Valerio La Gatta

Knowledge-informed Disinformation Mining: From Fact-Checking to Content Moderation

Tutor: prof. Vincenzo Moscato

Cycle: XXXVI Year: 2023

My background

- MSc degree in Computer Engineering from University of Naples Federico II
- Group: Pattern Analysis and Intelligence Computation for mUltimedia System (PICUSLab)
- PhD date: 01/11/2020 - 31/10/2023
- Scholarship type: Unina
- Period abroad: University of Sourthern California, Los Angeles, June – December 2022

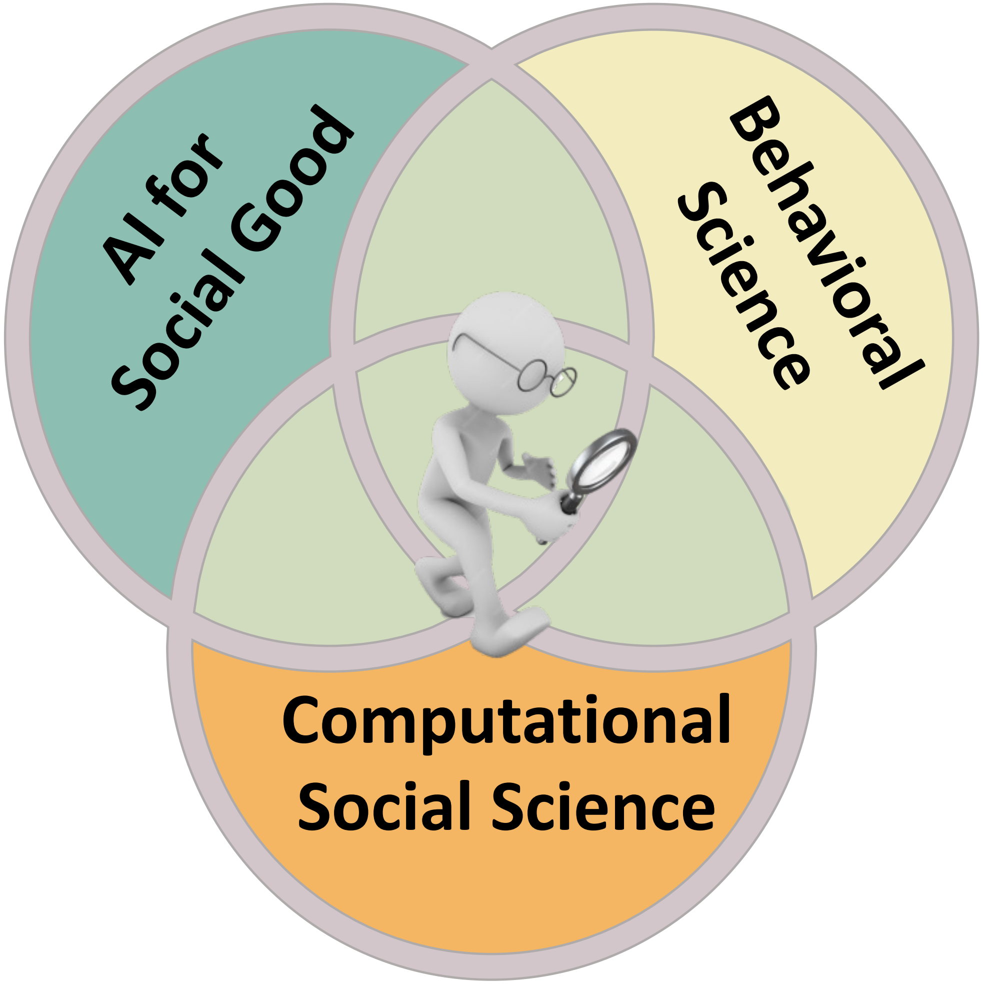
Summary of study activities

- Ad hoc PhD courses / schools
  - Scientific Programming and Visualization with Python
  - Statistical data analysis for science and engineering research
  - Data science for patient records analysis
  - Strategic Orientation for STEM research & writing
  - Big Data Architecture and Analytics
  - AIRO PhD School 2021 and 5th AIRO-Young Workshop
- Courses attended from MSc curricula
  - Natural Language Processing
  - Web and Real Time Communication Systems
- Attended Conferences
  - 34th ACM International Conference on Hypertext and Social Media (HT'2023)
  - The 2nd Italian Conference on Big Data and Data Science (ITADATA'2023).

Research area(s)

- Disinformation and Online Harms in Socio-Technical Systems
  - Fact-Checking, Disinformation Detection and Early Prevention
  - Malicious Behaviors on Social Media
  - Content Moderation across Social Media Platforms

Research Approach



Research products

[J1]	V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, *Covid-19 sentiment analysis based on Tweets,* **IEEE Intelligent Systems,** vol. 38 (3), pp. 51-55, 2023, DOI: 10.1109/MIS.2023.3239180
[J2]	T. Chakraborty, V. La Gatta, V. Moscato, G. Sperlì, *Information retrieval algorithms and neural ranking models to detect previously fact-checked information,* **Neurocomputing,** vol. 557, 2023, DOI: 10.1016/j.neucom.2023.126680
[J3]	A. Ferraro, A. Galli, V. La Gatta, M. Postiglione, *Benchmarking Open Source and Paid Services for Speech to Text: An Analysis of Quality and Input Variety,* **Frontiers in Big Data,** vol. 6, 2023, DOI: 10.3389/fdata.2023.1210559
[J4]	V. La Gatta, V. Moscato, M. Pennone, M. Postiglione, G. Sperlì, *Music Recommendation via Hypergraph Embedding,* **IEEE Transactions on Neural Networks and Learning Systems,** vol. 34 (10), pp. 7887-7899, 2022, DOI: 10.1109/TNNLS.2022.3146968

Research products

[J5]	A. Barducci, S. Iannaccone, V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, S. Zavota, *An end-to-end framework for information extraction from Italian resumes,* **Expert Systems with Applications,** vol. 210, 2022, DOI: 10.1016/j.eswa.2022.118487
[J6]	V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, *CASTLE: Cluster-aided space transformation for local explanations,* **Expert Systems with Applications,** vol. 179, 2021, DOI: 10.1016/j.eswa.2021.115045
[J7]	V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, *PASTLE: Pivot-aided space transformation for local explanations,* **Pattern Recognition Letters,** vol. 149, pp. 67-74, 2021, DOI: 10.1016/j.patrec.2021.05.018
[J8]	V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, *An Epidemiological Neural Network Exploiting Dynamic Graph Structured Data Applied to the COVID-19 Outbreak,* **IEEE Transactions on Big Data,** vol. 7 (1), pp. 45-55, 2020, DOI: 10.1109/TBDATA.2020.3032755

Research products

[C1]	V. La Gatta, L. Luceri, F. Fabbri, E. Ferrara *The Interconnected Nature of Online Harm and Moderation: Investigating the Cross-Platform Spread of Harmful Content between YouTube and Twitter,* **34th ACM International Conference on Hypertext and Social Media (HT2023),** Rome, Italy, Sept. 2023, ACM, DOI: 10.1145/3603163.3609058 *Nomination for the ACM Hypertext Ted Nelson Award*
[C2]	M. Postiglione, G. Esposito, R. Izzo, V. La Gatta, V. Moscato, R. Piccolo *Harnessing multi-modality and expert knowledge for adverse events prediction in clinical notes,* **International Conference on Image Analysis and Processing (ICIAP2023)** Udine, Italy, Sept. 2023
[C3]	V. La Gatta, C. Wei, L. Luceri, F. Pierri, E. Ferrara *Retrieving false claims on Twitter during the Russia-Ukraine conflict,* **Companion Proceedings of the ACM The Web Conference 2023 (WWW2023),** Austin, TX, USA, Apr. 2023, ACM, DOI: 10.1145/3543873.3587571

Research products

[C4]	G. Riccio, A. Romano, A. Korsun, M. Cirillo, M. Postiglione, V. La Gatta, A. Ferraro, A. Galli, V. Moscato *Healthcare Data Summarization via Medical Entity Recognition and Generative AI,* **The 2nd Italian Conference on Big Data and Data Science (ITADATA2023),** Naples, Italy, Sept. 2023, CEUR Workshop Proceedings
[C5]	A. Ferraro, A. Galli, V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, F. Amato *HEMR: Hypergraph Embeddings for Music Recommendation,* **Symposium on Advanced Database System (SEBD2023),** Galzignano Terme, Italy, July 2023, CEUR Workshop Proceedings
[C6]	A. Ferraro, A. Galli, V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, F. Moscato *Unsupervised Anomaly Detection in Predictive Maintenance using Sound Data,* **Symposium on Advanced Database System (SEBD2023),** Galzignano Terme, Italy, July 2023, CEUR Workshop Proceedings
[C7]	A. Ferraro, A. Galli, V. La Gatta, M. Postiglione *A Deep Learning pipeline for Network Anomaly Detection based on Autoencoders,* **IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE2022),** Rome, Italy, Oct. 2022, IEEE, DOI: 10.1109/MetroXRAINE54828.2022.9967598

Research products

[P1]	V. La Gatta, L. De Cegli, V. Moscato, G. Sperlì *From Single-Task to Multi-Task: Unveiling the Dynamics of Knowledge Transfers in Disinformation Detection,* Submitted to **ACM The Web Conference 2024 (WWW2024)**
[P2]	B. Grasso, V. La Gatta, V. Moscato, G. Sperlì *KERMIT: Knowledge-EmpoweRed Model In harmful meme deTection,* Submitted to **Information Fusion**
[P3]	A. Ferraro, A. Galli, M. Gallo, V. La Gatta, M. Postiglione, V. Moscato *ExpLusion: Explanation-driven Late Fusion for enhanced production process monitoring* Submitted to **Journal of Intelligent Manufacturing**
[P4]	R. Formisano, V. La Gatta, V. Moscato, G. Sperlì *A Multimodal Retrieval System for Previously Fact-checked Information Detection,* Submitted to **Information Systems**
[P5]	A. Galli, V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì *Interpretability in AI-based Behavioral Malware Detection Systems* Submitted to **Computers & Security**

Research products

[P6]	V. La Gatta, M. Postiglione, V. Moscato, G. Sperlì *An eXplainable Artificial Intelligence methodology on Big Data Architecture* Submitted to **Cognitive Computation**
[P7]	V. La Gatta, M. Postiglione, G. Sperlì *A novel augmentation strategy for credit scoring modeling* Submitted to **Engineering Applications of Artificial Intelligence**

PhD thesis – Research Context

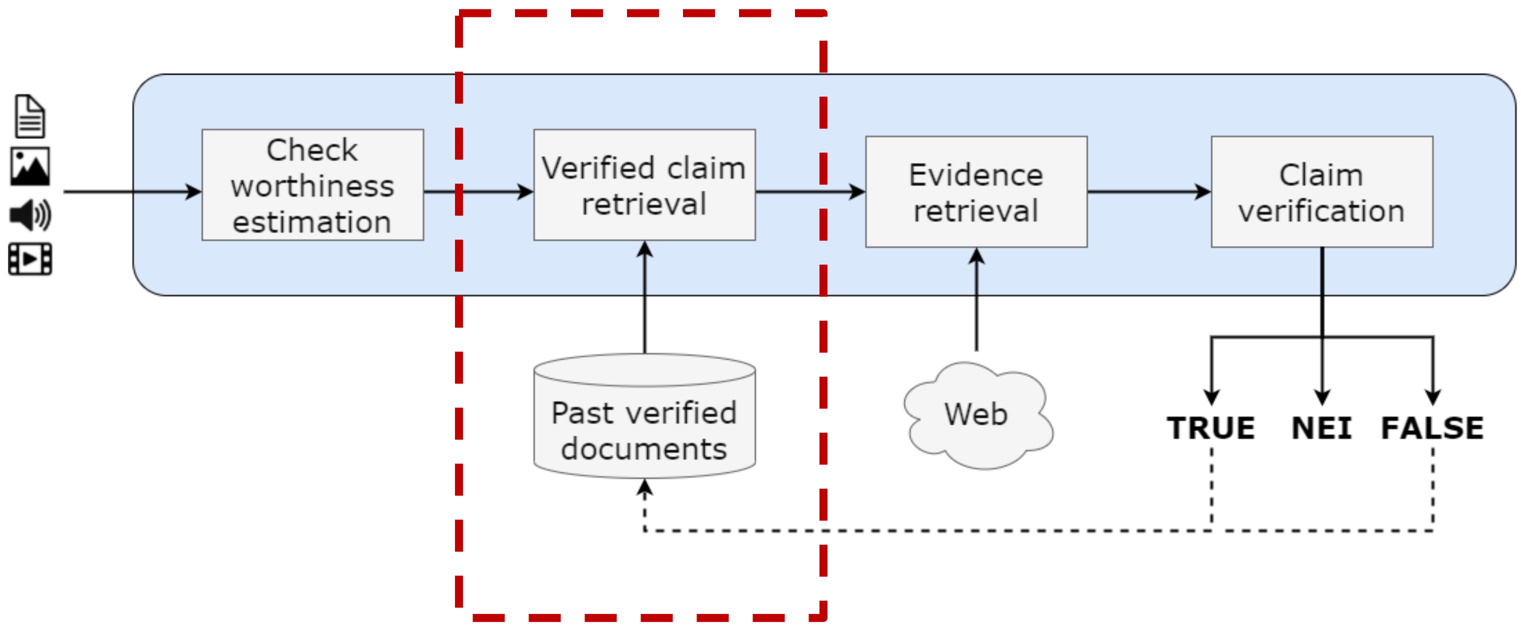
- ***Disinformation*** as *false, misleading*, and *potentially hateful* content across the digital information ecosystem
- ***Disinformation mining*** as a complex challenge intertwined with *human cognition*, *social dynamics*, and *emotional responses*
- ***Knowledge-informed strategies*** to contextualise and combat disinformation

PhD thesis – Research Thread(s)

- ***Fact-checked Databases*** to improve the fact-checking process
- ***Cultural and Common-sense Knowledge*** for disinformation detection
- ***Patterns from disinformation-related tasks*** to neutralize emotional appeal
- Platform Coordination and ***Collaborative Moderation***

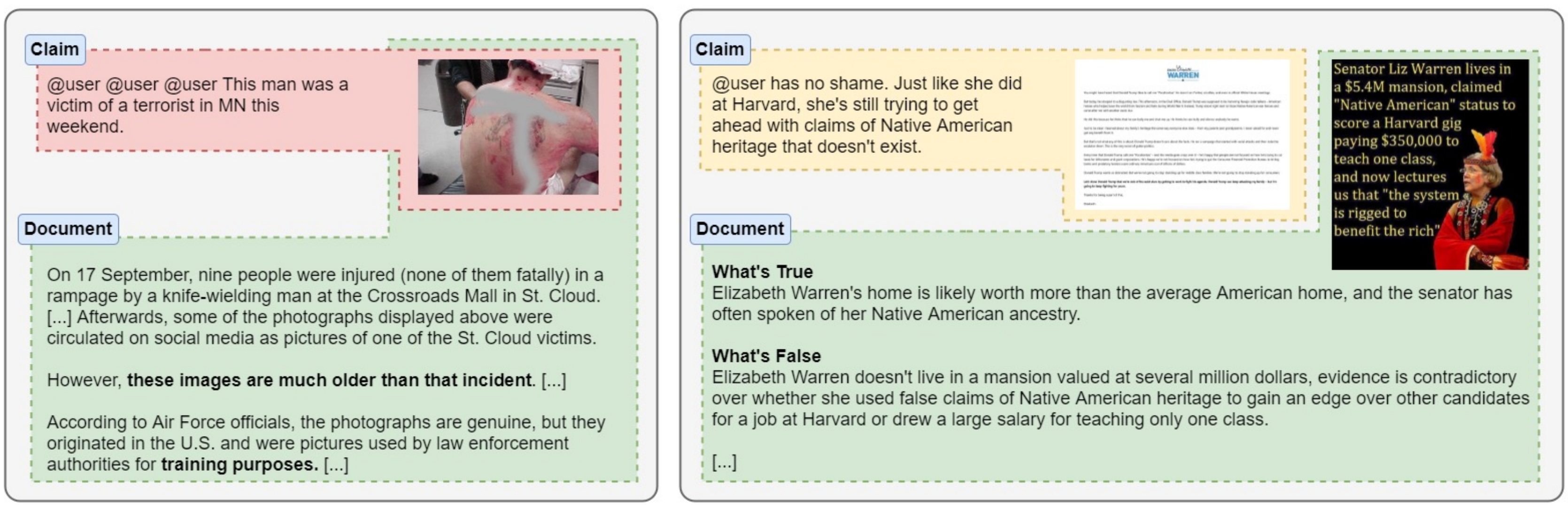
Fact-checked Databases
Objective

- Improving the fact-checking pipeline by ***detecting already-verified information***



Fact-checked Databases
Problem

- Retrieve a list of verified documents according to the relevance with an input claim.



Fact-checked Databases

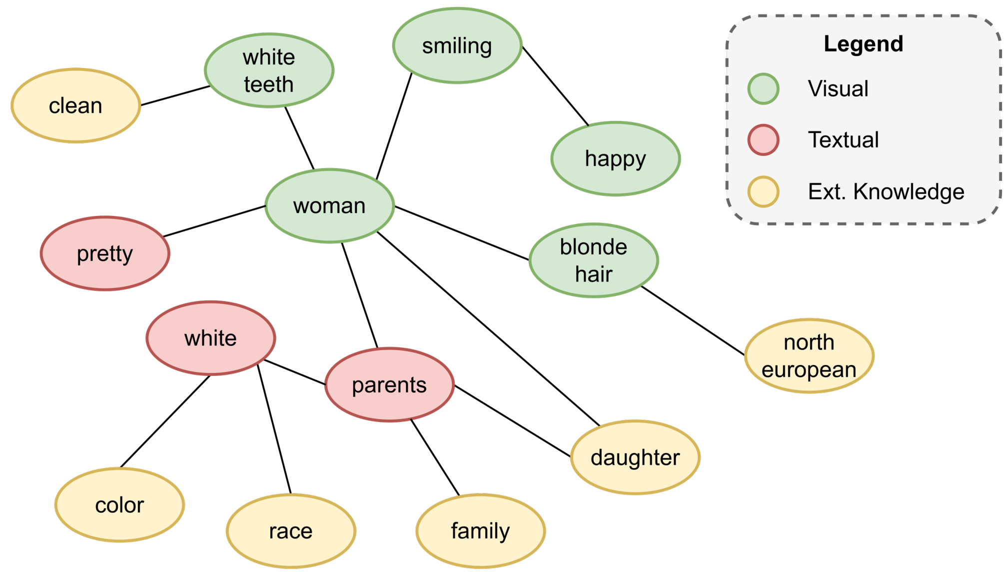
Contributions

- ***Benchmark*** semantic models and statistical approaches from related literature
- A ***novel information retrieval system*** to address the problem under ***multimodal*** settings
- A ***case study on the Ukraine-Russia conflict*** to show case the utility of the task in operational settings

Common-sense Knowledge

Objective

- Incorporating ***explicit content*** and ***implicit background knowledge*** for the analysis of complex information



Common-sense Knowledge
Problem

- Harmful Meme Detection
  - ***Text and image modalities*** within a meme are not always semantically consistent.
  - The understanding of a meme often relies on ***humans background knowledge***.

Common-sense Knowledge

Contributions

- ***KERMIT*** - A pioneering approach integrating meme content with external knowledge bases
- ***Knowledge-enriched network*** integrating meme's internal entities with external knowledge from ConceptNet
- ***Dynamic learning*** via memory-augmented neural networks & attention mechanisms

🔍 Disinformation-related Tasks 🔍

Objective

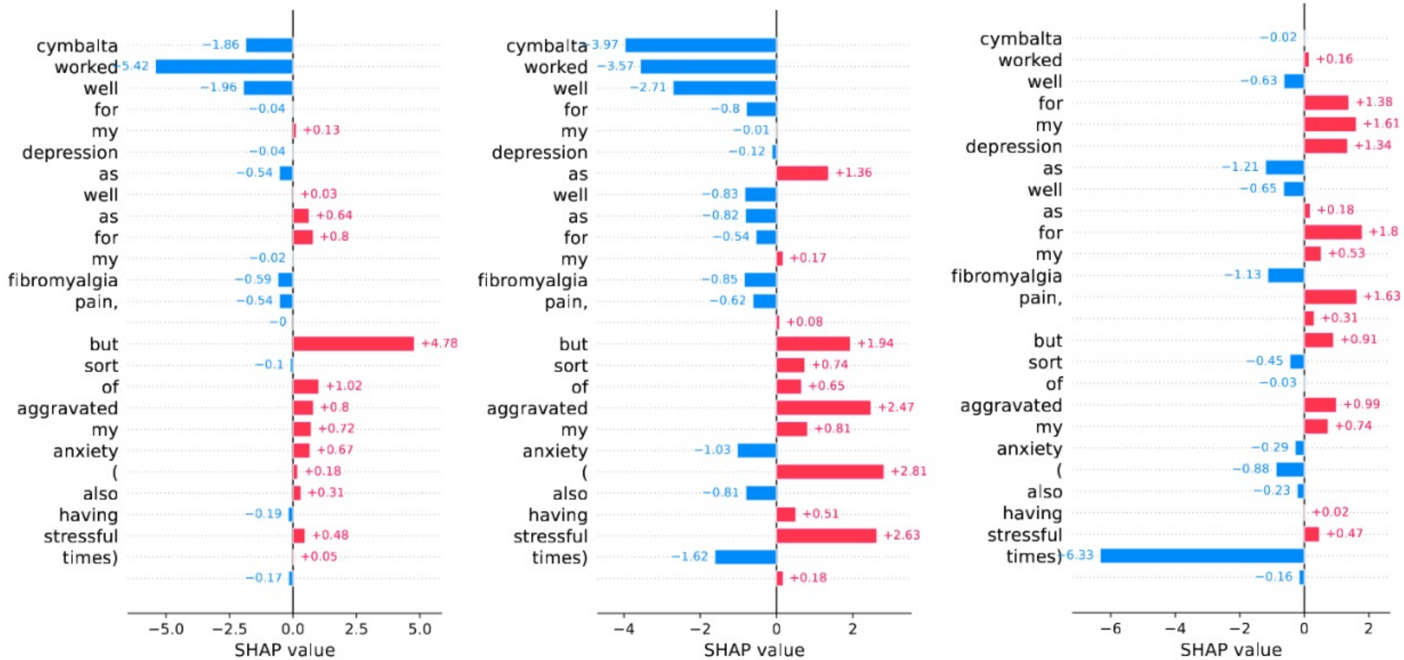
- Neutralising ***emotional triggers*** and ***human vulnerabilities*** in disinformation

🔍 Disinformation-related Tasks 🔍

Problem

- Understand ***positive and negative transfers*** in ***multi-task learning*** for disinformation mining

single-task | *multi-task positive transfer* | *multi-task negative transfer*



itee PhD
information technology
electrical engineering

Disinformation-related Tasks

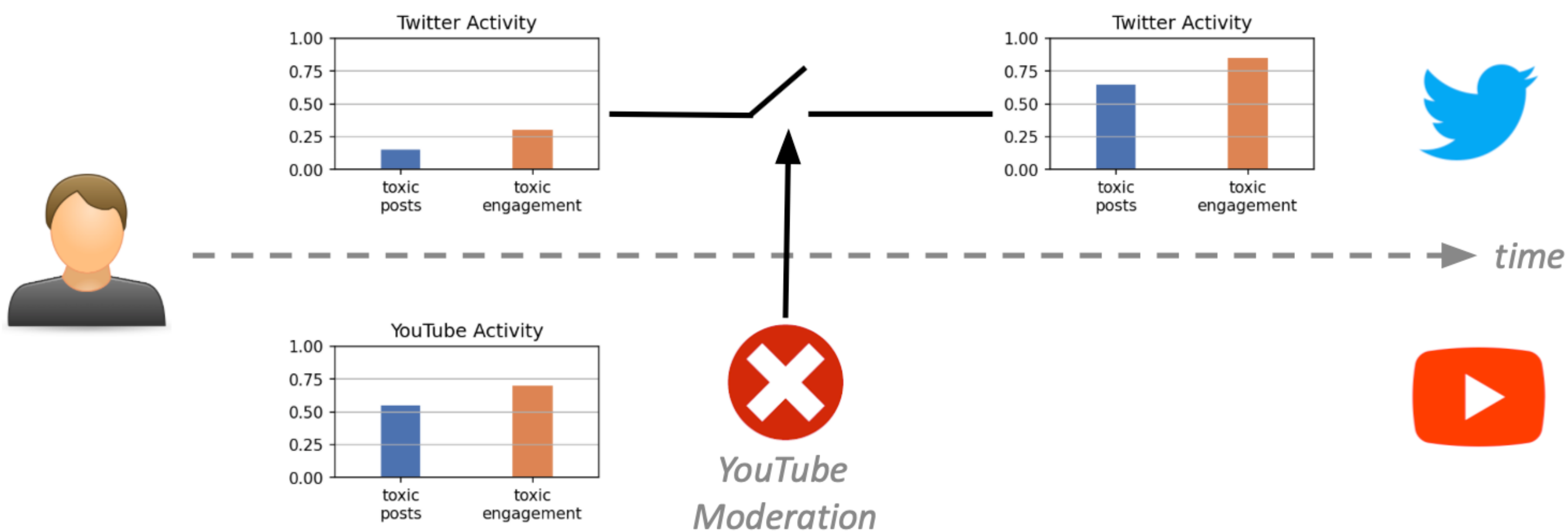
Contributions

- ***Interpretable Multi-task Framework*** to investigate ***positive and negative transfers*** among disinformation-related tasks
- ***Model Explanations*** to compare single-task vs. multi-task models' knowledge
- ***Dynamics of knowledge transfer*** between fake news detection, sentiment analysis, stance detection, topic detection

🤝 Collaborative Moderation 🤝

Problem

- The moderation within a platform affects other platforms as well



🤝 Collaborative Moderation 🤝

Objective

- Investigating whether content that has been deemed inappropriate on YouTube can ***inform moderation strategies*** on Twitter

RQ1. What is the prevalence, lifespan, and reach of moderated YouTube videos on Twitter?

RQ2. What are the characteristics of the mobilizers of moderated YouTube videos?

RQ3. Do the mobilizers of moderated YouTube videos receive significant engagement from the Twitter population?

🤝 Collaborative Moderation 🤝

Contributions

- 25% YouTube videos shared on Twitter are eventually moderated on YouTube
- Twitter users sharing moderated YT videos endorse ***extreme and conspiratorial ideas***
  - and are eventually suspended on Twitter!
- ***Sharing moderation interventions*** would benefit all entities within the information ecosystem.

Conclusions

- What we saw today
  - **Knowledge-informed disinformation mining** with *fact-checked information, cultural knowledge, cross-task patterns* and *collaborative moderation*
  - Motivations and **Contributions**
- What we didn't see today
  - Methodological background and theory (e.g., *memory-augmented neural network, attention mechanism, eXplainable AI*)
  - Technical results

Thank you for the attention!

Backup #1: Fact-checked Databases

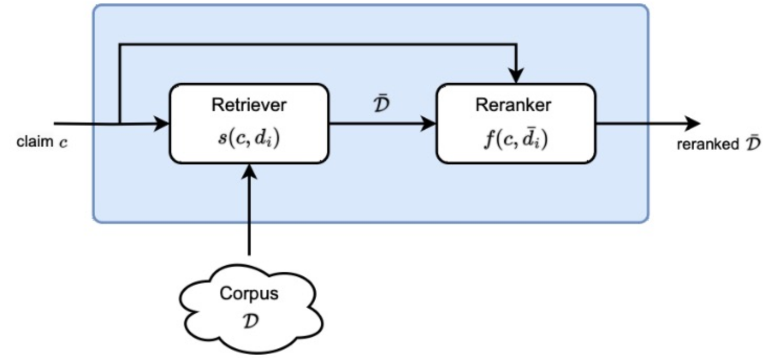
Benchmark semantic models and statistical approaches from related literature

Table 2: Performance of Neural Ranking Models (NRMs) (bold indicates the best results, underline the first runner up, * statistical significance at $p = 0.001$ w.r.t. the second best)

Category	Model	MRR	MAP@k				
		all	$k=1$	$k=3$	$k=5$	$k=10$	$k=20$
Interaction-based	BERT [52]	**0.968***	**0.942***	**0.968***	**0.968***	**0.968***	**0.968***
	ColBERT [24]	<u>0.903</u>	<u>0.847</u>	<u>0.893</u>	<u>0.901</u>	<u>0.902</u>	<u>0.903</u>
	MAN [45]	0.509	0.386	0.470	0.484	0.501	0.509
	MatchPyramid [33]	0.495	0.413	0.444	0.462	0.479	0.489
	KNRM [48]	0.319	0.212	0.272	0.287	0.298	0.307
	ConvKNRM [14]	0.744	0.677	0.721	0.729	0.738	0.742
Representation-based	ESIM [9]	0.507	0.370	0.451	0.482	0.498	0.504
	HAR [51]	0.602	0.331	0.508	0.557	0.557	0.560
Hybrid-based	DUET [27]	0.392	0.233	0.302	0.313	0.323	0.330

Table 1: Performance of retrievers (bold indicates the best results, underline the first runner up)

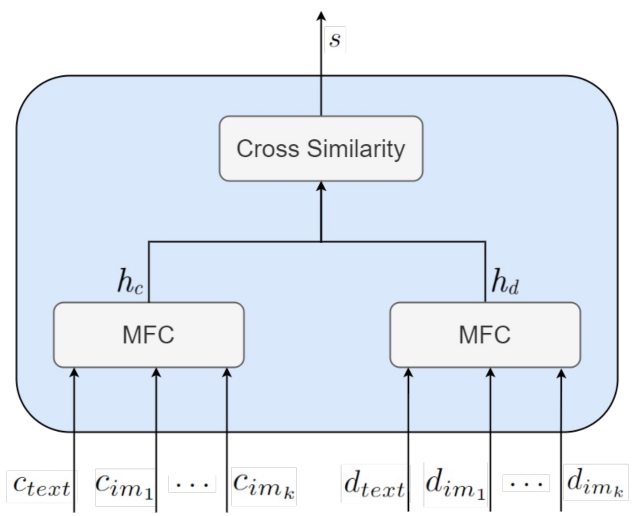
Category	Model	MRR	HasPositives@k						
		all	$k=1$	$k=3$	$k=5$	$k=10$	$k=20$	$k=50$	$k=100$
Classical	TF-IDF	0.681	0.593	0.739	0.789	0.829	0.869	0.914	0.924
	LM Dirichlet [50]	<u>0.799</u>	<u>0.770</u>	0.825	<u>0.860</u>	0.890	0.915	<u>0.95</u>	0.960
	BM25 [38]	**0.817**	**0.785**	**0.865**	**0.880**	**0.895**	**0.915**	**0.950**	**0.960**
Neural sparse	docT5query [32]	0.786	0.754	<u>0.834</u>	0.844	<u>0.894</u>	<u>0.919</u>	0.945	<u>0.960</u>
Term-based	ColBERT [24]	0.765	0.708	0.793	0.819	0.874	0.904	0.944	0.949
Document-level	SentenceBERT [37]	0.669	0.592	0.713	0.763	0.804	0.834	0.884	0.924
	DPR [23]	0.624	0.547	0.673	0.718	0.753	0.788	0.859	0.909



Retrieve and *re-rank* a corpus of documents according to the relevance with an input claim.

Backup #2: Fact-checked Databases

A ***novel information retrieval system*** to address the problem under ***multimodal*** settings



A powerful *vision-language model* followed by an efficient *vector similarity* module

Table 4: Re-ranking performance (bold indicates the best result, underline the first runner up)

		Politifact				
Method	MM	HIT@3	HIT@5	NDCG@1	NDCG@3	NDCG@5
BM25		.379	.433	.182	.292	.313
MatchPyramid		.455	.503	.294	.389	.408
KNRM		.636	.722	.422	.549	.585
BERT		.786	.856	.505	.675	.704
MAN	✓	.732	.786	.551	.654	.676
NSMN		.551	.679	.379	.477	.531
sentence-BERT		.139	.176	.059	.098	.113
Ours	✓	**.918**	**.922**	**.701**	**.712**	**.721**

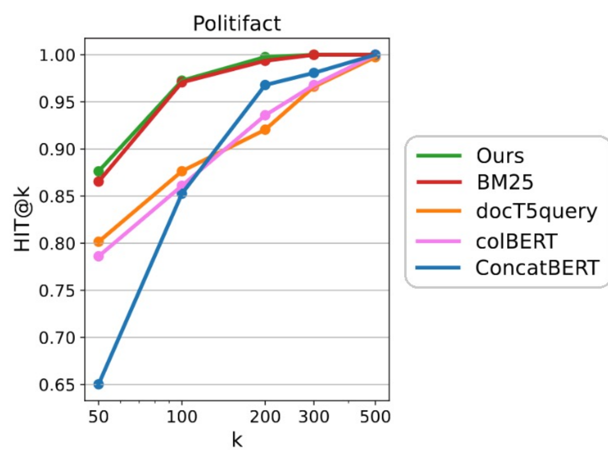


Figure 2: Retrieval hit ratios varying the number of retrieved documents from 50 to the full document corpus.

Backup #3: Fact-checked Databases

A case study on the ***Ukraine-Russia conflict*** to show case the utility of the task in operational settings

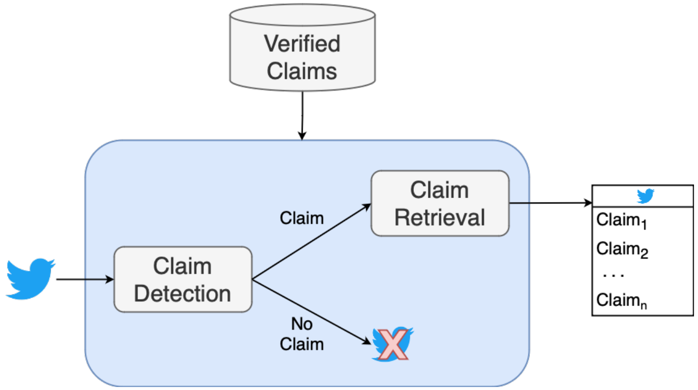
Table 1: Examples of some tweet-claim pairs annotated in the dataset

No.	Claim	Tweet
1	Russian President Vladimir Putin threatened India against getting involved in the Ukraine crisis.	Putin has warned India that don't try to interfere in their matter, otherwise be ready to face the consequences
2	The President Of Ukraine, Volodymyr Zelenskyy, Is On The Ground With His Fellow Troops	Volodymyr Zelenskyy the president of Ukraine has decided to stay behind and fight among his people against the Russian army send to kyiv [...]
3	The Russian armed forces are not striking at the cities of Ukraine; they are not threatening the civilian population.	It is clear that the Russian army does not want to harm civilians, its strikes were directed only at military targets, [...] life seems almost normal in Kiev.
4	The Russian armed forces are not striking at the cities of Ukraine; they are not threatening the civilian population.	Russian forces continue strikes in multiple cities [...]. This is premeditated mass murder and must be responded to as such.

Annotation of *8300 claim-tweet pairs* where the tweet either *supports, refutes or generally discusses* the claim.

Backup #4: Fact-checked Databases

A case study on the ***Ukraine-Russia conflict*** to show case the utility of the task in operational settings



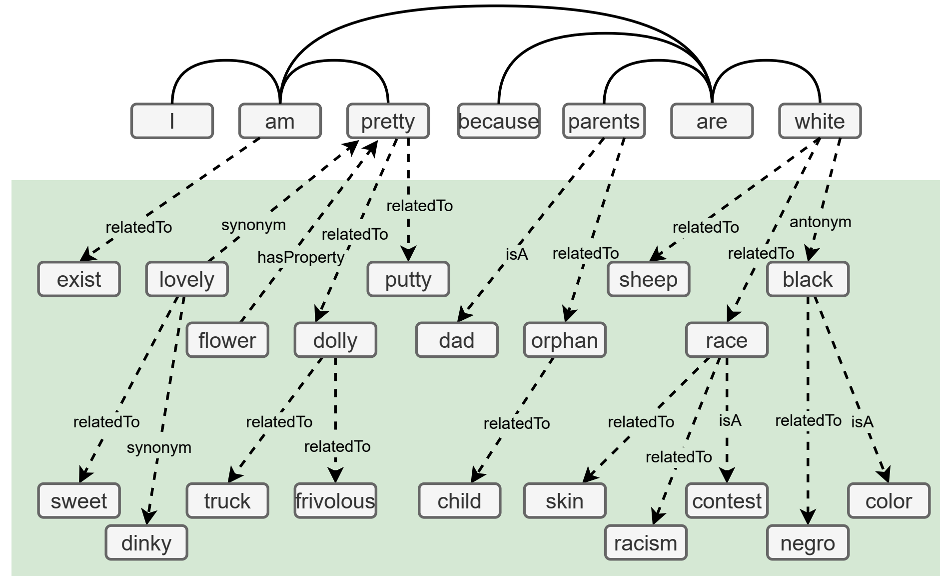
The *claim detection* model detects whether a tweet reports a fact-checked claim. If a claim is detected, the *claim retrieval* model retrieves the most relevant claims related to the tweet.

Table 5: Claim retrieval: performance comparison, and their 95% confidence interval, between the sentence-BERT baseline and our approach (bold indicates best on average, * indicates statistical significance ($p < 0.01$)

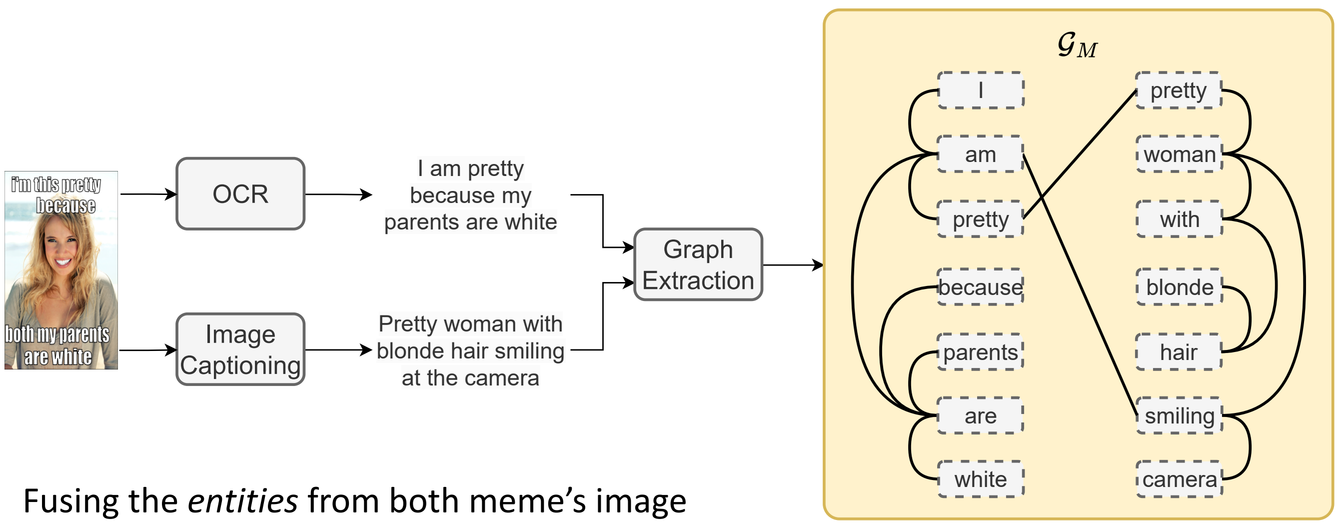
Setting	Model	HitRatio@k $k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$
LTO	Sentence-BERT	85.25% ± 2.07%	94.87% ± 1.69%	97.24% ± 0.96%	98.77% ± 0.43%1	99.27% ± 0.39%
	Ours	**86.05% ± 0.95%**	**96.35% ± 0.71%***	**98.04% ± 0.57%***	**99.27% ± 0.36%**	**99.78% ± 0.11%***
LCO	Sentence-BERT	77.60% ± 0.1196	95.68% ± 6.74%	98.01% ± 3.66%	99.63% ± 0.66%	99.78% ± 0.00%
	Ours	**82.25% ± 10.81%***	**96.42% ± 2.59%**	**98.26% ± 1.45%**	**99.88% ± 0.02%**	**99.96% ± 0.00%**

Backup #5: Common-sense Knowledge

Knowledge-enriched network
integrating meme's internal entities with external knowledge from ConceptNet



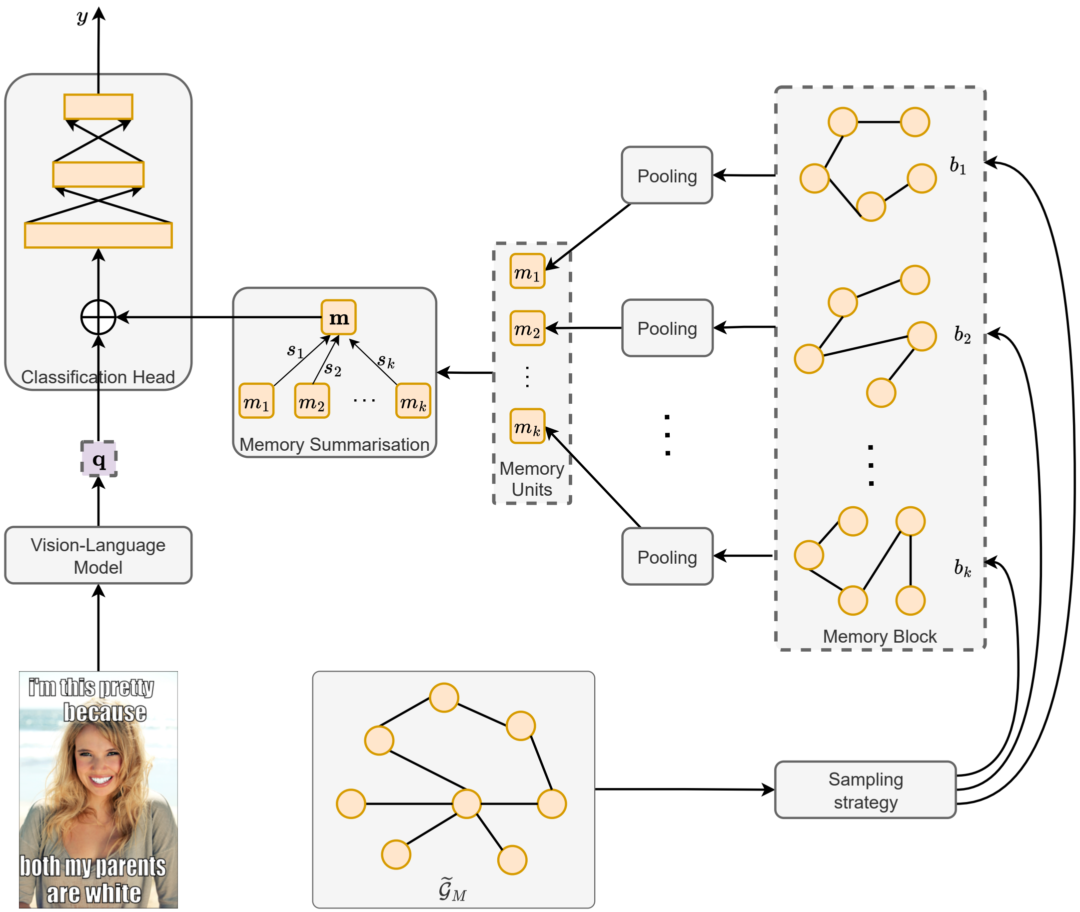
Recursively querying *ConceptNet* to obtain the *knowledge-enriched information network*



Fusing the *entities* from both meme's image and text modality to obtain the *meme graph*

Backup #6: Common-sense Knowledge

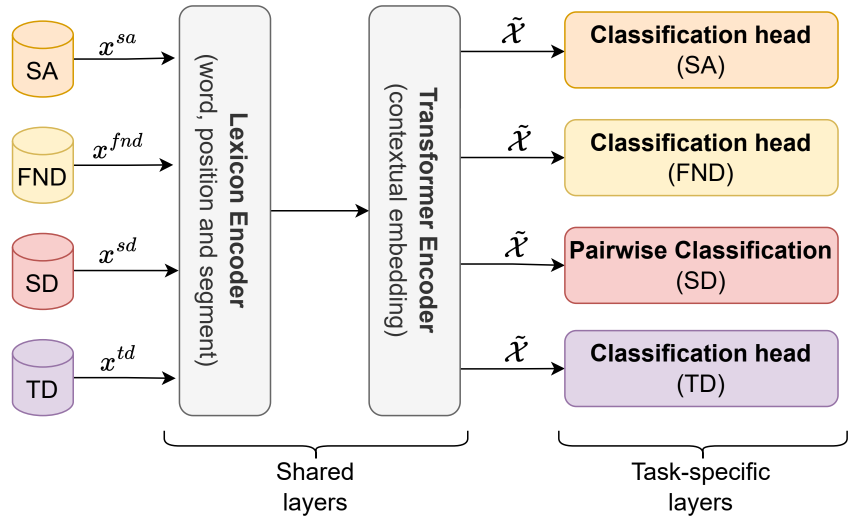
Dynamic learning via memory-augmented neural networks & attention mechanisms



The framework automatically learns *the most informative knowledge* to perform the meme classification

Backup #7: Disinformation-related Tasks

Interpretable Multi-task Framework to investigate ***positive and negative transfers*** among disinformation-related tasks



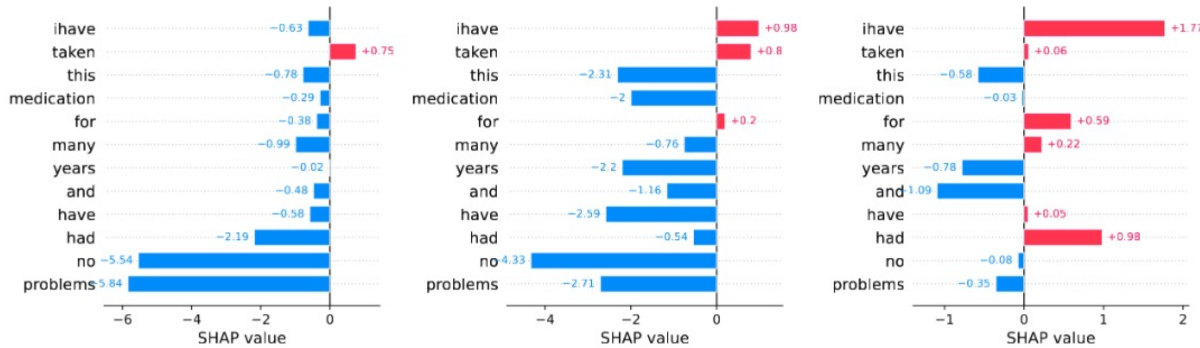
The *shared layers* capture common information across tasks, while task-specific layers learns custom feature for each task.

Table 3: Comparison, in terms of F1-score, with baselines, under both single-task (ST) and multi-task (MT) settings. GPT3.5 is configured under 0-shot settings. (bold indicates the best result, underline the first runner up)

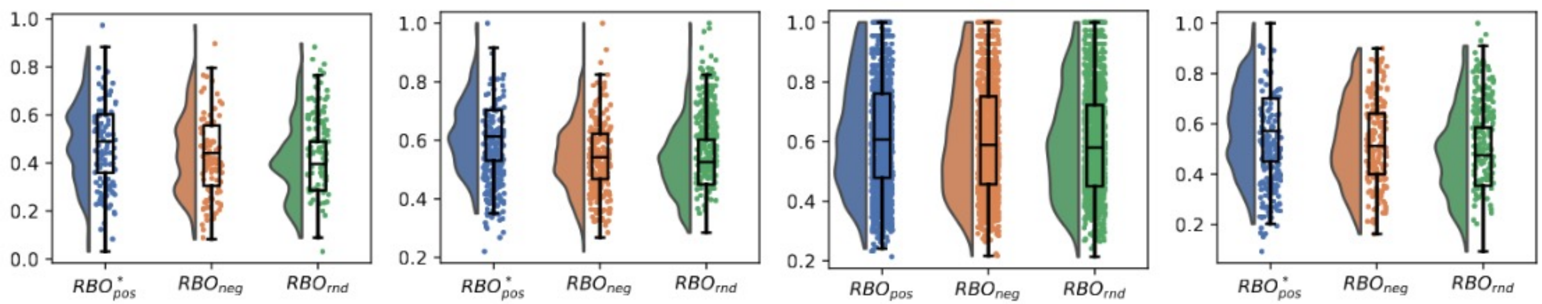
Method	Configuration	SA	FND	SD	TD
GPT3.5	0-shot	0.844	0.312	0.156	0.870
AdverMTL	ST	0.627	0.487	0.195	0.305
	MT	0.648	0.527	0.190	0.320
MaChAmp	ST	0.859	0.814	0.682	0.960
	MT	<u>0.879</u>	**0.834**	<u>0.729</u>	0.937
Ours	ST	0.861	0.768	0.728	<u>0.971</u>
	MT	**0.890**	<u>0.822</u>	**0.751**	**0.977**

Backup #8: Disinformation-related Tasks

Model Explanations to compare single-task vs. multi-task models' knowledge



Positive transfer has a regularization effect while *negative transfer* is equivalent to a random perturbation of feature importances.

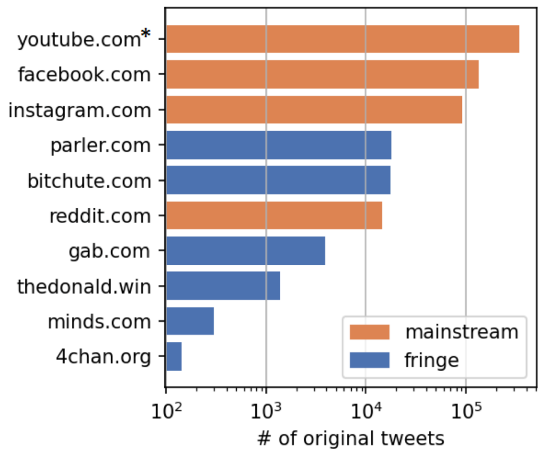


(a) Sentiment Analysis **(b) Fake News Detection** **(c) Stance Detection** **(d) Topic Detection**

Figure 3: The distribution of RBO_{pos}, RBO_{neg}, and RBO_{rnd} for each task. (* indicates statistical difference, at $p = .05$, with respect to RBO_{neg})

Backup #9: Collaborative Moderation

- 24.7% (130k out of 527k) YT videos shared on Twitter were moderated on YouTube
- Moderated videos are engaged more than non-moderated ones in their (shorter) lifespan!



(a) Number of original tweets containing a link to each social media platform (Log-scale)

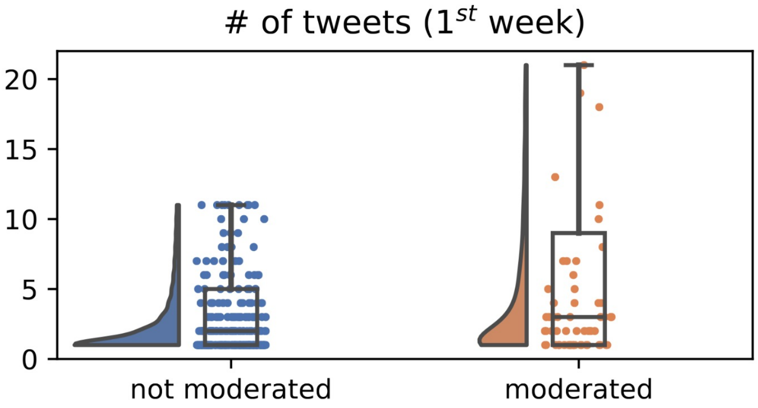
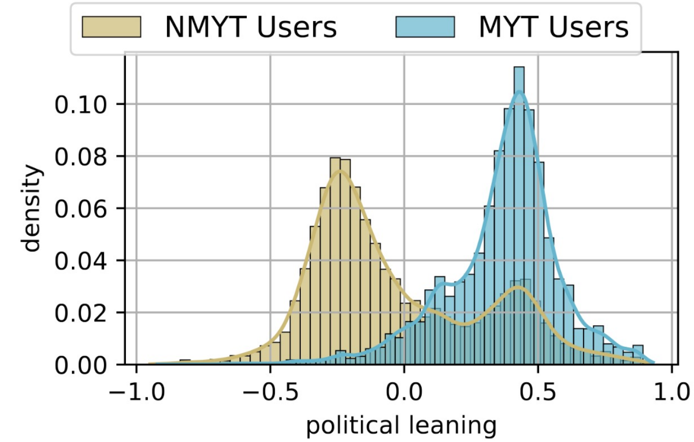


Figure 3: The distribution of the number of tweets sharing each video during the week after its first share

Backup #10: Collaborative Moderation

Twitter users sharing moderated YT videos endorse ***extreme and conspiratorial ideas*** and are eventually suspended on Twitter!



(b)

Figure 9: (a) The news outlet shared by each group of mobilizers; (b) the distribution of the political leaning within the two groups of mobilizers

	Total Accounts	Total Videos	Verified Accounts	Bot Accounts	Suspended Accounts	InfoOps Accounts
NMYT Mobilizers	25396	88451	268	2234	7984	569
MYT Mobilizers	14481	61884	19	586	7793	30