

# Tackling Covid-19 Conspiracies on Twitter using BERT Ensembles, GPT-3 Augmentation, and Graph NNs

Damir Korenčić, Ivan Grubišić, Gretel Liz De La Peña Sarracén,  
Alejandro Hector Toselli, Berta Chulvi, Paolo Rosso



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

Pattern Recognition and  
Human Language Technology  
Research Center



presenter: Damir Korenčić, PhD  
PRHLT Research Center, Universitat Politècnica de València

- 1 Introduction and Data
- 2 Task 1 – Fine-grained Text Classification
- 3 Task 2 – Graph-based User Classification
- 4 Task 3 – Text Classification w. Graph Data
- 5 Negative Results – Improved Classification Architecture
- 6 Negative Results – GPT-3 In-context Learning and Augmentation
- 7 Definitions for GPT: Zero-shot
- 8 Definitions for GPT: An Analysis of Use

- Twitter data: scraping, keyword-filtering, cleaning, annotation
- User graph: nodes are users, edges are user-user interactions
- Text-based detection of conspiracy theories
- Graph-based conspiracy spreader detection
- Evaluation: MCC metric  
(Chicco and Jurman 2020; Chicco, Tötsch, et al. 2021)
- Details: Pogorelov et al. 2023  
<https://github.com/konstapo/2022-Fake-News-MediaEval-Task>
- Proceedings with cited papers:  
<https://2022.multimediaeval.com>
- Paper: (Korenčić, Grubišić, Sarracén, et al. 2023)

- Conspiracy categories:  
suppressed cures, behaviour and mind control, antivax, fake virus, intentional pandemic, harmful radiation or influence, population reduction, new world order, and satanism
- Text-conspiracy relation:  
supports, mention, no-mention

The table contains statistics of annotated train and test sets – the number of texts and the quartiles of texts' length (in tokens).

Split	num. txt.	avg.	25%	50%	75%
Train	1913	45.68	43	46	49
Test	830	46.37	43	46	49

Given a text (tweet) and a conspiracy theory decide weather:

1. There is **no mention** of the conspiracy in the text
2. The text **mentions** the conspiracy but does not support it
3. The text **supports** the conspiracy

“This category collects narratives which propose that effective medications for COVID-19 were available, but whose existence or effectiveness has been denied by authorities, either for financial gain by the vaccine producers or some other harmful intent, including ideas from other conspiracy categories listed below. It thus refers to the treatment of COVID-19, irrespective of its origin.” (Langguth et al. 2023)

Example: There are other ways to reduce the impact of covid but it will not be profitable for the pharmaceutical companies. Plus the government won't be able to implement it's agenda of the new world order as stipulated by WEF. Vaccines should a very last resort. It's the nuclear option.

“In this category we collected narratives containing the idea that the pandemic is being exploited to control the behavior of individuals, either directly through fear, through laws which are only accepted because of fear, or through techniques which are impossible with today’s technology, such as mind control through microchips.” (Langguth et al. 2023)

Example: DEEP STATE ACTORS ..have their orders .. do not stop the population lock down 100% .. Keep the population in a limbo situation Indian Covid keep it going .. KEEP CONTROL AT ALL COSTS until the GREAT RESET IS READY TO IMPLIMENT .. BUILD BACK BETTER is the slogan

“We collect all statements that suggest that the COVID-19 vaccines serve some hidden nefarious purpose in this category. Examples include the injection of tracking devices, nanites or an intentional infection with COVID-19. This category does not include concerns about vaccine safety or efficacy, or concerns about the trustworthiness of the producers, since these are not conspiracies, even though they may contain misinformation. ...” (Langguth et al. 2023)

Example: Worse than you think. “Passports” use a QR code system. The 2.0 Covid vaccine is a subdermal vaccine microneedle patchbiometric ID digital banking. It also uses a QR code system pattern - Dystopian much? (“Quantum Dot Tattoo” / “SMART” Vaccine).



“Prominent narratives that surfaced early in the pandemic were that there is no COVID-19 pandemic or that the pandemic is just an over-dramatization of the annual flu season. Typically, the claimed intent is to deceive the population in order to hide deaths from other causes, or to control the behavior of the population through irrational fear.” (Langguth et al. 2023)

Example: Can you tell the world governments to stop spreading mis-information about Covid19! A virus that is no more harmful than the common cold and flu! Deaths have been falsified as well as mis-recording of positive case numbers using tests that are not fit for purpose! #Plandemic

“This straightforward narrative posits that the cause of the pandemic is purposeful human action pursuing some illicit goal. It thus produces a culprit for the situation. Note that this is distinct from asserting that COVID-19 is a bioweapon or discussing whether it was created in a laboratory since this does not preclude the possibility that it was released accidentally, which would not produce a culprit and thus not qualify as a conspiracy theory.”  
(Langguth et al. 2023)

Example: The 'Corona crisis' was planned years ago - to provide the pretext for Western governments to DELIBERATELY destroy their economies - and then to install a basically communist New World Order global Police State. There IS no 'crisis': this is all a scam!

“This class of conspiracy theories bundles all ideas that connect COVID-19 to wireless transmissions, especially from 5G equipment. This was done by claiming for example that 5G is deadly and that COVID-19 is a coverup, or that 5G allows mind control via microchips injected in the bloodstream. As 5G misinformation has already been studied separately, it was not the focus of this dataset but it is included nonetheless since it is related to other conspiracy theories.” (Langguth et al. 2023)

Example: 5G released in Wuhan Milan Italy and also South Korea which both are now Corona hotspots. The human body becomes an enhanced antenna for 5G because of Nano particulates raining down on people from chemtrails it enhances the effectiveness and Weaponry of 5G to us. #COVID2019

“Conspiracy theories on population reduction or population growth control suggest that either COVID-19 or the vaccines are being used to reduce population size, either by killing people or by rendering them infertile. In some cases, this is directed against specific ethnic groups. These narratives often use the term "population control" in the sense of population size control which needs to be distinguished from population behavior control covered in other conspiracy theories.” (Langguth et al. 2023)

Example: Covid is a cover up for the 21st century human cull (depopulation) 99.9% of us can't believe. You've been warned. Funniest of all 99.9% of us believe in over population but not the agenda to depopulate. Irony It's not a big or hard mental leap #covid1984 #covidcoverup

"New World Order (NWO) is a preexisting conspiracy theory which deals with the secret emerging totalitarian world government. In the context of the pandemic, this usually means that COVID-19 is being used to bring about this world government through fear of the virus or by taking away civil liberties, or some other, implausible ideas such as mind control." (Langguth et al. 2023)

Example: Brexit is now irrelevant - even if Boris hadn't already locked in a totally fake one. A far left permanent Police State is being installed in the UK with the Corona scam as the pretext. The agenda is to utterly trash the UK economy - & introduce the "New World Order".

“This category collects narratives in which the perpetrators are alleged to be some kind of satanists, perform objectionable rituals, or make use of occult ideas or symbols. . . While the concrete allegations differ, they have in common that they connect the alleged perpetrators to the representation of evil, and thus paint a picture of them as someone to be opposed at all cost.” (Langguth et al. 2023)

Example: With each attack launched by the Cabal (coronavirus MK Ultra shootings etc) there is an extremely dark underlying spiritual component to the attack. We are in a war against Satanists who draw power from rituals & sacrifices. For this reason Q continues to post Ephesians 6.

- Hard baseline (Pesquine, Alfarano, et al. 2021)
  - Fine-tuned pre-trained BERT
  - Multi-task learning
  - Pre-training on Covid-19 tweets
- Improvements
  - Ensembling
  - GPT3-based data augmentation
  - GPT3 in-context classification  
(Brown et al. 2020; Ornstein et al. 2022)

- What worked?
  - Ensembling - Yes
  - GPT3-based data augmentation - Yes, simple rephrasing
  - GPT3 in-context classification - No
- Details: Korenčić, Grubišić, Toselli, et al. 2023

model	train	std.dev	min	max	test
baseline	0.724	0.027	<u>0.592</u>	0.797	–
ensmbl	0.755	0.030	0.706	0.796	<b>0.738</b>
ensmbl.gpt3	0.750	0.028	0.704	0.783	<b>0.738</b>
gpt3	0.745	<u>0.021</u>	0.701	0.788	–
Xgboost.tm	0.446	0.046	0.394	0.526	0.376



Team	MCC Score
Korenčić, Grubišić, Toselli, et al. 2023	<b>0.738</b>
Peskine, Papotti, et al. 2023	<u>0.710</u>
Akbari 2023	0.702
Bocconi et al. 2023	0.596

Results achieved by the top 4 teams (out of 6 participating teams).

- The Setup  
An undirected graph derived from social network data is given: the vertices are users and the edges represent connections between them.
- 1.679.011 nodes, 268.694.698 edges, avg. 160 edges/node
- The Task  
Label the users as either **conspiracy spreaders** or **non-conspiracy spreaders**.
- Train set (1.913 users), Test set (830 users)

- Graph Neural Networks (Wu et al. 2020)
- Network Architecture
  - Two convolutional layers
  - Three fully connected layers
  - Classification layer
- Node features: user metadata
- Graph: full graph VS dev & train users only

- MCC up to 0.28 (SoA solution 0.43)
  - Small graph performs equally well
  - GNNs do not perform well
- SoA solution: Node2Vec + Xgboost (Jiménez et al. 2023)

Team	MCC Score
Jiménez et al. 2023	<b>0.434</b>
Peskine, Papotti, et al. 2023	<u>0.355</u>
Korenčić, Grubišić, Toselli, et al. 2023	0.283
Bocconi et al. 2023	0.110

Results achieved by the top 4 teams (out of 6 participating teams).

- The task: Goal is the same as for Subtask 1 – for each text, determine its relation to each conspiracy category (multi-label multi-class classification)  
The difference is that *graph data* from Subtask 2 can be used: for each text, ID of the author is given, and the author is a node in the graph of Twitter users.
- Train: 1.913 tweets, Test: *646 tweets*
- Graph: 1.679.011 user nodes, 268.694.698 edges

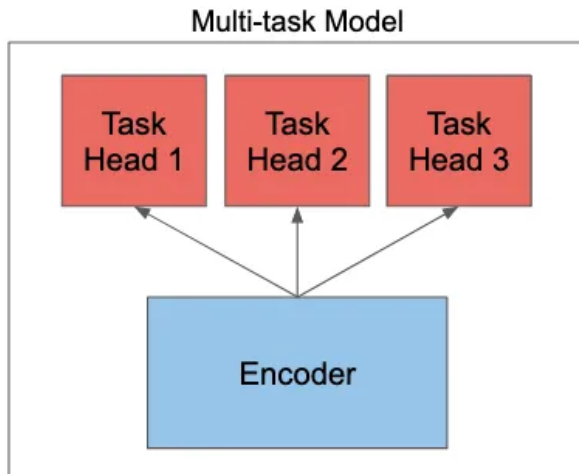
- GNN Architecture
  - Two convolutional layers
  - Three fully connected layers
  - Classification layer
- Graph: nodes are texts, edges via text-user-text
- Node features: BERT text representation

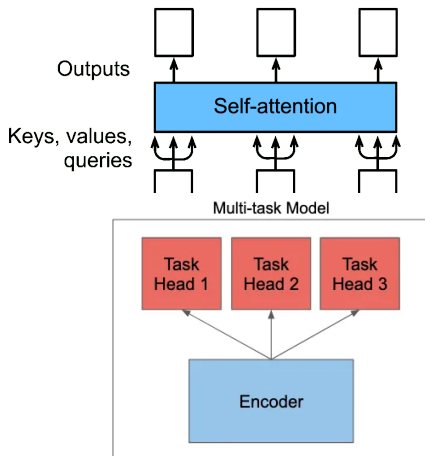
- GNN MCC up to 0.5
- Text-only solution MCC 0.7
- Similarly, in Peskine, Papotti, et al. 2023:  
graph+text MCC 0.690, text-only MCC 0.72
- Graph data seems detrimental



Team	MCC Score
Peskine, Papotti, et al. 2023	<b>0.719</b>
Korenčić, Grubišić, Toselli, et al. 2023	<u>0.698</u>
Akbari 2023	<u>0.698</u>
Bocconi et al. 2023	0.246

Results achieved by the top 4 teams (out of 6 participating teams).





- Results: MCC 0.68 – 0.7
- Without attention: MCC 0.74
- Why?

- Create a text prompt containing:
  - Task explanation
  - Labeled examples
  - Unlabeled example
- LLM is able to “understand” the task and generate the label
- More on the approach:  
Brown et al. 2020; Ornstein et al. 2022

Decide how a tweet relates to the conspiracy theory stating that covid-19 pandemic is used to control people.

TWEET: Y'all wanna know why you can't sue the makers of the Wuhan lab created Bat-Rona virus vaccine for the side effects?? Because they know that shit is poisonous sterilizing microchipped garbage they're injecting into all of you. They know and don't care!!

RELATION: support

TWEET: Someone here brought up Bill Gates wanting to microchip people that have Covid vaccine this week reasonable people from Denmark yet they got drawn into believing these unfounded conspiracy theories stating it's about population control...the irony started by fascists.

RELATION: mention

TWEET: Mr. President will sound crazy but what if Bill Gates Soros & Kissinger have teamed up with communist China in order to create this invisible World War III called COVID-19? They have the means and the technology to do this without nobody knowing.

RELATION: no mention

--

TWEET: lol what an idiot press secretary is they have not done a damn thing about the virus but whenever something else is going on like questions about qanon its but were in a pandemic hows that going for you seems like you havent done much about it

RELATION:

- Results
  - 0.175 MCC (SoA 0.74, non-neural 0.45)
  - ACC 0.463; F1 0.299; P 0.396; R 0.543
- Reasons for low performance
  - Hard problem: “specific” and mutually close classes
  - High proportion of false positives
- Improvements
  - Prompt structure?
  - More examples?

# Definitions Matter: Guiding GPT for Multi-label Classification

Youri Peskine, Damir Korenčić, Ivan Grubišić,  
Paolo Papotti, Raphael Troncy, Paolo Rosso



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



Pattern Recognition and  
Human Language Technology  
Research Center





- MediaEval 2022 FakeNews challenge
- Conspiracy categories:  
suppressed cures, behaviour and mind control,  
antivax, fake virus, intentional pandemic, harmful  
radiation or influence, population reduction, new  
world order, and satanism
- Text-conspiracy relation:  
mention (or support), no-mention

- Few-shot GPT-3 performs poorly
- Interesting problem: fine-grained classification for the presence of concepts in text
- Test-case for analyzing GPT
- Paper: (Peskin, Korenčič, et al. 2023)

- *Generating definitions from examples*  
and using them for zero-shot classification
- *Investigating how an LLM makes use of the definitions*
- Multi-label classification:
  - Nine conspiracies (labels)
  - Mention vs. no mention (concept detection)

Your task is to label tweets  
regarding the '[CONSPIRACY]' COVID-19 conspiracy theory.  
The available labels are:  
1) mentions the conspiracy, 2) does not mention the conspiracy.  
The definition of the '[CONSPIRACY]' conspiracy theory is the following:  
[CONSPIRACY definition]"

[TWEET]

Does the tweet:  
1) mention the '[CONSPIRACY]' conspiracy,  
2) do not mention the '[CONSPIRACY]' conspiracy?  
Please include the corresponding number in your answer.

- Phrasal descriptions (ZS):  
suppressed cures, behaviour and mind control,  
antivax, fake virus, intentional pandemic, harmful  
radiation or influence, population reduction, new  
world order, and satanism
- Human-generated (HW)
- Example-generated (EG)

You will be given two sets of tweets.  
The first set of tweets contains examples  
of texts that mention the same concept.  
The second set of tweets contains examples  
of texts that mention other concepts,  
but not the same concept like tweets from the first set.  
Your task is to provide the definition  
of the concept present in the first set

First set of tweets:  
[25x Tweets containing the conspiracy]

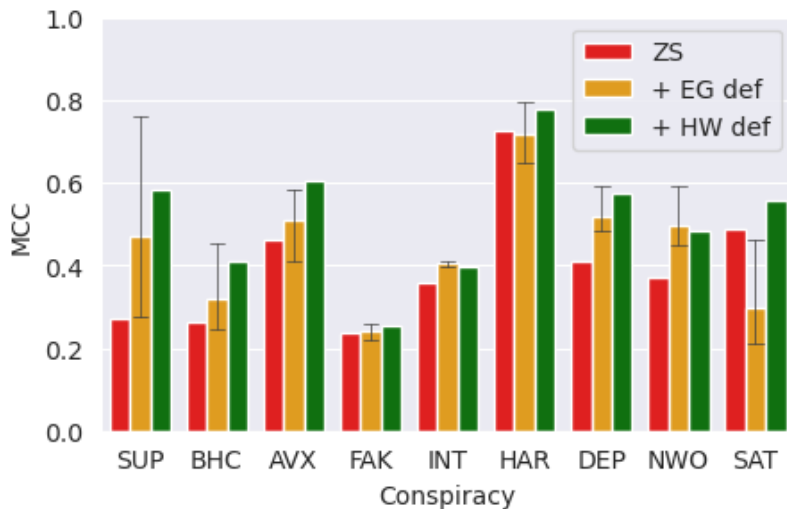
Second set of tweets:  
[25x Tweets not containing the conspiracy]

Given those two sets of tweets, what is the definition  
of the concept present in the first set that  
is not present in the second set of tweets?  
Start your answer with: 'The definition of the concept is'

- GPT-3.5 model used for generation  
25 positive and 25 negative examples
- 5 random seed for each category  $\Rightarrow$  45 definitions

Approach	MCC	Precision	Recall	F1
Zero-shot	0.398	0.331	<b>0.852</b>	0.440
w/ Example-generated definitions	0.442	0.371	0.831	0.485
w/ Human-written definitions	<b>0.516</b>	<b>0.464</b>	0.823	<b>0.555</b>
CT-BERT ensembling	0.780	0.779	0.849	0.810

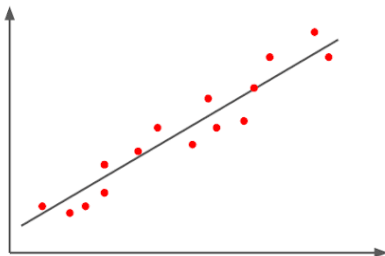




- Improved definitions lead to improved results
- EG definitions can perform well
- How to operationalize targeted creation of EG:  
*example selection*, generation model (and prompt)
- Few-shot learning with generated definitions (cost, stability)
- Other datasets, other LLMs
- Potential applications
  - Fixing the class imbalance
  - Correcting annotation errors (false negatives)

- Check if the GPT “applies” definitions “correctly”
  - “Understand” the definition
  - “Operationalize” the definition to classify texts

- How to test if the definitions are applied correctly?
- Similar definitions should produce similar results  
(outputs of the definition-based classification)
- Measure correlation
  - `definition1, definition2`
  - `semantic_similarity(definition1, definition2)`
  - `output_similarity(output1, output2)`



- Example-generated – Human definitions
  - semantic similarity of definition texts
  - classification performance  
(similarity between application of EG and HW (gold labels))
- Example-generated – Example-generated
  - semantic similarity of definition texts
  - similarity of two predictions (Cohen's  $\kappa$ )

	MCC	F1
Similarity (EG, HW)	0.375	0.390
	Cohen's $\kappa$	
Similarity (EG, EG)	0.407	

*"Fair" correlation.*

*Higher similarity between EG and HW definitions leads to more accurate classifications, which suggest that the model can translate better definitions into better predictions.*

*Higher similarity between two EG definitions correlates with higher agreement between their corresponding predictions, which suggest that the model translates similar definitions into similar predictions.*



- Very complex artifacts – unexpected effects?
- Alternative interpretations?
- Test on different datasets, LLMs

*Semantic abilities of LLMs are commonly evaluated on NL understanding and reasoning tasks, but work on targeted evaluation of fine-grained semantic properties is scarce. Sahu et al. (2022) propose to evaluate the LLM's comprehension of query-related concepts by using a knowledge graph. To the best of our knowledge, there is no previous work focused on the ability LLMs to understand and apply definitions.*


# Thank You for Your Attention!




Any Questions or Comments?








-  Chicco, Davide and Giuseppe Jurman (Jan. 2020). “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC Genomics* 21.1, p. 6.
-  Chicco, Davide, Niklas Tötsch, and Giuseppe Jurman (Feb. 2021). “The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation”. In: *BioData Mining* 14.1, p. 13.

-  Pogorelov, Konstantin, Daniel Thilo Schroeder, Stefan Brenner, Asep Maulana, and Johannes Langguth (2023). “(Task Description) Combining Tweets and Connections Graph for FakeNews Detection at MediaEval 2022”. In: *Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online*.
-  Korenčić, Damir, Ivan Grubišić, Gretel Liz De La Peña Sarracén, Alejandro Hector Toselli, Berta Chulvi, and Paolo Rosso (2023). “Tackling Covid-19 Conspiracies on Twitter using BERT Ensembles, GPT-3 Augmentation, and Graph NNs”. In.

-  Langguth, Johannes, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov (Apr. 2023). “COCO: an annotated Twitter dataset of COVID-19 conspiracy theories”. In: *Journal of Computational Social Science*.
-  Peskine, Youri, Giulio Alfarano, Ismail Harrando, Paolo Papotti, and Raphael Troncy (2021). “Detecting COVID-19-Related Conspiracy Theories in Tweets”. In: p. 3.
-  Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33, pp. 1877–1901.

-  Ornstein, Joseph T, Elise N Blasingame, and Jake S Truscott (2022). “How To Train Your Stochastic Parrot: Deep Language Models for Political Texts”. In: p. 30.
-  Korenčić, Damir, Ivan Grubišić, Alejandro Hector Toselli, Berta Chulvi, and Paolo Rosso (2023). “Tackling Covid-19 Conspiracies on Twitter using BERT Ensembles, GPT-3 Augmentation, and Graph NNs”. In: *Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online*.
-  Peskine, Youri, Paolo Papotti, and Raphaël Troncy (2023). “Detection of COVID-19-Related Conspiracy Theories in Tweets using Transformer-Based Models and Node Embedding Techniques”. In: *Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online*.

-  Akbari, Rohullah (2023). “Evaluating TF-IDF and Transformers-based models for Detecting COVID-19 related Conspiracies”. In: *Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online*.
-  Bocconi, Stefano, Alessandro Patruno, and Andrey Malakhov (2023). “Transformers and GNNs for Fake News Detection”. In: *Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online*.
-  Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip (2020). “A comprehensive survey on graph neural networks”. In: *IEEE transactions on neural networks and learning systems* 32.1, pp. 4–24.

-  Jiménez, Adrián Girón, Ángel Panizo-LLedot, Javier Torregrosa, and David Camacho (2023). “Representational learning for the detection of COVID related conspiracy spreaders in online platforms”. In: *Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online*.
-  Peskine, Youri, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso (Dec. 2023). “Definitions Matter: Guiding GPT for Multi-label Classification”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 4054–4063.