World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:8, 2024

# Automated Fact-Checking By Incorporating Contextual Knowledge and Multi-Faceted Search

Wenbo Wang, Yi-fang Brook Wu

Open Science Index, Cognitive and Language Sciences Vol:18, No:8, 2024 publications.waset.org/10013763.pdf

*Abstract*—The spread of misinformation and disinformation has become a major concern, particularly with the rise of social media as a primary source of information for many people. As a means to address this phenomenon, automated fact-checking has emerged as a safeguard against the spread of misinformation and disinformation. Existing fact-checking approaches aim to determine whether a news claim is true or false, and they have achieved decent veracity prediction accuracy. However, the state of the art methods rely on manually verified external information to assist the checking model in making judgments, which requires significant human resources. This study presents a framework, SAC, which focuses on 1) augmenting the representation of a claim by incorporating additional context using general-purpose, comprehensive and authoritative data; 2) developing a search function to automatically select relevant, new and credible references; 3) focusing on the important parts of the representations of a claim and its reference that are most relevant to the fact-checking task. The experimental results demonstrate that: 1) Augmenting the representations of claims and references through the use of a knowledge base, combined with the multi-head attention technique, contributes to improved performance of fact-checking. 2) SAC with auto-selected references outperforms existing fact-checking approaches with manual selected references. Future directions of this study include I) exploring knowledge graph in Wikidata to dynamically augment the representations of claims and references without introducing too much noises; II) exploring semantic relations in claims and references to further enhance fact-checking.

*Keywords*—Fact checking, claim verification, Deep Learning, Natural Language Processing.

## I. Introduction

FACT-CHECKING has emerged as a safeguard against the spread of misinformation and disinformation. There are plenty of existing studies focusing on automated fact-checking. Some of them have achieved decent detection accuracy in their studies, and there has also been some exploration in terms of justification production. However, the existing studies are not without issues. Based on two recent review papers, the current fact-checking studies are facing the following challenges: leveraging multi-lingual resources [1, 2], ambiguity in the claims [1], system bias [1, 2], lack of contextual information [1], multimodality [1, 2], choice of labels [2], sources and subjectivity [2], faithfulness of justification [2], and from debunking to early intervention and prebunking [2].

Among the identified issues, this study attempts to address the following: *1) Ambiguity in the claims [1]:* a poorly worded claim might lead to wrong interpretations. For example, on Dec 26, 2020, Snopes.com published a fact-checking article titled “Did a Woman Get Fired After Donating a Kidney on Her Boss’ Behalf?" [3] Donating a kidney on others’ behalf sounds farfetched, and the claim is unclear about the reasons for the woman’s kidney donation. Further down in this article, the fact-checker does provide the official claim “a woman donated her kidney to save the life of her boss and was later fired during her recovery" which sounds more plausible and eliminates the confusing wording. This example shows that even a fact-checking article might be ambiguous and the importance to address it. *2) Lack of contextual information [1]:* not enough contextual information is incorporated in automated fact-checking models. For claims, most are short and some are too short to give enough context. For example, on June 19, 2023, Snopes.com published a claim “Did WEF call for an Al-Written Bible to create new religions?” [4] WEF is an abbreviation that requires more context. In such instances, having contextual knowledge is crucial to accurately frame the claim.

Additionally, most automated fact-checking frameworks rely on available claims and references as training data [5–10]; we identify two other major issues in such a framework that we attempt to address. They are: *3) Labor intensive evidence generation:* the process of collecting references for claims often requires human involvement and domain-specific knowledge [5]. For example, one public fact-checking dataset PUBHEALTH is collected from various fact-checking websites, and those sites are maintained by experts [5]. Specifically, the claims in the dataset are collected from several fact-checking websites such as Snopes, Politifact, etc. Each claim undergoes a fact-checking process and is accompanied by a corresponding veracity label and relevant references, all curated manually [5]. *4) The handling of a) evolving facts for a claim and b) new claims:* for a claim that is evolving, such as whether there is a recession in the United States in 2023, without a most up-to-date reference, these models might not be able to effectively verify this claim. On the other hand, for new claims that are very different from readily available training data (claims and references), the fact-checking results are likely to be ineffective.

This study proposes a fact-checking framework called *SAC* (stands for Searching, Augmenting and Checking) to address the above four major issues. SAC focuses on: *I) augmenting the representations of claims to address issues of ambiguity and the lack of contextual information.* SAC incorporates a knowledge base (KB) Wikidata [11] to provide more contextual and semantic information to enhance the representation of both the claim and its reference. *II) Searching and selecting relevant, new and credible references to address issues of labor intensiveness and evolving/new claims.* We

Wenbo Wang and Yi-fang Brook Wu are with the Department of Informatics, New Jersey Institute of Technology, Newark, NJ, 07102 USA (e-mail: ww6@njit.edu, wu@njit.edu).

design a multi-faceted reference search method to find such references via a search engine, where the following indicators are considered: ranking of a reference in the search results, the number of days since it was published, and its PageRank [12] score. The importance of relevance is obvious. The number of days since publication is useful, especially for facts or claims that are evolving, such as COVID-19's long term effect on health [13]. PageRank can help us quickly obtain the 'importance' of a reference's source domain [14]. In this study, it is assumed that an important page, indicated by a higher page rank value, is more likely to be credible. *III) Achieving the previous two goals, while at least maintaining the same fact-checking effectiveness.* Additionally, our framework incorporates a multi-head attention (MHA) [15] mechanism. It allows the model to attend to different parts of the input sequence simultaneously and learn different representations [15], resulting in better capture of the complex relationships between the claim and its reference.

Our main contributions can be summarized as follows:

- With Information Augmentation module, the representations of both the claims and their corresponding references are augmented with the contextual information from Wikidata to address the issues of ambiguity and lack of context in the claims.
- The incorporation of an MHA-based neural network to capture the complex relationships between the claims and the references, where the focus is centered on parts of a claim and its reference which are more relevant to the fact-checking task.
- The proposed Multi-Faceted Search module, which considers relevance, recency and credibility of references, helps reduce human involvement in curating fact-checking datasets.

## II. Related Work

### A. Automated Fact-Checking Approaches

*Fact-checking approaches without references:* In earlier research on automated fact-checking, the focus was solely on the claim itself without utilizing any references. For example, Rashkin et al. [16] present a linguistic analysis on the claims only. The study identifies linguistic features and analyzes these features' contributions to understanding the differences between true and false claims. This approach is limited by the existing knowledge in the prediction model and could lead to a lack of sufficient background information to verify new or evolving claims.

*Fact-checking approaches with references from curated datasets:* To enhance the accuracy of fact-checking models, some studies use curated datasets which commonly include claims, corresponding references, and veracity labels [5, 17, 18]. These datasets are mostly collected from fact-checking websites such as Snopes, Politifact, etc. Although the methods in these studies share the same inputs, the designs are different. For example, the Hierarchical Attention Networks (HAN) proposed by Ma et al. [17] and the EVidence Inference Networks (EVIN) proposed by Wu et al. [18] both employ attention mechanisms to capture important evidence, while the methods (i.e. BERT, SciBERT and BioBERT) proposed by Kotonya et al. [5] use Sentence-BERT to select evidence sentences. Due to the use of manually curated datasets, these methods achieve decent performances. But they require human involvement and expert knowledge for collection. Furthermore, these approaches do not take into account the situation where claims do not have accompanying references, making them unable to handle new/evolving claims.

*Fact-checking approaches with references from the web:* Other studies extract relevant references from the web [19–23]. For example, in Augenstein et al.'s study [20], they conduct a crawl of all active fact-checking websites listed by Duke Reporters' Lab and the Fact Checking Wikipedia page, retrieving a total of 43,837 claims along with their corresponding metadata. To retrieve evidence pages, the authors use Google Search API. Each claim's text is directly used as a query to the API, without any modification. The retrieval process involves fetching the top 10 search results with the highest ranking for each claim. However, these references may contain unreliable content, requiring further filtering of these references.

### B. Augmenting the Representation of Claims and References

In information retrieval, automatic query expansion is used to enhance the representation of user queries with the goal of improving recall [24], e.g.: a query for "laptop computer" might be augmented with "notebook computer." This approach is especially useful, if a query does not yield many relevant results due to mismatched vocabulary.

Although claims are not exactly user queries, they can suffer the same issues: ambiguity and incompleteness. To enhance their representation, additional context to augment the representation of claims and references might be beneficial. WordNet [25] was a popular choice as a general purpose knowledge source. However, its most recent stable release was June 2011, making it impractical for claim representation. Wikidata [11], on the other hand, is an open project that relies on global community collaboration to enhance and maintain its knowledge base. This open and community-driven approach ensures rapid and comprehensive updates and improvements to the data, and it is kept up to date.

### C. Web Page Importance Algorithms

To reduce human involvement and handle the new/evolving claims, this study uses references from online search. To assess the credibility of a website, we consider several webpage importance algorithms, and most of them have varies issues. CheiRank [26] is based on outgoing links, which is not suitable because outgoing links can be manipulated. HITS [27] algorithm is query-dependent. TrustRank [28] is based on incoming links and requires a seed set, which has to be manually prepared, and it is a modification of PageRank [14]. Our preference is PageRank, which is based on incoming links and does not require manually prepared seed set.

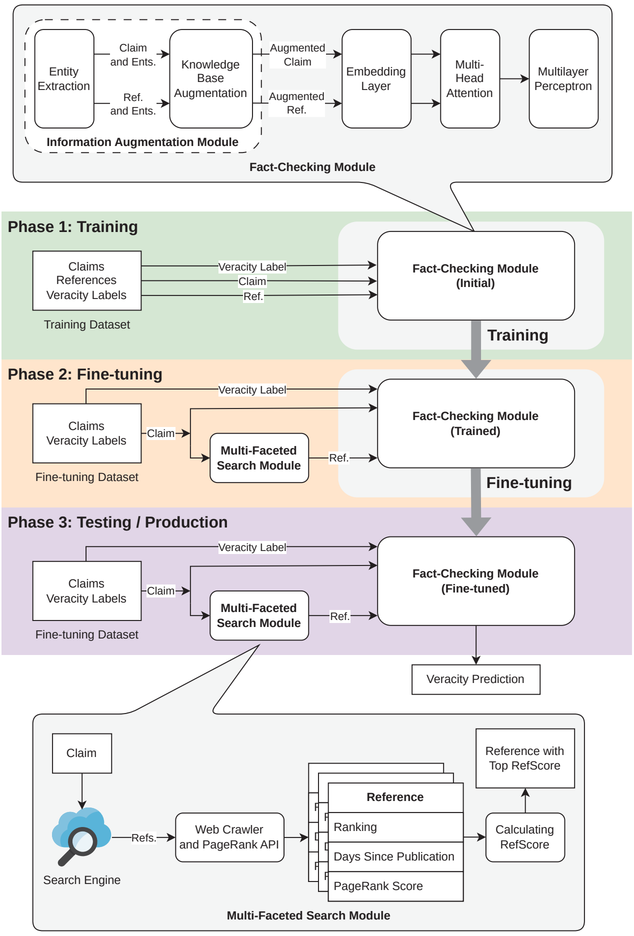

Fig. 1 Overview of SAC

## III. Framework

In this section we describe in detail the modules of our fact checking framework SAC.

As shown in Fig. 1, there are three phases in implementing SAC. *Phase 1*: Training. The fact-checking module (including Information Augmentation Module and MHA-based neural network) is trained using the claims and references in the curated dataset. *Phase 2*: Fine-tuning. The fact-checking module is fine-tuned using the references obtained using our Multi-Faceted Search Module. The purpose of such fine-tuning is to make the model adaptable to the references from the web. *Phase 3*: Testing/Production. The whole framework of SAC is tested using the claims only. Next, this section will introduce the key modules of SAC: Information Augmentation Module, Fact-Checking Module and Multi-Faceted Search Module.

### A. Fact-Checking Module

Fact-Checking module is designed to perform the verdict prediction task. Below is the problem definition of the task.

**Problem definition.** In automatic fact checking we are provided with a dataset of $D = \{(c_1, r_1, y_1), ..., (c_n, r_n, y_n)\}$, where $c_i$ corresponds to a textual claim, $r_i$ is reference, and $y_i$ is the associated veracity label to be predicted based on the claim and reference. Our target is to learn a function $f(y|c, r; \theta)$ to predict the veracity label of a claim, where $\theta$ represents all parameters of the model.

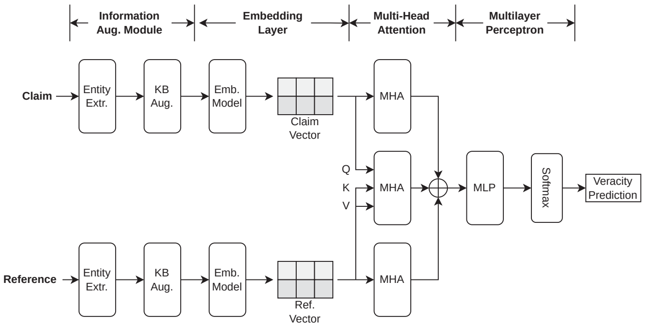

Fig. 2 Architecture of Fact-Checking Module

Fig. 2 shows the architecture of our Fact-Checking Module. The module takes claim $c$ and related reference $r$ as input, and the fact-checking results $y$ as output. It consists of the Information Augmentation Module, Embedding Layer, Multi-Head Attention Layer and Multilayer Perceptron Layer.

*1) Information Augmentation Module:* Information Augmentation module is utilized to provide additional information pertaining to the claim or reference. The first step is to extract the entities from the text of a given claim or reference. Then a KB is used to provide the information of the entities, i.e. descriptions. Finally, the entity and its description are concatenated together as the augmented information appended to the claim. For example, the claim "On March 21, 2023, Donald Trump was arrested" can be augmented as "On March 21, 2023, Donald Trump was arrested. Donald Trump : President of the United States from 2017 to 2021."

*2) Embedding Layer:* The Embedding Layer uses an embedding (Emb.) model to transform the augmented claim/reference into a fixed-length vector. Given the text of the claim or reference, the first step is to tokenize the text into subword units using WordPiece tokenization.

Then, a pre-trained BERT [29] model is utilized to extract embeddings for each token. The model is a deep neural network with 12 hidden layers. Following the common practice, the output is extracted from the last hidden layer, which will serve as the embeddings for each token. As a result, a sentence or text can be represented as a vector $X \in \mathbb{R}^{n*d_{input}}$, where $n$ represents the number of tokens and $d_{input}$ represents the dimension. Because of the nature of the neural network model, it can only handle fixed-length inputs and outputs. Therefore, $n$ is set as the maximum value among the token counts in all claims and references. To avoid unnecessary computations, we employ a masking technique to pad the texts with zeros for tokens with a count less than $n$. Given the augmented claim $c$ and augmented reference $r$, their embeddings can be represented as $X_{claim} \in \mathbb{R}^{n*d_{input}}$ and $X_{ref} \in \mathbb{R}^{n*d_{input}}$ respectively.

*3) Multi-Head Attention Layer:* Multi-head attention is favorable due to its capability to simultaneously focus on information from various representation subspaces and positions. The multi-head attention mechanism utilizes scaled

Open Science Index, Cognitive and Language Sciences Vol:18, No:8, 2024 publications.waset.org/10013763.pdf

dot-product attention, where it performs operations on a query $Q$, a key $K$, and a value $V$:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where $d_k = d_{input}/h$ is the key dimensionality, and $h$ is the number of heads. The multi-head attention mechanism acquires $h$ distinct representations of $(Q, K, V)$ - one for each head. It then calculates scaled dot-product attention for each representation, concatenates the outcomes, and passes the concatenation through a feed-forward layer. This process can be expressed using the following formulas:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \quad (2)$$

$$MHA(Q, K, V) = Concat(head_1, ..., head_n)W^O, \quad (3)$$

where $\{W_i^Q, W_i^K, W_i^V\} \in \mathbb{R}^{d_{model} \times d_v}, W^O \in \mathbb{R}^{hd_v \times d_{model}}$ are parameter matrices. Then three MHA components are deployed: one MHA takes $X_{claim}$ as input and is responsible for capturing the differences between the true claims and false claims; another MHA takes $X_{ref}$ as input and focuses on capturing crucial clues within the reference; the last MHA takes both $X_{claim}$ and $X_{ref}$ as input and aims to capturing inconsistencies between a claim and its reference. Below are the formulas for the three MHA components, and their respective outputs are denoted as $Y_{claim}$, $Y_{ref}$, and $Y_{diff}$, $\{Y_{claim}, Y_{ref}, Y_{diff}\} \in \mathbb{R}^{n \times d_{input}}$.

$$Y_{claim} = MHA(X_{claim}, X_{claim}, X_{claim}) \quad (4)$$

$$Y_{ref} = MHA(X_{ref}, X_{ref}, X_{ref}) \quad (5)$$

$$Y_{diff} = MHA(X_{claim}, X_{ref}, X_{ref}) \quad (6)$$

*4) Multi-Layer Perceptron Layer:* The purpose of MLP is to learn the probability distributions of the claim being True, False, Mixture or Unproven. It provides a nonlinear mapping between the representations of claims and references and the prediction vectors. It consists of three layers: one input layer, one hidden layer and one output layer. The computation of each layer is as follows:

$$v_i = tanh(w_i v_{i-1} + b_i), \quad (7)$$

$$tanh(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}}, \quad (8)$$

where $w_i$ and $b_i$ are weights and bias in the $i$th layer, and $tanh$ is the activation function. MLP takes the concatenation of the last output of $Y_{claim}, Y_{ref}$ and $Y_{diff}$ as the input. The results of verdict prediction can be presented as a probability distribution $P \in \mathbb{R}^l$, where $l$ is the number of veracity labels. The veracity label with the highest probability will be the final veracity prediction result. Below is the formula:

$$Y = Y_{claim} \oplus Y_{diff} \oplus Y_{ref} \quad (9)$$

$$P = Softmax(MLP(Y)) \quad (10)$$

### B. Multi-Faceted Search Module

Multi-Faceted Search module is used to search relevant, latest and credible reference given a claim. It considers the following three indicators: the ranking $S_{SR} \in [1, k]$ ($k$ is the number of top search results and is set to 10 in our experiments) of a reference, the number of days $D$ since it was published, and its PageRank score $S_{PR} \in [0, 10]$. We designed a formula to quantify the above three indicators into a single numerical value $S_{ref}$, called *RefScore*. To prevent any single indicator from exerting an excessive influence on $S_{ref}$, we also introduced additional weights to adjust the impact of each indicator. This approach allows us to fine-tune the importance of various factors and achieve a more balanced and accurate search process. Below is the formula:

$$S_{ref} = w_1 * S_{PR}/10 + w_2 * e^{-D/m} + w_3 * e^{-S_{SR}}, \quad (11)$$

where $w_1, w_2, w_3$ and $m$ are weights used to adjust the influence of each indicator on $S_{ref}$; $w_1, w_2$ and $w_3$ are floating numbers, and the sum of them is 1; $m$ is a positive integer. By default, $w_1 = 0.4$, $w_2 = 0.2$, $w_3 = 0.4$ and $m = 30$. The higher the $S_{ref}$ value of a reference, the more relevant, newer, and more credible it is. Note that we filtered out posts from social media sites from the top search results in our analysis due to: 1) PageRank is calculated for a website domain; 2) social media platforms have a large number of users whose posts inadvertently share the same PageRank score, which is not aligned with the design of our study. Also, search results from fact-checking websites such as snopes.com are removed to prevent our model from being overfitted.

## IV. Experiments

This section presents the experiments to evaluate the effectiveness of the proposed SAC framework.

### A. Datasets and Evaluation Metrics

Three datasets are used in the experiments. 1) Snopes and 2) PolitiFact, released by [9], are two widely used datasets, each containing 4341 and 3568 claims, along with relevant references collected from various websites. Note that Snopes has two veracity labels: *True* and *False*, while PolitiFact has six: *True*, *Mostly-true, Half-true, Mostly-false, False,* and *Pants-on-fire*. Following [17], we label *Mostly-true, Half-true* and *Mostly-false* as *Mixture*, and treat *False* and *Pants-on-fire* as *False*. 3) PUBHEALTH, released by [5], is a comprehensive dataset for automated fact-checking of public health claims. The dataset consists of 11,832 samples, each containing a claim, its corresponding veracity label (*True, False, Mixture,* and *Unproven*), and manually collected references. The distribution of veracity labels in these datasets are shown in Table I.

To evaluate the performance of fact-checking approaches, the following metrics are used: Accuracy, Precision, Recall, and F1 score.

TABLE I
DISTRIBUTION OF VERACITY LABELS IN THE DATASETS

| Dataset | True | False | Mixture | Unproven | Total |
|---|---|---|---|---|---|
| Snopes | 1,164 | 3,177 | / | / | 4,341 |
| PolitiFact | 520 | 1,114 | 1,934 | / | 3,568 |
| PubHealth | 6,176 | 3,570 | 1,526 | 299 | 11,832 |

### B. Baselines

The following five approaches are selected as the baselines: 1) *BERT, 2) SciBERT, and 3) BioBERT v1.1*, used in the study of Kotonya et al. [5], focus on using external references for the fact-checking task. The authors use Sentence-BERT [30] to encode contextualized representations for each of the reference sentences and then rank these sentences according to their cosine similarity with respect to the representation of the claim sentence. The top $k$($k$=5) sentences are selected as the input. Then they use BERT [29], SciBERT [31], BioBERT v1.1 [32] as the embedding layer, and a Softmax layer as the classifier to output the fact-checking results.

4) *X-Fact* [21] utilizes Google to obtain references related to claims and uses the top five snippets (along with metadata) from search results as evidence. The authors then designed an attention-based evidence aggregation model for fact-checking.

5) *EVIN* [18], includes a co-interactive shared layer and an evidence-aware coherence layer, is designed to capture the core semantic segments of claims and references and construct the conflicts between them. Then, the conflicts are used as evidence for fact-checking.

### C. Implementation Details

All modules of SAC are implemented using Python and its libraries, such as the deep learning library PyTorch, NLP library SpaCy, etc.

*Fact-Checking Module:* The embedding model used is BERT-base-uncased from Hugging Face [33]. The loss function used in the module is Cross Entropy [34]. The hyperparameters are automatically selected by an AutoML toolkit called Neural Network Intelligence (NNI) from Microsoft Research [35] to achieve the highest F1-score for SAC in the verdict prediction task.

*Information Augmentation Module* Information Augmentation module utilizes the Python-based package Spacy [36] to extract entities contained within the text. For each entity, it uses the Wikidata API [37] to retrieve the description of each entity from the knowledge base Wikidata [11].

*Multi-faceted Search Module:* Google Search API is used to search the references for a given claim. Since Google does not publish PageRank scores any more, we use Open Page Rank [38] to retrieve the PageRank score of each web page's domain. Readability(a Python library) is used to extract main text from a web page source code. AutoML [35] is used to automatically select the best weights of multi-faceted search formula, which can help SAC retrieve the highest F1 score. The values of $w_1, w_2, w_3, m$ that selected are 0.4, 0.2, 0.4 and 30 respectively.

For baselines BERT, SciBERT, BioBERT v1.1, and EVIN, the authors do not publish the code, we reproduce these approaches based on the details in their papers.

## V. RESULTS

In this section, we conduct several experiments to evaluate SAC and demonstrate the results.

### A. Experiments on Fact-Checking Module

To evaluate our Fact-Checking Module, we compare it to baselines on different datasets. On PubHealth dataset, following [5], we split claims as follows: 9,466 samples for training, 1,183 samples for validation and 1,183 samples for testing. On Snopes and PolitiFact datasets, following [18], we hold out 10% of the claims in the two datasets as development set for tuning the hyper-parameters, and conduct 5-fold cross-validation on the rest of the claims.

To evaluate the effectiveness of Information Augmentation module and MHA component, we also conduct ablation analysis on them. We first remove the Multi-Faceted Search module of SAC, namely SAC (w/o search). Then we set three different simplified versions of SAC (w/o search): 1) *SAC (w/o search) - Aug.* denotes that SAC (w/o search) removes Information Augmentation module; 2) *SAC (w/o search) - MHA* denotes that SAC (w/o search) removes MHA component; 3) *SAC (w/o search)* denotes that SAC (w/o search) use both Information Augmentation module and MHA component.

TABLE II
FACT CHECKING RESULTS ON PUBHEALTH DATASET

| Approach | Pre. | Rec. | F1 | Acc. |
|---|---|---|---|---|
| BERT | 0.77 | 0.55 | 0.64 | 0.66 |
| SciBERT | 0.76 | 0.66 | 0.71 | 0.70 |
| BioBERT | 0.75 | 0.62 | 0.67 | 0.69 |
| X-FACT | 0.74 | 0.71 | 0.72 | 0.71 |
| EVIN | 0.80 | 0.79 | 0.75 | 0.79 |
| SAC(w/o search) - Aug. | 0.80 | 0.81 | 0.80 | 0.81 |
| SAC(w/o search) - MHA | 0.80 | 0.80 | 0.80 | 0.80 |
| SAC(w/o search) | 0.86 | 0.84 | 0.83 | 0.84 |

Tables II and III present the experimental results. As shown in Table II, all three versions of SAC achieve better or same scores than the baselines in all evaluation metrics. The results indicate that both Information Augmentation and MHA individually can improve performance compare to the baselines. Moreover, when combined, they enhanced performance further. The incorporation of KB augmented the representations of the entities in claims and references effectively. Because of the more contextual information of an entity, the problem of ambiguity is alleviated. Additionally, MHA technique captures complex relationships within claims and references. The incorporation of these two modules results in an improvement in performance.

In Table III, we report Macro Precision, Recall, F1 Score, and Accuracy, as well as Precision, Recall, and F1 Score for the 'False' veracity label. SAC achieves the highest scores in all 'Macro' metrics. We observe that on the Snopes dataset, SAC and the baselines achieve higher scores under the 'False' label compared to 'Macro.' However, on the PolitiFact dataset, such a trend is not observed. This may be because neural networks, during the training process, attempt to minimize the overall loss function, and the loss associated with the majority class

TABLE III
FACT CHECKING RESULTS ON SNOPES AND POLITIFACT DATASETS

| Approach | Snopes | | | | | | | PolitiFact | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Macro | | | False | | | | Macro | | | False | | | |
| | Pre. | Rec. | F1 | Pre. | Rec. | F1 | Acc. | Pre. | Rec. | F1 | Pre. | Rec. | F1 | Acc. |
| BERT | 0.71 | 0.67 | 0.68 | 0.83 | 0.90 | 0.87 | 0.79 | 0.46 | 0.46 | 0.46 | 0.45 | **0.50** | 0.47 | 0.54 |
| SciBERT | 0.66 | 0.65 | 0.66 | 0.81 | 0.83 | 0.82 | 0.73 | 0.44 | 0.43 | 0.44 | 0.54 | 0.49 | **0.51** | 0.53 |
| BioBERT | 0.65 | 0.67 | 0.66 | **0.84** | 0.79 | 0.81 | 0.73 | 0.45 | 0.44 | 0.44 | 0.46 | 0.46 | 0.46 | 0.55 |
| X-FACT | 0.69 | 0.66 | 0.67 | 0.80 | 0.88 | 0.84 | 0.75 | 0.45 | 0.43 | 0.43 | 0.44 | 0.34 | 0.39 | 0.49 |
| EVIN | 0.77 | 0.78 | 0.76 | 0.76 | **0.94** | 0.84 | 0.75 | 0.51 | 0.52 | 0.49 | 0.40 | 0.38 | 0.39 | 0.52 |
| SAC (w/o search) | **0.79** | **0.81** | **0.80** | **0.84** | 0.92 | **0.88** | **0.81** | **0.58** | **0.59** | **0.54** | **0.64** | 0.32 | 0.43 | **0.59** |

contributes more significantly to the overall loss. Consequently, the model may tend to predict the majority veracity label, leading to a higher Accuracy. As shown in Table I, in the Snopes dataset, False claims are the most abundant, accounting for 73% of all claims, whereas in PolitiFact, False claims only make up 31%.

### B. Experiments on Multi-Faceted Search

To test the effectiveness of Multi-Faceted Search module, we conduct a fact-checking experiment on the references obtained using Multi-Faceted Search module. We employ the same dataset as in Experiment 1) and replace the references with those found using our Multi-Faceted Search module. Subsequently, we compare SAC with baselines in the verdict prediction task.

The references in the curated dataset, verified by experts, can effectively support or refute claims. However, references obtained automatically from the web by machines may include text that is not relevant to the claims, such as advertisements. It is difficult to completely filter out such content when automatically retrieving references. To enable fact-checking models to adapt to references obtained from the web, one solution is to fine-tune the models using these references.

TABLE IV
THE REFERENCE SOURCE FOR DIFFERENT EXPERIMENT SETTINGS.

| Settings | Training | Validation/Fine-tuning | Test |
|---|---|---|---|
| w/o Fine-tuning | Dataset | Dataset | MFS |
| w/ Fine-tuning | Dataset | MFS | MFS |

MFS stands for Multi-Faceted Search.

We design two experimental settings, as shown in Table IV, to validate the effectiveness of fine-tuning. The claims used in these two settings are identical to the experiment in Section V-A, with the only difference being the source of references for fine-tuning and testing phases. In the setting *w/o Fine-tuning*, SAC is trained and validated using manually collected references. While in the setting *w/ Fine-tuning*, SAC is trained using manually collected references but fine-tuned using the references obtained by Multi-Faceted Search. Finally, SAC under these two settings is tested using the references obtained by Multi-Faceted Search.

TABLE V
COMPARING SAC PERFORMANCE WITH AND WITHOUT FINE-TUNING

| Settings | Pre. | Rec. | F1 | Acc. |
|---|---|---|---|---|
| w/o Fine-tuning | 0.59 | 0.59 | 0.58 | 0.59 |
| w/ Fine-tuning | 0.77 | 0.77 | 0.77 | 0.77 |

Table V displays the experimental results of SAC under different settings. The results indicated that utilizing fine-tuning lead to improvements of 0.18 in Precision, Recall, and Accuracy, 0.19 in F1 score as compared to not using fine-tuning. Through fine-tuning, SAC learns how to mitigate the difference between the references from the web and the references from the curated dataset — resulting in higher performances.

Next, we conduct an ablation study of SAC using Multi-Faceted Search. We applied three versions of SAC: 1) *SAC - Aug.*: one reduced version without Information Augmentation; 2) *SAC - MHA*: another reduced version without MHA; and 3) *SAC*: the full version with Multi-Faceted Search, Information Augmentation and MHA.

TABLE VI
FACT-CHECKING RESULTS

| Approach | Pre. | Rec. | F1 | Acc. |
|---|---|---|---|---|
| BERT | 0.63(-0.14) | 0.64(+0.09) | 0.62(-0.02) | 0.64(-0.02) |
| SciBERT | 0.60(-0.16) | 0.62(-0.04) | 0.58(-0.13) | 0.62(-0.08) |
| BioBERT | 0.62(-0.13) | 0.63(+0.01) | 0.58(-0.09) | 0.63(-0.06) |
| X-Fact | 0.61(-0.13) | 0.64(-0.07) | 0.62(-0.10) | 0.70(-0.01) |
| EVIN | 0.72(-0.08) | 0.73(-0.06) | 0.69(-0.06) | 0.73(-0.06) |
| SAC - Aug. | 0.72(-0.08) | 0.70(-0.11) | 0.71(-0.09) | 0.70(-0.11) |
| SAC - MHA | 0.70(-0.10) | 0.68(-0.12) | 0.69(-0.11) | 0.68(-0.12) |
| SAC | 0.77(-0.09) | 0.77(-0.07) | 0.77(-0.06) | 0.77(-0.07) |

Table VI shows the results. The models are trained on the PUBHEALTH dataset, fine-tuned, and tested on the references retrieved using the Multi-Faceted Search module. The numbers in parentheses are the performance differences compared to Table II. From the table, all variations of SAC outperform the baselines under the same experimental settings. It is worth noting that compared to the performance of *SAC (w/o search)* in Table II, the performance of *SAC* in Table VI decreased. This is due to the only difference in this experiment, which is the use of automated searched references: in many cases, it returns useful results requiring subscriptions, making them impossible to be included as candidate references. In such a case, SAC moves on to the next reference with the next highest RefScore

When comparing the performance of *SAC* in Table VI with those of the baselines in Table II, *SAC* performs better than the baselines in all metrics except for BERT's Precision. Despite the decrease in performance caused when using automatically searched references, SAC still outperforms the baselines' original design, i.e. using manually curated references. SAC's automated search module still provides an alternative to partially alleviating human involvement in curating fact-checking datasets.

### *C. Experiments on Credibility*

The proposed Multi-Faceted Search in this study introduces PageRank as a metric to measure the credibility of a reference. To investigate the effect of PageRank in Multi-Faceted Search, we conduct an experiment by removing PageRank of our model. We employ two settings for selecting references: 1) *w/o Cred.*: selecting the top-1 reference in the search ranking (if the first reference is unavailable, proceeding to the next one sequentially until an available reference is encountered); 2) *w/ Cred.*: utilizing Multi-Faceted Search to retrieve the reference with the highest RefScore.

TABLE VII
FACT-CHECKING RESULTS OF SAC UNDER DIFFERENT REFERENCE SETTINGS

| Reference Settings | Fine-tuning | Pre. | Rec. | F1 | Acc. |
|---|---|---|---|---|---|
| w/o Cred. (top-1) | No | 0.58 | 0.57 | 0.55 | 0.57 |
| w/ Cred. (MFS) | No | 0.59 | 0.59 | 0.58 | 0.59 |
| w/o Cred. (top-1) | Yes | 0.74 | 0.69 | 0.70 | 0.69 |
| w/ Cred. (MFS) | Yes | 0.77 | 0.77 | 0.77 | 0.77 |

Table VII presents the experimental results. MFS stands for Multi-Faceted Search. Fine-tuning is Yes means that the model is fine-tuned with the corresponding searched references. When SAC is not fine-tuned, the results with credibility (*w/ Cred.*) are slightly better than those without credibility (*w/o Cred.*). When SAC undergoes fine-tuning and using credibility (*w/ Cred.*), there is a 0.03 increase in Precision and a 0.07 to 0.08 increase in Recall, F1 score, and Accuracy. These findings suggest that leveraging credibility can be beneficial for fact-checking, as search engines might adjust search result rankings for commercial purposes, potentially causing less credible web pages to rise to the top.

### *D. An Example of Fact-Checking on a Recent Claim*

To demonstrate how SAC performs on new emerging claims outside of the dataset, in this sub-section, we select a recent news claim from a fact-checking website and utilize Multi-Faceted Search module to find a relevant reference. Subsequently, we employ Fact-Checking module of SAC and baselines to fact-check the claim based on the reference. The models used are fine-tuned in Section V-B.

Table VIII shows an example of fake news that surfaced during the writing of this paper. The news claimed that *"On March 21, 2023, former U.S. President Donald Trump was arrested for his alleged involvement in hush-money payments made on his behalf"*. We discovered this claim on the fact-checking website snopes.com on March 21, 2023, the claim was labeled as 'false.' The reference found using multi-faceted search method was titled *"Barricades Go Up at Trump Tower, Manhattan Court as NYC Readies for Possible Protests"* and originated from nbcnewyork.com. In the fact-checking results, SAC and the other three baselines all correctly categorized this claim as 'false'. This demonstrates that utilizing automatically searched reference can enable fact-checking models to make correct judgments. The baseline models only correctly predicted the results, because they were provided with a reference discovered by our framework.

TABLE VIII
AN EXAMPLE OF FACT-CHECKING ON A RECENT CLAIM

| | | | | | | |
|---|---|---|---|---|---|---|
| Claim | On March 21, 2023, former U.S. President Donald Trump was arrested for his alleged involvement in hush-money payments made on his behalf. (False) | | | | | |
| Reference | Barricades Go Up at Trump Tower, Manhattan Court as NYC Readies for Possible Protests. www.nbcnewyork.com | | | | | |
| Results | BERT | SciBERT | BioBERT v1.1 | X-Fact | EVIN | SAC |
| | False | False | False | False | False | False |

## VI. LIMITATIONS AND DISCUSSION

### *A. Limitations*

With all its unique features, SAC has the following limitations: 1) Not all the top searched references are available due to web page formats and needing subscription. The retrieved web pages may contain file formats such as PDF that are not easily extractable automatically. This has a negative impact on the performance of our framework. 2) We discovered that in WikiData, the descriptions of entities are usually short. For example, Wikidata has lots of information about Donald Trump. However, what they choose as description for entity "Donald Trump" is just one sentence. This gives us a future direction on the need to further enhance the representation of entities. 3) The current SAC framework does not capture the semantic relations. Our framework solely focuses on the sequential structure of the text, ignores the semantic relations embedded within the text. The existing study has demonstrated that these semantic relations could be beneficial for claims [39].

### *B. Discussion*

*1) Credibility:* In this study, PageRank score is used as an alternative metric to measure the credibility of a reference. But it has a limitation where all web pages from the same domain share the same PageRank value. As a solution, the following search results are filtered out: I) Social media (e.g., Facebook) posts are filtered out due to having large numbers of users and posts of unequal credibility. Our choice in filtering out social media posts, though justified, might have a negative impact on our results. II) News articles from a news agency (e.g., CNN) share the same PageRank value, although they often have different levels of expertise and hence have different levels of credibility. The credibility score calculation is outside the scope of our study. However, the cases listed above might have a negative impact on our results.

*2) Multi-Faceted Search:* Automatic reference search can be used to assist fact checkers in quickly searching for references, saving them time. Despite being influenced by subscriptions or PDF files that affect automatic search for references, our multi-faceted search could be used to save the time on fact-checking and curating fact-checking datasets. It is worth further exploration in automatic reference search.

*3) Compare to the Baseline:* Although SAC performs similarly or slightly better compared to the best baselines like EVIN, SAC still has advantages. I) SAC has a simpler design. EVIN includes co-interactive layer, gate affine absorption, the gate G1, the gate G2, fine-grained conflict discovery layer, and evidence-aware coherence layer. SAC's Fact-Checking, on the other hand, is mainly composed of Information Augmentation and MHA. II) The authors of EVIN claim that their model is not suitable for early claim verification. The SAC's Multi-Faceted Search is able to automatically acquire the relevant, new and credible information. III) EVIN does not explicitly address the ambiguity issue and lack of context. In contrast, SAC's Information Augmentation module can provide contextual information to address it.

*4) System bias::* This is an issue present in existing fact-checking systems. It refers to these systems using datasets curated by a small group of people and often annotated by non-experts [1]. To minimize the system bias, SAC uses Multi-Faceted Search to automatically select relevant, new and credible references, without human intervention.

## VII. Conclusion and Future Directions

This study aims to achieve fully automated fact-checking by replacing a significant amount of manual effort with an automated search for references. The study presents a fact-checking framework called SAC. While utilizing the multi-faceted reference search method to retrieve reference, we enhance our fact-checking performance by incorporating a knowledge base to augment contextual information and a multi-head attention mechanism to capture representations of the claim and reference. The experimental results demonstrate that SAC with auto-selected references can retrieve decent performances. We will explore the following future directions.

First, we will further improve the augmentation of contextual information on claims and references beyond simply using an entity's corresponding description in its entirety in Wikidata. One direction is to utilize the knowledge graph in Wikidata and only selectively augment contextual information from nodes and paths in the knowledge graph that are closely related to the claim and description. Second, the automated reference selection function can be expanded to be a standalone module; when curating a fact-checking dataset, this module can be slightly re-designed to provide a short-listed relevant, new and credible references that support the verification results to minimize human experts' effort.

## References

[1] Preslav Nakov et al. "Automated Fact-Checking for Assisting Human Fact-Checkers". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Survey Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4551–4558. DOI: 10.24963/ijcai.2021/619. URL: https://doi.org/10.24963/ijcai.2021/619.

[2] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. "A survey on automated fact-checking". In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 178–206.

[3] *Did a woman get fired after donating a kidney on her boss' behalf?* https://www.snopes.com/fact-check/fired-kidney-donor/. July 2023.

[4] *Did WEF call for an AI-written bible to create new religions?* https://www.snopes.com/fact-check/wef-rewrite-bible. June 2023.

[5] Neema Kotonya and Francesca Toni. "Explainable Automated Fact-Checking for Public Health Claims". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7740–7754. DOI: 10.18653/v1/2020.emnlp-main.623. URL: https://aclanthology.org/2020.emnlp-main.623.

[6] Lianwei Wu et al. "DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1024–1035. DOI: 10.18653/v1/2020.acl-main.97. URL: https://aclanthology.org/2020.acl-main.97.

[7] Kai Shu et al. "defend: Explainable fake news detection". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 395–405.

[8] Zhiwei Yang et al. "A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2608–2621. URL: https://aclanthology.org/2022.coling-1.230.

[9] Kashyap Popat et al. "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*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 22–32. DOI: 10.18653/v1/D18-1003. URL: https://aclanthology.org/D18-1003.

[10] Nguyen Vo and Kyumin Lee. "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*. Online: Association for

Computational Linguistics, Apr. 2021, pp. 965–975. DOI: 10.18653/v1/2021.eacl-main.83. URL: https://aclanthology.org/2021.eacl-main.83.

[11] Denny Vrandečić and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase". In: *Communications of the ACM* 57.10 (2014), pp. 78–85.

[12] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine". In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.

[13] Victoria Higgins et al. "COVID-19: from an acute to chronic disease? Potential long-term health consequences". In: *Critical reviews in clinical laboratory sciences* 58.5 (2021), pp. 297–310.

[14] Monica Bianchini, Marco Gori, and Franco Scarselli. "Inside pagerank". In: *ACM Transactions on Internet Technology (TOIT)* 5.1 (2005), pp. 92–128.

[15] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[16] Hannah Rashkin et al. "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317. URL: https://aclanthology.org/D17-1317.

[17] Jing Ma et al. "Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2561–2571. DOI: 10.18653/v1/P19-1244. URL: https://aclanthology.org/P19-1244.

[18] Lianwei Wu et al. "Evidence Inference Networks for Interpretable Claim Verification". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.16 (May 2021), pp. 14058–14066. DOI: 10.1609/aaai.v35i16.17655. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17655.

[19] Ramy Baly et al. "Integrating Stance Detection and Fact Checking in a Unified Corpus". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 21–27. DOI: 10.18653/v1/N18-2004. URL: https://aclanthology.org/N18-2004.

[20] Isabelle Augenstein et al. "MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4685–4697. DOI: 10.18653/v1/D19-1475. URL: https://aclanthology.org/D19-1475.

[21] Ashim Gupta and Vivek Srikumar. "X-Fact: A New Benchmark Dataset for Multilingual Fact Checking". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 675–682. DOI: 10.18653/v1/2021.acl-short.86. URL: https://aclanthology.org/2021.acl-short.86.

[22] Xuming Hu et al. "CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3362–3376. DOI: 10.18653/v1/2022.naacl-main.246. URL: https://aclanthology.org/2022.naacl-main.246.

[23] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14940–14949.

[24] Claudio Carpineto and Giovanni Romano. "A survey of automatic query expansion in information retrieval". In: *Acm Computing Surveys (CSUR)* 44.1 (2012), pp. 1–50.

[25] George A Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[26] Alexander O Zhirov, Oleg V Zhirov, and Dima L Shepelyansky. "Two-dimensional ranking of Wikipedia articles". In: *The European Physical Journal B* 77 (2010), pp. 523–531.

[27] Jon M Kleinberg. "Authoritative sources in a hyperlinked environment". In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[28] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. "Combating web spam with trustrank". In: *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. 2004, pp. 576–587.

[29] Jacob Devlin et al. "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, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[30] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

[31] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. DOI: 10.18653/v1/D19-1371. URL: https://aclanthology.org/D19-1371.

[32] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[33] *Hugging Face – The AI community building the future.* https://huggingface.co/. July 2023.

[34] Sainbayar Sukhbaatar and Rob Fergus. “Learning from Noisy Labels with Deep Neural Networks”. In: *CoRR* abs/1406.2080 (2014).

[35] Microsoft. *Neural Network Intelligence*. https://github.com/microsoft/nni. Version 2.0. Jan. 2021.

[36] *spaCy · Industrial-strength Natural Language Processing in Python*. https://spacy.io/. July 2023.

[37] *MediaWiki API help*. https://www.wikidata.org/w/api.php. July 2023.

[38] Domcop.com. *Bringing back pagerank using Open Data [free API key]*. https://www.domcop.com/openpagerank/. July 2023. URL: https://www.domcop.com/openpagerank/.

[39] Wanjun Zhong et al. “Reasoning Over Semantic-Level Graph for Fact Checking”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 6170–6180. DOI: 10.18653/v1/2020.acl-main.549. URL: https://aclanthology.org/2020.acl-main.549.