Detection of Conspiracy vs. Critical Narratives and Their Elements using NLP

Notebook for the Lab at CLEF 2024

Aish Albladi[1,*], Cheryl D. Seals[1]

[1]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, USA

Abstract

The growing use of digital media or social media has given a platform to the people to deliver their ideas and viewpoints openly. It is facilitating the rapid spread of contrasting opinions openly. Ultimately, this has developed echo chambers or polarized platforms. These platforms either encourage conspiracy theories or critical narratives. So, nowadays, it has become crucial for people to be aware of being part of such narratives. Moreover, it is also necessary for the organisations to identify the conspiracy creators or the ones who do not possess a conspiracy mentality. This can be made possible by the use of natural language processing techniques to do this task automatically. This study presents the solutions to the two most pressing challenges in today's era. The subtask one of this study provides a solution for the identification of conspiracy narratives and critical narratives using the BERT model. Our model achieved an MCC score of 0.80. Secondly, the other subtask helps to recognize the span of the topic relevant to the important constituents of both categories of oppositional narratives, i.e., agent, facilitator, victim, campaigner, objective, negative effect. The overall accuracy of the model is 95%. These models can be used for the development of an algorithm that can automatically classify the conspiracy theories and critical narratives.

Keywords

BERT, Conspiracy Theory, Critical Thinking, RoBERTa, Social-media

1. Introduction

The advancement of IT industry has resulted in a notable transformation of traditional social interactions that were previously restricted by time and location [1], [2]. Social media has dominated people's professional and personal lives now-a-days [3]. This is how, social media sites have altered interpersonal interactions and created new communication patterns in recent years [4], [5]. People typically take in information that interests them, weed out information they do not, and form "echo chambers" with like-minded individuals around a common story [6]. This has given rise to the concept of oppositional thinking [7]. Oppositional thinking is concerned with people's tendency to perceive or approach conventional, mainstream and authoritative propositions, decisions or narratives in a negative, rebellious and critical manner [8]. However, oppositional thinking can be manifested in two ways in context of social media: conspiracy ideation and critical thinking [7].

Conspiracy theories appear to be closely linked to misinformation in reference of the internet and digital media [9]. These theories have many things in common, from dubious elements in the stories to the reasons that seem credible to prospective believers, particularly during an infodemic [10]. Understanding the spread of conspiracy narratives in online spaces, particularly in the context of a such a greater pool of knowledge, becomes crucial given that the propagation of such theories can have potentially detrimental effects on people and societies [11]. Conspiracy narratives are frequently the focal point around which echo chambers expand and deepen. Critical thinking is one form of civil disobedience that is characterized by the use of reason, analysis, evaluation, and proper questioning of information [12]. People who think critically and interact with any information use objective reasoning processes are receptive to having a change of mind once a new valid proof is introduced [13], [14].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

aza0266@auburn.edu (A. Albladi); sealscd@auburn.edu (C. D. Seals)

In the era of social media, critical thinking and conspiracy theories are fascinating social phenomena because they make us face the difficult problem of sometimes having multiple, contradictory interpretations of what actually happened [13]. Due to these reasons, there is a need to distinguish between critical and conspiracy narratives. Even though, a number of studies have been conducted previously to distinguish between these two categories of oppositional thinking through theoretical approaches [14]. However, the advent of machine learning techniques, deep learning algorithms and advanced NLP techniques have shown a new window for automatic analysis of oppositional thinking [14]. The hardest thing about using NLP models to detect conspiracy is figuring out how to tell the difference between conspiratorial and critical thinking in an automated manner [15]. This distinction is important because it could push people who were just asking queries into the conspiracy chambers if a message is labeled as conspiratorial when it is merely oppositional [16].

Our goal in this study is to analyze texts that exhibit critical or conspiratorial narratives and exhibit oppositional thinking [17]. We have addressed two pressing research questions for the NLP research: 1) differentiation of the two different types of oppositional narratives i.e., conspiracy narrative and critical thinking narrative, and 2) identification of the tokens or elements of a narrative that fuel the conflict in different groups during oppositional thinking in online messages. We utilized the Oppositional Thinking Analysis: Conspiracy vs Critical Narratives [18] of the PAN at CLEF 2024 [19] which included two subtasks mentioned above. We used the text in English corpora for the analysis of both tasks. Although, the dataset was also present in Spanish. The first challenge or subtask, an MCC score of 0.8050 was obtained indicating that there is a significant degree of agreement between the predicted and actual classifications produced by the customized RoBERTa model. For second challenge or subtask, we achieved a maximum accuracy of 95% in detecting the individual elements of each oppositional narrative in english text corpora to meet the above-mentioned challenges.

2. Previous Work

In this section, we reviewed the most recent studies on oppositional thinking analysis i.e., conspiracy theories and critical narratives [17]. examined the psycho-linguistic patterns and profiles of online users who spread conspiracy theory-debunking posts versus those who tend to spread conspiracy theory-supporting posts. Furthermore, they presented ConspiDetector, a convolutional neural network (CNN) based model that identified conspiracy propagators by fusing word embeddings with psycho-linguistic traits taken from user tweets. According to the F1-metric, the results demonstrate that ConspiDetector can detect conspiracy propagators with an 8.82% higher accuracy than the CNN baseline.

The authors' viewpoints on conspiracy narratives which can be expressed through description's individual constituents like the agent, action, or objective. However, these narratives can be implicitly expressed through acknowledgments to establish theories like chemtrails or the New World Order. This study developed a comprehensive system for categorizing conversations about conspiracies [20]. They trained a BERT-based model for online CT classification using human-labeled ground truth, and compared its performance with the GPT for online conspiratorial content detection. [21] suggested using sophisticated language models were fine-tuned to identify whether or not tweets are conspiratorial. The BERT model developed by Google served as the foundation for this model. By automating a manual process (identifying tweets that promote conspiracy theories), the classification method speeds up analysis. Using this method, they were able to identify tweets related to the COVID-19 origin conspiracy theory. Next, they used cybersecurity techniques on social media to examine the chambers, spreaders, and features of the various conspiracy narratives.

[23] presented an automated pipeline for the identification and characterization of real conspiracies covered by the media in addition to the creation of opposing narratives found in conspiracy theories that spread widely on digital media. This work is based on two independent, extensive repositories of news articles and blog posts detailing the popular conspiracy theories, which involved political conspiracies in New Jersey. They developed a graphical generative machine learning model, motivated by Greimas' qualitative narrative theory, in which nodes stand in for actors or actors, and repeated-edges and cyclic

Table 1
Summary of Relevant Work

Ref	Target	Input	Model	Pre-Processing	Outcome	Result
[22]	Conspiracy propagator detection	Tweets	CNN	Manual Annotation	Binary	F1: 0.74
[20]	Conspiracy Theory Classification	Reddit posts	RoBERTa	Data filtration, Annotation	Binary	AUC: 0.787
[21]	Covid-19 CT	Tweets	BERT	Tokenization	Binary	F1: 0.95
[23]	Narrative structures of CT	Blog posts and News Articles	Graphical generative ML model	Data cleaning, URLs removal, Pronouns Extracting	Identification of structural differences	N/A
[24]	Investigation of CT using NLG models	Data from Wikipedia and GPT-2	GPT-2	Temperature settings, Mechanical Turk	Analysis of model size and temperature effects	Larger model sizes and lower temperatures increased CTs

edges between nodes capture topic-specific relationships. The hidden narrative framework network's subgraphs were sampled from posts and news items.

[23] looked into LLM's ability to produce conspiracy theory text. The goal was to respond to the question: Is it possible to evaluate pretrained generative LLMs for the elicitation and memorization of conspiracy narratives without having exposure to the training set. They discussed this task in the context of memory, generalization, and hallucinations, emphasizing its challenges. Research has revealed that numerous conspiracy theories are deeply embedded within pretrained large language models. This was explored by using a new dataset comprising conspiracy narrative topics and machine-based conspiracy concepts. The tests show a connection between the tendency of models to produce conspiracy theory text and model parameters like size and temperature.

The previous research works done (given in Table 1) have focused only on conspiracy theory analysis and have developed various notable accomplishments in terms of their classification and understanding of conspiracy theories. However, studies have yet to be conducted to classify the text into conspiracy and critical narratives. This distinction is crucial because it may push people who were just raising questions toward conspiracy communities if a text is classified as conspiratorial when it is actually critical of mainstream beliefs.

Some of the recent studies targeted 'conspiracy spreaders' or 'conspiracy-related posts', which blurred the line between analyzing conspiracy theories and discrediting valid criticism. These studies mostly used binary, fine-grained classifiers to classify conspiracy and non-conspiracy theories without the proper identification of different constituent features in such narratives. Our study fills in these deficiencies by outlining a multifaceted approach on how to understand oppositional cognition, which distinguishes between conspiracy and critical frames using the customized RoBERTa. In this research, we propose a span-level annotation scheme that distinguishes agents, enablers, victims, campaigners, goals, and negative consequences in these narratives, in addition to locations and events, using the token-level BERT model. By using advanced NLP techniques, we identified and analyzed various degrees of oppositional thinking, i.e., conspiracy and critical narratives.

3. Methodology

The overall methodology consists of two subtasks. Subtask 1 is the binary classification of text into conspiracy and critical and subtask 2 is the detection of elements of the oppositional narratives. The

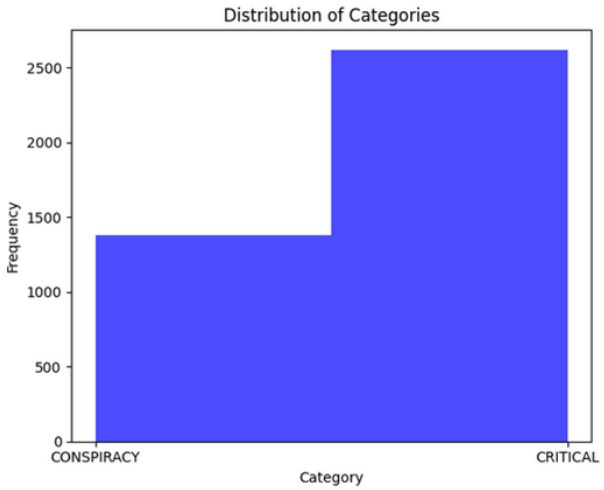

Figure 1: The distribution of two primary categories i.e., conspiracy and critical

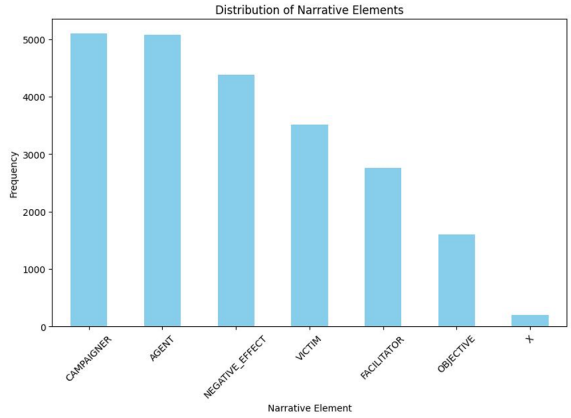

Figure 2: The distribution of narrative elements

step-by-step methodology is given below.

3.1. Dataset

The dataset used in this study is from Telegram posts related to Covid-19 pandemic discussions. This dataset has been collected and annotated for the PAN Conference 2024 specifically designed to distinguish between conspiracy; statements that portray the pandemic or decisions made about public health as the product of a sinister plot by a powerful, hidden group, and critical; critical messages that cast doubt on important public health decisions without endorsing conspiracy theories, in textual content.

3.1.1. For Subtask1

Each text entry is annotated with one of two classes i.e., Conspiracy and Critical as shown in Figure 1.

3.1.2. For Subtask 2

Each text entry is further annotated with spans that mark specific narrative elements i.e., agent, victim, facilitator, campaigner, objective and negative effect. Figure 2 gives the frequency of each element in the dataset.

3.2. Data Pre-Processing

3.2.1. Subtask 1: Binary Classification into Conspiracy and Critical

- **Tokenization** is the method of splitting up the text as input into a long string of characters for a computer, into smaller inputs or units known as tokens. RoBERTa tokenizer was used to facilitate the conversion of raw text into token IDs for model input.
- **Encoding** Encoding was done using the 'encode plus' function of the RoBERTa tokenizer to add further tokens and pads. The basic purpose was to truncate the text into the specified maximum length and generate attention masks.
- **Labelling** NLP utilizes data labelling to enable machines to process data efficiently, mimicking the understanding of human language. In order for NLP models to be trained, text input must be preprocessed with the right tags or categories specified. Conspiracy text was labelled as 1 whereas critical narratives were labelled as 0.
- **Splitting** In machine learning, data is usually split into three subsets: training, testing, and validation. This is necessary for model training, parameter tuning, and, in the end, performance evaluation. The dataset was split into 90% training and 10% validation sets.

3.2.2. Subtask 2: Detection of Elements of the Oppositional Narratives

- **Tokenization** BERT tokenizer was used to convert raw text into token IDs for model input. The function to align labels with BERT tokens was used to assign labels to tokens based on provided annotations. Text sequences were tokenized, and labels were aligned with the tokenized outputs. The label alignment process ensured that labels were correctly mapped to tokens while skipping special and padding tokens. A method was included to retrieve tokenized inputs and aligned labels as tensors, ensuring compatibility with PyTorch's data handling mechanisms.
- **Splitting** The data was split into 80% training and 20% validation.

3.3. Modelling

3.3.1. Subtask 1: Binary Classification into Conspiracy and Critical

For the classification of text into conspiracy and critical narratives, we used customized RoBERTa. RoBERTa is basically an enhanced version of BERT. The difference between BERT and RoBERTa is the larger training data. In addition, RoBERTa used dynamic masking modules and training on longer sequences. The primary improvements of RoBERTa over BERT are the predictions for the following sentences. We used a dropout rate of 30% to reduce overfitting.

3.3.2. Subtask 2: Detection of elements of the oppositional narratives

BERT model was used for subtask 2. Building on transformer networks, BERT pre-trains bidirectional representations by conditioning on both left and right contexts simultaneously in all layers. Predicting words in the input that are randomly masked and determining whether or not a sentence in the corpus is followed by the sentence are two ways that the representations are jointly optimized. BERT-based multi-task token classifier with a shared transformer backbone and independent classification heads was used to identify text spans that match the main ideas in opposing narratives.

3.4. Experimental Setup

The chosen equipment, including the NVIDIA RTX 4090 GPU and AMD EPYC 7R12 48-Core Processor, provides high computational power (1.8 TFLOPS and 24.0/192 CPU cores respectively) essential for intensive model calculations. The motherboard ROME2D32GM supports PCIe 4.0, enhancing data transfer speeds (22.8 GB/s), crucial for handling large datasets. With 516 GB of memory and a 4TB Predator SSD, the system ensures ample storage and quick data access (3830 MB/s), supporting efficient

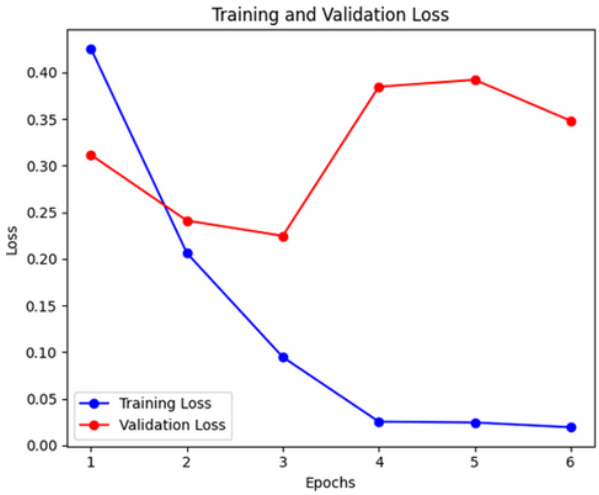

Figure 3: Performance graph indicating loss for subtask 1

Figure 4: Performance graph indicating accuracy for subtask 1

model training and analysis. The equipment's high-performance specifications are designed to optimize model development and execution.

4. Results and Discussion

4.1. Subtask 1: Binary Classification into Conspiracy and Critical

The primary aim of the subtask one was to classify the text data into conspiracy or critical narrative. We used customized RoBERTa model for this task. Figure 3 gives the performance graph indicating loss for subtask 1. The graph shows training and validation loss over six epochs. It depicts that training loss decreases progressively to almost zero while validation loss decreases at first but then rises, which is a sign of overfitting.

Figure 4 gives the performance graph indicating accuracy for subtask 1 accuracy for the task 1. The graph shows the training and validation accuracy over six epochs, with training accuracy in the blue line that increases to almost 100% within the first epoch, signifying good learning of the training data. The overall accuracy of our customized architecture was 91%.

Table 2 presents the performance metrics for a classification model for two categories: The categories that have been considered are "CRITICAL" and "CONSPIRACY". In the case of "CRITICAL", the model

Table 2
Classification report for subtask 1

Category	Precision	Recall	F1 Score
CRITICAL	0.94	0.93	0.93
CONSPIRACY	0.86	0.89	0.87

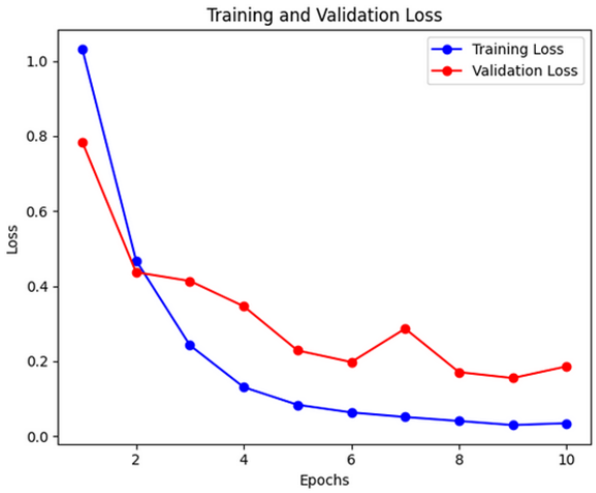

Figure 5: Performance graph indicating loss for subtask 2

has a high value of precision of 0. 94 and recall of 0. 93, which reflects a high level of F1 of 0. 93, which suggests that the model performed very well in terms of correctly predicting the "CRITICAL" class. For the "CONSPIRACY" model, the model has an accuracy of 0. 86 and recognition of 0. 89, which in turn results in a very high F1 score of 0. 87. These metrics demonstrate that the model is highly effective in differentiating between the two categories and has a slightly higher efficiency in the "CRITICAL" category, which indicates its ability to consistently and accurately identify instances of both types.

The discrete form of Pearson's correlation coefficient, known as the Matthews Correlation Coefficient (MCC), accepts values between 1 and 1. In the event of a complete correlation, the value is 1, 0 in the absence of correlation, and 1 in the presence of a negative correlation. By definition, only two categories are covered by the MCC. The MCC score is used for evaluation in binary classification even if the classes are imbalanced, i.e., of different sizes. The MCC score of our model was 0.8050. This value indicates that the model is maintaining a substantial concurrence between the predicted and the actual classifications.

4.2. Subtask 2: Detection of elements of the oppositional narratives

Figure 5 gives the performance graph indicating loss for subtask 2. The training loss (blue line) shows a smooth descending trend until it is almost zero, which implies that the model has learned the training data very effectively. The validation loss (red line) at first goes down steeply i.e., the model is learning well and can generalize well to new inputs, but then it oscillates and converges to a higher value than the training loss.

The accuracy of the subtask 2 is given in Figure 6, throughout ten epochs. The model showed a good level of generalization to the data. The learning curves of training and validation accuracy indicate that the model does not overfit, and the performance on both datasets remains good, with a slight fluctuation in the validation accuracy in later epochs. In conclusion, the proposed model is able to learn from the data and generalize the learned knowledge. The overall accuracy of the model was 95%.

Table 3 gives the classification report for the subtask2. The F1 score for the narrative elements is given.

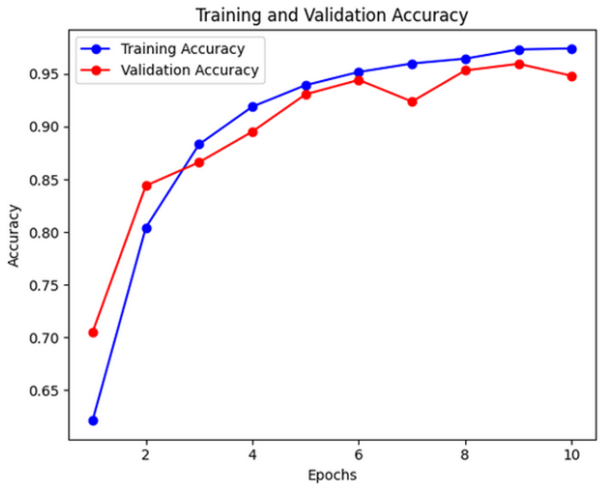

Figure 6: Performance graph indicating accuracy for subtask 2

Table 3
Classification report for subtask 2. The F1 score for the narrative elements is given.

Category	Precision	Recall	F1 Score
OBJECTIVE	0.13	0.66	0.22
AGENT	0.48	0.70	0.57
FACILITATOR	0.16	0.68	0.26
CAMPAIGNER	0.42	0.80	0.55
NEGATIVE EFFECT	0.22	0.58	0.32
VICTIM	0.46	0.74	0.57
0	0.99	0.96	0.97

5. Conclusion

The work reveals the use of the recent NLP approaches to identify conspiracy theories from the critical narratives in social media posts. The researchers using a specifically fine-tuned RoBERTa model got an MCC score of 0. 80 for identifying conspiracy narratives, thus showing how well the model is able to discern conspiratorial narratives from critical ones. In addition, the BERT model provided a high overall accuracy of 95% and was capable of detecting narrative elements like agents, victims and objectives. Thus, these findings imply that NLP models can be used to enhance the automation of the detection and differentiation of conspiracy theories from critical thinking in order to control information flow and promote informed debate.

References

[1] K. Y. A. McKenna, J. A. Bargh, Causes and consequences of social interaction on the internet: A conceptual framework, Media Psychol 1 (1999) 249–269. doi:10.1207/S1532785XMEP0103_4.

[2] E. Kross, P. Verduyn, G. Sheppes, C. K. Costello, J. Jonides, O. Ybarra, Social media and well-being: Pitfalls, progress, and next steps, Trends Cogn Sci 25 (2021) 55–66. doi:10.1016/J.TICS.2020.10.005.

[3] H. K. Azzaakiyyah, The impact of social media use on social interaction in contemporary society, Technology and Society Perspectives (TACIT) 1 (2023) 1–9. doi:10.61100/TACIT.V1I1.33.

[4] S. S. S. Ahmad, A. N. B. S. Osman, H. Basiron, The impact of social media on human interaction in an organisation based on real-time social media data, International Journal of Data Science 4 (2019) 260. doi:10.1504/IJDS.2019.102793.

[5] D. Felmlee, R. Faris, Interaction in social networks, in: Handbooks of Sociology and Social Research, 2013, pp. 439–464. doi:10.1007/978-94-007-6772-0_15/FIGURES/00153.
[6] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, Echo chambers on social media: A comparative analysis, arXiv preprint arXiv:2004.09603 (2020). URL: https://arxiv.org/abs/2004.09603v1, accessed: May 29, 2024.
[7] D. Choi, S. Chun, H. Oh, J. Han, T. T. Kwon, Rumor propagation is amplified by echo chambers in social media, Scientific Reports 10 (2020) 1–10. doi:10.1038/s41598-019-57272-3.
[8] A. D. Schrift, Nietzsche and the critique of oppositional thinking, Hist Eur Ideas 11 (1989) 783–790. doi:10.1016/0191-6599(89)90266-0/ASSET//CMS/ASSET/4B6CAA08-357B-4283-ABBC-F44DA71E2B0C/0191-6599(89)90266-0.FP.PNG.
[9] K. M. Douglas, R. M. Sutton, A. Cichocka, The psychology of conspiracy theories, Curr Dir Psychol Sci 26 (2017) 538–542. URL: https://doi.org/10.1177/0963721417718261. doi:10.1177/0963721417718261.
[10] K. M. Douglas, et al., Understanding conspiracy theories, Polit Psychol 40 (2019) 3–35. doi:10.1111/POPS.12568.
[11] K. Y. L. Ku, Q. Kong, Y. Song, L. Deng, Y. Kang, A. Hu, What predicts adolescents' critical thinking about real-life news? the roles of social media news consumption and news media literacy, Think Skills Creat 33 (2019) 100570. doi:10.1016/J.TSC.2019.05.004.
[12] A. D. Knochel, Assembling visuality: Social media, everyday imaging, and critical thinking in digital visual culture, Visual Arts Research 39 (2013) 13–27. doi:10.5406/VISUARTSRESE.39.2.0013.
[13] What is actually true? approaches to teaching conspiracy theories and alternative narratives in history lessons, 2024. URL: https://journals.uio.no/adnorden/article/view/8377/7355, accessed: May 30, 2024.
[14] S. Steinert, L. Marin, S. Roeser, Feeling and thinking on social media: emotions, affective scaffolding, and critical thinking, Inquiry (2022). doi:10.1080/0020174X.2022.2126148.
[15] J. Du, et al., Using machine learning-based approaches for the detection and classification of human papillomavirus vaccine misinformation: Infodemiology study of reddit discussions, J Med Internet Res 23 (2021) e26478. doi:10.2196/26478.
[16] H. Guo, A. Ash, D. Chung, G. Friedland, Detecting conspiracy theories from tweets: Textual and structural approaches (2020). URL: http://arxiv, accessed: May 30, 2024.
[17] B. A. Galende, G. Hernandez-Penaloza, S. Uribe, F. A. Garcia, Conspiracy or not? a deep learning approach to spot it on twitter, IEEE Access 10 (2022) 38370–38378. doi:10.1109/ACCESS.2022.3165226.
[18] D. Korenčić, B. Chulvi, X. B. Casals, M. Taulé, P. Rosso, F. Rangel, Overview of the oppositional thinking analysis pan task at clef 2024, in: G. Faggioli, N. Ferro, P. Galuvakova, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.
[19] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of pan 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification - condensed lab overview, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association CLEF-2024, 2024.
[20] A. Diab, R. Nefriana, Y.-R. Lin, Classifying conspiratorial narratives at scale: False alarms and erroneous connections (2024). URL: https://arxiv.org/abs/2404.00141v1, accessed: May 30, 2024.
[21] J. D. Moffitt, C. King, K. M. Carley, Hunting conspiracy theories during the covid-19 pandemic, Social Media and Society 7 (2021). doi:10.1177/20563051211043212/ASSET/IMAGES/LARGE/10.1177_20563051211043212-FIG8.JPEG.
[22] A. Giachanou, B. Ghanem, P. Rosso, Detection of conspiracy propagators using psycho-linguistic characteristics, J Inf Sci 49 (2023) 3–17. doi:10.1177/0165551520985486/ASSET/IMAGES/LARGE/10.1177_0165551520985486-FIG6.JPEG.

[23] T. R. Tangherlini, S. Shahsavari, B. Shahbazi, E. Ebrahimzadeh, V. Roychowdhury, An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web, PLoS One 15 (2020) e0233879. doi:10.1371/JOURNAL.PONE.0233879.

[24] S. Levy, M. Saxon, W. Y. Wang, Investigating memorization of conspiracy theories in text generation, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 4718–4729. doi:10.18653/v1/2021.findings-acl.416.