# A Survey on Automated Fact-checking

Zhijiang Guo, Michael Schlichtkrull, Andreas Vlachos
TACL 2022

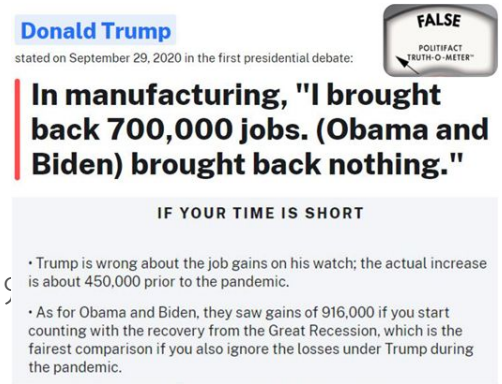
발제자: 김한성
23-03-24

# Abstract

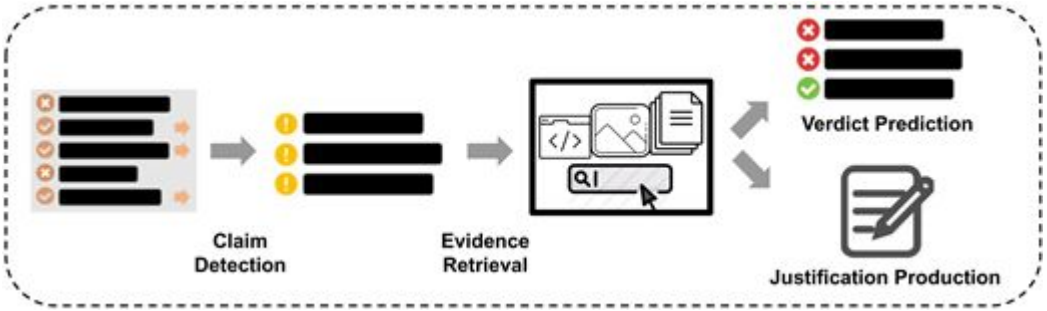

- Fact-checking은 더 중요해지고 있는 정보화 사회에서 NLP,ML,Knowledge representation를 활용하여 어떻게 하면 자동화된 fact-checking을 할 수 있을까하는 연구가 많이 진행되고 있다.
- 본 저자는 fact-checking을 자동화하는 연구에 대한 조사와 discussion을 남긴다.

# Introduction

- 미국에서 운영하는 비영리기구로 "Truth-O-Meter" rating을 기준으로 해당 진술이 사실인지 아닌지 모아놓은 사이트
  - 여러 가십이 사실인지 아닌지 분별
  - 의회에서 나눈 회의가 참인지 아닌지 판별
- 진위 여부를 알기 위한 노력
  - Rumor Checking
  - Stance detection
  - fake news detection
- 현재 연구의 보완점
  - Checking은 다양한 방향 (루머(SNS), 허위(WiKi)등으로 이뤄지고 있
  - Checking의 근거가 Explainable하지 않다.
- 저자는 본 서베이에서 automated fact-checking에 대한 framework를 제안 및 이후 연구에 기여하고자하는 목적



# Task Definition




1. Claim Detection : 입력 받은 문장이 주장인지 아닌지 판단
2. Evidence Retrieval : 진위여부를 알 수 있는 근거를 수집
3. Verdict Prediction : 진위여부 판단
4. Justification Production : 근거에 대한 설명문을 생성

# Claim Detection

- Check worthy : binary decision for each potential claim

"Over six million Americans had COVID-19 in January"
Check-worthy

"warter is wet"
No Check-worthy

  - importance-ranking of claims
    - subjective task
    - COVID-19 vs. 21.Spanish flu

by several questions, to help annotators think about different aspects of check-worthiness. Annotators were asked to answer the following three questions for each tweet (using a scale of 1–5):

- Checkable
  - checkable : claim make assertion about the world
  - "i woke up at 7 am today" : No checkable

# Evidence Retrieval

- Evidence
  - text, table, knowledge graph, images, relevant metadata
- define : information that can be retrieved from source and veracity as coherence with the evidence
- Source
  - Wikipedia(FEVER)
  - provided search engine
  - legal documents

  task
- stance detection, Information Retrieval

# Verdict Prediction

- binary classification or Multiple label classification
  - supported/refuted

# Justification Production

- 의사결정 자체는 믿을 수 있으나 설명 가능하지 않다 보니 의사결정과정을 이해할 수가 없다.이후 black-box로 정의하고 이를 해결하기 위해 justification production을 제안한다.
- 이에 evidence와 claim, claim result 등을 활용하여 generation하는 태스크가 제안됨

- 전략
  - highlight the salient parts of evidence (token별 attention weights).
  - Decision making process (Decision Tree)
  - 생성모델

# Dataset _ Input

- claim detection을 위한 것 : 주로 sns의 올라온 텍스트 형태의 주장을 입력으로 한다.
  1. Danish dataset Reddit
     1. CheckThat
     2. Anotator로 하여금 check worthy있는 지 확인하는 것으로 진행
  2. document of multiple claims (e.g 회의 기록, 법정 진술)
     1. 토론 등 주장이 오가는 논쟁 문서를 기준으로 check worthy가 있는 주장인지 분류하는 작업
  3. Only Politic for checkable set
  4. crawling

| Dataset | Type | Input | #Inputs | Evidence | Verdict | Sources | Lang |
|---|---|---|---|---|---|---|---|
| CredBank (Mitra and Gilbert, 2015) | Worthy | Aggregate | 1,049 | Meta | 5 Classes | Twitter | En |
| Weibo (Ma et al., 2016) | Worthy | Aggregate | 5,656 | Meta | 2 Classes | Twitter/Weibo | En/Ch |
| PHEME (Zubiaga et al., 2016) | Worthy | Individual | 330 | Text/Meta | 3 Classes | Twitter | En/De |
| RumourEval19 (Gorrell et al., 2019) | Worthy | Individual | 446 | Text/Meta | 3 Classes | Twitter/Reddit | En |
| DAST (Lillie et al., 2019) | Worthy | Individual | 220 | Text/Meta | 3 Classes | Reddit | Da |
| Suspicious (Volkova et al., 2017) | Worthy | Individual | 131,584 | ✗ | 2/5 Classes | Twitter | En |
| CheckThat20-T1 (Barrón-Cedeño et al., 2020) | Worthy | Individual | 8,812 | ✗ | Ranking | Twitter | En/Ar |
| CheckThat21-T1A (Nakov et al., 2021b) | Worthy | Individual | 17,282 | ✗ | 2 Classes | Twitter | Many |
| Debate (Hassan et al., 2015) | Worthy | Statement | 1,571 | ✗ | 3 Classes | Transcript | En |
| ClaimRank (Gencheva et al., 2017) | Worthy | Statement | 5,415 | ✗ | Ranking | Transcript | En |
| CheckThat18-T1 (Atanasova et al., 2018) | Worthy | Statement | 16,200 | ✗ | Ranking | Transcript | En/Ar |
| CitationReason (Redi et al., 2019) | Checkable | Statement | 4,000 | Meta | 13 Classes | Wikipedia | En |
| PolitiTV (Konstantinovskiy et al., 2021) | Checkable | Statement | 6,304 | ✗ | 7 Classes | Transcript | En |

Table 1: Summary of claim detection datasets. Input can be a set of posts (aggregate) or an individual post from social media, or a statement. Evidence include text and metadata. Verdict can be a multi-class label or a rank list.

# Dataset _ evidence

- wiki data만 사용한 것이 아니라 이질성()을 확보하기 위해 많은 데이터 추가함.
  1. Source 다양
     1. Crime
     2. Climate
     3. Science
     4. Social issue
  2. evidence
     1. Knowledge graph
     2. text
     3. other meta data

| Dataset | Input | #Inputs | Evidence | Verdict | Sources | Lang |
|---|---|---|---|---|---|---|
| CrimeVeri (Bachenko et al., 2008) | Statement | 275 | ✗ | 2 Classes | Crime | En |
| Politifact (Vlachos and Riedel, 2014) | Statement | 106 | Text/Meta | 5 Classes | Fact Check | En |
| StatsProperties (Vlachos and Riedel, 2015) | Statement | 7,092 | KG | Numeric | Internet | En |
| Emergent (Ferreira and Vlachos, 2016) | Statement | 300 | Text | 3 Classes | Emergent | En |
| CreditAssess (Popat et al., 2016) | Statement | 5,013 | Text | 2 Classes | Fact Check/Wiki | En |
| PunditFact (Rashkin et al., 2017) | Statement | 4,361 | ✗ | 2/6 Classes | Fact Check | En |
| Liar (Wang, 2017) | Statement | 12,836 | Meta | 6 Classes | Fact Check | En |
| Verify (Baly et al., 2018) | Statement | 422 | Text | 2 Classes | Fact Check | Ar/En |
| CheckThat18-T2 (Barrón-Cedeño et al., 2018) | Statement | 150 | ✗ | 3 Classes | Transcript | En |
| Snopes (Hanselowski et al., 2019) | Statement | 6,422 | Text | 3 Classes | Fact Check | En |
| MultiFC (Augenstein et al., 2019) | Statement | 36,534 | Text/Meta | 2-27 Classes | Fact Check | En |
| Climate-FEVER (Diggelmann et al., 2020) | Statement | 1,535 | Text | 4 Classes | Climate | En |
| SciFact (Wadden et al., 2020) | Statement | 1,409 | Text | 3 Classes | Science | En |
| PUBHEALTH (Kotonya and Toni, 2020b) | Statement | 11,832 | Text | 4 Classes | Fact Check | En |
| COVID-Fact (Saakyan et al., 2021) | Statement | 4,086 | Text | 2 Classes | Forum | En |
| X-Fact (Gupta and Srikumar, 2021) | Statement | 31,189 | Text | 7 Classes | Fact Check | Many |
| cQA (Mihaylova et al., 2018) | Answer | 422 | Meta | 2 Classes | Forum | En |
| AnswerFact (Zhang et al., 2020) | Answer | 60,864 | Text | 5 Classes | Amazon | En |
| NELA (Horne et al., 2018) | Article | 136,000 | ✗ | 2 Classes | News | En |
| BuzzfeedNews (Potthast et al., 2018) | Article | 1,627 | Meta | 4 Classes | Facebook | En |
| BuzzFace (Santia and Williams, 2018) | Article | 2,263 | Meta | 4 Classes | Facebook | En |
| FA-KES (Salem et al., 2019) | Article | 804 | ✗ | 2 Classes | VDC | En |
| FakeNewsNet (Shu et al., 2020) | Article | 23,196 | Meta | 2 Classes | Fact Check | En |
| FakeCovid (Shahi and Nandini, 2020) | Article | 5,182 | ✗ | 2 Classes | Fact Check | Many |

Table 2: Summary of factual verification datasets with natural inputs. KG denotes knowledge graphs. ChectThat18 has been extended later (Hasanain et al., 2019; Barrón-Cedeño et al., 2020; Nakov et al., 2021b). NELA has been updated by adding more data from more diverse sources (Nørregaard et al., 2019; Gruppi et al., 2020, 2021)

# Dataset _ verdict

- 최종 라벨을 위함.

| Dataset | Input | #Inputs | Evidence | Verdict | Sources | Lang |
|---|---|---|---|---|---|---|
| KLinker (Ciampaglia et al., 2015) | Triple | 10,000 | KG | 2 Classes | Google/Wiki | En |
| PredPath (Shi and Weninger, 2016) | Triple | 3,559 | KG | 2 Classes | Google/Wiki | En |
| KStream (Shiralkar et al., 2017) | Triple | 18,431 | KG | 2 Classes | Google/Wiki | En |
| UFC (Kim and Choi, 2020) | Triple | 1,759 | KG | 2 Classes | Wiki | En |
| LieDetect (Mihalcea and Strapparava, 2009) | Passage | 600 | ✗ | 2 Classes | News | En |
| FakeNewsAMT (Pérez-Rosas et al., 2018) | Passage | 680 | ✗ | 2 Classes | News | En |
| FEVER (Thorne et al., 2018a) | Statement | 185,445 | Text | 3 Classes | Wiki | En |
| HOVER (Jiang et al., 2020) | Statement | 26,171 | Text | 2 Classes | Wiki | En |
| WikiFactCheck (Sathe et al., 2020) | Statement | 124,821 | Text | 2 Classes | Wiki | En |
| VitaminC (Schuster et al., 2021) | Statement | 488,904 | Text | 3 Classes | Wiki | En |
| TabFact (Chen et al., 2020) | Statement | 92,283 | Table | 2 Classes | Wiki | En |
| InfoTabs (Gupta et al., 2020) | Statement | 23,738 | Table | 3 Classes | Wiki | En |
| Sem-Tab-Fact (Wang et al., 2021) | Statement | 5,715 | Table | 3 Classes | Wiki | En |
| FEVEROUS (Aly et al., 2021) | Statement | 87,026 | Text/Table | 3 Classes | Wiki | En |
| ANT (Khouja, 2020) | Statement | 4,547 | ✗ | 3 Classes | News | Ar |
| DanFEVER (Nørregaard and Derczynski, 2021) | Statement | 6,407 | Text | 3 Classes | Wiki | Da |

Table 3: Summary of factual verification datasets with artificial inputs. Google denotes Google Relation Extraction Corpora, and WSDM means the WSDM Cup 2017 Triple Scoring challenge.

# Dataset _ justification

- 데이터가 고착화 되지는 않음. 다만, 이러한 데이터 구축을 위한 시도 존재
- “Where is your Evidence: Improving Fact-checking by Justification Modeling” Alhindi et al., EMNLP 2018

**Statement:**“Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations.”
**Speaker:** Florida Democratic Party
**Context:** TV Ad
**Label:** half-true
**Extracted Justification:** A TV ad by the Florida Democratic Party says Scott ”cut education to pay for even more tax breaks for big, powerful, well-connected corporations.” However, the ad exaggerates when it focuses attention on tax breaks for ”big, powerful, well-connected corporations.” Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.

Table 1: Excerpt from the LIAR-PLUS dataset

# Modeling Strategies

기존 연구되었던 모델의 프레임 워크의 구성요소를 비교하며

각 subtask에 적절한 모델을 소개한다.

# Claim Detection

Sequece model을 사용하거나 GNN계열 모델을 활용.

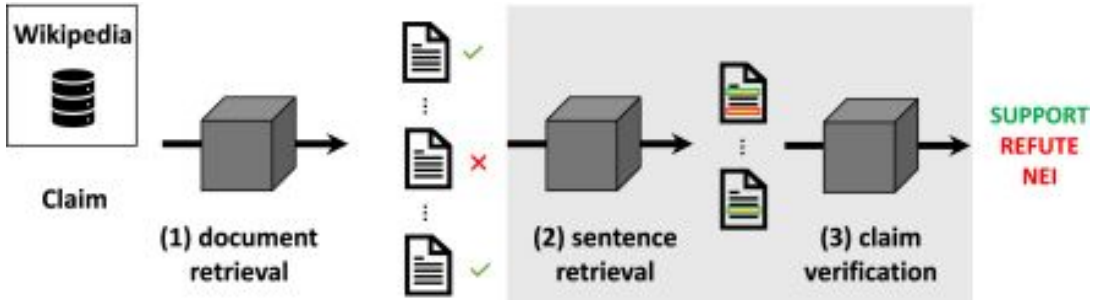
Rumor Detection, Detecting Scientific etc.

**Claim Detection**

- SciTweets - A Dataset and Annotation Framework for Detecting Scientific Online Discourse (Hafid et al., 2022) [Paper] [Code] **CIKM 2022**
- Zoom Out and Observe: News Environment Perception for Fake News Detection (Sheng et al., 2022) [Paper] [Code] **ACL 2022**
- DDGCN: Dual Dynamic Graph Convolutional Networks for Rumor Detection on Social Media (Sun et al., 2022) [Paper] **AAAI 2022**
- Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks (Lin et al., 2021) [Paper] **EMNLP 2021**
- STANKER: Stacking Network based on Level-grained Attention-masked BERT for Rumor Detection on Social Media (Rao et al., 2021) [Paper] [Code] **EMNLP 2021**
- Inconsistency Matters: A Knowledge-guided Dual-inconsistency Network for Multi-modal Rumor Detection (Sun et al., 2021) [Paper] [Code] **Findings EMNLP 2021**
- Active Learning for Rumor Identification on Social Media (Farinneya et al., 2021) [Paper] **Findings EMNLP 2021**
- Towards Propagation Uncertainty: Edge-enhanced Bayesian Graph Convolutional Networks for Rumor Detection (Wei et al., 2021) [Paper] [Code] **ACL 2021**

# Evidence Retrieval and Claim Verification

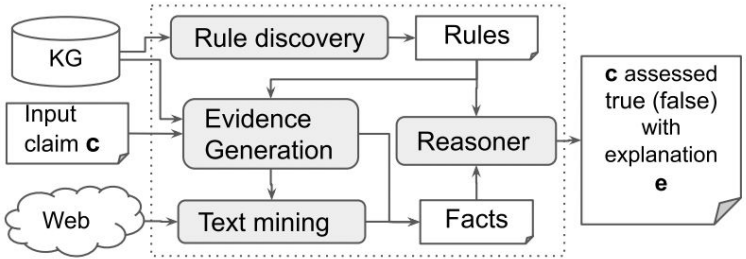
evidence retrieval과 Claim verification은 같은 파이프라인을 구성한 케이스가 다수 (e.g FEVER)



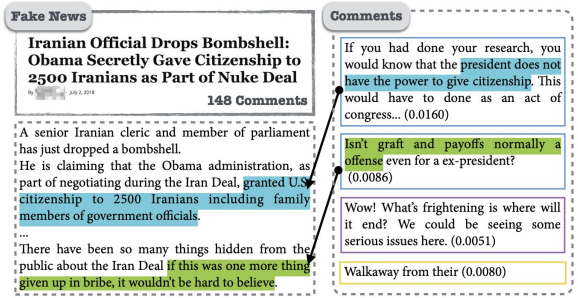

https://dl.acm.org/doi/full/10.1145/3485127?casa_token=6qBLQSdeSsAAAAAA%3AJTC-vBi08E7iCttwI2EG4zfG2GeDPnLxsWcbR7Alv4Nf_MdNR9oN4PPnXAsrJ0F2YjvPtfC8YHGNQw

# Justification Production

- highlight the salient parts of evidence(token별 attention weights).
- Decision making process
- 생성모델




DecisionTree : Explainable Fact Checking with Probabilistic Answer Set Programming

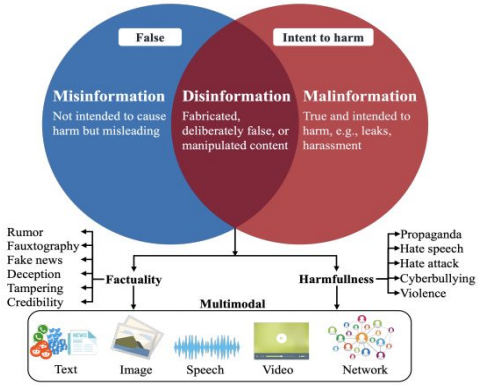



dEFEND: Explainable Fake News Detection

# Related Task

Misinformation and Disinformation

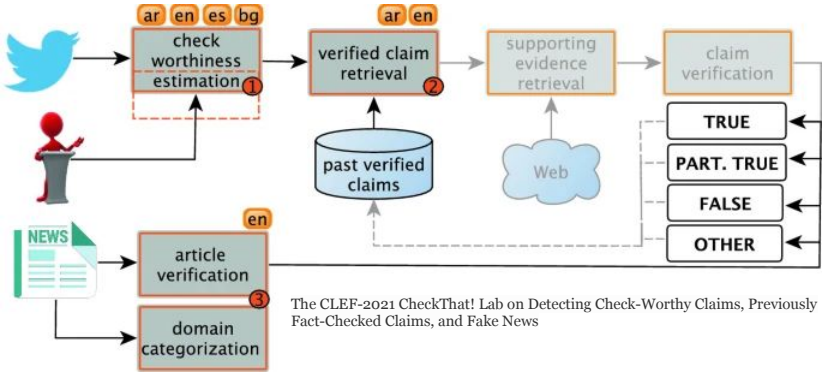
- Fact-checking can help misinformation but not distinguish it from disinformation
- disinformation : 의도가 가미된 조작된 정보




A Survey on Multimodal Disinformation Detection

Detecting Previously Fact-checked Claim

- 이전 정보를 기억




The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News

# Research Challenges

**Choice of Labels** : misleading of trend

**Source and Subjectivity** : Source별 각기 다른 요소(잡지, 기사는 최신성 필요 but 위키는 신뢰성 필요)

**Dataset Artefacts and Biases**

: FEVER -> dependency on claim word (adverserial example에 약함)

**Multimodality** : text뿐 아니라 다양한 컨텐츠에도 연결성이 필요

**Multilinguality** : Gupta and Srikumar(2021) a multilingual dataset (cover 25 lan)

**Faithfulness** : justification production의 평가지표의 신뢰부족(생성모델).

**From Debunking to Early Intervention and Prebunking**

: misinformation의 개입 시기성 논의 -> misinformation이 reference가 되기 전

# Conclusion

본 저자는 automated fact-checking과 관련한 연구를 조사하고 평가.

태스크를 공식화하고 각 태스크별 제안된 방법론들을 다양하게 소개.

또한 각각의 활용가능성있는 데이터셋과 모델의 아키텍쳐를 참조함.

추후 연구에 있어 분야의 발전을 기여함.

https://github.com/Cartus/Automated-Fact-Checking-Resources#claim-detection-dataset

# (추가)Explainable Claim Verification

DTCA

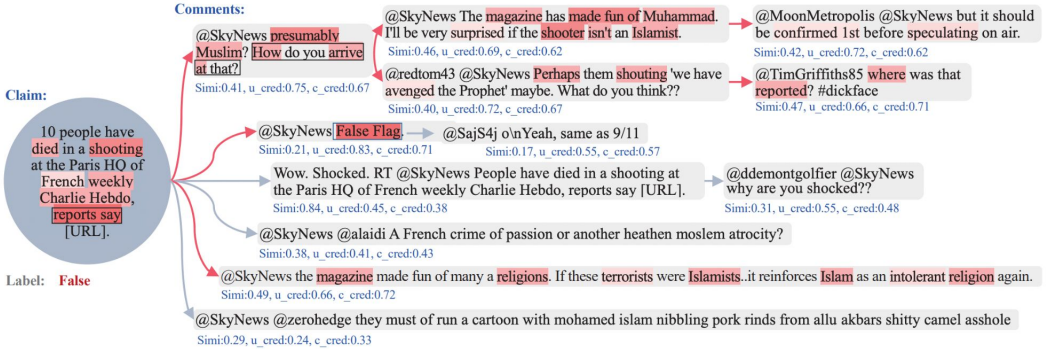
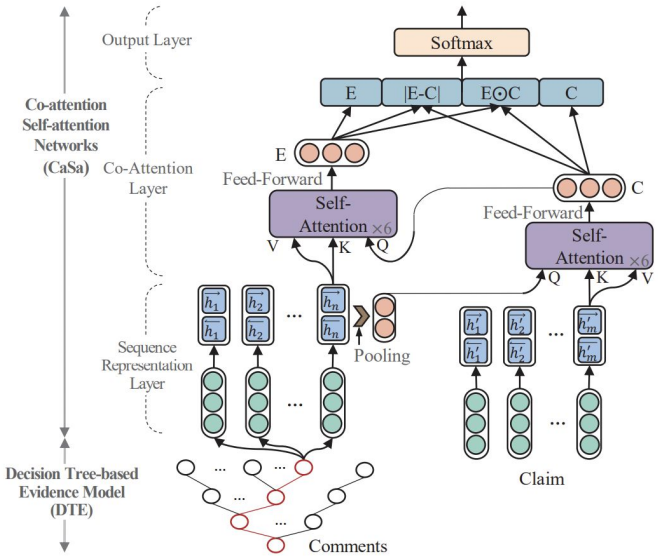



Figure 4: The visualization of a sample (labeled false) in PHEME by DTCA, where the captured evidence (red arrows) and the specific values of decision conditions (blue) are presented by DTE, and the attention of different words (red shades) is obtained by CaSa.

유관된 comments를 DTE에서 추출 및 명시(selected Tree)



# (추가)Generating Fact Checking Explanations

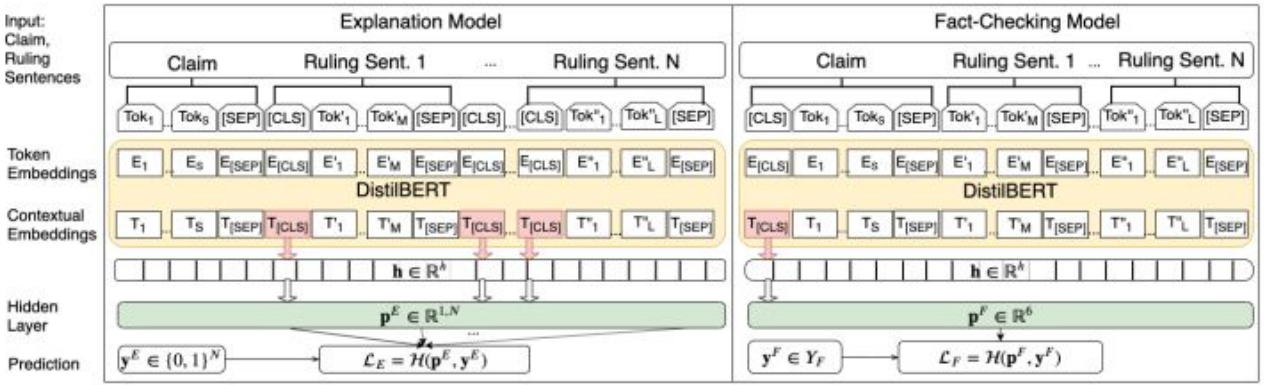



Figure 1: Architecture of the *Explanation* (left) and *Fact-Checking* (right) models that optimise separate objectives.

**Claim**: The last major oil spill from a drilling accident in America happened over 40 years ago in 1969.

**Ruling Comments**: (...) The last major oil spill from a drilling accident in America happened over 40 years ago in 1969.
(...) The largest in volume was the Santa Barbara spill of 1969 referenced by Murdock and Johnson, in which an estimated 100,000 barrels of oil spilled into the Pacific Ocean, according to the API. The Santa Barbara spill was so big it ranked seventh among the 10 largest oil spills caused by marine well blowouts in the world, the report states. Two other U.S. spills, both in 1970, rank eighth and 10th. Fourteen marine blowouts have taken place in the U.S. between 1969 and 2007. Six of them took place after 1990 and spilled a total of nearly 13,700 barrels.
(...) We interviewed three scientists who said that the impact of a spill has little to do with its volume. Scientists have proven that spills far smaller than Santa Barbara's have been devastating.

**Justification**: While the nation's largest oil well blowout did take place in 1969, it's not factually correct to call it the "last major oil spill". First of all, two of the largest blowouts in the world took place in the U. S. the following year. More importantly, experts agree that spills far smaller in volume to the 1969 disaster have been devastating. From a scientific perspective, Johnson's decision to single out the 1969 blowout as the last "major" one makes no sense.

**Ruling**: Half-True

Table 1: Example instance from the LIAR-PLUS dataset, with oracle sentences for generating the justification highlighted.

Ruling sent를 같이 넣어서 Extractive Summary

감사합니다.