This involves deleting ranked tokens from each article.

2 Completeness

Preservation Check Acc-reduced



Determining to what degree aspects of the black box behaviour are shown in the explanation output

The Preservation Check replaces progressively everything except the top features based on the local importance of features. This involves adding ranked tokens from each article.

3 Continuity

Fidelity for Slight Variations

The generalisability and stability of an explanation.

ENR	0%	5%	10%	15%	20%
Accuracy:	0.99 (0.976-0.996)	0.984 (0.972-0.994)	0.982 (0.964-0.991)	0.982 (0.968-0.992)	0.978 (0.962-0.988)
Recall:	0.98 (0.955-0.995)	0.967 (0.938-0.988)	0.967 (0.932-0.984)	0.963 (0.938-0.983)	0.955 (0.917-0.975)
Precision:	1.0 (-)	1.0 (-)	0.996 (0.983, 1.0)	1.0 (-)	1.0 (-)
F1 score:	0.99 (0.979, 0.996)	0.983 (0.971, 0.992)	0.981 (0.967, 0.991)	0.981 (0.968, 0.994)	0.977 (0.966, 0.99)
EMNAD	0%	5%	10%	15%	20%
Accuracy:	0.728 (0.69-0.764)	0.722 (0.688-0.76)	0.698 (0.66-0.732)	0.670 (0.621-0.706)	0.672 (0.635-0.713)
Recall:	0.491 (0.426-0.545)	0.479 (0.43-0.545)	0.434 (0.365-0.479)	0.382 (0.329-0.442)	0.386 (0.339-0.451)
Precision:	1.0 (-)	1.0 (-)	1.0 (-)	1.0 (-)	1.0 (-)
F1 score:	0.658 (0.604-0.711)	0.648 (0.591-0.706)	0.606 (0.557-0.66)	0.553 (0.503-0.609)	0.557 (0.497-0.62)

Fidelity for Slight Variations compares original and slightly perturbed inputs in order to assess whether a model's results lead to similar predictions

4 Consistency

Jaccard Similarity of LIME

Consistency deals with how deterministic an XAI method is.

Jaccard Similarity of LIME, n = 500.

Model:	jy46604790	winterForestStump		
Dataset:	ISOT	EUvsDISINFO	ENR	EMNAD
Round #1	–	–	–	–
Round #2	0.99322	0.99736	0.99921	0.99151
Round #3	0.99309	0.99724	0.99928	0.99177
Round #4	0.99317	0.99736	0.99922	0.99070
Round #5	0.99312	0.99639	0.99911	0.99092
Round #6	0.99232	0.99732	0.99914	0.99093
Round #7	0.99253	0.99700	0.99931	0.99156
Round #8	0.99261	0.99728	0.99916	0.99083
Round #9	0.99334	0.99700	0.99908	0.99123
Round #10	0.99302	0.99736	0.99927	0.99059
Time:	18 m	9 m	15 m	16 m

When randomly reinitialising a classification model multiple times, the Jaccard Similarity between feature score importance is used to assess the model's implementation invariance.

Lime is the only XAI method of our portfolio that produces inconsistent weights for the selected features.

5 Contrastivity

KL Divergence

The ability of an XAI method to make its explanations discriminative or separable with respect to different events or objectives.

KL Divergence (in *bits*, best metrics in *bold*), n = 500.

Model:	jy46604790	winterForestStump		
Dataset:	ISOT	EUvsDISINFO	ENR	EMNAD
SHAP				
Random #1	0.07884	0.02470	0.00686	0.04290
Random #2	0.07894	0.02498	0.00692	0.04290
Random #3	0.07886	0.02502	0.00692	0.04282
LIME				
Random #1	0.00031	0.00016	0.00011	0.00019
Random #2	0.00029	0.00019	0.00013	0.00020
Random #3	0.00029	0.00018	0.00011	0.00021
Vanilla Gradients				
Random #1	0.01355	0.04720	0.02416	0.02435
Random #2	0.01356	0.04720	0.02412	0.02435
Random #3	0.01358	0.04712	0.02414	0.02433

KL Divergence allows us to quantify contrastivity in regard to the importance of feature distributions. This shows how well a model can differentiate two classes (e.g., positive and negative) from each other.

The smaller the KL Divergence and the closer it is to zero the more alike two distributions are.

- SHAP is the most advanced/suitable method for long-text IDD tasks
 - But it is only partially effective in identifying supporting features
 - Struggles with unseen data and is slow

- Vanilla Gradients performed well as a starting point and preliminary tool
 - Fast and consistent explanations
 - But suffers from the gradient saturation problem, which limits standalone use

- LIME is not recommended for this domain
 - Excessive resource demand
 - Technical limits and boundaries for the domain of long-text IDD

Information Disorders

Misinformation, Disinformation, Malinformation

FALSENESS

INTENT TO HARM

Misinformation

Unintentional mistakes such as inaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.

Disinformation

Fabricated and deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.

Malinformation

Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of Genuine content.



<https://theonion.com/>



<https://waterfordwhispersnews.com/>



Mark Nicolas Gruensteidl¹, Sabrina Kirrane¹
¹Institute for Complex Networks, WU, Vienna, Austria

A Knowledge Graph of Information Disorder Actors, Agendas, and Events

https://penni.wu.ac.at/papers/ESWC2026_poster_GruensteidlKirrane.pdf

Information Disorder Knowledge Graph

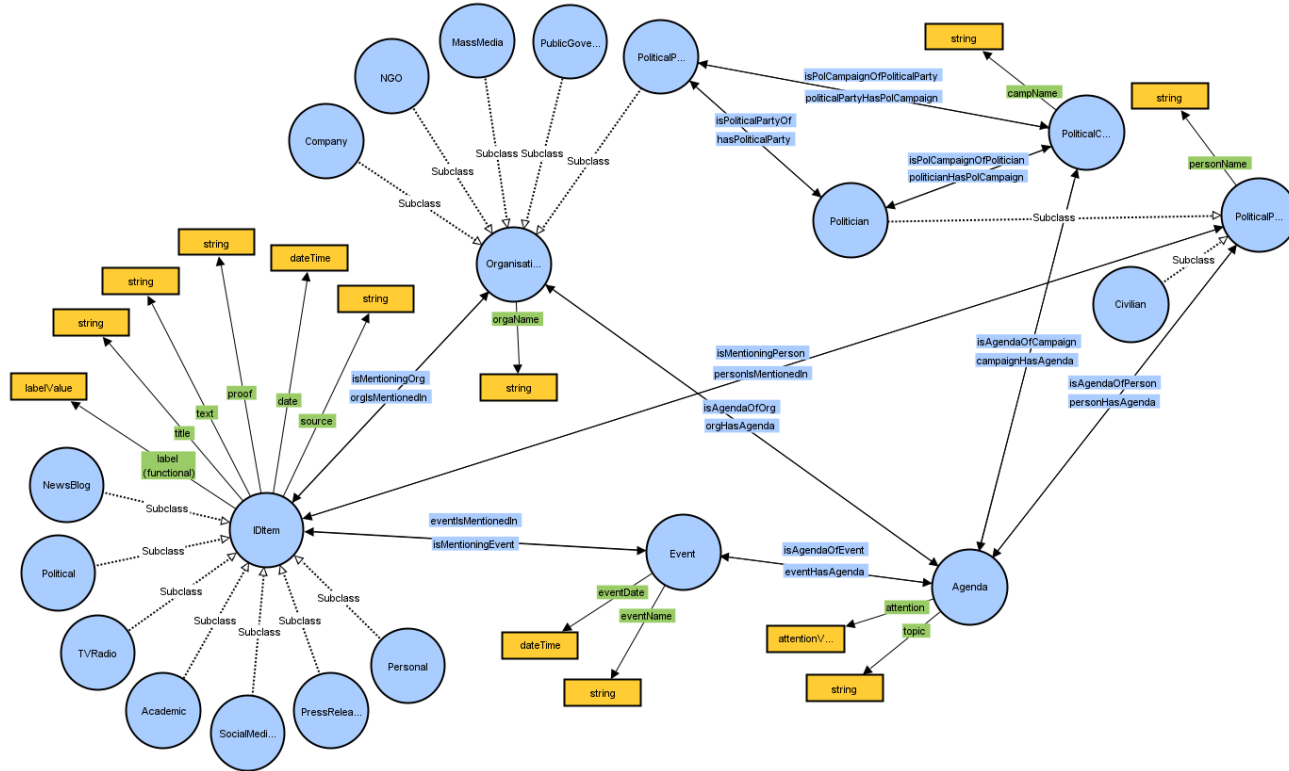
Data

Dataset Name	News Articles	Text	Theme or Domain	Size	Year
PolitiFact-Oslo Corpus [7]	Multi-label	Title+Text	Political/world news	2,745	2019-22
LIAR2 [10]	Multi-label	Stateme	Political/world news	22,962	2007-23
BuzzFeed [6]	Multi-label	Title+Text	U.S. Pres. Elections	1,627	2016

- *PolitiFact-Oslo Corpus* & *LIAR2*: True, Mostly True, Half True, Mostly False, False, and Pants on Fire
- *BuzzFeed*: mostly true, mostly false, mixture of true and false, and no factual content

Information Disorder Knowledge Graph

Actors, Agendas, and Events



Information Disorder Knowledge Graph

Competency Questions

CQ Pool

No	ID	Aa CQ	Reject	Accept
358		Which factcheck labels are most frequently assigned to statements in [genre X]?	Reject	Accept
357		What is the original body text of the claim published at [URL X] in [genre Y]? 1	Reject	Accept
356		Which sources published statements on [date X] that were later factchecked?	Reject	Accept
355		What is the original headline of the claim evaluated with the factcheck headline [headline X]? 1	Reject	Accept
354		Which claims made between [date X] and [date Y] have a factcheck label of [label Z]?	Reject	Accept
353		What is the combined label for the claim published in [genre X] with the original headline [headline Y]? 3	Reject	Accept
352		Which URLs point to original statements made by [source X]? 2	Reject	Accept
351		What original genre is most commonly associated with claims rated as [label X]? 2	Reject	Accept



<https://scalewise.com/resources/chatgpt-vs-gemini-vs-claude/>

- Builds a more comprehensive information disorder ontology describing (political) actors, agendas, and events
- Automate the knowledge graph population using structured and unstructured information
- Dig deeper into:
 - Intent or motives of relevant actors
 - Role of campaigns and events

Excuses to come visit us in Vienna

- 22nd International Conference on Artificial Intelligence and Law (ICAAIL) July 5-9, 2027
- Declarative AI September 15-22, 2027
 - The 11th International Joint Conference on Rules and Reasoning
 - The 23rd Reasoning Web Summer School
- 22nd International Conference on Neurosymbolic Learning and Reasoning September 13-15, 2028





VIENNA UNIVERSITY OF
ECONOMICS AND BUSINESS

**Department of Information Systems
& Operations Management**

Institute for Complex Networks
Welthandelsplatz 1, 1020 Vienna, Austria

Dr. Sabrina Kirrane

T +43-1-313 36-4494
F +43-1-313 36-90 4494
sabrina.kirrane@wu.ac.at
www.wu.ac.at
www.sabrinakirrane.com



This work is partially funded by
the FWF Austrian Science Fund
[10.55776/COE12]

Existing Model Performance (Experiment 4)

DEBERTa (Covid19 statements)

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
DeBERTa [73]					
Sample size:	4000	1342 (max)	2000	2000	2000
Accuracy:	0.518 (0.501-0.533)	0.103 (0.089-0.12)	0.448 (0.426-0.472)	0.675 (0.654-0.693)	0.718 (0.695-0.735)
Recall:	0.0 (-)	0.0 (-)	0.042 (0.031-0.053)	0.002 (0.0-0.009)	0.002 (0.0-0.007)
Precision:	0.0 (-)	0.0 (-)	0.702 (0.598-0.801)	0.167 (0.0-0.743)	0.167 (0.0-0.701)
F1 score:	0.0 (-)	0.0 (-)	0.078 (0.054-0.097)	0.003 (0.0-0.012)	0.004 (0.0-0.019)

- Used for maximum accuracy for complex classification, question answering, or natural language inference tasks
- The original fine-tuning data consists of short COVID-19 news statements or tweets

Existing Model Performance (Experiment 4)

Longformer (ISOT text)

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
Longformer [77]					
Sample size:	500	500	500	500	500
Accuracy:	1.0 (-)	0.654 (0.616-0.692)	0.458 (0.414-0.502)	0.484 (0.451-0.527)	0.504 (0.464-0.553)
Recall:	1.0 (-)	0.796 (0.682-0.88)	0.878 (0.832-0.916)	0.407 (0.358-0.455)	0.484 (0.427-0.533)
Precision:	1.0 (-)	0.21 (0.165-0.277)	0.444 (0.402-0.486)	0.736 (0.676-0.803)	0.759 (0.705-0.819)
F1 score:	1.0 (-)	0.332 (0.26-0.406)	0.59 (0.545-0.63)	0.524 (0.476-0.573)	0.591 (0.544-0.629)

- Longformer is as effective as the RoBERTa model on seen data
- The Longformer model's evaluation numbers, except when applied on BuzzFeed articles, are a bit better overall

Model performance on seen and unseen data (best accuracy & F1 metrics in **bold**, confidence intervals in brackets: *confidence level 0.95 and number of resamples 200*).

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
BERT [61]					
Sample size:	4000	500	500	500	500
Accuracy:	0.215 (0.204-0.227)	0.152 (0.122-0.18)	0.434 (0.394-0.472)	0.614 (0.57-0.648)	0.558 (0.518-0.596)
Recall:	0.0 (-)	0.087 (0.059-0.115)	0.173 (0.137-0.217)	0.106 (0.068-0.15)	0.115 (0.065-0.179)
Precision:	0.0 (-)	0.696 (0.561-0.804)	0.475 (0.382-0.57)	0.216 (0.133-0.311)	0.124 (0.075-0.196)
F1 score:	0.0 (-)	0.155 (0.108-0.199)	0.253 (0.192-0.313)	0.142 (0.092-0.214)	0.119 (0.074-0.177)
RoBERTa-1 [64]					
Sample size:	500	500	500	500	500
Accuracy:	1.0 (-)	0.78 (0.739-0.82)	0.522 (0.474-0.556)	0.564 (0.518-0.604)	0.488 (0.437-0.534)
Recall:	1.0 (-)	0.814 (0.776-0.851)	0.579 (0.523-0.633)	0.682 (0.617-0.763)	0.592 (0.507-0.685)
Precision:	1.0 (-)	0.931 (0.9-0.952)	0.569 (0.493-0.624)	0.377 (0.321-0.434)	0.275 (0.227-0.321)
F1 score:	1.0 (-)	0.868 (0.844-0.893)	0.574 (0.527-0.624)	0.486 (0.431-0.549)	0.376 (0.324-0.441)
RoBERTa-2 [68]					
Sample size:	500	500	500	500	500
Accuracy:	1.0 (-)	0.396 (0.35-0.445)	0.456 (0.413-0.498)	0.626 (0.588-0.668)	0.646 (0.608-0.682)
Recall:	1.0 (-)	0.336 (0.294-0.383)	0.076 (0.048-0.114)	0.172 (0.118-0.239)	0.115 (0.069-0.177)
Precision:	1.0 (-)	0.962 (0.927-0.987)	0.583 (0.429-0.76)	0.296 (0.194-0.393)	0.195 (0.119-0.297)
F1 score:	1.0 (-)	0.498 (0.457-0.553)	0.134 (0.087-0.191)	0.218 (0.151-0.302)	0.145 (0.089-0.215)
RoBERTa-2-EU [69]					
Sample size:	500	500	4000	500	500
Accuracy:	0.858 (0.82-0.884)	0.552 (0.511-0.593)	0.443 (0.429-0.456)	0.554 (0.514-0.592)	0.52 (0.486-0.564)
Recall:	1.0 (-)	0.522 (0.477-0.569)	0.001 (0.0-0.003)	0.431 (0.359-0.524)	0.439 (0.367-0.516)
Precision:	0.775 (0.722-0.822)	0.955 (0.929-0.979)	0.5 (0.0-1.0)	0.322 (0.247-0.378)	0.255 (0.192-0.317)
F1 score:	0.873 (0.839-0.898)	0.675 (0.628-0.709)	0.002 (0.0-0.007)	0.368 (0.302-0.431)	0.322 (0.271-0.39)
ALBERT [70]					
Sample size:	500	500	500	500	500
Accuracy:	0.644 (0.608-0.686)	0.474 (0.433-0.514)	0.472 (0.418-0.508)	0.544 (0.497-0.587)	0.562 (0.519-0.608)
Recall:	0.291 (0.241-0.347)	0.422 (0.377-0.469)	0.331 (0.281-0.391)	0.331 (0.262-0.417)	0.339 (0.27-0.426)
Precision:	0.934 (0.816-0.974)	0.974 (0.943-0.995)	0.541 (0.47-0.62)	0.283 (0.216-0.352)	0.249 (0.191-0.315)
F1 score:	0.444 (0.383-0.515)	0.588 (0.547-0.627)	0.411 (0.357-0.459)	0.305 (0.244-0.362)	0.287 (0.226-0.345)

DeBERTa and Longformer models' performance on seen and unseen data (best accuracy & F1 metrics in *bold*, confidence intervals in brackets: *confidence level 0.95 and number of resamples 200*).

Datasets:	ISOT	BuzzFeed	LIAR	NELA-GT-2022	NELA-GT-2018
DeBERTa [73]					
Sample size:	4000	1342 (max)	2000	2000	2000
Accuracy:	0.518 (0.501-0.533)	0.103 (0.089-0.12)	0.448 (0.426-0.472)	0.675 (0.654-0.693)	0.718 (0.695-0.735)
Recall:	0.0 (-)	0.0 (-)	0.042 (0.031-0.053)	0.002 (0.0-0.009)	0.002 (0.0-0.007)
Precision:	0.0 (-)	0.0 (-)	0.702 (0.598-0.801)	0.167 (0.0-0.743)	0.167 (0.0-0.701)
F1 score:	0.0 (-)	0.0 (-)	0.078 (0.054-0.097)	0.003 (0.0-0.012)	0.004 (0.0-0.019)
Longformer [77]					
Sample size:	500	500	500	500	500
Accuracy:	1.0 (-)	0.654 (0.616-0.692)	0.458 (0.414-0.502)	0.484 (0.451-0.527)	0.504 (0.464-0.553)
Recall:	1.0 (-)	0.796 (0.682-0.88)	0.878 (0.832-0.916)	0.407 (0.358-0.455)	0.484 (0.427-0.533)
Precision:	1.0 (-)	0.21 (0.165-0.277)	0.444 (0.402-0.486)	0.736 (0.676-0.803)	0.759 (0.705-0.819)
F1 score:	1.0 (-)	0.332 (0.26-0.406)	0.59 (0.5450.63)	0.524 (0.476-0.573)	0.591 (0.544-0.629)