ACTI at EVALITA 2023: Automatic Conspiracy Theory Identification Task Overview

Giuseppe Russo[1], Niklas Stoehr[1] and Manoel Horta Ribeiro[2]

[1]ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
[2]EPFL, Rte Cantonale, 1015 Lausanne, Switzerland

Abstract

English. Automatic Conspiracy Theory Identification (ACTI) is a new shared task proposed for the first time at the EVALITA 2023 evaluation campaign. ACTI is based on a new, manually labeled dataset of comments scraped from conspiratorial Telegram channels and consists of two subtasks: (1) identifying conspiratorial content (conspiratorial content classification); and (2) classifying content into specific conspiracy theories (conspiratorial category classification). A total of 15 teams participated in the task with 81 submissions. In this task summary, we discuss the data and task, and outline the best-performing approaches that are largely based on large language models. We conclude with a brief discussion of the application of large language models to counter the spread of misinformation on online platforms.

Keywords

Conspiracy Theory, Content Moderation, Large Language Models, Computational Social Science

1. Introduction

From ancient tales of secret societies, [2] to speculation on whether the moon landing happened [3], belief in conspiracy theories has been prevalent throughout human history [4] and has inflicted harm upon individuals and groups falsely accused of wrongdoing [5]. For example, in the middle ages, the Blood Libel conspiracy theory falsely accused Jews of murdering Christian boys, fostering their persecution [6].

Fast-forward to the digital age, the Internet has emerged as the prominent medium through which individuals are exposed to conspiracy theories [7, 8]. Indeed, mainstream and fringe platforms have served as *de-facto* incubators of online conspiracies [9]. Notably, the impact of online conspiracy theories has been far-reaching, inciting real-world violence and influencing public health. The QAnon conspiracy, which gained momentum during the Trump administration, was pivotal in planning the 2021 invasion of the US Capitol [10, 11]. At the same time, the conspiracy theories associated with COVID-19 fueled anti-vaccination sentiments and skepticism towards public health measures [12, 13].

Mainstream platforms limit the diffusion of conspiratorial content through interventions that range from banning online communities [14, 15] to telling users that the information presented may be inaccurate [16]. While these interventions may help curb the proliferation of conspiracy theories in online spaces [17], they require a fundamental technology: ways to identify conspiratorial content accurately and at scale across various languages and cultural contexts [18].

In this context, we propose the *Automatic Conspiracy Theory Identification (ACTI)* task. Considering a dataset with over 25 thousand posts in Italian extracted from five Telegram channels, the ACTI consists of two subtasks: (i) a binary classification task where the goal is to determine if a given text piece is conspiratorial or not; and (ii) a multi-class classification task to recognize specific conspiracy theories.

2. Task Description

The ACTI shared task comprises two subtasks, which we describe below.

A: Conspiratorial Content Classification. The first subtask is determining whether a Telegram post is conspiratorial. We consider conspiratorial texts as those that either: (i) express the belief that influential people create major events (e.g., COVID-19) to protect their interests or (ii) interpret events in a way that supports the narrative of a conspiracy theory.

Note that this definition of "conspiratorial" is broad, as *texts* may be defined as conspiratorial if they undermine commonly accepted views on societal issues. For example, the text "il cancro femminista sta prendendo piene" should be classified as conspiratorial, as it subtly supports a broader theory claiming that women's rights are destroying the stability of Western societies.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT [1]
russog@ethz.ch (G. Russo); niklas.stoehr@inf.ethz.ch (N. Stoehr); manoel.hortaribeiro@epfl.ch (M. H. Ribeiro)

B: Conspiracy Category Classification. The second subtask is determining which conspiracy theory a post belongs to. In particular, we consider four possible conspiracy theories.

- **COVID-19**: Text concerning vaccine production, 5G, and non-pharmacological interventions as a tool of control over people. Texts denying the pandemic was a real event or minimizing its importance.
- **QAnon**: Texts associated with the QAnon theory. According to QAnon, a group of Satanic cannibalist sex abusers conspired against former U.S. President Donald Trump during his term in office. This theory extended far over its original scope embodying other beliefs that support (among the others) the idea that women are enemies (hate against women) and that a powerful elite (led by public figures like Pope Francis, Queen Elizabeth, and Hillary Clinton) is trying to organize a New World Order.
- **Flat-Earth**: Texts associated with the claim that the earth is flat and that influential organizations hide this fact from laypeople. Usually, the flat-earth conspiracy theory is supported by pseudo-scientific evidence.
- **Pro-Russia** : Texts associated with conspiratorial beliefs promoting Russian interests, e.g., that nazists control Ukraine's governments and army.

3. Data Collection

To gather the necessary data for the ACTI task, we employ a customized web crawler using the *Selenium* and *BeautifulSoup* libraries in Python. Our web crawler targets specific sources known for hosting conspiratorial content on the Telegram platform.

Specifically, we focus on a selection of Telegram channels that gained notoriety for promoting far-right ideologies and disseminating conspiracy theories. The channels we collect data from include: *Qlobal-Change Italia, Basta Dittatura, Studi Scientifici Vaccini, Terra Piatta,* and *Dentro La Notizia.* For example, the channel "Basta Dittatura" has been actively involved in various events, including the siege of a trade union headquarters, indicating its strong affiliation with conspiratorial movements.

Our data collection process spanned from January 1, 2020, to June 30, 2020, during which we capture and retain comments written in Italian. To ensure sufficient text for analysis, we filtered out comments with less than ten words. We gathered a dataset comprising $25,612$ posts extracted from these five Telegram channels. We summarize statistics about our dataset in fig. 1

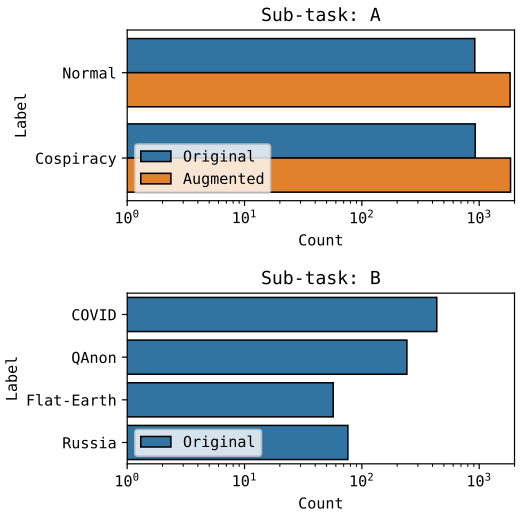

Figure 1: Distribution of labels for Subtask A and Subtask B.

3.1. Annotation Process

The data collection process for our study on conspiratorial content in online channels involved several steps to ensure the accuracy and relevance of the collected data. One of the main challenges we encountered was the presence of non-conspiratorial content within the channels. While some comments discussed conspiratorial topics, others contained valid points or critiques regarding conspiratorial perspectives. Additionally, some comments were deemed meaningless and needed to be filtered out to maintain the integrity of the dataset.

To address this, we employed two human annotators who were responsible for labeling the comments according to three categories: "Not Relevant," "Non-Conspiratorial," and "Conspiratorial." The "Not Relevant" label was assigned to comments that did not contribute to the discussion, while the "Non-Conspiratorial" label was used for comments that did not involve conspiratorial content. The "Conspiratorial" label indicated comments that contained or supported conspiratorial discussions. For the comments labeled as "Conspiratorial," we further categorized them into four subcategories: "QAnon", "Covid19", "Russia", and "Flat-Earth". These subcategories allowed us to analyze specific conspiracy theories in greater detail. The definitions of conspiratorial content are based on established studies in the field [19, 20], ensuring consistency and clarity in our annotation process.

To assess the agreement between the annotators, we calculated inter-annotator agreement rates using Cohen's κ coefficient. The two annotators achieved high agree-

ment levels, with a Cohen's κ of 0.93 for the first task and 0.86 for the second task, demonstrating the reliability of the annotation process. To maintain data integrity, we excluded comments that did not receive the same classification from both annotators. Additionally, comments labeled "Not Relevant" were discarded from the dataset to focus solely on relevant conspiratorial content.

Our data collection process yielded 2,301 comments for the first subtask and 1,110 comments for the second subtask. This resulted in a curated dataset that provides a solid foundation for research on conspiratorial content in online discussions.

4. Evaluation Measures

We chose different evaluation metrics for subtasks A and B because of the distribution of the labels provided by the annotators. In particular

A: Conspiratorial Content Classification. The systems submitted by participants are evaluated using the standard accuracy measures and ranked accordingly.

B: Conspiracy Category Classification. Given the class imbalance for the four types of conspiracy theories we identified, we opt for using as a metric the F1-Score. For a multi-class classification problem, we calculate the F1-score per class in a one-vs-rest manner. We rate each class separately, computing the F1-score for each conspiracy theory in our dataset. To obtain a single score, we then average the per-class F1-scores.

Baselines. We follow the same methodological approach to provide a baseline for both subtasks. Specifically, the baselines for subtasks A and B are a Random Forest trained on a bag-of-words representation of the comments. In particular, we trained the random forest with 500 estimators and validated it using a five-fold cross-validation. These baselines achieve 0.63 accuracy for the first and 0.68 for the second subtask, respectively.

5. Results

A total of fifteen teams submitted from seven institutions participated in the two tasks. Specifically, eight teams submitted for the conspiracy content classification and seven for the conspiracy category classification. In total, we obtain 81 submissions. In Tables 1 and 2, we show the results for both submissions.

Rank	Team Name	Score
1	UPB	0.85712
2	extremITA	0.85647
3	HFI	0.84469
4	Flavio Giobergia	0.83709
5	Michael Vitali	0.82297
6	Giacomo Cignoni	0.82284
7	sCambiaMenti	0.79182
8	Mario Graff	0.78207

Table 1
Conspiratorial Content Classification: Ranking of the eight teams joining the task. The best performing approach was obtained via Contrastive Training

5.1. Conspiratorial Content Classification

Table 1 reports the results of the Conspiratorial Content Classification subtask, which received 40 submissions. The "UPB" team from the University Politehnica of Bucharest achieves the highest accuracy of 0.85 with five submissions. Their methodology consists of an Italian language Sentence Transformer model trained it using contrastive learning. Due to imbalanced data, the participants integrated a data augmentation step in their classification pipeline. Specifically, their methodology generates synthetic data via a Large Language Model (LLM). These synthetic data are then used for training the model. Figure 2 provides an overview of their methodology. The second best-performing team submitted a LLM-based model as well. Specifically, the participants tested extremIT5 (an encoder-decoder model) and extremITLLaMA (an instruction-tuned Decoder-only Large Language Model) designed for handling Italian instructions. While LLM-based approaches performed best, other participants developed methods based on transformers and ensembles, which achieved an accuracy of over 0.80.

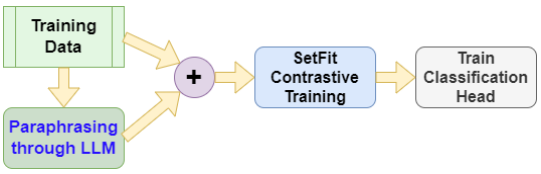

Figure 2: Overview of the methodology used by the 'UPB' team from the University Politehnica of Bucharest, who achieved the highest accuracy of 0.85 with five submissions

5.2. Conspiracy Category Classification

Table 2 reports the results of the Conspiracy Category Classification task, which received 41 submissions in total. Once again, the "UPB" team from the University Politehnica of Bucharest achieved the highest F1-score

Rank	Team Name	Score
1	UPB	0.91225
2	Michael Vitali	0.89826
3	HFI	0.89476
4	Giacomo Cignoni	0.88534
5	extremITA	0.85562
6	Flavio Giobergia	0.83600
7	sCambiaMenti	0.67507

Table 2
Conspiratorial Category Classification: Ranking of the eight teams joining the task. The best-performing approach was obtained via Contrastive Training

(0.91). Interestingly, the data augmentation process used for subtask A did increase the model performance. Indeed, the participants submitted the same transformer-based model trained with contrastive learning, excluding the data augmentation block. The second-best performing team from Tor Vergata University (Michael Vitali) achieved an F1-Score of 0.89. They fine-tuned two BERT models, one in Italian and one multilingual, and combined them in an ensemble. Numerous teams performed well in this task, achieving F1-scores beyond 0.80. Only one participant obtained a result slightly inferior to the provided baseline.

6. Discussion

A comprehensive analysis of the submitted systems reveals that most participants opted for LLMs-based models. Within this context, we emphasize two distinct approaches the participants employ: (i) prompting and (ii) data augmentation. Upon thorough analysis, we find that while prompting does result in positive outcomes, the predictive capabilities of zero-shot LLMs are still inferior to systems that have been finetuned for a specific task.

6.1. Prompting Large Language Models

Prompting consists of providing information to a trained model to predict output labels for a task. It is a task-agnostic approach, making it versatile and widely applicable [21]. This is achieved through concise instructions, referred to as prompts, which guide the model's behavior.

The power and flexibility of prompting LLMs are well exemplified by team ExtremITA's approach: adopting a Large Language Model (LLM) to address all EVALITA tasks simultaneously. For the ACTI task, ExtremITA is prompted with simple questions such as *"Does this text talk about a conspiracy? Answer yes or no"* and *"Which conspiracy theory is discussed in this text: Covid, QAnon, Flat Earth, or Russia?"* This approach ranks second in subtask A of ACTI, achieving a score of 0.86 F1-score. However, it significantly drops in performance in subtask B, ranking fifth with a score of 0.85 F1-score.

In other EVALITA tasks, ExtremITA showed significant variability in predictive capacity. It ranked first in eight out of twenty-five tasks, but in the remaining subtasks, it performed poorly, ranking between fifth and eleventh. These results confirm LLMs' high potential and applicability in real-world scenarios. However, the high variability of results shows that LLMs need help improving over models fine-tuned on specific tasks. Future research should focus on refining prompting techniques to improve the predictive capacity of LLMs at the single-task level.

6.2. Augmenting Data with LLMs

The winning team of subtasks A and B ("UPB") used an approach based on data augmentation via Large Language Models and the training of sentence transformers with contrastive learning. This approach tackles the challenge of the acquisition of conspiratorial data. Indeed, collecting and labeling conspiratorial data requires substantial efforts by domain specialists. This approach tested the possibility of leveraging LLMs to generate synthetic data and use it to train systems for automatically detecting conspiratorial content based. However, it is essential to note that validating the quality of data generated by LLM is an open issue within the NLP community. While LLMs can effectively produce synthetic content, assessing its authenticity and alignment with real-world conspiratorial beliefs is crucial. The lower performance of the model augmented with synthetic data suggests that the quality of the generated data drastically impacts the overall model performance. Therefore, human evaluation is mandatory to evaluate the effectiveness of these approaches.

7. Conclusion

A recent position paper [22] asks whether EVALITA has reached its end in light of the increasing use of LLMs. However, based on the outcomes presented in this report, it becomes evident that the answer remains negative. The challenges posed by EVALITA tasks persist as a crucial asset in comprehending and advancing language resources and tools specifically for the Italian language. This fact is exemplified by transformer-based models' differing rankings, demonstrating the evaluation campaign's diversity and significance. However, the performance achieved by LLMs is undoubtedly pushing the limits of some tasks, especially text classification tasks. In conclusion, while LLMs have shown great potential, EVALITA remains an essential platform for improving language tools for the Italian language.

Acknowledgments

We thank the data annotators for their careful and valuable work. Niklas Stoehr acknowledges funding from the Swiss Data Science Center (SDSC) fellowship.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] J. Roisman, The rhetoric of conspiracy in ancient Athens, Univ of California Press, 2006.

[3] Macey, Richard, One giant blunder for mankind: how nasa lost moon pictures, 2006.

[4] J.-W. Van Prooijen, K. M. Douglas, Conspiracy theories as part of history: The role of societal crisis situations, Memory studies 10 (2017) 323–333.

[5] K. M. Douglas, J. E. Uscinski, R. M. Sutton, A. Cichocka, T. Nefes, C. S. Ang, F. Deravi, Understanding conspiracy theories, Political psychology 40 (2019) 3–35.

[6] E. M. Rose, The Murder of William of Norwich: The Origins of the Blood Libel in Medieval Europe, Oxford University Press, 2015.

[7] C. Sunstein, # Republic: Divided democracy in the age of social media, Princeton university press, 2018.

[8] T. Goertzel, Belief in conspiracy theories, Political psychology (1994) 731–742.

[9] Anti-Defamation League, ADL statement on Facebook's decision to finally ban QAnon content from platform, https://www.adl.org/news/press-releases/adl-statement-on-facebooks-decision-to-finally-ban-qanon-content-from-platform, 2020.

[10] BuzzFeed News, 'the rioters who took over the capitol have been planning online in the open for weeks', 2021. URL: https://www.buzzfeednews.com/article/janelytvynenko/trump-rioters-planned-online.

[11] B. Collins, B. Zadrozny, Facebook bans qanon across its platforms, https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339, 2020.

[12] N. Puri, E. A. Coomes, H. Haghbayan, K. Gunaratne, Social media and vaccine hesitancy: new updates for the era of covid-19 and globalized infectious diseases, Human vaccines & immunotherapeutics 16 (2020) 2586–2593.

[13] J. Friedrichs, N. Stoehr, G. Formisano, Fear-anger contests: Governmental and populist politics of emotion, Online Social Networks and Media 32 (2022) 100240. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468696422000428. doi:10.1016/j.osnem.2022.100240.

[14] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, Proceedings of the ACM on Human-Computer Interaction 1 (2017) 1–22.

[15] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[16] S. Zannettou, " i won the election!": An empirical analysis of soft moderation interventions on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 15, 2021, pp. 865–876.

[17] G. Russo, M. Horta Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, in: Proceedings of the 15th ACM Web Science Conference 2023, 2023, pp. 251–259.

[18] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Helping a friend or supporting a cause? disentangling active and passive cosponsorship in the U.S. congress, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2952–2969. URL: https://aclanthology.org/2023.acl-long.166.

[19] C. R. Sunstein, A. Vermeule, Conspiracy theories: Causes and cures, Journal of political philosophy 17 (2009) 202–227.

[20] V. Swami, R. Coles, S. Stieger, J. Pietschnig, A. Furnham, S. Rehim, M. Voracek, Conspiracist ideation in britain and austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories, British Journal of Psychology 102 (2011) 443–463.

[21] C. Lefebvre, N. Stoehr, Rethinking the event coding pipeline with prompt entailment, in: arXiv, volume 10.48550, 2022. URL: https://arxiv.org/pdf/2210.05257.pdf.

[22] V. Basile, et al., Is EVALITA done? on the impact of prompting on the italian nlp evaluation campaign, in: CEUR Workshop Proceedings, volume 3287, Debora Nozza, Lucia C. Passaro, Marco Polig-

nano, 2022, pp. 127–140.