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

ABSTRACT

The ease of spreading false information online enables individuals with malicious intent to manipulate public opinion and destabilize social stability. Recently, fake news detection based on evidence retrieval has gained popularity in an effort to identify fake news reliably and reduce its impact. Evidence retrieval-based methods can improve the reliability of fake news detection by computing the textual consistency between the evidence and the claim in news. In this paper, we propose a framework for fake news detection based on **MU**lti-**S**tep **E**vidence **R**etrieval enhancement (MUSER), which simulates the steps of human beings in the process of reading news, summarizing, consulting materials, and inferring whether the news is true or fake. Our model can explicitly model dependencies among multiple evidences, and perform multi-step associations for the evidence required for news verification through multi-step retrieval. In addition, our model is able to automatically collect existing evidence through paragraph retrieval and key evidence selection, which can save the tedious process of manual evidence collection. We conducted extensive experiments on real-world datasets in different languages, and the results demonstrate that our proposed model outperforms state-of-the-art baseline methods for detecting fake news by at least 3% in F1-Macro and 4% in F1-Micro. Furthermore, it provides interpretable evidence for end users.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Information systems** → *Data mining*.

KEYWORDS

Evidence-based Fake News Detection, Multi-step Retrieval, Explainability

ACM Reference Format:

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. MUSER : A MUlti-Step Evidence Retrieval Enhancement Framework for Fake News Detection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (KDD '23).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The explosive growth of fake news has exerted serious negative consequences across society affecting areas such as politics, economy and public health [1]. The phenomenon is characterized by sensationalism and alarmism, which just caters to the mindset of netizens and is easily exploited by the "headline party" [12]. To get more attention, people are inclined to share news or retweet tweets with catchy headlines without proper evaluation. This makes the rapid spread of fake news on social media platforms, outpacing that real news [33]. An overwhelming amount of fake news on social media has made it difficult for individuals to identify true from falsehood, thereby posing a huge threat to social stability [52, 56]. In light of these challenges, automated fake news detection has drawn widespread attention.

Fake news detection is an important task, as it aims to timely and accurately identify fake news on social platforms, and reduce the harm induced by the spread of fake news in the cradle. Meanwhile, fake news detection can help netizens improve their ability to distinguish between true and fake news, and improve the health and ecology of social networks. Despite the efforts of websites and social media platforms to combat fake news, such as Meta encourages users to report untrustworthy posts, and Sina Weibo provides a channel for debunking rumors [45]. Fact-checking sites like FactCheck[1], PolitiFact[2], and Full Fact[3] have also begun to hire

[1]https://www.factcheck.org/
[2]https://www.politifact.com/
[3]https://fullfact.org/

KDD '23, August 06–10, 2023, Long Beach, CA, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

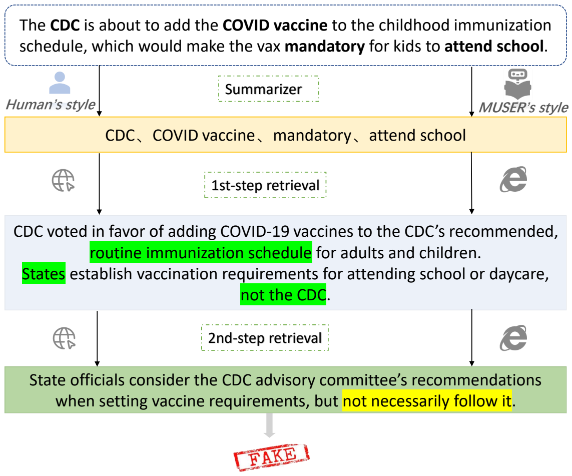

Figure 1: A motivating example of MUSER model. Our model simulates a human evaluating news through three steps: (1) Summarization of the key information, (2) Retrieval and evaluation of relevant evidence: the model assesses the sufficiency and quality of the evidence, determining if additional inquiries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

professional fact-checkers to conduct fact-checking. The increasing volume of news data, with its diversity and complexity, makes manual verification a time-consuming and unscalable process.

To tackle this problem, data mining and machine learning techniques were introduced to detect fake news [3, 43]. These methods typically rely on textual features, such as sentence semantics and news entities, for binary classification through supervised learning [10, 36, 54]. Though effective, these content-based methods exhibit some limitations, as fake news often resembles real news in textual features and lacks important information, such as social context [14]. To overcome these limitations, multi-modal fake news detection frameworks have been proposed, which consider social context by analyzing news propagation patterns on social media, such as retweet relationship networks [24, 25], and user-friend relationships [5, 29]. Fake news can spread rapidly and become difficult to control once it has reached a wide audience [48]. The above-mentioned social context information-based methods require a collection of a substantial amount of social context information, which may not curb the spread of fake news in a timely manner. In addition to the temporal delay issue of detection, methods based on social context face the challenge of preserving user privacy. Thus, recent studies have turned towards evidence-based verification to distinguish between true and false information in fake news detection. These methods treat fake news detection as an inference process where external evidence are provided to probe the accuracy of the assertion. The aim is to extract and incorporate relevant information from the given evidence for claim verification to improve the interpretability of fake news detection. This approach demonstrated promising results in recent studies [26, 35, 47, 50].

Despite substantial advancements over these years, fake news detection still confronts numerous challenges. Evidence-based detection methods suffer from the assumption that evidence is easily accessible, ignoring a large amount manual effort required for evidence collection. Furthermore, prior work has inadequately explored complex, long-range semantic dependencies in evidence, neglecting the intricate relationships between information.

Inspired by brain science [4], we propose a fake news inference framework MUSER, which augmented with **MU**lti-**S**tep **E**vidence **R**etrieval. The human cognitive process typically involves three steps when evaluating information [11] as shown in Figure 1: First, a summary of the key findings or claims in the text is read. Second, supporting evidence for the claims is located and evaluated for quality, which may include sources such as website data, official experiments, or research. Finally, conclusions are drawn based on the evaluated evidence. By following these steps, it is possible to ascertain the sources of information, the evidence used, evidence quality, and limitations, thus helping readers to make informed judgments about the validity of the information. MUSER[4] automatically retrieves existing evidence from Wikipedia through paragraph retrieval and key evidence selection, eliminating the need for manual evidence collection. Evidence needed for news verification are correlated through multi-step retrieval. Furthermore, our model can perform early detection without relying on social context information and provides reasons for the authenticity of the news through retrieved evidence. Although the social media can provide external information for early fake news detection. But there are two drawbacks - privacy concerns related to user comments and the presence of noisy information among user posts. Our main contributions can be summarized as follows:

- We propose an automatic fact-checking framework for fake news detection that is based on multi-step evidence retrieval. Our framework can explicitly model dependencies among multiple pieces of evidence and retrieves the evidence necessary for news verification through multi-step retrieval. The framework simulates the searching behavior of people when verifying news content on the Internet, making it possible to narrow the gap between computers and human experts in fake news detection.
- The implementation of our proposed model includes three core modules: text summarization, multi-step retrieval, and text reasoning. In the multi-step retrieval module, we empoly the method of key evidence selection to control the number of hops, realizing adaptive retrieval steps control.
- We conduct extensive experiments on three real-world datasets , and the results demonstrate the effective of our model in terms of improved interpretability and good performance when compared with state-of-the-art models.

2 RELATED WORK

2.1 Fake News Detection

In recent years, researchers have collaborated with the news ecosystem to better define and characterize fake news through news content and social feedback from web users. We briefly introduce

[4]Code is available at https://anonymous.4open.science/r/MUSER-6FB3/

related work from the following aspects: 1) content-based; 2) social context-based; 3) evidence-based.

Content-based: Content-based methods detect fake news by exploiting news text, writing style, or external knowledge about news entities. Some works detect fake news by extracting news text features, e.g., n-gram distribution and/or utilize Linguistic Inquiry and Word Count (LIWC) [32] features and sentence relationships based on Rhetorical Structure Theory (RST) [39]. The stylistic feature-based approach distinguishes between real and fake news by capturing the specific writing style and emotion usually present in the textual content of fake news [37]. KAN [8] directly evaluates the authenticity of news by comparing news knowledge with knowledge entities in the knowledge graph. Content-based methods are often used in the early detection of fake news to curb the spread of rumors in the early stages of news dissemination.

Social context-based: Social media plays an important role in detecting fake news research [57]. It has been used to improve the performance of fake news detection by integrating contextual information on social platforms, such as user characteristics, comments, and positions [41]. It has been found that the communication structure of real news and fake news is very different [48], so the method based on communication structure becomes popular. Network structure-based methods extract network features by constructing specific networks, such as user interaction networks, user social structures, participation patterns, and news dissemination networks [15, 28, 29, 51].

Evidence-based: The semantic similarity (conflict) in the claim-evidence pairs can be used to determine the veracity of the news by searching Wikipedia or fact-checking websites according to the claims in the news. Early research approaches employ sequence models to embed semantics and apply attention mechanisms to capture claim-evidence semantic relations. For example, DeClar[35] uses BiLSTM to embed the semantics of the evidence, and calculates the evidence score through the attention interaction mechanism. MAC[47] proposes a multi-level multi-head attention network combining word attention and evidence attention to detect fake news. GET[50] models claims and evidence as graph-structured data, proposing a unified evidence-graph-based fake news detection method for the first time. Evidence-checking-based methods can reveal false parts of claims, provide users with evidence that news is true or fake, and improve the interpretability of fake news detection. Although the above studies have achieved great success, they all assume that the evidence declared in the news already exists, but the collection and arrangement of evidence in the actual process often requires a lot of manual operations.

Different from the aforementioned studies, we propose a fake news inference framework augmented by multi-step evidence retrieval. Our model can automatically retrieve existing evidence through Wikipedia, conduct evidence collection, and capture dependencies among evidence through multi-step retrieval.

2.2 Retrieval Enhancement

Recent work has shown that retrieving additional information can improve the performance of various downstream tasks [20]. Such tasks include open-domain question answering, fact checking, fact completion, long form question answering, Wikipedia article generation, and dialogue. In the classic and simplest form of fact-checking, with claims as query conditions, the k relevant passages $K_S = \{P_1, P_2, \ldots, P_{|K_S|}\}$ needed to verify the claims are obtained. Evidence may be contained within a paragraph, or even within a sentence. Retrieve multiple relevant passages $P_i \in K_S$ by a given query Q, and let the reading comprehension model extract the answer from P_i [13, 18]. These studies all used a single-step search. Contrary to the case of single-step retrieval, evidence for some types of queries cannot be obtained through one retrieval and requires multiple iterative queries. The ability to retrieve information with multiple iterations is known in the literature as multi-step retrieval [9]. In multi-step retrieval, evidence may need to be obtained with additional information from a previous search, which might otherwise be interpreted as not being fully relevant to the question and no evidence could be found. We extend the capability of multi-step retrieval to fake news claim verification, querying relevant evidence passages in an iterative retrieval manner.

2.3 Natural Language Inference

Given a statement and selected evidence sentences, the task of NLI is to predict their relation labels y. The advent of large annotated datasets, such as SNLI [2], CreditAssess [34], FEVER [46], has facilitated the development of many different neural NLI models, facilitating model development for this task [30, 31]. The fact verification task related to natural language inference aims to classify a pair of claims and evidence extracted from Wikipedia into three categories: entailment, contradiction, or neutrality. NSMN [30] uses a connected system of three homogeneous neural semantic matching models that jointly perform document retrieval, sentence selection, and claim verification for fact extraction and verification. Soleimani et al. [44] retrieve and validate claims using a BERT [6] model. With the popularity of graph neural networks, graph-based models are also used for semantic reasoning. EVIN [27] proposes evidence reasoning network, which extracts core semantic conflicts of claims as evidence to explain verification results. Our work differs from prior research in that we focus on classifying news claims as true or fake on a comprehensive examination of relevant evidence.

3 PROBLEM STATEMENT

In this section, we first define the problem of fake news detection based on evidence retrieval enhancement. We transform fake news detection into the process of human beings verifying a piece of news. First, we read the news content and summarize the key information expressed in the news (content summary), then query the evidence in multiple steps based on the summary (multi-step retrieval), and finally infer true or fake of news (i.e., Natural Language Inference). So our problem is defined as follows: the input is only news text A, and then the news key statement C is obtained through the text summarization module. Retrieve relevant passages in Wikipedia through C to get $P = \{P_1, P_2, P_3, \ldots\}$, and then perform evidence extraction to obtain $E = \{e_1, e_2, e_3, \ldots\}$. The output is the predicted probability of news authenticity $\hat{y} = f(C, E)$, f is the natural language inference verification model, $y \in \{0, 1\}$, $y = 0$ means fake news, $y = 1$ means true news.

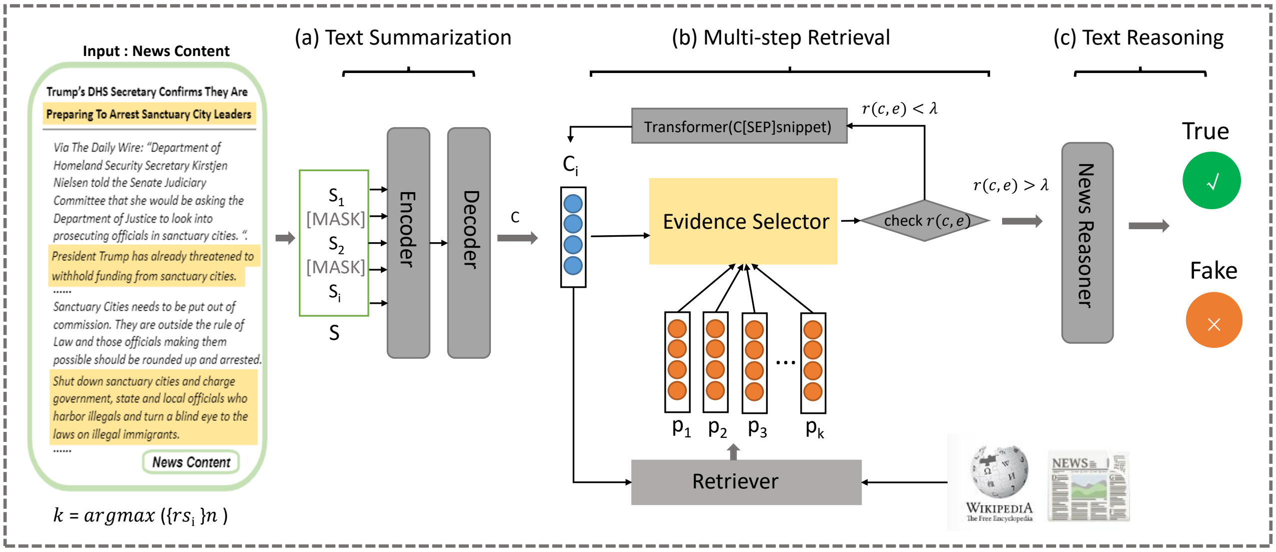

Figure 2: Our framework unfolds in three steps: (a) Summarization of the initial news text to obtain the key statement C, corresponding to human process of summarizing key information, (b) Evidence finding through multi-step retrieval, corresponding the human process of querying external relevant information based on the news claim. The retriever sends the first k paragraphs to the evidence selector, which evaluates whether the evidence meets the requirements. The correlation coefficient between C and evidence snippets is represented by $r(c, e)$, and a settable correlation score threshold, λ, is used to judge the quality of the evidence, and (c) The textual reasoner infers the consistency of evidence and claims, corresponding to the human process of judging news based on evidence.

4 THE PROPOSED MODEL

In this section, we propose a framework for fake news detection based on **MU**lti-**S**tep **E**vidence **R**etrieval augmentation(MUSER). Figure 2 illustrates the overall architecture of MUSER. Our model mainly consists of three modules:

Part 1: Text summarization module: Simulating the human behavior of reading news and summarizing key news information, the proposed module extracts the key information in the news and filters out the interference of redundant or unimportant information in the news.

Part 2: Multi-step retrieval module: Simulating the behavior of humans querying external relevant information in response to news statements, we incorporate a retrieval module into our model. To handle situations where the initially retrieved paragraph may not contain the answer, we adopt a multi-step iterative retrieval method. This process starts by generating a new query vector based on the key information and the current query vector. The retriever module then uses this new query vector for re-retrieval, enabling deeper exploration of relevant evidence.

Part 3: Text reasoning module: Simulating the behavior of humans to judge true or fake news based on the supplementary information queried, this module can extract semantic links between news claims and evidence, and then classify news into two categories: true news and fake news. Through the method of evidence retrieval enhancement, the interpretability of news true and fake judgment is improved, and the heavy labor of the method based on manual evidence extraction is avoided.

4.1 Text Summarization Module

A person tends to pay more attention to the main content expressed in the news when reading the news. For example, "More than 6 million Americans were infected with COVID-19 in January" is more worth checking than "The water is wet." In order to simulate the ability of humans to summarize news information, we first pre-train a text summarization module. The purpose of this module is to extract the key information in the news, and extract the statements worth checking. Although pre-trained language models, such as BERT [6] and UniLM [7], have achieved remarkable results in NLP scenarios, the word and subword mask language models used in the models may not be suitable for generative text summarization tasks. The reason is that the summarization tasks requires a coarser-grained semantic understanding, such as sentence and paragraph semantic level understanding, for effective summary generation.

Inspired by the recent success in masking words and continuous spans, in this work we pre-train a transformer-based encoder-decoder model on a large text corpus for news summarization generation [55]. To leverage a large text corpus for pre-training, we design a sequence-to-sequence self-supervised objective without abstract summarization. We mask sentences from news text and generate an output sequence from the remaining sentences for extracting news summaries. To enhance the relevance of the generated summaries, we select sentences that are deemed important or central to the news.

A piece of news A contains multiple sentences, that is, $A = \{s_i\}_i^N$, where N is the number of sentences. We select the set S of m sentences with the highest scores based on importance. As a proxy for importance, we compute ROUGE1-F1 [21] between the sentence and the rest of the news.

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

$A \setminus \{S \cup s_i\}$ represents the remaining sentences, and S is initially an empty set. Then select important sentences according to the importance score rs_i:

$$k = argmax(\{rs_i\}_n) \tag{2}$$

$$S = S \cup s_k \tag{3}$$

The corresponding position of each selected sentence is replaced by a mask token [MASK] to inform the model. Making m selections, at the end we select the masked m sentences from the document and concatenate the sentences into a pseudo-summary. the module then generate an output sequence from the remaining sentences, producing the masked sentences. We pre-train the model on the open source news dataset [5] to achieve a better summary generation results. The Mask sentences ratio (MSR) which refers to the ratio of the number of selected gap sentences to the total number of sentences in the document, is an important hyperparameter, similar to the mask rate in other works [55]. A low MSR reduces the difficulty and computational efficiency of pre-training. On the other hand, masking a large number of sentences at high MSR loses the contextual information necessary for guidance generation. In our experiments, we found an MSR of 30% to effective.

4.2 Multi-step Retrieval Module

The purpose of this module is to perform retrieval enhancement based on the key information in the news extracted in the previous step, which is similar to humans looking up data, and finding supplementary information to assist in the identification of true and fake news. Single-step retrieval may lead to insufficient auxiliary information retrieved. Therefore, we adopt a multi-step iterative retrieval method to improve information sufficiency [9]. Through iterative retrieval and supplementation, relevant information can be extracted more comprehensively, so as to better assist in judging the authenticity of news. When implementing this module, it is important to consider how to effectively extract the retrieved key information and how to maintain the sufficiency of information during the multi-step iterative retrieval process.

The multi-step retrieval problem we attempt to address is divided into three steps. In the first step, the news statement C is used to retrieve the relevant paragraph P from the Wikipedia corpus. The second step is to extract evidence from the retrieved long paragraphs and extract the key evidence of the paragraphs. Finally, in the case where no evidence is found in the retrieved paragraphs, the information retrieved this step is fused with statement C to generate a new statement for the retrieval iteration. The search terminates when evidence is found in the retrieved passages.

Paragraphs retrieval: Paragraphs retrieval is the selection of Paragraphs on Wikipedia that are relevant to a given statement. The paragraph retrieval module is based on BERT [6] and creates dense vectors for paragraphs by computing their average token embedding. The relevance of paragraph p to statement c is given by their dot product:

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

$\varphi(\cdot)$ is an embedding function used to map paragraphs and statements to a dense vector. Dot product search can use the approximate nearest neighbor index implemented by the FAISS library to improve search efficiency [16]. For the embedding function $\varphi(\cdot)$, we use the average token embedding of the BERT-base language model, which has been fine-tuned on several tasks:

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

where $BERT(p, i)$ is the embedding of the i-th token in paragraph p, and $|p|$ is the number of tokens in p.

Key evidence selection: Key evidence selection is to extract evidence-related key sentences from the retrieved relevant passages. Similar to paragraph retrieval, sentence selection can also be seen as performing semantic matching between each sentence in a paragraph and a statement query to select the most plausible evidence interval. Since the search space has been reduced to a controllable size via the paragraph retrieval in the previous step, we can directly traverse all relevant paragraphs to find key evidence. In this paper, we employ two approaches for key evidence selection: a relevance score-based approach and a context-aware approach.

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 $P = \{s_1, s_2, \ldots, s_k\}$ whose relevance score $r(c, s_i)$ is greater than a certain threshold λ set experimentally.

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 [23], minimizing the cross-entropy loss:

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

where N is the number of examples in the training batch, l_i is the number of non-padding tokens of the i-th example, and $p\theta(y_i^j)$ is the estimated softmax probability of the correct label for the j-th token of the i-th example. We train this model on Factual-NLI [40] with batch size 64, Adam optimizer and initial learning rate 5×10^{-5} until convergence.

Multi-step retrieval: In the process of selecting key evidence, we assess the sufficiency of the evidence's relevance using a threshold λ. When the evidence is insufficient, we use iterative retrieval to supplement information. To prioritize the most significant fragments in the paragraph, We rank the selected fragments based on their scores. Similar to human behavior of recursively querying external sources like Wikipedia step by step until the desired information is found, only the fragments with the highest scores will be

[5]http://atp-modelzoo-sh.oss-cn-shanghai.aliyuncs.com/release/tutorials/generation/en_train.tsv

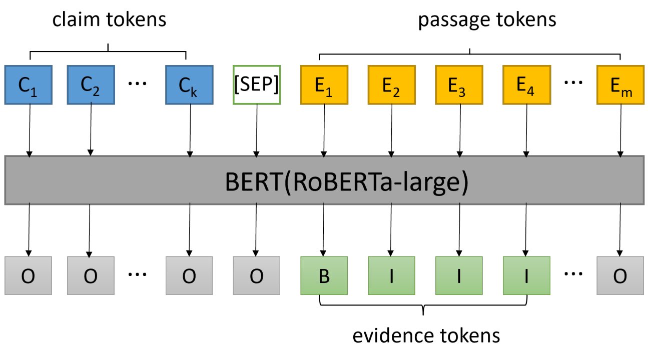

Figure 3: The context-aware sentence selection method uses a BERT-based sequence tagging model.

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

kept. The fragment with the highest score, referred to as the "winner," is then incorporated into the current query $[C[SEP]snippet]$. A reformulated query will be generated by combining the current query with current relevant paragraph information and updating it through a transformer.

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

The reformulated query is fed back to the retriever, which uses it to reformulate and rank the passages in the corpus. C_i fully interacts with the snippet through the transformer, avoiding information loss during the embedding process. The new query C_{i+1} is again subjected to paragraph retrieval and key evidence selection, achieving the effect of multi-step iterative retrieval. This multi-step iterative approach allows our model to combine the multi-step information needed to validate claims from multiple Wikipedia pages.

4.3 Text Reasoning Module

The last step of our model is to infer whether the news is true or false through multi-step retrieved evidence and news statements. This step aligns with human behavior, where individuals gather information from external sources and then evaluate the credibility of the news based on that information. Given a news claim C and relevant evidence E retrieved through multi-step retrieval process, our text reasoning module performs a logical inference from the evidence to the claim. The textual reasoning model acts as an evaluator to judge whether a statement is logically consistent with the retrieved evidence, thus identifying a pair of claims and related evidence as true or false. Thus, the training task of a text reasoning model can be considered as a binary classification task, where the goal is to minimize the binary cross-entropy loss function for each news item and its associated evidence. The cross-entropy loss is defined as follows:

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

N is the number of samples in the current batch, $y = 1$ means that claim C and evidence E are logically consistent, and $y = 0$ means that C and E are contradictory. V is a pre-trained language model that can perform discriminative classification tasks, such as BERT [6] , ALBERT [19] and RoBERTa [23]. In this work we choose BERT as the discriminator, we concatenate the claim C and the evidence E as the input of the discriminator, the input is [CLS] C [SEP] E [SEP], the batch size N is 64, Adam optimizer and an initial learning rate of 5×10^{-5} until convergence.

5 EXPERIMENTS

To verify the effectiveness of our proposed MUSER model, we conduct extensive experimental studies on three real-world datasets. Four research questions are addressed through comprehensive experimentation:

- RQ1: Is our MUSER model able to achieve improved fake news detection performance compared to previous fake news detection baseline methods?
- RQ2: How does the impact of the number of steps in multi-step retrieval on model performance?
- RQ3: How does each module of the model contribute to improved fake news detection performance?
- RQ4: Is the evidence retrieved by our model meaningful and explainable through multi-step retrieval?

5.1 Experimental Setup

5.1.1 Datasets. We conduct experiments on three real-world fake news detection datasets, including two English datasets (PolitiFact and GossipCop) and one Chinese dataset (Weibo). The English datasets PolitiFact and GossipCop are collected through FakeNewsNet [42]. The Weibo dataset is a hot news topic obtained through crawler tools [22]. Their key statistics are shown in Table 1.

PolitiFact: In this dataset, the news is divided into real news and fake news, taking into account journalists' and experts' reviews of political news on websites.

GossipCop: In this dataset, entertainment news with ratings are collected from various media.

Weibo: The data in this dataset are hot news topics from the Sina Weibo platform, and news is marked as rumors and non-rumors.

Above datasets contain labeled news content and associated social information. However, we study fake news detection in the early stage of news dissemination, thus only utilizing the news text with the social information excluded. This scenario resembles the situations where fake news detection must be performed before social information becomes available. Our focus here lies on curbing the spread of fake news at the early stage after its release.

5.1.2 Baselines. To verify the effectiveness of MUSER, we compare it with several existing methods, including content-based and evidence-based verification, as described below:

Content-based methods

- **TextCNN (EMNLP'14)** [17]: TextCNN combines convolutional neural networks and news content, which can automatically extracting text features through multiple convolutional hidden layers,
- **TextRNN (ACL'16)** [53]: TextRNN uses LSTM to encode the textual information in the last output of the recurrent neural network.
- **TCNNURG (IJCAI'18))** [38]: TCNNURG utilizes two convolutional neural networks and a conditional variational autoencoder for classification.
- **BERT (NAACL'19)** [6]: BERT uses the Transformer-based architecture to pre-train deep bidirectional representations of unlabeled text.

Evidence-based methods

- **DeClarE (EMNLP'18)** [35]: They use BiLSTM to embed the semantics of evidence and compute evidence scores through an attention interaction mechanism.
- **HAN (ACL'19)** [26]: HAN adopts GRU embedding and two modules of topic consistency and semantic entailment based on sentence-level attention mechanism to simulate claim-evidence interaction.
- **EHIAN (IJCAI'20)** [49]: EHIAN discusses the questionable parts of claims for interpretable claim verification through an evidence-aware hierarchical interactive attention network to explore more plausible evidence semantics.
- **MAC (ACL'21)** [47]: MAC combines multi-head word-level attention and multi-head document-level attention, which facilitates interpretation for fake news detection at both word-level and evidence-level.
- **GET (WWW'22)** [50]: GET models claims and evidences as graph-structured data to explore complex semantic structures and reduces information redundancy through the semantic structure refinement layer.

5.1.3 Implementation Details. In the fake news detection task, a binary classification is a common approach, and the commonly used evaluation indicators are F1, Precision, Recall, F1-Macro, and F1-Micro for model performance evaluation [50]. We randomly select 75% of the data as the training set, and the remaining 25% as the test set. We use the Adam optimizer with a learning rate of 5×10^{-5} for all three datasets. We set the number of training epochs to 20 for both our model and the baseline method. The hyperparameters for the baselines were set based on the corresponding papers, and key hyperparameters are carefully tuned for optimal performance (e.g., learning rate and embedding size). We conduct all experiments on Linux servers equipped with GeForce RTX 3080 GPUs (32GB memory each) using PyTorch 1.8.0. The implementation details can be found in the appendix and repository.

Table 2: Performance comparison of Our model w.r.t. baselines. We repeat the experiment 10 times, and calculate the average performance of the experimental results. "F1-Ma" and "F1-Mi" denote the metrics F1-Macro and F1-Micro, respectively. "-T" represents "True News as Positive" and "-F" denotes "Fake news as Positive" in computing the precision and recall values. The best results are highlighted in bold.

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**

Table 3: Performance comparison of on GossipCop.

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**

Table 4: Performance comparison of on Weibo.

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**

5.2 Performance Results (RQ1)

We compare our model, MUSER, to 9 baselines, including 4 content-based methods and 5 evidence-based methods. The results are reported in Tables 2, 3, and 4, and we have the following observations:

Firstly, it is worth noting that evidence-based methods tend to predict more correctly than content-only methods (i.e., the first four methods in the tables), indicating the extra value of incorporating additional evidential information, which can well make up for the insufficiency of news content features alone. The evidence-based methods rely on external evidence to verify the validity of the claims, reducing excessive reliance on textual schemas.

Secondly, in comparison to three recent evidence-based methods (GET, EHIAN, MAC), our proposed MUSER achieves superior results (MUSER > GET > EHIAN > MAC). In particular, MUSER improves the performance by 3% on F1-Macro and F1-Micro compared to the current SOTA baseline GET on the three datasets, which can better reflect the overall detection ability of the model. Furthermore, for more fine-grained evaluation, we computed "True news as Positive" and "Fake news as Positive" separately. MUSER also achieved the suprior results in F1, Precision, and Recall scores

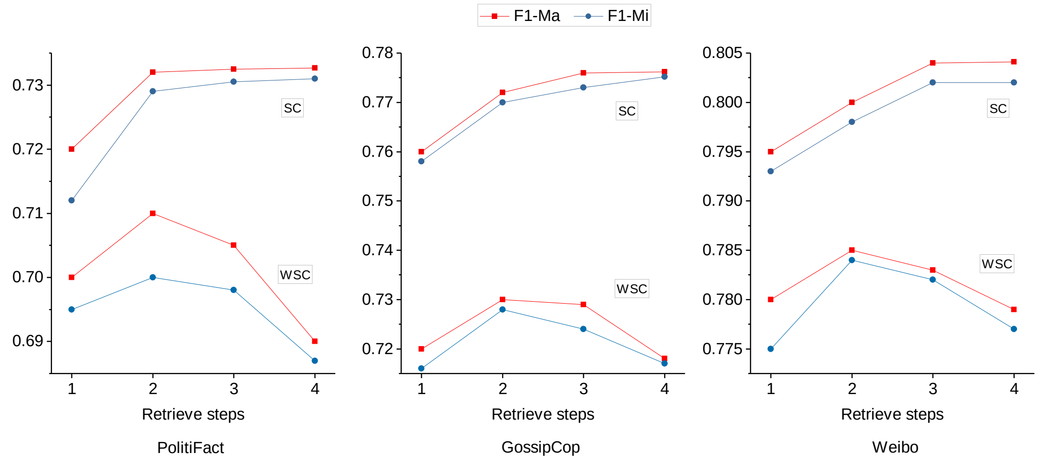

Figure 4: Results of retrieve step comparison study. The term SC (Step Control) means that the key evidence selection function is activated, while WSC (Without Step Control) means that the key evidence selection function is not included.

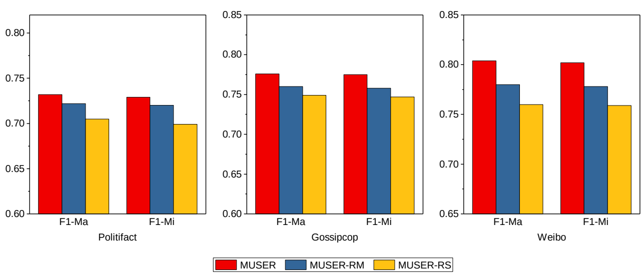

Figure 5: Results of ablation study. MUSER represents the complete model performance, MUSER- RM represents the removal of the multi-step retrieval module, and MUSER-RS represents the removal of the text summary module.

on the three datasets. Accuracy is equivalent to F1-Macro and thus omitted in the evaluation.

Finally, our results demonstrate that MUSER outperforms all baseline methods in fake news detection as positive detection metric. For instance, as far as GossipCop is concerned, the F1-False, Precision-False, and Recall-False values have been increased by 5%, 0.4%, and 11%, respectively. Similar obvious improvements can be observed on other datasets. These results show that our method can more accurately judge news labeled as "fake". MUSER conducts evidence retrieval and supplement through multi-step iterative retrieval, and can extract relevant information more comprehensively, so as to better assist the judgment of news authenticity.

5.3 Retrieve Steps Comparison (RQ2)

Next, we investigate the performance improvement of the number of retrieval steps in the multi-step retrieval module. The evaluation was conducted using the commonly used F1-Macro and F1-Micro scores on each dataset and results are presented in Figure 4. In order to examine the effectiveness of key evidence selection in the multi-step retrieval process, we remove it and use a fixed number of retrieval steps to conduct experiments, and then compare it with the model that has the key evidence selection function.

Firstly, we can find that in experiments where key evidence selection is not enabled, as the number of retrievals increases, the performance decreases instead. This is because there is no evidence screening for the retrieved paragraphs, which contains too much redundant information, which leads to a decrease in performance.

Secondly, we find that after the key evidence selection is enabled, the performance has been improved compared with that without the key evidence selection enabled. Since we will judge whether the current retrieval results contain key evidence in the key evidence selection stage, when key evidence is retrieved, our model will stop iterative retrieval to reduce the interference of redundant information. This shows that the strategy selection is to explore first, and the increase in the number of retrieval steps does not lead to an increase in redundant information. And even as the number of retrieval steps grows, the performance does not degrade, suggesting that if the retrieval module has already found good evidence, it will choose to utilize them instead of continuing to retrieve.

The key takeaway from this experiment is that multiple retrieval steps consistently improve performance compared to single-step retrieval. That is, even if relevant evidence passages are not retrieved in the initial step, the retriever will retrieve relevant passages in the subsequent iterative retrieval process. The performance peaks around 2 to 3 steps, and increasing the number of steps further does not provide much benefit and degrades the performance of the model. However, with the difficulty level of the datasets varies, the optimal number of steps remains consistent.

5.4 Ablation Study (RQ3)

In this part, in order to verify the effectiveness of each module, we conduct the comparative performance experiments in Figure 5. MUSER represents the complete model performance, MUSER-RS represents the removal of the text summary module, and MUSER-RM represents the removal of the multi-step retrieval module.

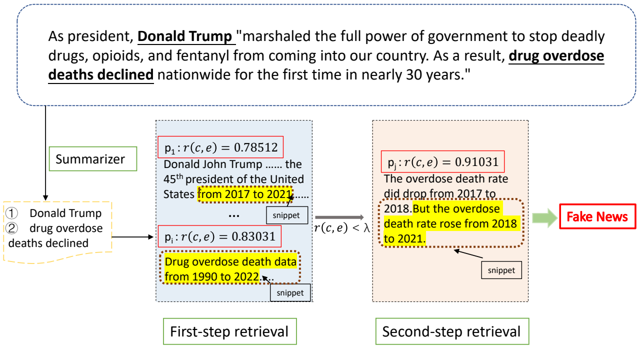

Figure 6: A verification example generated by MUSER in the Case study. The evidence correlation score $r(c, e)$ obtained by the first step of retrieval is smaller than the threshold λ we set. Then proceed to the second step of retrieval to obtain more sufficient evidence.

The results in Figure 5 show that MUSER has better performance than MUSER-RM, which proves our idea that through multi-step iterative retrieval for evidence retrieval and supplementation, relevant information can be extracted more comprehensively, so as to better news verification. And we find that the text summarization module also plays an important role, which also proved that extracting key statements in the news can reduce the interference of unimportant texts in the news, thereby achieving more accurate predictions. Furthermore, MUSER performs better than MUSER-RS and MUSER-RM, showing that removing any of them leads to model performance degradation, which demonstrates the effectiveness and relevance of our main components.

5.5 Explainability Study (RQ4)

5.5.1 Case Study. In this case study, we showcase the effectiveness of our model in enabling a deeper understanding of how the model works during multi-step retrieval. In particular, we demonstrate the capability of our model through an example of evaluating the authenticity of a news story about US President Donald Trump "marshaled the full power of government to stop deadly drugs, opioids, and fentanyl from coming into our country. As a result, drug overdose deaths declined nationwide for the first time in nearly 30 years." Our model, through key evidence extraction and multi-step search for supplementary evidence, successfully identify the news as fake. This case highlights the ability of MUSER to effectively evaluate the authenticity of news through a multi-step search to find critical evidence reasons.

Specifically, Figure 6 shows the steps in the verification processes by MUSER, starting with text summarization to extract key information from the news. The first step of retrieval is then performed, and relevant paragraph data is obtained from the corpus. Evidence extraction identifies information related to Donald Trump and data on drug overdose deaths in the United States. The calculated $r(c, e)$ from the key evidence selection is less than the preset limit value λ, indicating the need for another retrieval step. In the second step, the snippet information retrieved is carried forward and the statement "The overdose death rate did drop from 2017 to 2018. But the overdose death rate rose from 2018 to 2021." was obtained. Finally, the reasoning module judges the news to be fake. Evidence from multi-step retrieval makes it easier for users to understand the judgments made by the model on the authenticity of news.

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%**

5.5.2 User Study. In this user study, we aim to determine if real-world users are able to accurately assess the veracity of news articles based on the evidence retrieved by our proposed MUSER model. Specifically, we conduct a user study in which there are 60 news articles randomly selected from PolitiFact, GossipCop, and Weibo, including 10 fake and 10 real news articles from each dataset. We compare the evidence retrieved by MUSER with the evidence obtained by the GET model after refinement by semantic structure, and ask 8 participants to score the evidence. For each piece of news, we will give the relevant evidence of MUSER or GET, and then ask the participant to determine whether the news is true or fake according to the given evidence within three minutes, as well as give about her/his adjustment according to the 5-point Likert scale. In order to ensure fairness in our user study experiment, each participant is given the news articles to be judged in a randomized manner and participate in the experiment independently.

Table 5 shows the results of the experiments. By comparing the labels given by different participants, we find that the conclusions drawn by the participants have a high level consistency with the predicted labels produced by the MUSER model. This indicates that by observing the multi-step retrieval evidences generated by MUSER, human participants can much more accurately decide whether a news article is fake or not.

6 CONCLUSION

In this paper, we propose a framework for fake news detection based on multi-step evidence retrieval enhancement—MUSER. Our model leverages a three-phase methodology inspired by human verification processes, including summarization, retrieval and reasoning. Through text summarization, key information is extracted from the news, reducing irrelevant information. The multi-step retrieval phase enables evidence association for news verification, increasing the dependency between multiple pieces of evidence. Finally, the semantic connection between the news statement and the evidence is analyzed for news classification into two categories: true news and fake news. The results of our experiments on three real-world demonstrated the effectiveness of MUSER. Moreover, our results also show that evidence association via multi-step retrieval enhances the interpretability of the fake news detection task, making it easier for users to assess the credibility of information and form their own valid judgments.

7 ACKNOWLEDGMENTS

Thanks to Dr. Jianxun Lian and the anonymous reviewers for their valuable comments and constructive feedback. The authors acknowledge financial support from the National Natural Science Foundation of China (Grant Nos. 62276171 and 62072311), Shenzhen Fundamental Research-General Project (Grant Nos. JCYJ20190808162601658, 20220811155803001, 20210324094402008 and 20200814105901001), CCF-Baidu Open Fund (Grant No. OF2022028), and Swiftlet Fund Fintech funding. Kai Shu is supported by NSF SaTC-2241068. Kai Shu and Xing Xie are the corresponding authors.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36. https://doi.org/10.1257/jep.31.2.211

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 632–642. https://doi.org/10.18653/v1/D15-1075

[3] Meeyoung Cha, Wei Gao, and Cheng-Te Li. 2020. Detecting fake news in social media: an Asia-Pacific perspective. *Commun. ACM* 63, 4 (2020), 68–71. https://doi.org/10.1145/3378422

[4] Robert C. Coghill, John G. McHaffie, and Yi-Fen Yen. 2003. Neural correlates of interindividual differences in the subjective experience of pain. *Proceedings of the National Academy of Sciences* 100, 14 (2003), 8538–8542. https://doi.org/10.1073/pnas.1430684100

[5] Mansour Davoudi, Mohammad R. Moosavi, and Mohammad Hadi Sadreddini. 2022. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Systems with Applications* 198 (2022), 116635. https://doi.org/10.1016/j.eswa.2022.116635

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

[7] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 13063–13075. https://doi.org/10.5555/3454287.3455457

[8] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. KAN: Knowledge-aware Attention Network for Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 1 (2021), 81–89. https://doi.org/10.1609/aaai.v35i1.16080

[9] Yair Feldman and Ran El-Yaniv. 2019. Multi-Hop Paragraph Retrieval for Open-Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2296–2309. https://doi.org/10.18653/v1/P19-1222

[10] Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2021. Detecting fake news with capsule neural networks. *Applied Soft Computing* 101 (2021), 106991. https://doi.org/10.1016/j.asoc.2020.106991

[11] Andrew Gordon, Jonathan C.W. Brooks, Susanne Quadflieg, Ullrich K.H. Ecker, and Stephan Lewandowsky. 2017. Exploring the neural substrates of misinformation processing. *Neuropsychologia* 106 (2017), 216–224. https://doi.org/10.1016/j.neuropsychologia.2017.10.003

[12] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378. https://doi.org/10.1126/science.aau2706

[13] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. https://doi.org/10.1162/tacl_a_00454

[14] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*. 228–243. https://doi.org/10.1007/978-3-319-13734-6_16

[15] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5746–5754. https://doi.org/10.1609/aaai.v36i5.20517

[16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[17] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1746–1751. https://doi.org/10.3115/v1/D14-1181

[18] Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A Multi-Level Attention Model for Evidence-Based Fact Checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2447–2460. https://doi.org/10.18653/v1/2021.findings-acl.217

[19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=H1eA7AEtvS

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Article 793. https://doi.org/10.5555/3495724.3496517

[21] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. 74–81. https://aclanthology.org/W04-1013

[22] Qiang Liu, Feng Yu, Shu Wu, and Liang Wang. 2018. Mining Significant Microblogs for Misinformation Identification: An Attention-Based Approach. *ACM Trans. Intell. Syst. Technol.* 9, 5 (2018). https://doi.org/10.1145/3173458

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019).

[24] Yang Liu and Yi-Fang Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018), 254–261. https://doi.org/10.1609/aaai.v32i1.11268

[25] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 505–514. https://doi.org/10.18653/v1/2020.acl-main.48

[26] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2561–2571. https://doi.org/10.18653/v1/P19-1244

[27] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. 2561–2571. https://doi.org/10.18653/v1/P19-1244

[28] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. In *Proceedings of the ACM Web Conference 2022*. 1148–1158. https://doi.org/10.1145/3485447.3512163

[29] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1165–1174. https://doi.org/10.1145/3517214

[30] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. *in Proceedings of the AAAI Conference on Artificial Intelligence* 33 (07 2019), 6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859

[31] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2249–2255. https://doi.org/10.18653/v1/D16-1244

[32] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report. https://doi.org/10.13140/RG.2.2.23890.43205

[33] Kashyap Popat. 2017. Assessing the Credibility of Claims on the Web. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 735–739. https://doi.org/10.1145/3041021.3053379

[34] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2173–2178. https://doi.org/10.1145/2983323.2983661

[35] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 22–32. https://doi.org/10.18653/v1/D18-1003

[36] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 231–240. https://doi.org/10.18653/v1/P18-1022

[37] Piotr Przybyla. 2020. Capturing the Style of Fake News. *in Proceedings of the AAAI Conference on Artificial Intelligence* 34 (Apr. 2020), 490–497. https://doi.org/10.1609/aaai.v34i01.5386

[38] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence.* 3834–3840. https://doi.org/10.5555/3304222.3304302

[39] Victoria L. Rubin and Tatiana Lukoianova. 2015. Truth and Deception at the Rhetorical Structure Level. *J. Assoc. Inf. Sci. Technol.* 66, 5 (2015), 905–917. https://doi.org/10.1002/asi.23216

[40] Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2020. Latent Retrieval for Large-Scale Fact-Checking and Question Answering with NLI training. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI).* 941–948. https://doi.org/10.1109/ICTAI50040.2020.00147

[41] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. DEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 395–405. https://doi.org/10.1145/3292500.3330935

[42] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[43] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36. https://doi.org/10.1145/3137597.3137600

[44] Amir Soleimani, Christof Monz, and Marcel Worring. 2020. *BERT for Evidence Retrieval and Claim Verification.* 359–366. https://doi.org/10.1007/978-3-030-45442-5_45

[45] Chenguang Song, Kai Shu, and Bin Wu. 2021. Temporally evolving graph neural network for fake news detection. *Information Processing & Management* 58, 6 (2021), 102712. https://doi.org/10.1016/j.ipm.2021.102712

[46] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 809–819. https://doi.org/10.18653/v1/N18-1074

[47] Nguyen Vo and Kyumin Lee. 2021. Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* 965–975.

[48] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[49] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2020. Evidence-Aware Hierarchical Interactive Attention Networks for Explainable Claim Verification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20.* 1388–1394. https://doi.org/10.24963/ijcai.2020/193

[50] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-Aware Fake News Detection with Graph Neural Networks. In *Proceedings of the ACM Web Conference 2022.* 2501–2510. https://doi.org/10.1145/3485447.3512122

[51] Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement Subgraph Reasoning for Fake News Detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2253–2262. https://doi.org/10.1145/3534678.3539277

[52] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5644–5651.

[53] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 1480–1489. https://doi.org/10.18653/v1/N16-1174

[54] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-Modal Knowledge-Aware Event Memory Network for Social Media Rumor Detection. In *Proceedings of the 27th ACM International Conference on Multimedia.* 1942–1951. https://doi.org/10.1145/3343031.3350850

[55] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning.* 12 pages.

[56] Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* 57, 2 (2020), 102025. https://doi.org/10.1016/j.ipm.2019.03.004

[57] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 36 pages. https://doi.org/10.1145/3161603

A APPENDIX ON REPRODUCIBILITY

A.1 Experimental Environment

This experiment runs on GeForce RTX 3080 GPUs (32GB memory each) and CentOS 7 servers. The code is implemented with PyTorch 1.8.0.

A.2 Code Resources

We compare the proposed framework, MUSER, with 9 baseline methods discussed in Section 5.2, the content-based methods including TextCNN, TextRNN, TCNNURG , BERT, and the evidence-based methods including DeClarE, HAN, EHIAN, MAC and GET. The implementation details of our proposed framework, including code and settings, are available through the following link: https://anonymous.4open. science/r/MUSER. Other codes were obtained as follows:

- **TextCNN:** we use the publicly available implementation at: https://github.com/FinIoT/text_cnn
- **TextRNN:** we use the publicly available implementation at: https://github.com/luchi007/ RNN_Text_Classify
- **TCNNURG:** we use the publicly available implementation at: https://github.com/text_classify
- **BERT:** we use the publicly available implementation at: https://github.com/google-research/bert
- **DeClarE:** we use the publicly available implementation at: https://github.com/atulkumarin/DeClare
- **HAN:** we use the publicly available implementation at: https://github.com/majingCUHK/Claim_Verification
- **EHIAN:** we use the publicly available implementation at: https://github.com/evidence-inference
- **MAC:** we use the publicly available implementation at: https://github.com/nguyenvo09/EACL2021
- **GET:** we use the publicly available implementation at: https://github.com/CRIPAC-DIG/GET

A.3 Corpus processing

In this article, we use Wikipedia data as the retrieval corpus. The download address of Wikipedia Chinese corpus is: https://dumps.wikimedia.org/zhwiki/latest/, and the download address of Wikipedia English corpus is: https://dumps.wikimedia.org/enwiki/latest/.

We extract the Wikipedia corpus through WikiExtractor, which can extract the main article content of the corpus ending with .bz downloaded from Wikipedia. The download address of the tool is: https://github.com/attardi/wikiextractor.