# MUSER: A MUlti-Step Evidence Retrieval Enhancement Framework for Fake News Detection

Hao Liao
Shenzhen University
Shenzhen, China
haoliao@szu.edu.cn

Jiahao Peng
Shenzhen University
Shenzhen, China
2070276145@email.szu.edu.cn

Zhanyi Huang
Shenzhen University
Shenzhen, China
huangzhanyi2020@email.szu.edu.cn

Wei Zhang
Shenzhen University
Shenzhen, China
2210275010@email.szu.edu.cn

Guanghua Li
Shenzhen University
Shenzhen, China
2210275050@email.szu.edu.cn

Kai Shu*
Illinois Institute of Technology
Chicago, USA
kshu@iit.edu

Xing Xie*
Microsoft Research Asia
Beijing, China
xingx@microsoft.com

KDD2023

code:https://github.com/Complex-data/MUSER

**Reported by Xiaoke Li**

# Introduction

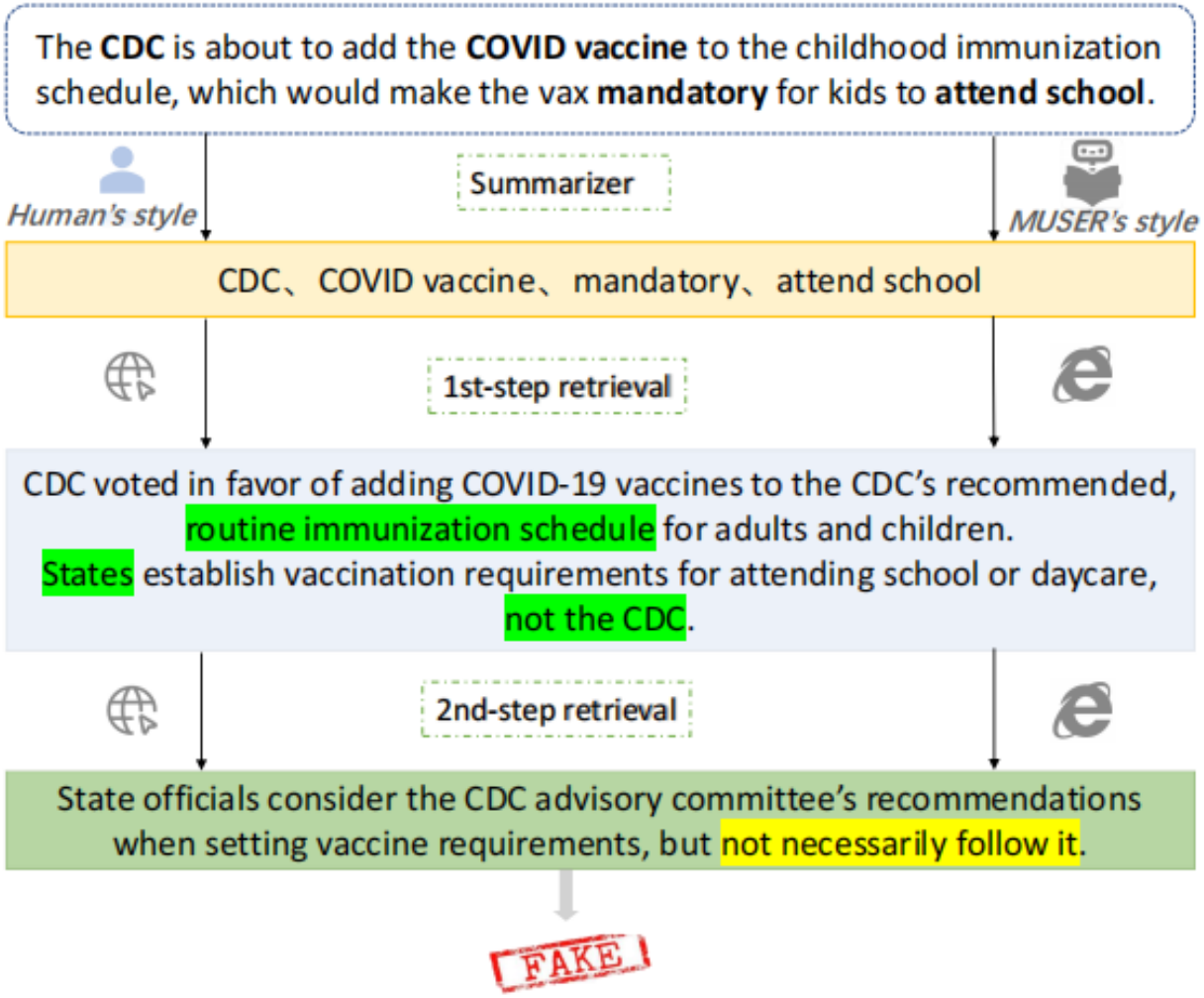
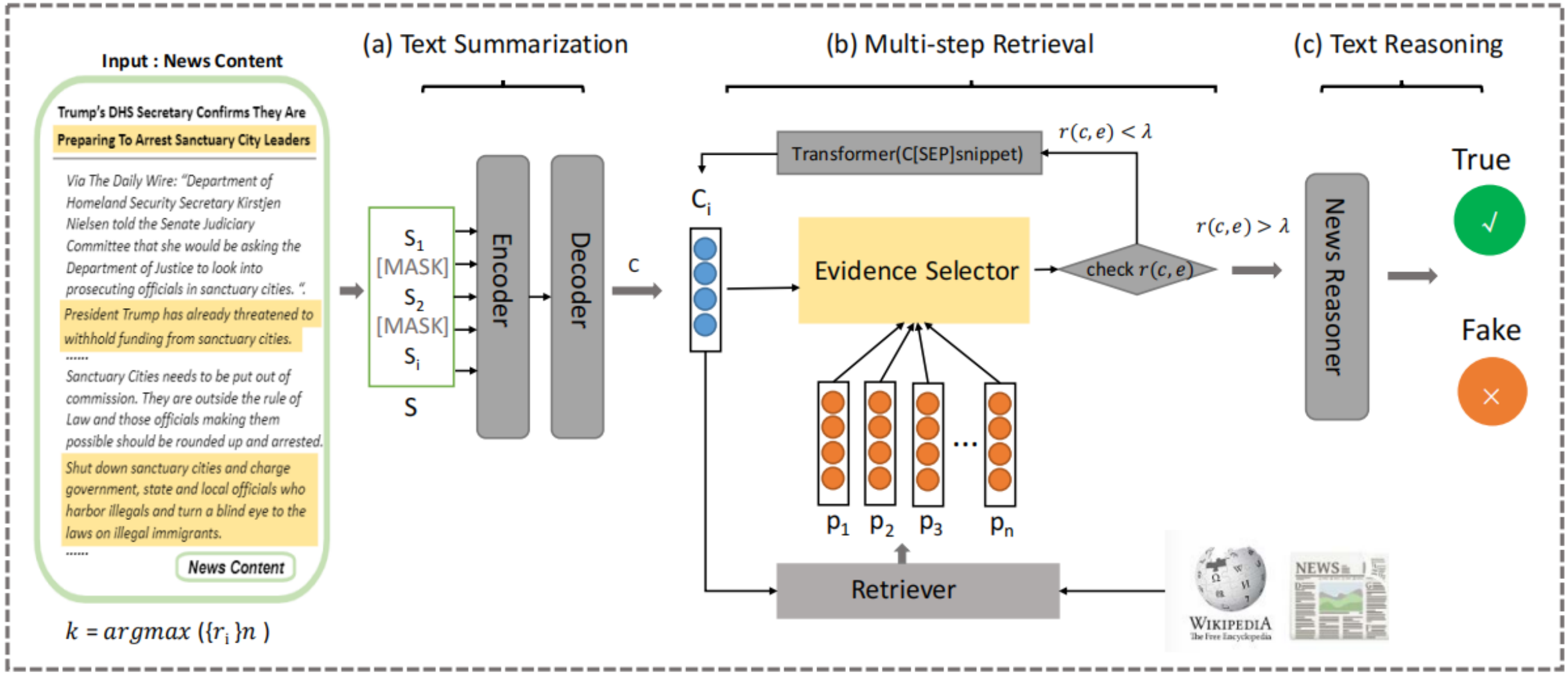



**Figure 1: A motivating example of MUSER model. Our model simulates a human evaluating news through three steps: (1) Summarization of the key information, (2) Retrieval and evaluation of relevant evidence: the model assesses the sufficiency and quality of the evidence, determining if additional inquiries are necessary, (3) Conclusion regarding the truthfulness of the news based on the gathered evidence.**


Input : News Content
Trump's DHS Secretary Confirms They Are Preparing To Arrest Sanctuary City Leaders
Via The Daily Wire: "Department of Homeland Security Secretary Kirstjen Nielsen told the Senate Judiciary Committee that she would be asking the Department of Justice to look into prosecuting officials in sanctuary cities. ".
President Trump has already threatened to withhold funding from sanctuary cities.
.....
Sanctuary Cities needs to be put out of commission. They are outside the rule of Law and those officials making them possible should be rounded up and arrested.
Shut down sanctuary cities and charge government, state and local officials who harbor illegals and turn a blind eye to the laws on illegal immigrants.
.....
News Content
$k = argmax(\{r_i\}n)$
(a) Text Summarization
S1 [MASK] S2 [MASK] Si
S
Encoder
Decoder
c
(b) Multi-step Retrieval
Ci
Transformer(C[SEP]snippet)
$r(c,e) < \lambda$
Evidence Selector
check $r(c,e)$
$r(c,e) > \lambda$
p1 p2 p3 ... pn
Retriever
WIKIPEDIA
The Free Encyclopedia
NEWS
(c) Text Reasoning
News Reasoner
True
Fake

**Input : News Content**

**Trump's DHS Secretary Confirms They Are Preparing To Arrest Sanctuary City Leaders**

*Via The Daily Wire: "Department of Homeland Security Secretary Kirstjen Nielsen told the Senate Judiciary Committee that she would be asking the Department of Justice to look into prosecuting officials in sanctuary cities. ".*

*President Trump has already threatened to withhold funding from sanctuary cities.*

.....

*Sanctuary Cities needs to be put out of commission. They are outside the rule of Law and those officials making them possible should be rounded up and arrested.*

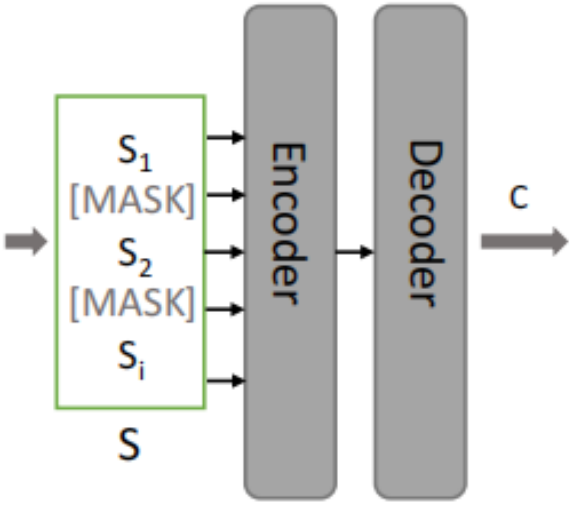
*Shut down sanctuary cities and charge government, state and local officials who harbor illegals and turn a blind eye to the laws on illegal immigrants.*

.....

News Content

$$k = argmax\ (\{r_i\}n\ )$$

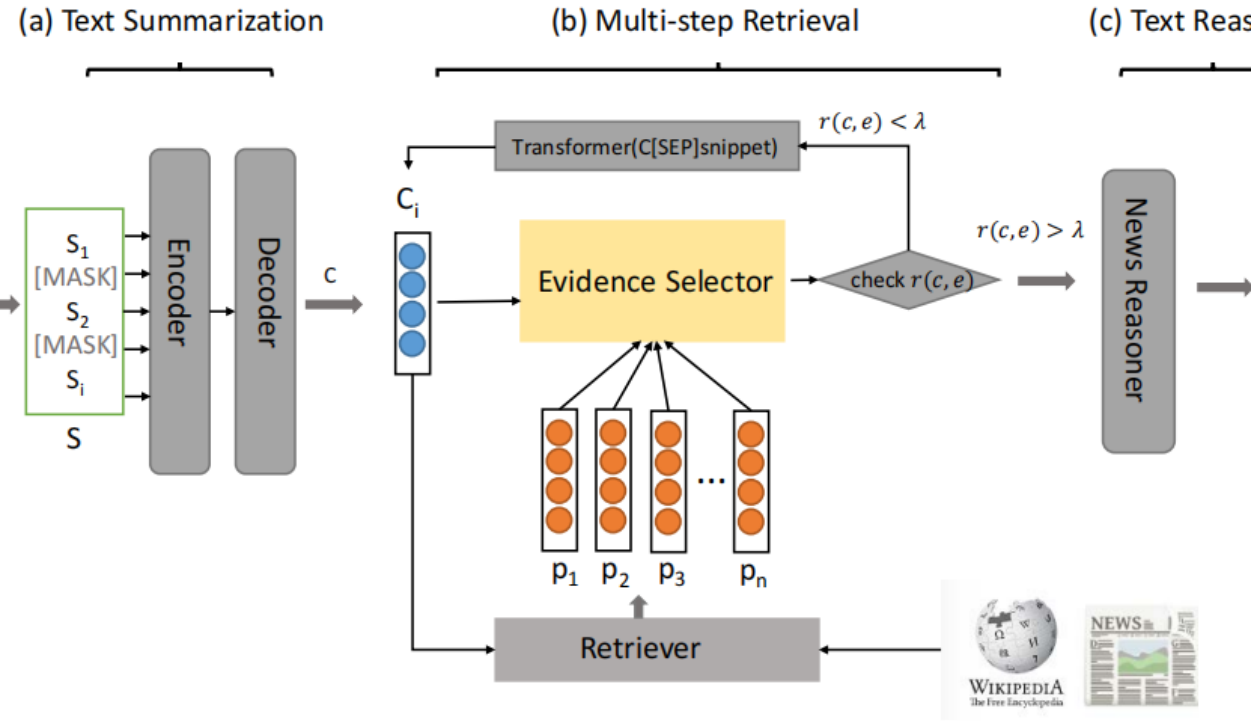
(a) Text Summarization




$$r_i = rouge(S \cup s_i, A \backslash \{S \cup s_i\}), \quad \forall i,\ s_i \notin S \qquad (1)$$

$$k = argmax(\{r_i\}_n) \qquad (2)$$

$$S = S \cup s_k \qquad (3)$$




$$r(c,p) = \varphi(c)^T \varphi(p) \tag{4}$$

$$\varphi(p) = \frac{1}{p} \sum_{i=1}^{|p|} BERT(p,i) \tag{5}$$

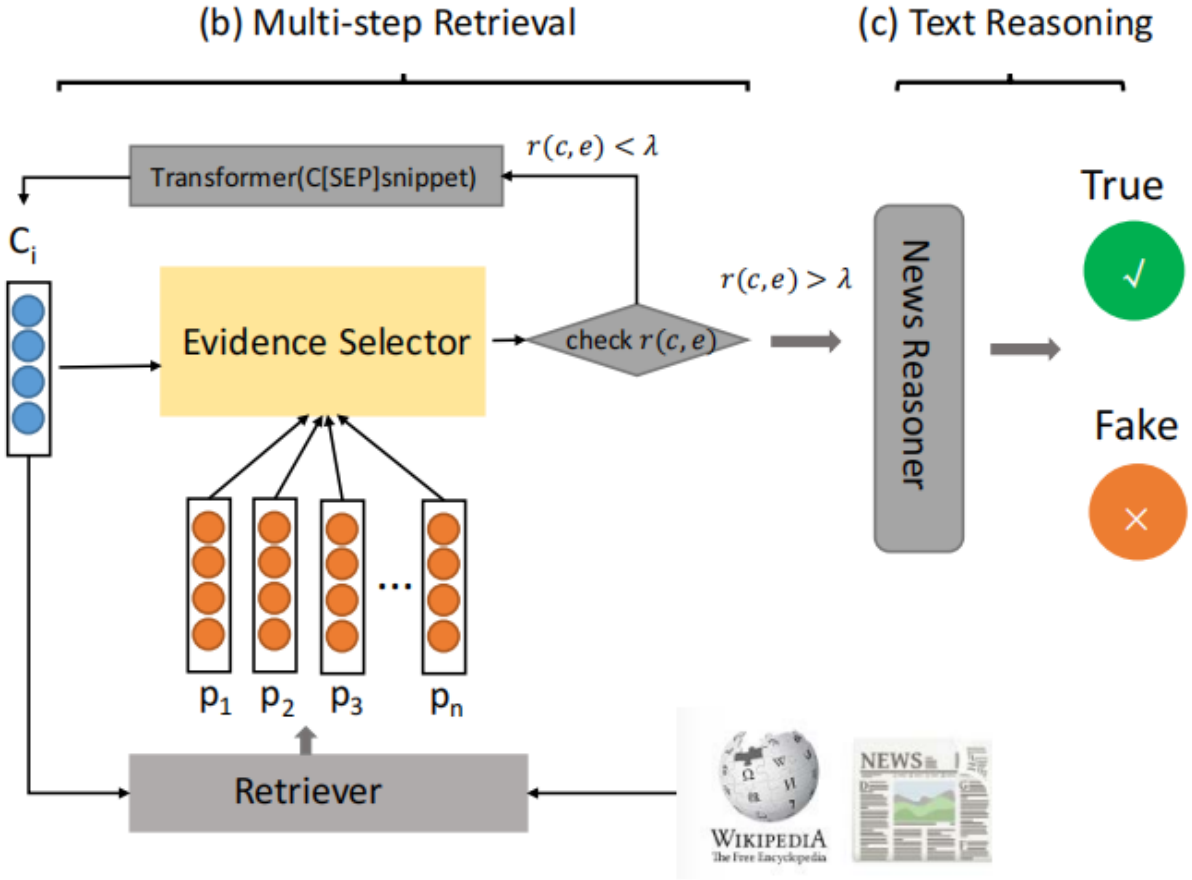
Relevance score-based selection methods rely on vector representations of statements and sentences in paragraphs. For a given statement $C$, we select sentences $s_i$ from the retrieved relevant passages $P = \{s_1, s_2, \ldots, s_n\}$ whose relevance score $r(c, s_i)$ is greater than a certain threshold $\lambda$ set experimentally. Details on setting lambda values can be found in Appendix A.2.3.

The context-aware sentence selection method uses a BERT-based sequence tagging model. We take as input the concatenation of statement claim $C = \{c_1, c_2, ..., c_k\}$ and passages $P = \{p_1, p_2, ..., p_m\}$ and separate them using special tokens: $[CLS]C[SEP]P[EOS]$. For the output of the model, we adopt the BIO token format, which classifies all irrelevant tokens as O, the first token of an evidence sentence as B evidence, and the remaining tokens of an evidence sentence as I evidence. We train a RoBERTa-large based model [50], minimizing the cross-entropy loss:

$$\mathcal{L}_\theta = -\sum_{i=1}^{N} \sum_{j=1}^{l_i} log(p\theta(y_i^j)) \tag{6}$$

$$C_{i+1} = Transformer([C_i[SEP]snippet]) \tag{7}$$

# Method




$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(V(C_i, E_i)) + (1 - y_i) log(1 - V(C_i, E_i)) \quad (8)$$

# Experiments



**Table 1: Statistics of three datasets.**

| Platform | PolitiFact | GossipCop | Weibo |
|---|---|---|---|
| #Real News | 399 | 4,219 | 436 |
| #Fake News | 345 | 3,393 | 311 |
| #Total | 744 | 7,612 | 747 |

# Experiments



| Method | PolitiFact | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | F1-Ma | F1-Mi | F1-T | P-T | R-T | F1-F | P-F | R-F |
| TextCNN | 0.601 | 0.602 | 0.608 | 0.641 | 0.579 | 0.594 | 0.564 | 0.615 |
| TextRNN | 0.610 | 0.609 | 0.616 | 0.650 | 0.586 | 0.603 | 0.572 | 0.636 |
| TextURG | 0.621 | 0.619 | 0.637 | 0.651 | 0.624 | 0.601 | 0.587 | 0.617 |
| BERT | 0.597 | 0.598 | 0.608 | 0.619 | 0.599 | 0.586 | 0.577 | 0.597 |
| DeClarE | 0.654 | 0.651 | 0.656 | 0.689 | 0.673 | 0.651 | 0.613 | 0.664 |
| HAN | 0.661 | 0.660 | 0.679 | 0.676 | 0.682 | 0.643 | 0.650 | 0.637 |
| EHIAN | 0.664 | 0.663 | 0.674 | 0.680 | 0.651 | 0.650 | 0.628 | 0.627 |
| MAC | 0.678 | 0.675 | 0.700 | 0.695 | 0.704 | 0.653 | 0.655 | 0.645 |
| GET | 0.694 | 0.692 | 0.725 | 0.712 | 0.770 | 0.669 | 0.720 | 0.665 |
| MUSER | **0.732*** | **0.729*** | **0.757*** | **0.735*** | **0.780*** | **0.702*** | **0.728*** | **0.681*** |

| Method | GossipCop | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | F1-Ma | F1-Mi | F1-T | P-T | R-T | F1-F | P-F | R-F |
| TextCNN | 0.628 | 0.624 | 0.658 | 0.671 | 0.646 | 0.590 | 0.604 | 0.576 |
| TextRNN | 0.629 | 0.628 | 0.636 | 0.667 | 0.609 | 0.620 | 0.591 | 0.651 |
| TextURG | 0.644 | 0.643 | 0.650 | 0.684 | 0.619 | 0.636 | 0.605 | 0.637 |
| BERT | 0.617 | 0.613 | 0.635 | 0.664 | 0.649 | 0.578 | 0.635 | 0.562 |
| DeClarE | 0.660 | 0.657 | 0.686 | 0.677 | 0.694 | 0.629 | 0.638 | 0.619 |
| HAN | 0.702 | 0.700 | 0.722 | 0.721 | 0.716 | 0.678 | 0.676 | 0.680 |
| EHIAN | 0.705 | 0.702 | 0.731 | 0.713 | 0.749 | 0.673 | 0.694 | 0.654 |
| MAC | 0.729 | 0.727 | 0.725 | 0.742 | **0.756** | 0.705 | 0.713 | 0.697 |
| GET | 0.733 | 0.731 | 0.751 | 0.749 | 0.727 | 0.712 | 0.710 | 0.715 |
| MUSER | **0.776*** | **0.775*** | **0.784*** | **0.843*** | 0.734 | **0.768*** | **0.714*** | **0.830*** |

# Experiments



| Method | Weibo | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | F1-Ma | F1-Mi | F1-T | P-T | R-T | F1-F | P-F | R-F |
| TextCNN | 0.722 | 0.721 | 0.740 | 0.742 | 0.736 | 0.703 | 0.706 | 0.700 |
| TextRNN | 0.741 | 0.737 | 0.771 | 0.730 | 0.812 | 0.701 | 0.756 | 0.654 |
| TextURG | 0.709 | 0.704 | 0.741 | 0.712 | 0.628 | 0.667 | 0.707 | 0.759 |
| BERT | 0.699 | 0.698 | 0.719 | 0.720 | 0.716 | 0.678 | 0.676 | 0.680 |
| DeClarE | 0.746 | 0.745 | 0.765 | 0.758 | 0.771 | 0.724 | 0.732 | 0.717 |
| HAN | 0.689 | 0.687 | 0.711 | 0.706 | 0.716 | 0.662 | 0.668 | 0.657 |
| EHIAN | 0.753 | 0.752 | 0.770 | 0.768 | 0.772 | 0.734 | 0.754 | 0.731 |
| MAC | 0.734 | 0.732 | 0.709 | 0.722 | 0.697 | 0.755 | 0.745 | 0.766 |
| GET | 0.756 | 0.754 | 0.776 | 0.760 | 0.794 | 0.730 | 0.761 | 0.712 |
| MUSER | **0.804*** | **0.802*** | **0.824*** | **0.812*** | **0.837*** | **0.791*** | **0.806*** | **0.778*** |

# Experiments

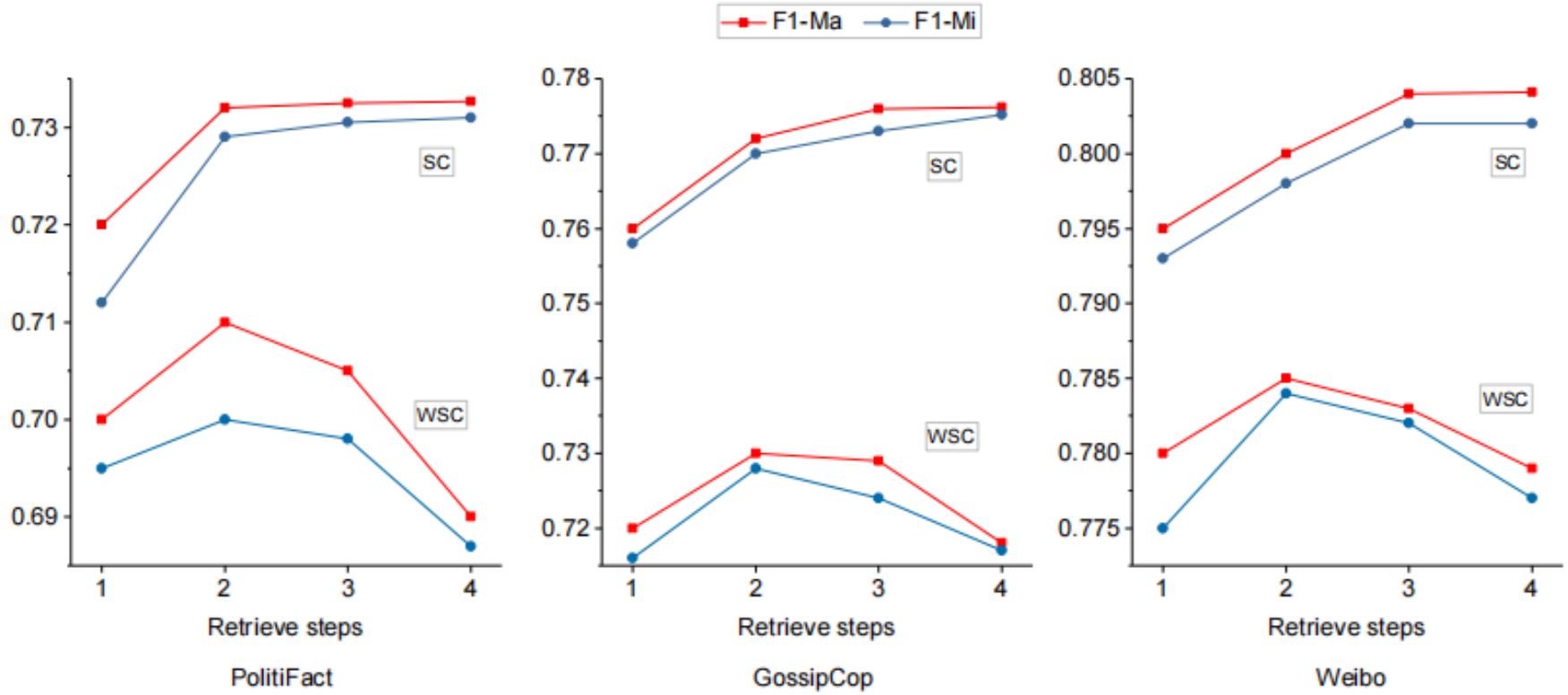



**Figure 3: Results of retrieve step comparison study. The term SC (Step Control) means that the key evidence selection function is activated, while WSC (Without Step Control) means that the key evidence selection function is not included.**

# Experiments

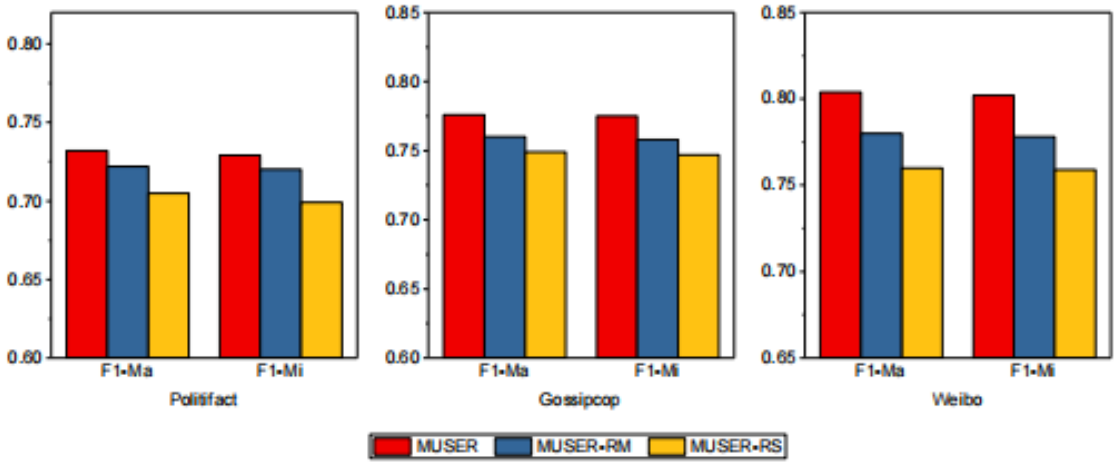



**Figure 4: Results of ablation study. MUSER represents the complete model performance, MUSER-RM represents the removal of the multi-step retrieval module and MUSER-RS represents the removal of the text summary module.**

**Table 5: Results of the user study. The agreement measure means the proportion of concurrence between the user's judgment and the model's judgment.**

| Method | F1 | Precision | Agreement |
|---|---|---|---|
| GET | 0.690 | 0.667 | 70% |
| MUSER | **0.758** | **0.733** | **76.7%** |

# Experiments

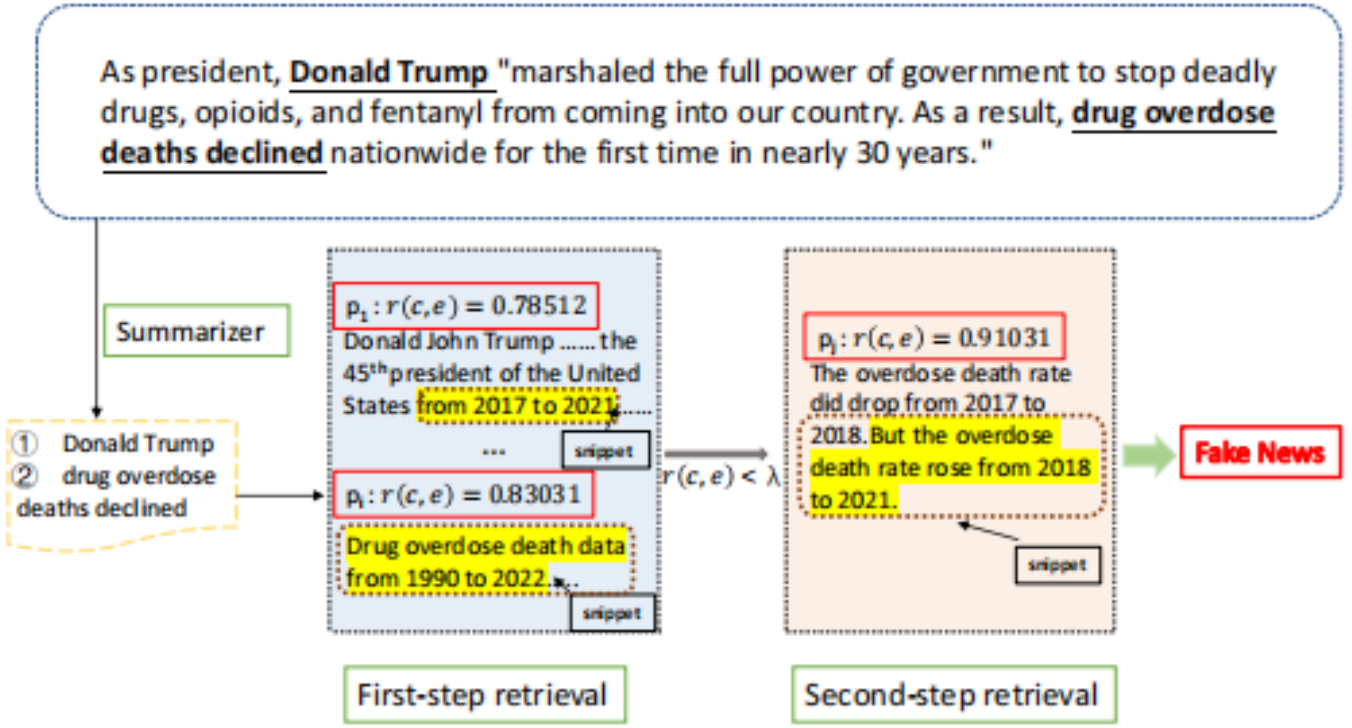



**Figure 5: A verification example generated by MUSER in the Case study. The evidence correlation score $r(c, e)$ obtained by the first step of retrieval is smaller than the threshold $\lambda$ we set. Then proceed to the second step of retrieval to obtain more sufficient evidence.**

# Thanks