Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources (Supplementary Material)

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz
CISPA Helmholtz Center for Information Security
{sahar.abdelnabi,rakibul.hasan,fritz}@cispa.de

In the supplementary material, we first discuss more implementation details in Section 1 and present additional experiments in Section 2. Then we present some examples that were selected as 'hard to verify' in the user study in Section 3. We present other qualitative examples in Section 4. Finally, we discuss societal aspects and potential risks in Section 5.

1. Implementation Details

We elaborate on some implementation details of our framework.

- **Sentence representation.** We preprocessed the crawled captions to remove some artefacts (e.g., HTML tags). When using BERT+LSTM, we used the pre-trained 'bert-base-uncased' model, whose dimension is 768. We set a maximum length of 150 tokens for the captions. Items (i.e., query captions, evidence captions, and entities) are padded to the maximum sequence length in this item's batch. When using the sentence transformer model, we used the 'paraphrase-mpnet-base-v2' model[1]. For both, we used the Hugging Face library[2]. We used the PyTorch framework[3] for all our experiments.

- **Memory.** The items in each memory (images, entities, and captions) are padded to the maximum number of evidence items in this memory's batch.

- **CLIP.** We used the pre-trained ViT-B/32 CLIP model[4], where the text length is truncated at 77 tokens.

- **Training details.** When fine-tuning CLIP, we follow the implementation details in [5], we used a learning rate of 5e-5 for the linear classifier and 5e-7 for other layers of the CLIP model itself, in addition to using the Adam optimizer [4]. We used a batch size of 64 and trained the model for 100 epochs. For training *CCN*, we used a batch size of 32, the Adam optimizer, and a cyclical learning rate [8] with a maximum value of 6e-5. We trained the model for 30 epochs. We used a dropout [9] value of 0.05 to the input representations, 0.25 to domain embeddings, and 0.25 to the memory representations. Experiments were done on one NVIDIA A100 GPU. With precomputing the representations, the training takes roughly 5 hours. When training using BERT without precomputing, training takes roughly 30 hours.

2. Additional Experiments

Evidence-only classification. We examine whether claims (and consequently, the evidence) are having different characteristics (and thus, unwanted biases or naive give-aways) between pristine and falsified classes. The NewsCLIPpings dataset avoided linguistic biases in creating falsified examples by using real news **captions** mismatched with real news **images**, instead of introducing manipulations in the captions. Also, to avoid text bias, each **caption** (and consequently, its **visual evidence** in our dataset) appears twice (within the same split), once as pristine and once as falsified. Therefore, we hypothesize that the evidence websites for both classes are similar. To confirm, we ran an *evidence-only* model, which achieved 53.4% (*basically chance level*), showing that *reasoning against the query* is the distinguishing factor.

Additional ablation studies. We include further experiments related to the fusion of the different components in

Conc.	Avg-pool	Max-pool	Multiply
83.9	82.46	82.48	77.1

Table 1. Accuracy (%) vs. aggregation strategies.

[1] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[2] https://huggingface.co/

[3] https://pytorch.org/

[4] https://github.com/openai/CLIP

our model (visual reasoning, textual reasoning, and CLIP). We tried a late fusion by having a separate classifier on top of each branch and aggregating the decision, however, this performed worse than the current intermediate fusion we employ. We also tried other strategies (Table 1) to combine visual and textual memories before concatenating with CLIP, where we found that concatenation had the highest performance.

Finally, we found that changing the dimension of the penultimate layer had a relatively small effect; e.g., increasing the dimension to 2048 increased the accuracy by 0.3 percentage points.

3. User Study: 'Hard to Verify' Examples

In Figure 1, we show some examples that were selected as 'hard to verify' in the user study. This is possibly due to: 1) the **captions** could contain specific context information (e.g., locations such as *'Denver'* or *'Massachusetts'*) that is hard to verify with the **image** alone, 2) the lack of **textual** evidence returned by the search ⇄●; the **images** were not found by the inverse image search, so there are no **captions/titles** found. Moreover, the **entities** are generic descriptions of the **image**, or not at all related (the first example). The performance of the model on these examples is possibly dependent on how similar the **visual** evidence is to the query **image** ⇄●.

Another possible reason is having falsified pairs that are highly similar in context to the original ones (and, therefore, to the evidence as well). For instance, the last example shows a 'hard to verify' falsified example (that was also misclassified by our model); the **image** shows the same people mentioned in the **caption**, and thus, they also appeared in the **visual** evidence. Additionally, the **caption** mentions the band name *'One Direction'* that is also mentioned in the **textual** evidence, without strong contradictions. Meanwhile, the actual **image** of this **caption** showed the band performing on a stage, however, this was not clearly emphasized by the **caption**; that is possibly why the **visual** evidence is generic.

4. Qualitative Examples

In Figure 2, we show more qualitative examples. *CCN* predicted many examples correctly despite not having a one-to-one matching with the evidence in the case of pristine examples and having close similarity to the evidence in the case of falsified examples.

For instance, in the first three examples (pristine), we observed that the model highly attended to supporting evidence such as persons' and countries' names, topics, and events. Additionally, in the third example, we observed that the model prioritized the **image** that is from the same scene and the evidence **caption** that contains a subset from the query **caption** (*'soon to be a Trump International Hotel'*).

The fourth and fifth examples (falsified) suggest that the model does not simply rely on having any similarity or overlap between the query and evidence in order to identify pristine examples. Despite having the same persons in the evidence, they were correctly predicted as falsified, possibly as they have contradicting location information and different scene details (e.g., lighting, stage setup, or colours), indicating a different context or event. The last falsified example also indicates that both **textual** and **visual** evidence is helpful, as the evidence **images** are clearly different from the falsified one (showing a different building and place).

5. Limitations and Societal Aspects

Nowadays, with the spread and reliance on social media to digest and get updated with news, misinformation (e.g., on Twitter) can reach hundreds of millions of users [12]. This crucially motivates the need to fact-check and verify the credibility of online content, especially during critical times such as a pandemic or political instabilities. On the other hand, manual fact-checking is usually time-consuming, needing from less than one hour to many days to verify a claim [10]. Therefore, automating fact-checking can be extremely beneficial to alleviate the burden upon fact-checkers and journalists.

However, completely or overly relying on automated tools might give an unwanted sense of security and could have many dangerous consequences. These include the dangers of flagging many true examples as falsified due to the real-life class imbalance, and missing out challenging falsified examples that require more fine-grained and complex reasoning. In addition, a currently active and much-needed research direction in the textual domain shows that fact-verification models might be partially relying on dataset biases without in-depth understanding and reasoning [7]. They might also be brittle to complex claims that require multi-hop reasoning [3]. Additionally, as facts are continuously evolving, we face the danger of relying on old retrieved evidence [6] or even possibly outdated world knowledge that is implicitly stored in pre-trained language models during training [7].

In addition to their inherent limitations in reasoning and interpretation, several works have shown that textual verifications models are also vulnerable to adversarial attacks [11], such as inserting trigger words [1], introducing lexical variations [3], or paraphrasing [11]. As we have a multi-modal task, our model might also be vulnerable to image-based adversarial attacks [2]. Another potential misuse scenario is using the fact-checking model as an adversarial filter in order to curate hard examples that might be misclassified by fact-checking models in general.

As a conclusion, we believe that automating fact-checking is strongly beneficial and that there have been many encouraging advancements to improve and harden it in the textual domain and the multi-modal domain, as we propose. However, due to their limitations and vulnerabilities to active attacks and manipulation, they should be used to assist humans and speed up the process, while still keeping them in the loop to avoid such dangers and consequences. In this regard, in our framework, we show that the model can filter and select the most important evidence, which would enable quicker inspection of the evidence items.

References

[1] Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. Generating label cohesive and well-formed adversarial claims. In *EMNLP*, 2020. 2

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2014. 2

[3] Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. Deseption: Dual sequence prediction and adversarial examples for improved fact-checking. In *ACL*, 2020. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[5] Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *EMNLP*, 2021. 1

[6] Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin c! robust fact verification with contrastive evidence. In *NAACL-HLT*, 2021. 2

[7] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In *EMNLP-IJCNLP*, 2019. 2

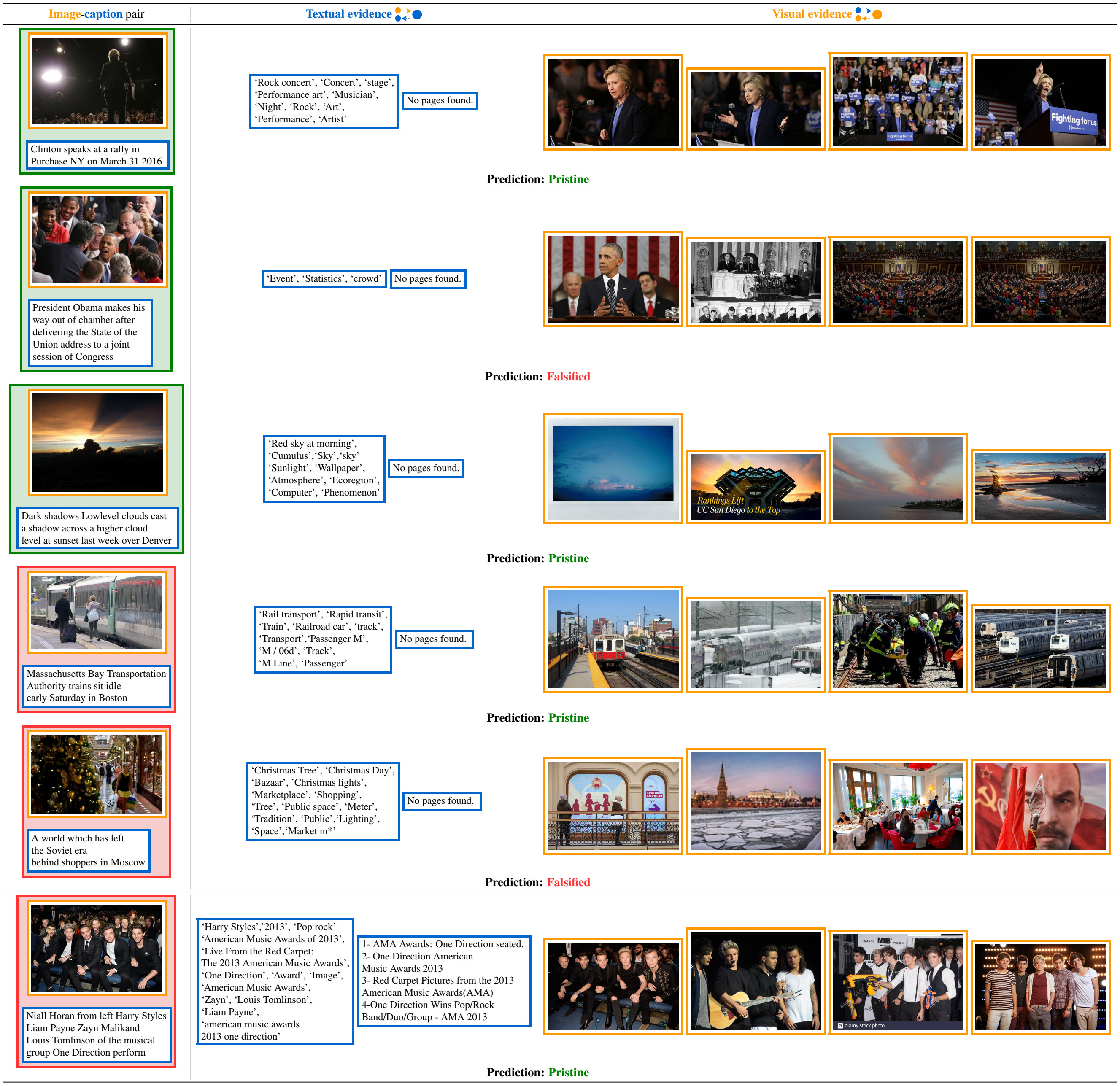
[8] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. 1

[9] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014. 1

[10] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *International Conference on Computational Linguistics*, 2018. 2

[11] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Evaluating adversarial attacks against multiple fact verification systems. In *EMNLP-IJCNLP*, 2019. 2

[12] Nguyen Vo and Kyumin Lee. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *EMNLP*, 2020. 2



Clinton speaks at a rally in Purchase NY on March 31 2016

'Rock concert', 'Concert', 'stage', 'Performance art', 'Musician', 'Night', 'Rock', 'Art', 'Performance', 'Artist'

No pages found.

Prediction: Pristine

President Obama makes his way out of chamber after delivering the State of the Union address to a joint session of Congress

'Event', 'Statistics', 'crowd'

No pages found.

Prediction: Falsified

Dark shadows Lowlevel clouds cast a shadow across a higher cloud level at sunset last week over Denver

'Red sky at morning', 'Cumulus','Sky','sky' 'Sunlight', 'Wallpaper', 'Atmosphere', 'Ecoregion', 'Computer', 'Phenomenon'

No pages found.

Prediction: Pristine

Massachusetts Bay Transportation Authority trains sit idle early Saturday in Boston

'Rail transport', 'Rapid transit', 'Train', 'Railroad car', 'track', 'Transport','Passenger M', 'M / 06d', 'Track', 'M Line', 'Passenger'

No pages found.

Prediction: Pristine

A world which has left the Soviet era behind shoppers in Moscow

'Christmas Tree', 'Christmas Day', 'Bazaar', 'Christmas lights', 'Marketplace', 'Shopping', 'Tree', 'Public space', 'Meter', 'Tradition', 'Public','Lighting', 'Space','Market m*'

No pages found.

Prediction: Falsified

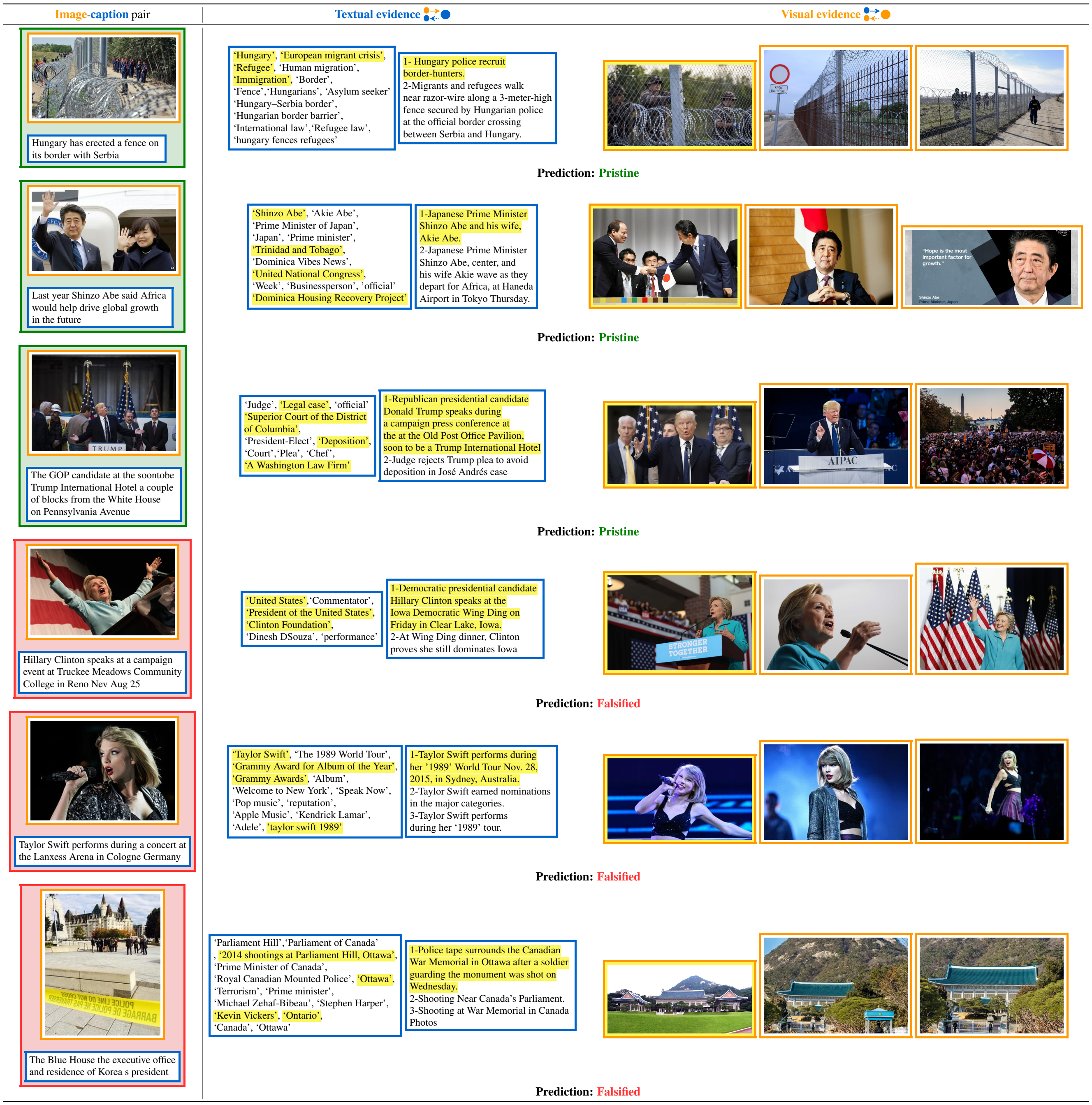
Niall Horan from left Harry Styles Liam Payne Zayn Malikand Louis Tomlinson of the musical group One Direction perform

'Harry Styles','2013', 'Pop rock' 'American Music Awards of 2013', 'Live From the Red Carpet: The 2013 American Music Awards', 'One Direction', 'Award', 'Image', 'American Music Awards', 'Zayn', 'Louis Tomlinson', 'Liam Payne', 'american music awards 2013 one direction'

1- AMA Awards: One Direction seated.
2- One Direction American Music Awards 2013
3- Red Carpet Pictures from the 2013 American Music Awards(AMA)
4-One Direction Wins Pop/Rock Band/Duo/Group - AMA 2013

Prediction: Pristine

Figure 1. Some of the examples that were selected as **'hard to verify'** in the user study. The ground truth is indicated by the pairs' background colour; examples with green background are pristine, red background are falsified. The model's prediction is indicated below each example's set; **green** for predicting pristine and **red** for predicting falsified. The first group of examples does not have textual evidence retrieved, making it harder to verify the context. The last example shows a falsified example with a similar context to the evidence, therefore, the evidence is highly similar to the query.



Hungary has erected a fence on its border with Serbia

‘Hungary’, ‘European migrant crisis’, ‘Refugee’, ‘Human migration’, ‘Immigration’, ‘Border’, ‘Fence’,‘Hungarians’, ‘Asylum seeker’ ‘Hungary–Serbia border’, ‘Hungarian border barrier’, ‘International law’,‘Refugee law’, ‘hungary fences refugees’

1- Hungary police recruit border-hunters.
2-Migrants and refugees walk near razor-wire along a 3-meter-high fence secured by Hungarian police at the official border crossing between Serbia and Hungary.

Prediction: Pristine

Last year Shinzo Abe said Africa would help drive global growth in the future

‘Shinzo Abe’, ‘Akie Abe’, ‘Prime Minister of Japan’, ‘Japan’, ‘Prime minister’, ‘Trinidad and Tobago’, ‘Dominica Vibes News’, ‘United National Congress’, ‘Week’, ‘Businessperson’, ‘official’ ‘Dominica Housing Recovery Project’

1-Japanese Prime Minister Shinzo Abe and his wife, Akie Abe.
2-Japanese Prime Minister Shinzo Abe, center, and his wife Akie wave as they depart for Africa, at Haneda Airport in Tokyo Thursday.

Prediction: Pristine

The GOP candidate at the soontobe Trump International Hotel a couple of blocks from the White House on Pennsylvania Avenue

‘Judge’, ‘Legal case’, ‘official’ ‘Superior Court of the District of Columbia’, ‘President-Elect’, ‘Deposition’, ‘Court’,‘Plea’, ‘Chef’, ‘A Washington Law Firm’

1-Republican presidential candidate Donald Trump speaks during a campaign press conference at the at the Old Post Office Pavilion, soon to be a Trump International Hotel
2-Judge rejects Trump plea to avoid deposition in José Andrés case

Prediction: Pristine

Hillary Clinton speaks at a campaign event at Truckee Meadows Community College in Reno Nev Aug 25

‘United States’,‘Commentator’, ‘President of the United States’, ‘Clinton Foundation’, ‘Dinesh DSouza’, ‘performance’

1-Democratic presidential candidate Hillary Clinton speaks at the Iowa Democratic Wing Ding on Friday in Clear Lake, Iowa.
2-At Wing Ding dinner, Clinton proves she still dominates Iowa

Prediction: Falsified

Taylor Swift performs during a concert at the Lanxess Arena in Cologne Germany

‘Taylor Swift’, ‘The 1989 World Tour’, ‘Grammy Award for Album of the Year’, ‘Grammy Awards’, ‘Album’, ‘Welcome to New York’, ‘Speak Now’, ‘Pop music’, ‘reputation’, ‘Apple Music’, ‘Kendrick Lamar’, ‘Adele’, ’taylor swift 1989’

1-Taylor Swift performs during her ’1989’ World Tour Nov. 28, 2015, in Sydney, Australia.
2-Taylor Swift earned nominations in the major categories.
3-Taylor Swift performs during her ‘1989’ tour.

Prediction: Falsified

The Blue House the executive office and residence of Korea s president

‘Parliament Hill’,‘Parliament of Canada’ , ‘2014 shootings at Parliament Hill, Ottawa’, ‘Prime Minister of Canada’, ‘Royal Canadian Mounted Police’, ‘Ottawa’, ‘Terrorism’, ‘Prime minister’, ‘Michael Zehaf-Bibeau’, ‘Stephen Harper’, ‘Kevin Vickers’, ‘Ontario’, ‘Canada’, ‘Ottawa’

1-Police tape surrounds the Canadian War Memorial in Ottawa after a soldier guarding the monument was shot on Wednesday.
2-Shooting Near Canada’s Parliament.
3-Shooting at War Memorial in Canada Photos

Prediction: Falsified

Figure 2. Other qualitative examples. The ground truth is indicated by the pairs’ background colour; examples with green background are pristine, red background are falsified. The model’s prediction is indicated below each example’s set; **green** for predicting pristine and **red** for predicting falsified. Highlighted items are the ones with the highest attention.