UPB at ACTI: Detecting Conspiracies using fine tuned Sentence Transformers

Andrei **Paraschiv**[1], Mihai **Dascalu**[1]

[1]*University Politehnica of Bucharest 313 Splaiul Independetei, Bucharest, Romania*

Abstract

English. Conspiracy theories have become a prominent and concerning aspect of online discourse, posing challenges to information integrity and societal trust. As such, we address conspiracy theory detection as proposed by the ACTI @ EVALITA 2023 shared task. The combination of pre-trained sentence Transformer models and data augmentation techniques enabled us to secure first place in the final leaderboard of both sub-tasks. Our methodology attained F1 scores of 85.71% in the binary classification and 91.23% for the fine-grained conspiracy topic classification, surpassing other competing systems.

Keywords

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

1. Introduction

Conspiracy theories distort the shared understanding of reality and erode trust in crucial democratic institutions. By substituting reliable, evidence-based information with dubious, implausible, or blatantly false claims, these theories foster a climate of disagreement regarding facts and give undue weight to personal opinions and anecdotal evidence over established facts and scientifically validated theories. Aaronovitch [2] defines conspiracy theories as 'the attribution of deliberate agency to something more likely to be accidental or unintended; therefore, it is the unnecessary assumption of conspiracy when other explanations are more probable.' Due to the rapid spread of information across the internet, coupled with the alarming speed at which false information can proliferate [3], we find ourselves amidst what some have dubbed a "golden age" of conspiracy theories [4]. Being a distinct form of misinformation, conspiracy theories exhibit unique characteristics. Brotherton et al. [5] identified five key attributes commonly found in modern conspiracy theories: government malfeasance, extraterrestrial cover-up, malevolent global conspiracies, personal well-being, and information control.

While embracing conspiracy theories can give individuals a sense of reclaiming power or accessing hidden knowledge, these beliefs can sometimes have negative and dangerous consequences. One recent example is the violent insurrection on the US Capitol on 6 January 2021 driven by conspiracy theories surrounding QAnon and election fraud [6]. Additionally, these theories can serve as powerful tools in the hands of nefarious groups, politicians, or state actors who exploit susceptible communities, manipulating them into taking or endorsing actions that can result in significant and dramatic social repercussions [7, 8].

Building upon the importance of addressing conspiracy theories, efforts have been made to research and develop automated methods for detecting conspiratorial content on various platforms and languages. For instance, as part of the EVALITA 2023 workshop, the organizers of the ACTI shared task introduced a novel approach: the automatic identification of conspiratorial content in Italian language Telegram messages. This initiative aimed to enhance our ability to quickly recognize and respond to conspiracy theories, enabling the promotion of critical thinking and media literacy by providing reliable sources and encouraging evidence-based discourse. Leveraging such advancements can effectively limit the influence of conspiracy theories while fostering a more informed and resilient society.

This paper presents our contribution to the ACTI @ EVALITA 2023 shared task [9]. We focused on employing the power of pretrained Italian language sentence Transformers. To further enhance the performance and address potential biases, we employed Large Language Models (LLMs) to augment the training data, resulting in a more balanced and comprehensive training set. This combination of leveraging pre-trained models and data augmentation techniques formed the foundation of our methodology, enabling us to achieve first place in the final leaderboard of both sub-tasks with F1 scores of 85.71% and respectively 91.23%.

2. Related Work

Recently, online platforms have often banned—entirely deactivated—communities that breached their increasingly comprehensive guidelines. In 2020 alone, Reddit banned around 2,000 subreddits (the name a community receives on the platform) associated with hate speech. Similarly, Facebook banned 1,500 pages and groups related to the QAnon conspiracy theory [10]. While these decisions are met with enthusiasm [e.g., see Anti-Defamation League [11]], the efficacy of "de-platforming" these online communities has been questioned [12, 13]. When mainstream platforms ban entire communities for their offensive rhetoric, users often migrate to alternative *fringe platforms*, sometimes created exclusively to host the banned community [14, 15]. Banning, in that context, would not only strengthen the infrastructure hosting these fringe platforms [12] but allow these communities to become more toxic elsewhere [16]. In order to improve the efficacy of such moderation policies identifying and tracking the propagation of problematic content like conspiracy theories is crucial. For example the Zika virus outbreak in 2016, coupled with the influence of social networks and the declaration of a public health emergency by the WHO, showed the harm the dissemination of conspiracy theories can generated [17, 18].

The COVID-19 pandemic had a profound impact, emphasizing the dangers associated with the proliferation of conspiracy theories. These theories encompassed a wide range of topics, including the virus's origin, its spread, the role of 5G networks, and the efficacy and safety of vaccines. With COVID-related lockdowns in place, people became more reliant on social networking platforms such as Twitter, Facebook, and Instagram, which increased their exposure to disinformation and conspiracy theories. MediaEval 2020 [19] focused on a 5G and COVID-19 conspiracy tweets dataset, proposing two shared tasks to address this issue. The first task involved detecting conspiracies based on textual information, while the second task focused on structure-based detection utilizing the retweet graph. Various systems were proposed to tackle these tasks, employing different approaches such as methods relying on Support Vector Machine (SVM) [20], BERT [21], and GNN [22] . In their study, Tyagi and Carley [23] employed an SVM to classify the stance of Twitter users towards climate change conspiracies. Their findings revealed that individuals who expressed disbelief in climate change tend to share a significantly higher number of other types of conspiracy-related messages compared to those who believe in climate change. Furthermore, Amin et al. [24] manually labeled 598 Facebook comments as Covid-19 vaccine conspiracy or neutral and used a BERT-based model in conjunction with Google Perspective API to classify these messages, providing valuable insights into the prevalence of vaccine conspiracy theories on social media platforms.

Tunstall et al. [25] presented a new approach based on Sentence Transformers[26] called SetFit that focused on data-efficient fine-tuning of sentence embeddings, particularly for binary labels. The training of SetFit follows a two-step process. First, it fine-tuned the sentence embeddings in a contrastive manner. This step helped in optimizing the embeddings for the specific classification task. Subsequently, a classification head was trained using finetuned sentence embeddings, enabling effective classification on the training labels. Their approach aimed to enhance the efficiency and performance of fine-tuning sentence embeddings in scenarios with limited data. The efficacy and power of Sentence-Transformers has been shown in multiple tasks spanning from text generation [27, 28] to sentence classification tasks.[29, 30, 31]. These models capture the semantic and contextual information of sentences or paragraphs, enabling nuanced representations of textual data. Leveraging such models, Bates and Gurevych [32] used SetFit to propose LAGONN, a hate speech and toxic messages classification framework for content moderation.

3. Method

3.1. Task Description

The ACTI @ EVALITA 2023 organizers put forth two sub-tasks for participants to address. The first sub-task [9] involved binary classification, where participants were provided with a dataset consisting of 1,842 training samples and 460 test samples. The objective was to classify messages as either conspiratorial or non-conspiratorial. The second sub-task focused on fine-grained conspiracy topic classification. Participants were required to classify messages into one of four specific conspiracy topic classes: Covid, QAnon, Flat-Earth, or Russia-conspiracy. A training set of 810 records was provided for this sub-task, while the evaluation test set contained 300 samples. Table 1 shows the class distribution for both sub-tasks.

	Classes	Count
Sub-Task A	Non Conspiratorial	917
	Conspiratorial	925
Sub-Task B	Covid	435
	QAnon	242
	Flat-Earth	76
	Russian	57

Table 1
ACTI Dataset distribution for the training sets on Sub-task A and B.

The macro F1 score was adopted as a criterion to evaluate the two sub-tasks. During the competition, 30% of the test dataset was immediately evaluated on the Public Leaderboard, giving participants an initial indication of their model's performance. However, the final evaluation was conducted on the remaining 70% of private entries. These final evaluation scores were then used to compile the Private Leaderboard made public after the conclusion of the competition.

3.2. Sentence Transformer and Data Augmentation

We considered an Italian language Sentence Transformer model for our submissions and trained contrastive with SetFit[1] as described by Tunstall et al. [25]. Since the training dataset is highly imbalanced between the conspiratorial classes (see in Table 1), we integrated a data augmentation step in our classification pipeline, as seen in Figure 1.
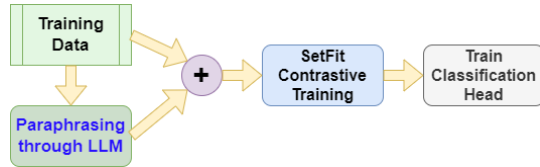

Figure 1: End-to-End training Pipeline.

In the data augmentation step, we used an LLM to create paraphrases for our training data using the prompt "riformulare questo testo: [*comment_text*]" and different seeds to create variations of the answers. In our experiments, we used "text-davinci-003" from the GPT-3 family[2] and the mT5 model finetuned on Italian language paraphrases[3]. We set a high temperature (t=0.9) for the LLMs to ensure diverse text generation. The distribution for the augmented dataset is shown in Table 2.

	Classes	Count
Sub-Task A	Non Conspiratorial	1,822
	Conspiratorial	2,524
Sub-Task B	Covid	779
	QAnon	672
	Flat-Earth	362
	Russian	322

Table 2
Class distribution on the augmented training sets used for Sub-task A and B

---

[1]https://github.com/huggingface/setfit
[2]https://platform.openai.com/docs/models
[3]https://huggingface.co/aiknowyou/mt5-base-it-paraphraser

Sentence-Transformers are pretrained Transformer models finetuned in a Siamese network, such that semantically similar sentences or paragraphs are projected near each other in the embedding space; in contrast, the distance in the embedding space is maximized for sentence pairs that are different. In our experiments, we used several Italian pretrained Sentence Transformers from the Huggingface Hub[4], as mentioned in Table 3. The first step in the SetFit training process involves generating positive and negative triplets. Positive triplets consist of sentences from the same class, while negative triplets contain sentences from different classes. The training data is expanded by including positive and negative triplets, providing a more comprehensive and diverse training set. The Sentence Transformer captures the contextual and semantic information of the messages, providing a powerful feature representation. In the second step, a fully connected classification head is trained on top of the Sentence-Transformer to distinguish between the available classes.

4. Results

Besides experimenting with different pre-trained models, as shown in Table 3, we also performed grid search tuning with several key hyper-parameters, namely the number of iterations, the learning rate, and the number of epochs for training. The number of iterations determined the quantity of generated triplets during training. By adjusting this parameter, we controlled the training data's size, potentially influencing the model's ability to generalize and capture important patterns. We set the maximum sequence length for the tokenizer to 512 for all of our experiments. We withheld 20% of the training data to evaluate the performance of the trained models during the development time.

The best-performing model differed between the subtasks. The best-performing model in the binary classification sub-task was based on "efederici/sentence-BERTino". This model was trained on the "text-davinci-003" augmented dataset for 1 epoch. We used 5 iterations and a learning rate of 1e-05. In contrast, the larger "nickprock/sentence-bert-base-italian-xxl-uncased" model performed best for the fine-grained conspiracy topic classification sub-task. We trained this model on the same dataset for 1 epoch. The learning rate used was 1e-05, and the number of iterations was set to 10. This model yielded the best results in both Leaderboards (see Table 4).

We conducted an ablation study after the competition ended to assess the impact of data augmentation. We trained the best-performing models under different

---

[4]https://huggingface.co/models

Model	Embedding Size
efederici/sentence-BERTino	768
efederici/sentence-bert-base	768
efederici/sentence-BERTino-3-64	64
efederici/mmarco-sentence-BERTino	768
efederici/sentence-it5-base	512
efederici/sentence-it5-small	512
nickprock/sentence-bert-base-italian-uncased	768
nickprock/sentence-bert-base-italian-xxl-uncased	768
aiknowyou/aiky-sentence-bertino	768

Table 3
Sentence-Transformer models considered in our experiments.

	Public Leaderboard	Private Leaderboard
Sub-Task A	85.36%	85.71%
Sub-Task B	87.62%	91.23%

Table 4
Best model performance on the Public and Private leaderboard

conditions: a) using the original training data with 20% reserved for development evaluation, b) considering the entire original training data, and c) employing the dataset that was augmented with the mT5 paraphrasing LLM. The results in Tables 5 and 6 show the importance of the augmentation step.

	Public Leaderboard	Private Leaderboard
No augmented data	75.80%	81.29%
No augmented data with development set	79.36%	83.83%
mT5 augmented Data	78.15%	82.25%

Table 5
Ablation Study for Sub-Task A.

	Public Leaderboard	Private Leaderboard
No augmented data	83.32%	93.67%
No augmented data with development set	83.39%	89.67%
mT5 augmented Data	83.24%	87.07%

Table 6
Ablation Study for Sub-Task B.

In the case of sub-task A, the additional data substantially influenced both the Public and Private test results. The augmented dataset led to significant improvements in performance. However, we see a decline in the Private Leaderboard for the fine-grained task results as the amount of data increased, despite the Public Leaderboard performance keeping the same. This performance decline could be attributed to an unusual distribution difference between the Public and Private test rows. Furthermore, the quality of the paraphrases used in the augmentation process played a crucial role in both sub-tasks. The poor performance achieved by the mT5 model suggests that the quality of the generated paraphrases has a notable impact on the overall model performance. Similarly, a drastic decrease in performance was observed for the second sub-task private leaderboard, arguing for the questionable quality of the paraphrases.

5. Conclusion

In this paper, we described our approach addressing the two sub-tasks in the ACTI @ EVALITA 2023 competition. The challenge focuses on automatically detecting conspiratorial Telegram messages and the classification into four conspiracy topics: Covid, QAnon, Flat-Earth, and Russian conspiracies. Through the utilization of text augmentation techniques and the training of Sentence-Transformers with contrastive learning, we developed robust classifiers. Our best models achieved first place in the Private Leaderboard on both tasks with F1 scores of 85.712% in the binary classification and 91.225% for the fine-grained conspiracy topic classification. This paper contributes to the growing body of research on conspiracy theory detection and emphasizes the effectiveness of leveraging pre-trained models and data augmentation techniques. Our results argue the potential of these approaches in addressing the challenges posed by conspiracy theories and their propagation in online platforms.

Acknowledgement

124812, within POC.

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] D. Aaronovitch, Voodoo histories: The role of the conspiracy theory in shaping modern history, NY: Riverhead (2010).

[3] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[4] H. W. Hanley, D. Kumar, Z. Durumeric, A golden age: Conspiracy theories' relationship with misinformation outlets, news media, and the wider internet, arXiv preprint arXiv:2301.10880 (2023).

[5] R. Brotherton, C. C. French, A. D. Pickering, Measuring belief in conspiracy theories: The generic conspiracist beliefs scale, Frontiers in psychology 4 (2013) 279.

[6] A. Seitz, Mob at u.s. capitol encouraged by online conspiracy theories, The Associated Press (2021). URL: https://apnews.com/article/donald-trump-conspiracy-theories-michael-pence-media-social-media-daba3f5dd16a431abc627a5cfc922b87.

[7] W. Audureau, Why conspiracy theorists and the kremlin echo each other's disinformation, Le Monde (2023). URL: https://www.lemonde.fr/en/les-decodeurs/article/2023/03/02/conspiracy-theorists-the-kremlin-echo-each-other-s-disinformation_6017960_8.html.

[8] I. Yablokov, Russian disinformation finds fertile ground in the west, Nature Human Behaviour 6 (2022) 766–767.

[9] G. Russo, N. Stoehr, M. H. Ribeiro, Acti at evalita 2023: Overview of the conspiracy theory identification task, arXiv preprint arXiv:2307.06954 (2023).

[10] 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.

[11] 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.

[12] E. Zuckerman, C. Rajendra-Nicolucci, Deplatforming our way to the alt-tech ecosystem, Knight First Amendment Institute at Columbia University, January 11 (2021).

[13] 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.

[14] C. Dewey, Washington Post — These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy, and why it banned them, https://wapo.st/3AO7pbl, 2016.

[15] 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, WebSci '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 251–259. URL: https://doi.org/10.1145/3578503.3583608. doi:10.1145/3578503.3583608.

[16] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, R. West, Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–24.

[17] A. Ghenai, Y. Mejova, Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter, in: 2017 IEEE International Conference on Healthcare Informatics (ICHI), 2017, pp. 518–518. doi:10.1109/ICHI.2017.58.

[18] M. J. Wood, Propagating and debunking conspiracy theories on twitter during the 2015–2016 zika virus outbreak, Cyberpsychology, behavior, and social networking 21 (2018) 485–490.

[19] K. Pogorelov, D. T. Schroeder, L. Burchard, J. Moe, S. Brenner, P. Filkukova, J. Langguth, Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020., in: MediaEval, 2020.

[20] M. Moosleitner, B. Murauer, G. Specht, Detecting conspiracy tweets using support vector machines., in: MediaEval, 2020.

[21] A. Malakhov, A. Patruno, S. Bocconi, Fake news classification with BERT, in: Working Notes Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020, volume 2882 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2882/paper38.pdf.

[22] A. Paraschiv, G.-E. Zaharia, D.-C. Cercel, M. Dascalu, Graph convolutional networks applied to fakenews: corona virus and 5g conspiracy, UPB Scientific Bulletin, Series C: Electrical Engineering

83 (2021) 71–82.

[23] A. Tyagi, K. M. Carley, Climate change conspiracy theories on social media, arXiv preprint arXiv:2107.03318 (2021).

[24] M. H. Amin, H. Madanu, S. Lavu, H. Mansourifar, D. Alsagheer, W. Shi, Detecting conspiracy theory against covid-19 vaccines, arXiv preprint arXiv:2211.13003 (2022).

[25] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055 (2022).

[26] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[27] A. Amin-Nejad, J. Ive, S. Velupillai, Exploring transformer text generation for medical dataset augmentation, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 4699–4708.

[28] G. Russo, N. Hollenstein, C. C. Musat, C. Zhang, Control, generate, augment: A scalable framework for multi-attribute text generation, ArXiv abs/2004.14983 (2020).

[29] J. Hong, J. Park, D. Kim, S. Choi, B. Son, J. Kang, Empowering sentence encoders with prompting and label retrieval for zero-shot text classification, 2023. `arXiv:2212.10391`.

[30] G. Piao, Scholarly text classification with sentence bert and entity embeddings, in: Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25, Springer, 2021, pp. 79–87.

[31] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Disentangling active and passive cosponsorship in the u.s. congress, ArXiv abs/2205.09674 (2022).

[32] L. Bates, I. Gurevych, Like a good nearest neighbor: Practical content moderation with sentence transformers, arXiv preprint arXiv:2302.08957 (2023).