



Network-aware Fake News mitigation on social media

Scientific Report #2

Mitrica Octavian

S.L. Dr. Ing. Ciprian-Octavian Truică
Conf. Dr. Ing. Elena-Simona Apostol

Faculty of Automation and
Computers Politehnica
University of Bucharest
Romania

Introduction

Task: We want to detect fake news in social media posts containing both text and image.

Classification: **0 - Fake / 1 - Real**

Introduction



Maryland drive
gets probation for
Delaware crash
that killed 5 NJ
family members

True



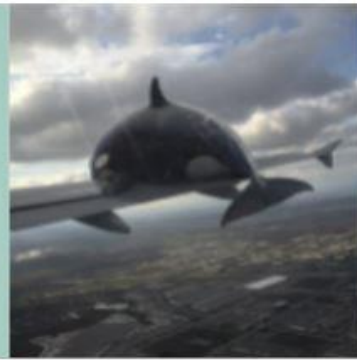
New 'Natural
Feeding' trend has
parents puking on
babies

Satire/Paroday



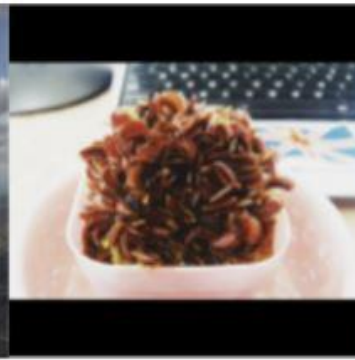
Neuroscience
Says Doing This 1
Thing Makes You
Just as Happy as
Eating 2,000
Chocolate Bars

**Misleading
Content**



My plane hit an
orca right after
takeoff

**Manipulated
Content**



Bowl of mussels

**False
Connection**



I just thought that
was sitting in the
deli

**Imposter
Content**

Dataset sample [1]

Introduction

The main *objective* of this scientific report is to test the effectiveness of LLMs and prompt engineering tactics in the context of social media fake news detection, through experiments. We will also test the importance of adding visual information to the decision process.

Related work

Hu et al. [2] present their takes on the typical strategies like Zero-Shot and Few-Shot and the CoT [3] approach. They also introduce their hybrid model that passes the reasoning further to a trained BERT model, showing significant results.

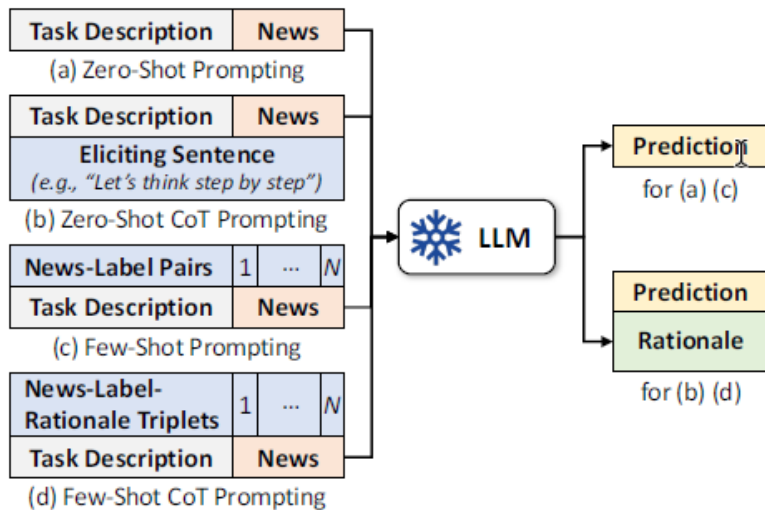


Figure 1: Prompting approaches for LLMs [2]

Model	Usage	Chinese	English
GPT-3.5-turbo	Zero-Shot CoT	0.677	0.666
	from Perspective TD	0.667	0.611
	from Perspective CS	0.678	0.698
BERT	Fine-tuning	0.753	0.765
Ensemble	Majority Voting	0.735	0.724
	Oracle Voting	0.908	0.878

Figure 2: Results for article [2]

Related work

Zhang et al. [4] explored a step-by-step type of prompting in order to separate a claim into several subclaims and then verify them via progressive question answering. One of the fundamental problems of LLMs is hallucination, and the authors tackle this issue by providing real-time contextual information.

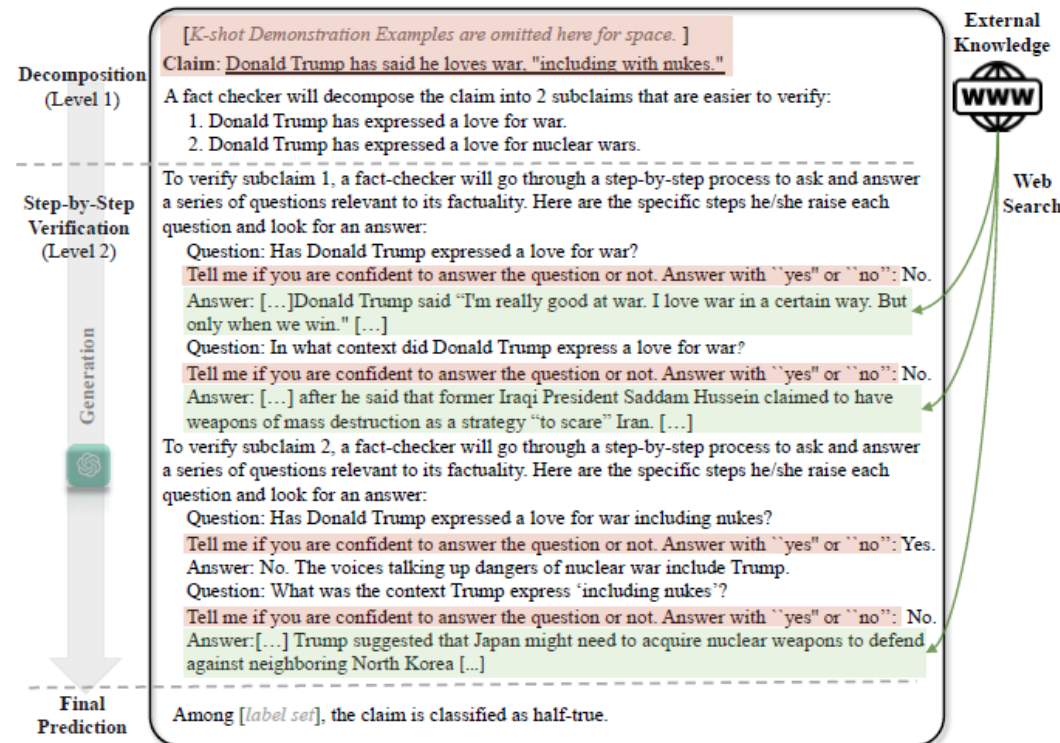


Figure 3: Step-by Step pipeline [4]

Related work

Li et al. [5] published another collection of experiments similar to the ones documented in [3]. Their contribution lies in a new way of gathering contextual information by using specific tools and leveraging the step-by-step method to extract the final answer.

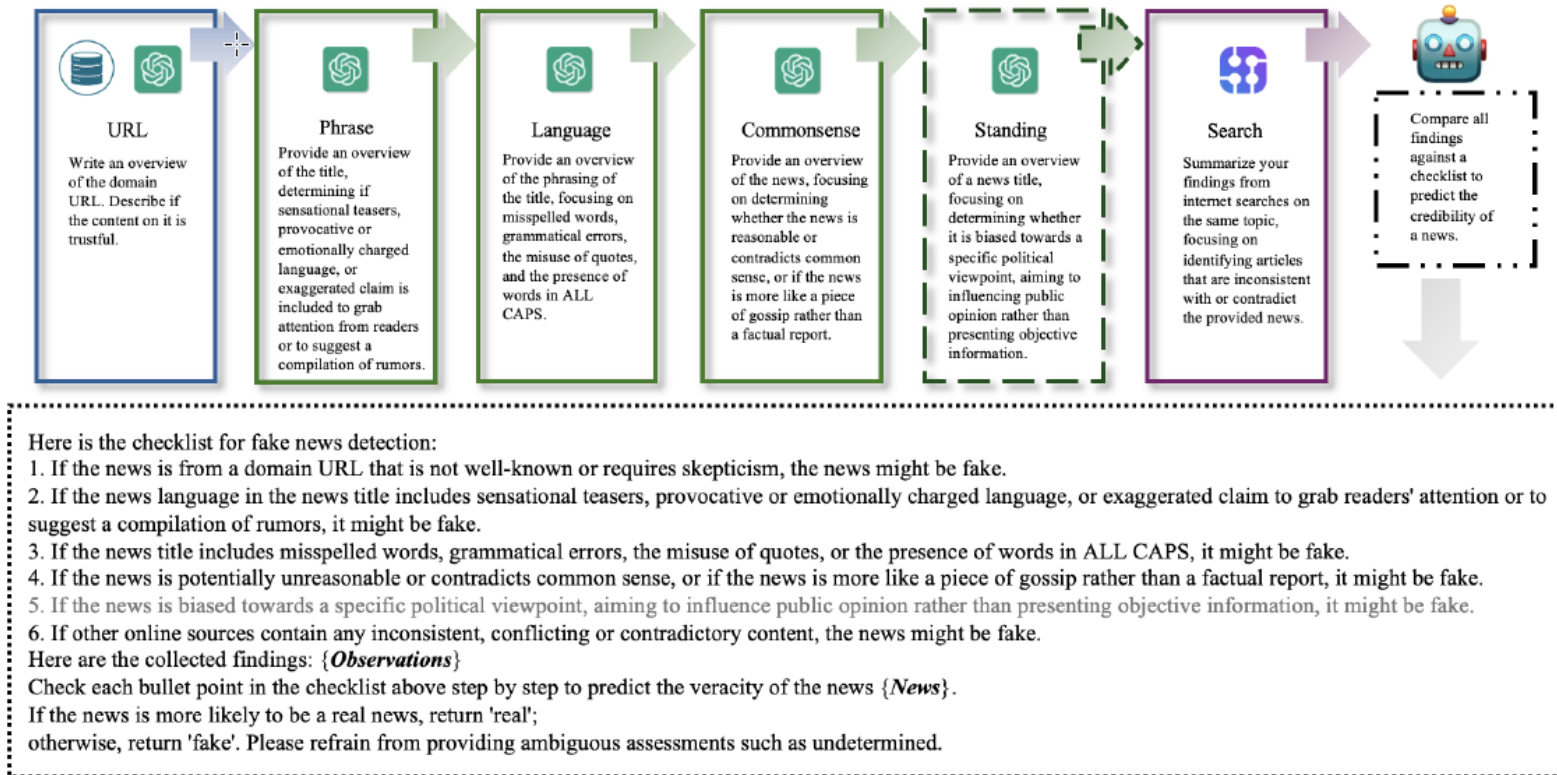


Figure 4: Agent method pipeline [5]

Related work

This article [6], presents a unique way of leveraging the image context and up-to-date web searching for an accurate prediction of fake news in the domain of politics.





Claim	Image	Generated Image Context	Image QA
The man next to Mike Pompeo in a November 2020 photo is the guy the Trump administration helped get out of jail in 2018 and who is now the 'president' of Afghanistan.		The image shows Mike Pompeo standing next to a man wearing traditional Afghan attire and a face mask. The setting appears to be a formal meeting room with a painting on the wall, two gold-colored chairs, and a table with water bottles. Both individuals are standing apart, likely adhering to social distancing guidelines prevalent during the COVID-19 pandemic. The background and attire suggest the meeting is official in nature and likely related to diplomatic or governmental affairs. Claim stated in 2021.	Q. Who is the man standing next to Mike Pompeo in the November 2020 photo? A. Mullah Abdul Ghani Baradar
Says Joe Biden took a knee during the national anthem.		The image shows Joe Biden taking a knee, but it does not provide evidence that this act occurred during the national anthem. Without additional context, the claim cannot be confirmed solely based on this image. The time frame of the claim is 'Claim stated in 2020'.	Is there a specific date and location associated with the image of Joe Biden taking a knee? A. FILE - In this Monday, Oct. 5, 2020 file photo, Democratic presidential candidate former Vice President Joe Biden and his wife Jill Biden pose for a photo with dancers as they visit Little Haiti Cultural Complex in Miami.
The Trump administration worked to free 5,000 Taliban prisoners.		The image shows individuals, presumed to be Taliban prisoners, inside a bus with a guard standing nearby, which potentially correlates to the release of Taliban prisoners. The context suggests this may represent a prisoner release process.	Q. Were the individuals shown in the provided image actually Taliban prisoners being released as part of the agreement? A. 'Taliban prisoners are released from Pul-e-Charkhi jail in Kabul, Afghanistan, Thursday, Aug. 13, 2020
These were not chemical irritants' used to clear a crowd. Pepper balls are 'not a chemical irritant.		The image shows law enforcement in protective gear amidst a haze that is consistent with the use of some form of crowd control substance, such as a chemical irritant. Visible smoke and the dispersing crowd strongly suggest the use of a substance to clear the area, counter to the claim that no chemical irritants were used. The presence of pepper balls would depend on identifying specific items or equipment in the scene that are known to dispense pepper balls.	Can we identify the specific equipment or methods used by law enforcement in the image to determine if pepper balls or another substance was deployed? A. ... The caption snippets suggest that teargas was used to clear Lafayette Park for a photo opportunity...

Figure 5: Example with claims, images and QA [6]

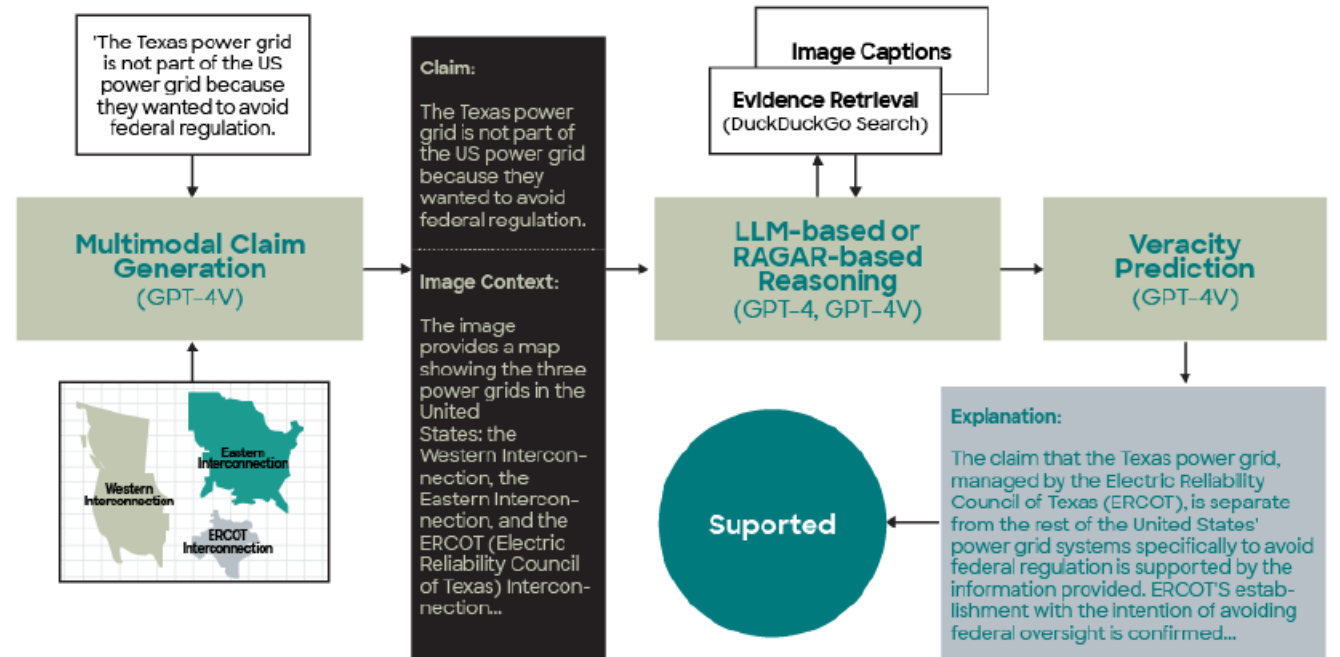


Figure 6: Detailed overview of the pipeline [6]

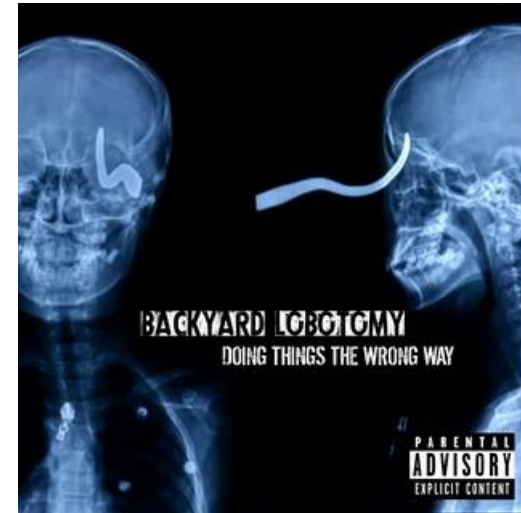
Datasets

The dataset used for testing was Fakeddit [1], along with Twitter15 and Twitter16 [13] for baselines. Fakeddit is a multimodal dataset containing over 1 million samples and 2-way, 3-way or 6-way classification labels.

More examples:



At my local community center they make disabled people get out the car to move a cone before they can park



Backyard Lobotomy - Doing Things the Wrong Way

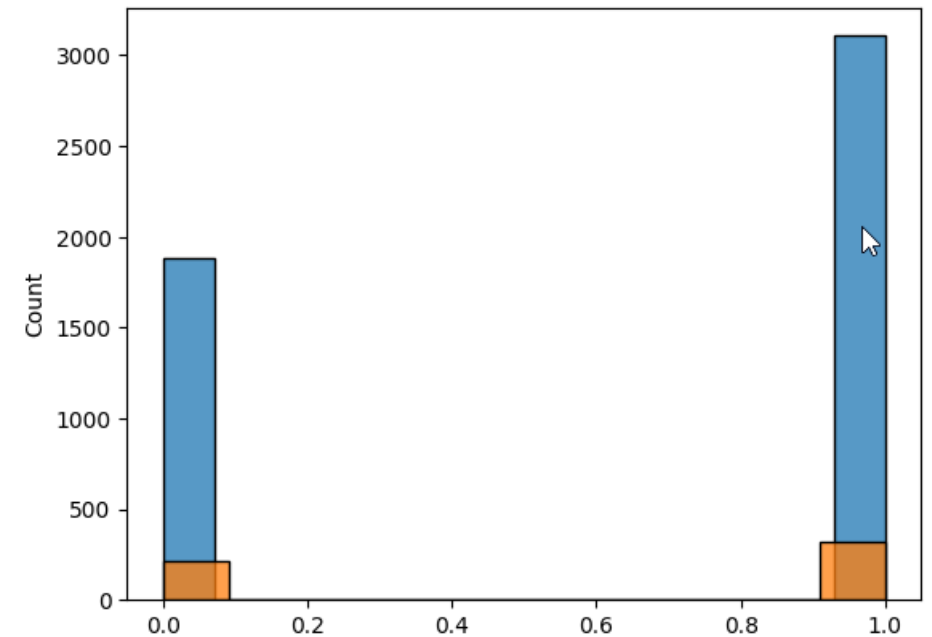
Datasets

Dataset Statistics	
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 - 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

Fakeddit statistics [1]

Dataset	Articles	Fake news	Real news	Unverified
Fakeddit	5633	2214	3419	0
Twitter15	1340	332	670	335
Twitter16	740	187	370	181

Used datasets statistics



Label distribution (2-way labels)

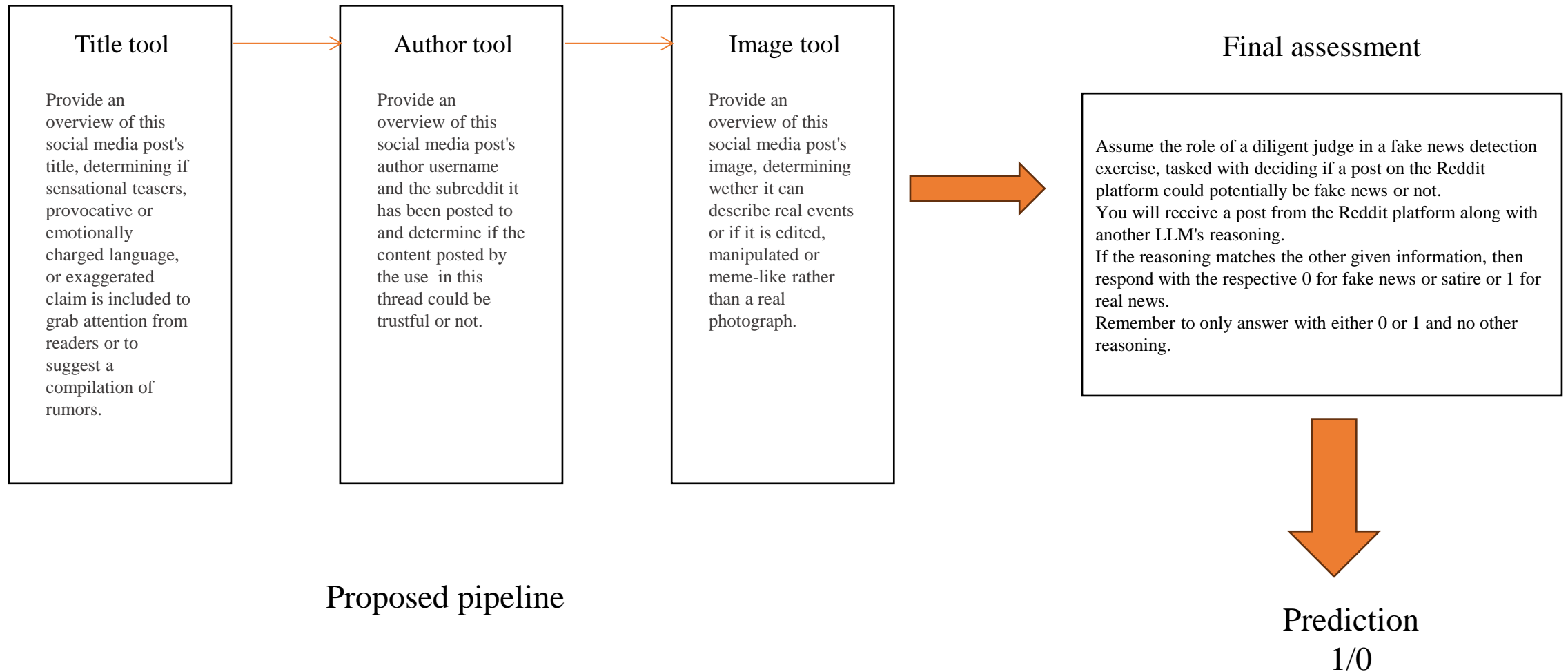
Proposed solution

We propose a Few-Shot Agent Prompting method that leverages tools similar to the ones proposed by Li et al. [5] and the power of example that comes with few-shot prompting. The final assessment will be simple role-playing prompt, not using the step-by-step checklist presented in [5], due to its weak performance on our specific task.

Tools:

1. **Title tool:** a tool designed to determine if a social media post's title contains sensational teasers, provocative or emotionally charged language, or exaggerated claims to grab attention from readers or to suggest a compilation of rumors.
2. **Author tool:** a tool that provides insight into the author's username and the type of content posted on the respective subreddit that contains the post (can it be trustful?).
3. **Image tool:** a tool that analyzes a given image, trying to determine if it presents signs of alteration or meme-like characteristics or if it describes real events.

Proposed solution



Proposed solution

Examples

Title: EU is convinced it has found new way for young people to fall in love with neighboring countries free train rides...kicked off DiscoverEU project Thursday to send up to 30,000 18-year-olds chugging across continent this year, giving them free rail tickets to broaden their horizons.

Author: madazzahatAter

Domain: abcnews.go.com

Number of Comments: 7.0

Score: 75

Subreddit: upliftingnews

Upvote ratio: 0.89

1

The post is from a reputable domain (abcnews.go.com) and is posted in the subreddit upliftingnews, which typically shares positive and uplifting stories. The title also seems plausible and in line with positive news stories. The high upvote ratio of 0.89 and a score of 75 indicate that the post is well-received by the community.

Title: Robert Mueller resigns from the Department of Justice (2019)

Author: RayBrower

Domain: i.imgur.com

Number of Comments: 0.0

Score: 30

Subreddit: fakehistoryporn

Upvote ratio: 0.9

0

The post is from the subreddit 'fakehistoryporn' which suggests that the content is likely satire or parody. Additionally, the title states that Robert Mueller resigned from the Department of Justice in 2019, which is not factually accurate. The high upvote ratio could indicate that users found the post entertaining rather than informative.

Proposed solution

Tools output

(Label: Fake)

Title: We've got a new radiator
Author: Justrolledinshop_SS
Domain: imgur.com
Number of Comments: 20.0
Score: 100
Subreddit: subredditsimulator
Upvote ratio: 0.97

Title tool: The title \"We've got a new radiator\" appears to be straightforward and lacks sensational teasers, provocative language, or exaggerated claims. It does not suggest a compilation of rumors; instead, it seems to convey a simple update or announcement about a new radiator. The title is likely intended to inform rather than to provoke or sensationalize, making it more neutral in tone. Overall, it does not employ tactics commonly used to grab attention in a sensational manner.

Author tool: The author of the post, \"Justrolledinshop_SS,\" appears to be associated with a username that suggests a connection to a shop or business, possibly related to automotive or repair services, as \"Just Rolled In\" is a common phrase used in that context. However, without specific content from the post, it's difficult to assess the credibility of the author. The subreddit \"subredditsimulator\" is a unique community that simulates the experience of browsing various subreddits. Posts in this subreddit may not always reflect genuine content or discussions, as the nature of the subreddit is to create a simulated environment rather than a platform for serious discourse. Given these factors, the trustworthiness of the content posted by the user in this thread would likely be low. The context of the subreddit suggests that the posts may be more for entertainment or simulation rather than factual or reliable information. To make a more accurate assessment, it would be necessary to review the specific content of the post.

Image tool: The image depicts a car engine compartment with a visible battery and some tools or items nearby. The battery appears to be a Duralast brand, which is a common automotive battery. The overall condition of the engine bay looks somewhat dirty or neglected, suggesting it may be an older vehicle or one that hasn't been maintained recently. Based on the description, this image likely represents a real event, possibly related to car maintenance or repair. There are no obvious signs of manipulation or editing that would suggest it is a meme or altered photograph. However, without further context, it's difficult to determine the specific circumstances surrounding the image.

Final result: 0

Results

Model	Accuracy	Precision	Recall	F1
DistilBERT BiGruCNN	76.73	76.66	75.11	
Fine-tuned ViT	74.32	73.77	73.15	
Multimodal	77.70	77.27	76.72	
Few-Shot Agent	85.68			

Comparison with supervised learning models

Results

Model	Accuracy	Precision	Recall	F1-Score
Zero-Shot Prompting (GPT3.5)	76.52	0	0	0
Zero-Shot CoT Prompting (GPT3.5)	78.05	0	0	0
Few-Shot Prompting (GPT3.5)	67.17	0	0	0
Few-Shot CoT Prompting 4ex (GPT3.5)	78.96	0	0	0
Few-Shot CoT Prompting 8ex (GPT3.5)	83.74	0	0	0
Zero-Shot Tools RAG (GPT4mini)	70.71	0	0	0
Few-Shot Tools Prompting 8ex (GPT4mini)	85.68	0	0	0

Comparison between our LLM experiments

Results

Model	Accuracy	Precision	Recall	F1	
MCWDST [4]	76.90	77.27	76.89	76.82	
Fakeddit [8]	89.09	-	-	-	
DEAP-FAKED [3]	89.55	-	-	-	
(BERT+Dense)+Xception [9]	91.87	93.39	93.29	93.25	
Multimodal transformers [10]	92.51	93.83	93.74	93.79	(Best supervised)
LLM+SLM [3]	78.60	78.40	81.40	80.40	
FactAgent with Expert Workflow [5]	88.00	88.00	89.00	88.00	(Best LLM)
ToRAG + CoTVP + CoVe [6]	84.00	85.00	86.00	85.50	
Ours	85.68	-	-	-	

Comparison with state-of-the-art

Conclusion

In conclusion, we confirmed that large language models have a substantial capability to detect fake news. The multimodal aspect of the task seemed to help the model achieve better reasoning and comparing the context of the image with the news title can be a future area to improve on.

The main upside of working with LLMs is that they don't need training and can run at a relatively low cost. As a solution for a social media company that can either have its own LLM or a subscription to one already, it proves to be essential to take into consideration when mentioning the automatic fake news detection task.

Of course, all of this does not come without its flaws. The somewhat randomness of the model's answers and the web searches are the main points of concern. Also, we need to take into consideration the hallucination problem.

Further experiments can improve this kind of detection and we will continue with it.

References

1. Nakamura, K., Levy, S., & Wang, W.Y. (2019). r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. ArXiv, abs/1911.03854.
2. B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi, “Bad actor, good advisor: Exploring the role of large language models in fake news detection,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, p. 22105–22113, Mar. 2024.
3. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” 2023.
4. X. Zhang and W. Gao, “Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method,” 2023.
5. X. Li, Y. Zhang, and E. C. Malthouse, “Large language model agent for fake news detection,” 2024.
6. M. A. Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Miletic, “Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models,” 2024.
7. M. Mayank, S. Sharma, and R. Sharma, “DEAP-FAKED: knowledge graph based approach for fake news detection,” CoRR, vol. abs/2107.10648, 2021.
8. C. -O. Truică, E. -S. Apostol, R. -C. Nicolescu and P. Karras, "MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media," in IEEE Access, vol. 11, pp. 125861-125873, 2023, doi: 10.1109/ACCESS.2023.3331220.

References

9. V. Slovikovskaya, “Transfer learning from transformers to fake news challenge stance detection (fnc-1) task,” 2019.
10. Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, “Novel visual and statistical image features for microblogs news verification,” *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 598–608, 2017.
11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.
12. P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, “Multi-modal fake news detection via bridging the gap between modals,” *Entropy*, vol. 25, no. 4, 2023.
13. X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, “Real-time rumor debunking on twitter,” in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, (New York, NY, USA), p. 1867–1870, Association for Computing Machinery, 2015.