EVALITA 2026 — February 26th, 2026

# FadeIT at EVALITA 2026

## *Fallacy Detection in Italian Social Media Texts Task*

ALAN **RAMPONI** and SARA **TONELLI**

*Fondazione Bruno Kessler, Italy*

*Funded by the European Union's Horizon Europe research and innovation programmes under GA No. 101070190 (AI4Trust) and under the Marie Skłodowska-Curie GA No. 101073351 (HYBRIDS).*

# Fallacies

Arguments that seem valid but are not

*– Aristotle*

**Hasty generalization**

Alice got the flu after the influenza vaccine. Vaccines are really useless.

Either used ***intentionally*** (for persuading) or ***unintentionally***

- Not only logical: also *structural*, from *diversion*, due to *language use* [1]
- Difficult to spot: closely follow the patterns of valid arguments [2]
- Impactful: can mislead a wide audience → spread of misinformation



[1] ***Tindale, 2007***. *"Fallacies and Argument Appraisal". Critical Reasoning and Argumentation; CUP.*
[2] ***Musi et al., 2022***. *"Developing Fake News Immunity: Fallacies as Misinformation Triggers During the Pandemic". OJCMT.*

# Fallacy detection

**Why is it useful?**

Recognizing fallacies in everyday argumentation plays
a key role in **developing individuals' critical thinking skills**,
contributing to **mitigate faulty argumentation at its root**

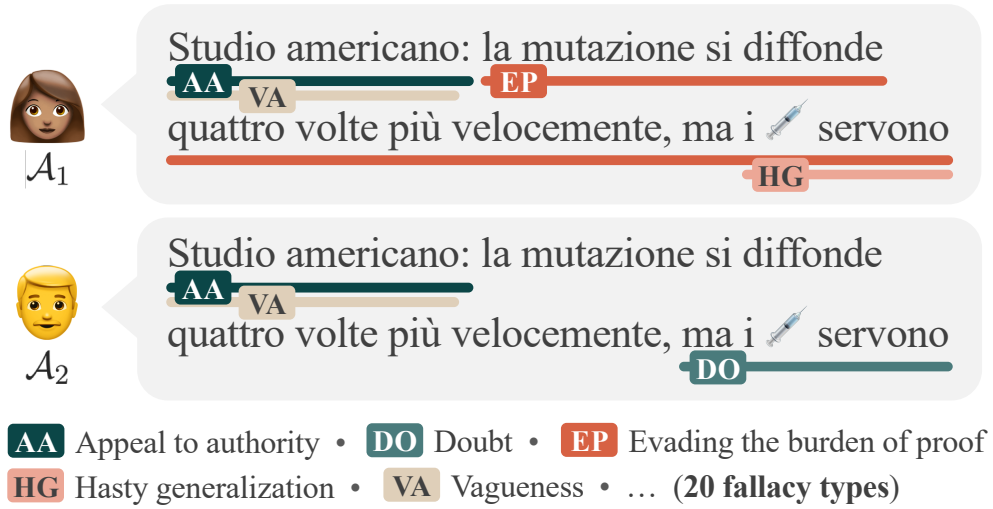
NLP can support this, but current datasets have some limitations:

- Coarse-grained annotations (e.g., *text-level*), few fallacy types, or few annotations
- One fallacy for each text or no overlaps between fallacies in the *span-level* case
- Single ground truth: genuine disagreement is not taken into account!

# 🐻 FAINA dataset

Dataset for **fine-grained** fallacy detection with **human label variation**

- Focuses on **Italian** social media posts
- Embraces **multiple plausible answers** and **natural disagreement**
- Large inventory of **20 fallacy types**
- Fine-grained annotation at the **span-level** with potential **overlaps**






**en**: *American study: mutation spreads four times faster, but 💉 are needed*

# 🐻 FAINA dataset: *data collection* and *annotation*

Public discourse on Twitter/X ***minimizing temporal, topic, and stylistic biases***

- **Multi-year**: ⏳ 4-year time frame (2019-01 — 2022-12)
- **Multi-topic**: 🔄 migration, 🌱 climate change, and 🏥 public health
- **Author diversity**: resample posts by same authors after their most impactful one

*by like+retweet [4] in each month/topic combination*

Minimize errors whilst keeping signals of ***human label variation***

- **5 rounds** of *manual* annotation/discussion (resolve *errors*, keep *genuine disagreement*)
- **2 expert annotators** with different sociodemographics and backgrounds



[4] ***Nakov et al., 2022***. "Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets". CLEF.

# 🐻 FAINA dataset: *fallacy categories*

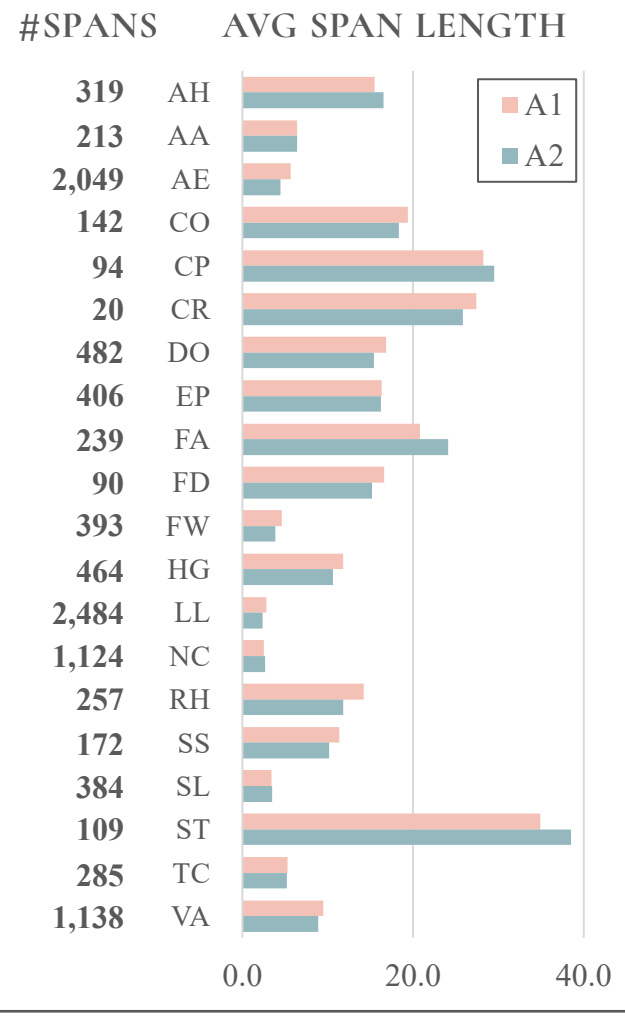
- **[AH]** Ad hominem
- **[AA]** Appeal to authority
- **[AE]** Appeal to emotion
- **[CO]** Causal oversimplification
- **[CP]** Cherry picking
- **[CR]** Circular reasoning
- **[DO]** Doubt
- **[EP]** Evading the burden of proof
- **[FA]** False analogy
- **[FD]** False dilemma
- **[FW]** Flag waving
- **[HG]** Hasty generalization
- **[LL]** Loaded language
- **[NC]** Name calling or labeling
- **[RH]** Red herring
- **[SS]** Slippery slope
- **[SL]** Slogan
- **[ST]** Strawman
- **[TC]** Thought-terminating cliché
- **[VA]** Vagueness

# 🐻 FAINA dataset: some *statistics*

- **58,490 tokens** in 1,440 social media posts
- **11,064 spans** – $5{,}532_{\pm 253}$ spans/annotator
- **Avg token length of 7.6**$_{\pm 9.3}$ – from 2.5 [LL] to 36.3 [ST]
- **Dense annotation** – $3.8_{\pm 0.2}$ spans/post
- **Frequent overlaps** – up to $23\%_{\pm 2\%}$ for TC and AE

**Data splits**: 80% `train/dev` set, 20% `test` set

- Topics, time, and labels **equally distributed** across sets



# FADEIT: *task description* and *evaluation*

**Task description**

Given the text of a social media post, detect the fallacies expressed in it

- **SUBTASK A (*post-level fallacy detection*)**: detect which fallacy types (if any) are expressed in the post
- **SUBTASK B (*span-level fallacy detection*)**: detect all the (potentially overlapping) text segments in the post that express fallacies, and give each of them a type

✨ **Scores on individual test sets are macro-averaged**

**Metrics**

Micro-averaged P, R, F1
Macro-averaged P, R, F1

span-level micro-averaged P, R, F1
span-level macro-averaged P, R, F1
(in *strict* and *soft* evaluation modes)

**accounts for the varying severity of errors**
(0.5 if *pred* is an immediate parent of *gold*)

# FADEIT: *participation*

FADEIT attracted interest from both **academia and industry**, with teams from institutions across **five different countries**: 🇮🇳 🇮🇹 🇯🇵 🇳🇱 🇻🇳

- Each team was allowed to submit **up to 3 runs per subtask**

We received a total of **25 runs** by **7 teams**

- **SUBTASK A**: 16 runs by 6 teams
- **SUBTASK B**: 9 runs by 3 teams
- **SUBTASK A** and **SUBTASK B**: 6 runs by 2 teams

# SUBTASK A: *results*

| | Team | Run | micro-averaged P | micro-averaged R | micro-averaged $F_1$ | macro-averaged P | macro-averaged R | macro-averaged $F_1$ |
|---|---|---|---|---|---|---|---|---|
| 1 | TiGRO | 3 | 53.24 | 59.99 | **56.39** | 34.47 | 34.74 | 33.35 |
| 2 | MALTO | 2 | 55.75 | 53.60 | 54.63 | 41.95 | 30.95 | 32.63 |
| 3 | TiGRO | 2 | 53.43 | 51.48 | 52.41 | 35.91 | 27.21 | 27.94 |
| 4 | UNICA | 1 | 55.22 | 45.23 | 49.71 | **47.13** | 37.28 | **36.75** |
| 5 | TiGRO | 1 | **62.52** | 39.85 | 48.65 | 27.46 | 18.11 | 20.57 |
| 6 | UNICA | 3 | 51.19 | 44.26 | 47.45 | 35.08 | 24.95 | 26.14 |
| 7 | Kenji-Endo | 1 | 49.38 | 45.55 | 47.37 | 14.25 | 17.17 | 13.71 |
| 8 | UNICA | 2 | 58.70 | 38.44 | 46.44 | 35.68 | 19.53 | 22.76 |
| 9 | MALTO | 1 | 46.56 | 43.65 | 45.05 | 28.11 | 23.33 | 23.17 |
| | *MVML-ALB* | | *64.29* | *34.41* | *44.82* | *37.80* | *15.42* | *19.68* |
| 10 | RBG-AI | 2 | 33.09 | 57.78 | 42.07 | 26.54 | 46.21 | 29.32 |
| * | MALTO | 3 | 37.45 | 45.04 | 40.88 | 23.53 | 30.02 | 25.64 |
| 11 | RBG-AI | 3 | 30.65 | 57.62 | 40.00 | 31.08 | 54.90 | 31.31 |
| 12 | Label | 3 | 27.60 | **68.08** | 39.26 | 22.01 | **56.97** | 29.04 |
| 13 | Label | 1 | 52.82 | 30.94 | 38.96 | 14.50 | 10.13 | 10.31 |
| 14 | RBG-AI | 1 | 36.35 | 41.11 | 38.57 | 32.77 | 29.60 | 23.42 |
| 15 | Label | 2 | 52.76 | 30.11 | 38.32 | 14.62 | 10.17 | 10.82 |
| | *MVML-UMB* | | *38.53* | *14.28* | *20.84* | *15.13* | *3.45* | *5.10* |

## Overall insights

- **Encoder-based models** confirm to be extremely competitive in the task – 1°, 2°, 3°, and 5° $F1_{(m)}$
- **Multi-task learning** is effective – 1° and 3° $F1_{(m)}$, 2° and 7° $F1_{(M)}$
- **Decoder-based models** (esp. *closed-source* ones) capture well under-represented labels, positively impacting $F1_{(M)}$

# SUBTASK B: *results*

| | | | STRICT MODE | | | | | | SOFT MODE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | micro-averaged | | | macro-averaged | | | micro-averaged | | |
| | Team | Run | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| 1 | PuDy | 1 | 27.68 | 37.83 | 31.97 | 26.24 | 22.39 | 21.11 | 43.76 | 60.88 | **50.92** |
| 2 | PuDy | 2 | 25.96 | **40.80** | 31.73 | 19.24 | 23.73 | 18.95 | 40.30 | **64.79** | 49.69 |
| 3 | PuDy | 3 | 29.90 | 32.96 | 31.36 | 20.99 | 18.83 | 18.20 | 45.75 | 52.36 | 48.83 |
| 4 | TiGRO | 3 | **47.82** | 37.67 | **42.13** | **35.69** | 23.23 | 25.79 | **51.01** | 40.25 | 44.98 |
| 5 | TiGRO | 1 | 38.33 | 40.35 | 39.30 | 31.52 | **25.30** | **26.05** | 42.50 | 45.20 | 43.80 |
| 6 | TiGRO | 2 | 38.23 | 40.05 | 39.11 | 29.85 | 24.16 | 24.76 | 42.68 | 44.87 | 43.74 |
| | *MVMD-ALB* | | *48.83* | *26.87* | *34.66* | *36.13* | *16.42* | *20.87* | *52.98* | *29.48* | *37.89* |
| 7 | RBG-AI | 3 | 19.47 | 25.25 | 21.99 | 17.68 | 16.43 | 15.18 | 24.18 | 32.44 | 27.71 |
| 8 | RBG-AI | 2 | 19.21 | 18.24 | 18.71 | 17.52 | 12.98 | 12.66 | 24.67 | 23.96 | 24.31 |
| 9 | RBG-AI | 1 | 17.13 | 11.01 | 13.41 | 16.66 | 7.67 | 9.55 | 20.70 | 13.27 | 16.17 |
| | *MVMD-UMB* | | *60.94* | *3.05* | *5.80* | *10.51* | *3.21* | *3.80* | *65.97* | *3.28* | *6.25* |

## Overall insights

- **Encoder-based models** lead the rank, exhibiting different strengths according to *strict/soft* $F1_{(M)}$ *vs* $F1_{(m)}$ metrics

- **Decoder-based models** seem to lag behind due to the intrinsic difficulty of the task; however, they are useful for **contextual data augmentation** purposes

# *Analysis* and *discussion*

**Models**: all teams used transformer-based language models

- Encoder-based: AlBERTo, UmBERTo, mmBERT (also in a *multi-task learning* framework)
- Decoder-based: Gemma3 12B, Llama3.1 8B, Mixtral 8x7B, GPT-5, GPT-5.1, Gemini (also in *few-shot*)

**Label variation & extra-linguistic information**: little-explored signals

- Human label variation has been explored by 2 teams, extra-linguistic info (e.g., topic) by 1 team only

**Data augmentation**: diverse approaches/assumptions, different results

- E.g., Full paraphrases (ChatGPT / GPT-5.1), span context perturbation (Gemini), back-translation

# Conclusions

- FADEIT has attracted **notable interest** from the research community
- **Encoder-based fine-tuning** is competitive, esp. w/ **multi-task learning**
- **LLMs** struggle w/ *span-level* detection, but useful for **data augmentation**
- Results indicate that there is **ample room for further research**

✨ Check the great work by teams at **POSTER SESSION A** (11:15 – 12:15)! ✨

EVALITA 2026 — February 26th, 2026

# FadeIT at EVALITA 2026

## *Fallacy Detection in Italian Social Media Texts Task*

ALAN **RAMPONI** and SARA **TONELLI**

*Fondazione Bruno Kessler, Italy*

*Funded by the European Union's Horizon Europe research and innovation programmes under GA No. 101070190 (AI4Trust) and under the Marie Skłodowska-Curie GA No. 101073351 (HYBRIDS).*