DisFact: Fact-Checking Disaster Claims

Ademola Adesokan[1,a][0000−0003−3803−5906], Haiwei Hu[2,b][0009−0004−0761−1485], and Sanjay Madria[1,c✉][0000−0002−2768−3660]

[1] Missouri University of Science and Technology, Rolla, MO 65401, USA
[2] University of Pittsburgh, Pittsburgh, PA 15260, USA
[a] aaadfg@mst.edu, [b] haiwei.hu@pitt.edu, [c] madrias@mst.edu

Abstract. The rapid proliferation of false information on the internet poses a significant challenge before, during, and after disasters, emphasizing the critical need for domain-specific automatic fact-checking systems. In this study, we introduce DisFact, a new fact-checking pipeline, and a dataset of disaster-related claims generated from the Federal Emergency Management Agency (FEMA) press releases and disaster declarations. Our retrieval method involves no model training, making it more efficient and less resource-intensive. It starts by breaking a lengthy document into sentences; we further apply embeddings to calculate the relevancy score between a claim and document pairs and then compute the similarity score between claims and sentences to rank the retrieved evidence(s). For claim verification, we utilize a deep learning approach that comprises a transformer-based embedding with a feedforward neural network. The experimental findings demonstrate that our fact-checking models achieve top performance on our custom disaster dataset. Furthermore, our models outperform other state-of-the-art models on FEVER and SciFact shared tasks, underscoring the effectiveness of our approach and its adaptability in handling longer documents and generalizing across diverse fact-checking datasets. DisFact signifies a pivotal advancement in automated fact-checking, emphasizing simplicity, accuracy, and computational efficiency. DisFact dataset and code are available on GitHub.[3]

Keywords: Fact-checking · Disaster · Retrieval · Claim Verification

1 Introduction

The emergence of the internet has transformed the way information is spread, offering numerous advantages such as easy access to information and increased public awareness. However, it also presents a notable challenge in the form of misinformation [10]. Although the risks of online misinformation have received considerable attention in areas like sports [11], politics [12], and journalism [13], there has been little to no emphasis on fact-checking within disaster management. The dissemination of inaccurate information before, during, or after critical disaster occurrences hinders collaborative efforts for community safety, obstructs the effective allocation of emergency resources, and disrupts business continuity.

The consequences of misinformation in disaster management are substantial and extensive [14]. They significantly impact all phases, from the preparedness stage to the recovery phase, as they are juxtaposed with information provided by reputable sources such as the Federal Emergency Management Agency (FEMA).

[3] DisFact Dataset and Code - https://github.com/abdul0366/DisFact

One crucial method for combating misinformation involves verifying claims through trusted sources supported by relevant evidence [15], playing a crucial role in all stages of disaster management. In the context of disaster management, fact-checking involves evaluating the accuracy of textual claims, often commencing with the assembly of a comprehensive dataset. This manual process is exemplified in datasets such as Fact Extraction and VERification (FEVER) [16] and SciFact [17], reflecting the challenging nature of manually verifying textual information within disaster management due to the considerable volume of data, leading to labor-intensive and time-consuming procedures. Moreover, the limitations in resources further impede the effective verification of information.

Recently, the field of fact-checking has undergone a significant transformation, focusing on the potential of deep learning models, including transformers. For instance, [18] employed Bidirectional Encoder Representations from Transformers (BERT) in a pairwise approach to ranking, whereas [20] utilized graph neural networks. In contrast, [19] developed a performance-optimized pipeline encompassing document retrieval, point-wise sentence selection, and claim classification. This shift has sparked a new wave of exploration and innovation in fact-checking.

[18] and [21], have demonstrated the effectiveness of transformer methods in their research. Nonetheless, most fact-checking methods, including transformers, depend on the training or fine-tuning of models to enhance contextual comprehension and adaptability to specific subjects. This intricate undertaking demands substantial computational resources, time, and extensive labeled datasets. Furthermore, these models necessitate frequent updates to accommodate new data, rendering the process resource-intensive.

Traditional transformers such as BERT [22] and Robustly Optimized BERT Pretraining Approach (RoBERTa) [23] face a particular challenge when processing longer texts, as they are limited to sequences of up to 512 tokens. This constraint is insufficient for long documents containing substantial evidence to support a claim. To tackle this limitation, [21] introduced a novel method utilizing sparse attention. This new approach can handle longer texts by initially predicting a score for each token in a document and subsequently aggregating these scores at the sentence level to facilitate sentence retrieval. Stammbach's [21] method for binary

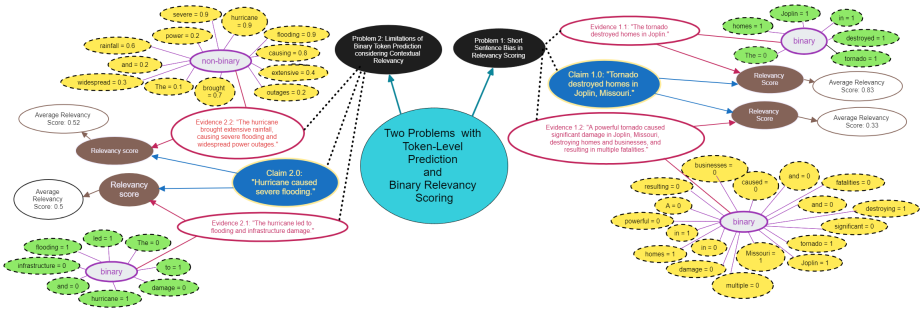

Fig. 1: ***Problem with Token-Level Binary Prediction.***

prediction at the token level encounters two primary challenges. Firstly, there is a bias in relevance scoring towards shorter sentences with fewer relevant tokens,

resulting in higher relevancy scores than longer sentences containing both relevant and irrelevant tokens. As depicted in Figure 1 on the right-hand side, *evidence 1.1* receives a higher relevancy score (0.83) than *evidence 1.2* (0.33), despite *evidence 1.2* encompassing more relevant tokens. This bias manifests in the form of inflated scores for shorter evidence. Secondly, the binary token prediction approach fails to consider the contextual fluctuations in word significance. Assigning identical scores to all tokens neglects to discern their varying relevance to the argument. As exemplified in Figure 1 on the left-hand side, employing binary scoring for *evidence 2.1* yields relevancy score of 0.5, while utilizing non-binary scoring for *evidence 2.2* leads to a higher score of 0.52. This discrepancy highlights the deficiency of binary token prediction in capturing contextual similarities, which better encapsulates the diverse importance of tokens in the given context.

In response to these challenges, we introduce DisFact, a novel and simple approach for fact-checking disaster claims. The method involves the automatic generation of a fact-checking dataset from the FEMA website, where claims are matched with relevant documents (articles) to obtain relevant evidence. This evidence is subsequently utilized to ascertain whether it corroborates or contradicts the claim. In contrast to prior methodologies, our approach harnesses embeddings to enhance precision and accommodate lengthier documents without necessitating training or fine-tuning. Furthermore, our claim verification model integrates embeddings and a straightforward neural network to categorize claims as either substantiated or debunked.

Our contributions are as follows:

- We provide the first domain-specific fact-checking disaster dataset from FEMA Press Releases and Declarations. It contains over 40K pairs of textual claims and documents, with ground truth evidence labeled as supporting or refuting the claims.
- We introduce DisRev, a cost-effective, simple, and novel retrieval method called *"embedding is all you need,"* which uses pre-trained embeddings to retrieve relevant evidence(s) for claim-document pairs without requiring training or fine-tuning.
- Our claim verification model, DisC, follows a supervised learning approach that combines a transformer-based embedding with a feedforward neural network to classify claims as either supported or refuted based on DisRev retrieved evidence.
- Evaluation results show that DisRev achieved 82% precision, 97% recall, 81% F1 score, and MRR of 0.82, while DisC achieved 83% across all metrics.
- To test the generalizability of our models, we applied DisRev and DisC to two well-known fact-checking datasets, FEVER and SciFact. We found that our models effectively outperformed state-of-the-art models.

2 Related Works

In fact-checking, there has been a proliferation of works and recent progress. This section delves into related works in fact verification, focusing on diverse methodologies and their advancements.

Finetuning and Multi-task Learning Approaches: [19] introduced BEVERS. This highly optimized fact verification system achieved state-of-the-art performance on FEVER and SciFact by finetuning each pipeline component without novel improvements. Similarly, [5] proposed a paragraph-level multi-task learning model employing BERT to jointly optimize rationale selection and stance prediction for scientific claim, thereby improving performance on the SciFact dataset. [14] introduced a unified model for multi-task learning in fact verification, integrating document retrieval, evidence extraction, and claim validation into a unified framework, yielding robust performance across multiple datasets.

Transformer and Neural Network-Based Approaches: [6] leveraged the T5 model in their VERT5ERINI system for abstract retrieval, sentence selection, and label prediction, significantly advancing state-of-the-art scientific claim verification. [10] developed a joint model for document-level relation extraction using dynamic pruning and sentence-level attention mechanisms to enhance accuracy. [18] utilized BERT for evidence retrieval and claim verification, achieving state-of-the-art results on the FEVER task with both pointwise and pairwise training approaches. [17] introduced the SciFact dataset to facilitate scientific claim verification, providing a challenging benchmark with annotated claims and supporting evidence from scientific literature and proposing a baseline model combining retrieval and textual entailment.

Knowledge Graphs and Reasoning Frameworks: [15] discussed a method that uses a multi-hop reasoning framework leveraging contextualized representations for fact verification, achieving notable improvements in handling complex claims. [20] integrated knowledge graphs with attention mechanisms to enhance fact verification systems, significantly boosting verification accuracy. [3] utilized logical reasoning and structured proofs for fact verification, providing transparent and interpretable verification results and outperforming existing baselines.

Graph-Based Approaches: [1] presented a graph-based reasoning approach for fact-checking, using semantic role labeling and graph convolutional networks to improve accuracy and achieve state-of-the-art performance on the FEVER dataset. [9] proposed a distillation-based method for improving recall in scientific claim verification, achieving better performance on the SciFact dataset.

Novel Methodologies and Combined Approaches: [12] demonstrated significant improvements in accuracy and efficiency with a neural network-based approach for automatic fact verification over traditional methods. [19] highlighted the effectiveness of combining traditional and neural approaches for improved accuracy and recall in various fact-verification systems. [21] proposed a token-level prediction approach for evidence selection, outperforming sentence-level approaches in the FEVER dataset.

In contrast to the works mentioned above, which primarily rely on training or finetuning their fact-checking models, DisFact diverges by introducing an "Embedding is All You Need" method for evidence retrieval. This method leverages pre-trained embedding models for evidence retrieval without the necessity for extensive training or finetuning. It effectively integrates token-level and

sentence-level information, providing a holistic view of evidence and demonstrating simplicity and generalizability in fact-checking tasks.

3 Methodology

This section outlines our approach to DisFact, depicted in Figure 2, encompassing the creation of datasets, document retrieval, claims generation, evidence retrieval, and claims classification. Each of the steps depicted in Figure 2 will be briefly explained in the subsequent subsections.

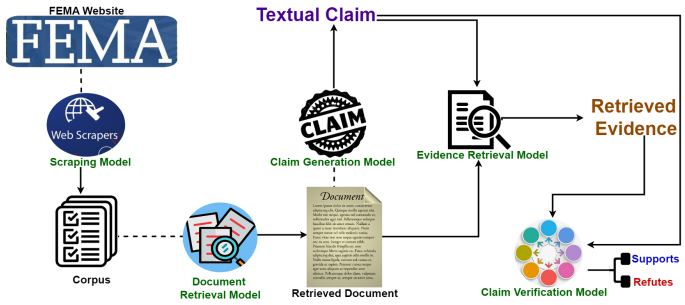

Fig. 2: ***Our DisFact Approach.***

3.1 Document Retrieval

In order to facilitate fact-checking, an automated document retrieval system was developed to systematically scrape FEMA Press Releases, FEMA Disaster Declarations, news media, and reports/notices from the FEMA website. The retrieved document, as shown in Table 1, consists of 40,648 articles. For each article, the full text, title, and date were scraped. The earliest article was published in 2001, while the latest was published in 2024.

The document retrieval system comprises collecting detailed and relevant disaster-related information, thereby forming a robust dataset for analysis and verification. To ensure the efficiency and effectiveness of the scraping process, various components, including user agents and asynchronous requests, were employed. This approach enabled comprehensive data collection for our work while maintaining a coherent flow of information. In order to systematically access and navigate the disaster-related pages on the FEMA website, we employed a base URL and disaster declarations URL. To enhance the robustness of the scraping process and evade detection, we utilized a range of user agents to simulate different browsers. Additionally, random headers were allocated to each request to maintain access and further avoid detection.

Furthermore, we implemented asynchronous fetch using asynchronous I/O to efficiently send HTTP GET requests and retrieve webpage content by concurrently handling multiple requests. Employing BeautifulSoup for HTML parsing, our scraping function extracted relevant data such as titles, publication dates, and main text content from individual articles. To simulate human browsing behavior and reduce the risk of being blocked, random delays were introduced between requests.

Concurrently, the scraping function gathered links to articles from a given page and initiated scraping tasks for each article, significantly accelerating data collection through parallel processing. Moreover, the scrape disaster page function navigated through disaster declaration pages, identified links to specific declarations, and coordinated the scraping of associated articles and reports/notices. To ensure data integrity and efficient processing, we utilized batch processing, looping through multiple pages, and saving collected data incrementally.

Throughout the entire scraping process, we integrated error handling mechanisms to capture and log issues, ensuring a smooth continuation of the scraping process in the event of encountering problems with certain pages or articles.

3.2 Claim Generation

Claims play a crucial role in the process of fact-checking as they serve to authenticate the accuracy and veracity of statements, thereby safeguarding the public from misinformation. In our pursuit to automatically generate claims from textual documents, particularly focusing on claims related to disasters, we employed the Text-to-Text Transfer Transformer (T5) model [7]. Through training on the FEVER dataset, T5 understood the typical construction of claims, which we leveraged to produce claims from our FEMA dataset. The selection of T5 was based on its capacity to craft abstractive claims, which encapsulate the essence of a sentence rather than merely extracting the exact text, resulting in more intricate claims. In contrast to extractive models, which extract verbatim text from the source, abstractive models generate new text that effectively conveys the original input's intended meaning. The process commences with compiling a dataset comprising articles containing titles, text, and publication dates. A crucial step involved the utilization of two pre-trained T5 models: one dedicated to producing affirming statements and the other specialized in refuting claims. These models were meticulously fine-tuned using the FEVER dataset to enhance their comprehension of fact-checking claims. Text tokenization was achieved using the NLTK library to segment the articles into individual sentences. Subsequently, two random sentences were selected from each article containing at least two sentences, with each sentence being modified to incorporate the respective article title for contextual reference.

Due to computational limitations, the text was processed in batches, comprising up to 64 sentences. The text was inputted into the T5 models following tokenization to generate supporting and refuting claims. These claims were then decoded back into text format.

Finally, the generated claims were matched with their respective sentences and additional metadata such as article text, title, and date. Each sentence was linked to both a supportive and a refuted claim. The resulting dataset, which includes the generated claims, evidence sentences, and relevant article metadata marked as either "Supported" or "Refuted," can be accessed on GitHub, and detailed dataset statistics are provided in Table 1.

3.3 Evidence Retrieval

Our retrieval task (DisRev) is inspired by [21]; however, the method presented therein encounters the two issues outlined in Figure 1 of Section 1. To address

Label	Train	Test
Supports	16,248	4,076
Refutes	16,245	4,079
Total	32,493 (80%)	8,155 (20%)

Table 1: ***Statistics of DisFact Annotated Dataset***

these challenges and enhance the performance of existing models, we introduce a novel method known as "Embedding is All You Need" for evidence retrieval in the context of automatic fact-checking. This technique involves processing a set of claims $\{c_1, c_2, \ldots, c_n\}$ and their associated documents $\{d_1, d_2, \ldots, d_n\}$. Our method utilizes a no-training approach, which computes relevancy and similarity scores between pairs of claims and their corresponding documents to extract pertinent evidence. This is achieved through a combination of tokenization, embedding models, and similarity measures, as shown in Figure 3. To prepare

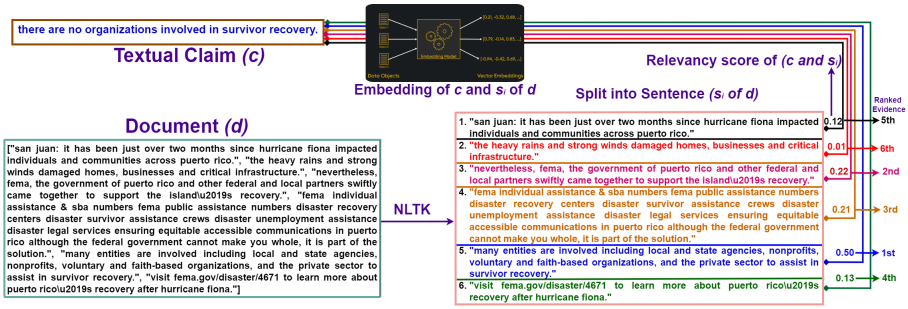

Fig. 3: ***Evidence Retrieval Approach (DisRev).***

our input data for the model, we tokenize each claim c and document d using the BigBirdTokenizer. Furthermore, we ensure uniform sequence length for batch processing by implementing padding and attention masks. It is important to note that contextual document information is provided in the form of a list of sentences. To tokenize each document into its constituent sentences, we employ the pre-trained nltk.tokenize.punkt tokenizer.

The tokenization process for a claim c and a document d is defined as:

$$\boldsymbol{T(c)} = \{t_1, t_2, \ldots, t_n\}$$
$$\boldsymbol{T(d)} = \{s_1, s_2, \ldots, s_m\}$$

Where t_i and s_j represent the tokens from the claim and document, respectively.

The tokenized claim $T(c)$ and document $T(d)$ are encoded and depicted by:

$$\boldsymbol{input\text{-}ids} = [\text{CLS}] + T(c) + [\text{SEP}] + T(d)$$

Note: *CLS is the Classification Token and SEP is the Separator Token.*

The utilization of the RoberTa model in our sentence selection framework serves to complement BigBird by leveraging its pre-trained knowledge, which has demonstrated strong performance in natural language processing (NLP) tasks and offers robust language understanding capabilities. Extending the Roberta Model to incorporate BigBird addresses the challenge posed by the limitation of sequence length in most transformer models, particularly when handling long sequences. Drawing inspiration from [21]'s approach, we opted to integrate BigBird into our model due to its efficient handling of long sequences, with the ability to

accommodate up to 4096 tokens through the use of sparse attention mechanisms, thus making it well-suited for processing lengthy documents. This is a crucial feature as it obviates the necessity for document truncation, thereby preserving a more comprehensive context for accurate evidence retrieval.

In our approach, we employ a multi-head attention mechanism to augment the model's capacity to focus on various segments of the input sequence, thereby enhancing the accuracy of evidence retrieval. Subsequently, relevancy scores for each pair are outputted. We utilize 16 heads in the attention mechanism. The multi-head attention is represented by:

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

Where Q, K, and V denote the query, key, and value matrices derived from the input sequences and d_k is the dimensionality of the key vectors. This integration of the RoberTa and BigBird models, combined with a multi-head attention mechanism, enhances the effectiveness of our approach in processing and analyzing lengthy texts for evidence retrieval.

In our retrieval model, we consider the relevance of each token or sentence in the document in relation to the claim. To generate token and sentence embeddings, we employ the Sentence Transformer. This particular model can produce high-quality sentence embeddings that capture the semantic meaning of the text. It offers a method to gauge the similarity between sentences and claims without additional training. Additionally, we utilize cosine similarity as a measure to assess the similarity between the claim and context sentences due to its simplicity and efficiency in computing the direct measure of similarity between claim and sentence embeddings. This method enables the comparison of sentence and token embeddings to determine relevance without complex training procedures.

To be succinct, we compute three kinds of relevancy scores for each input: token-level, sentence-level, and combined-level.

a. Token-Level Relevancy Score - Recall that document d contains sentences sn, each sentence s in the document d is tokenized, and token t embeddings are generated using the Sentence Transformer model. Subsequently, the cosine similarity between each token embeddings of s in d and embedding of the claim is computed, as demonstrated by:

$$\textbf{\textit{cosine-similarity}}(c_i, t((s_i)) = \frac{c_i \cdot t(s_i)}{\|c_i\|\|t(s_i)\|}$$

Moreover, the average token relevancy similarity score for aggregated tokens at the sentence level is computed, as depicted by:

$$\textbf{\textit{avg-token-score}}(s) = \frac{1}{n}\sum_{i=1}^{n} cosine\text{-}similarity(t_i, claim)$$

Where t_i represents the tokens in sentence s_i. This average score indicates the relevance of each token in each sentence in the d_i to the claim.

b. Sentence-Level Relevancy Score - At the sentence level, an embedding is generated for the entire claim c and each sentence s in the document d. The cosine similarity between each sentence s embedding in d and the embedding of

c is calculated similar to cosine-similarity in the token-level. This similarity score is employed to measure the relevance of each s in d to c.

$$\textbf{\textit{cosine-similarity}}(c_i, t((s_i)) = \frac{c_i \cdot t(s_i)}{\|c_i\|\|t(s_i)\|}$$

c. Combined-Level Relevancy Score - As previously mentioned, the calculated average token-level relevancy scores for each sentence in d and the sentence-level relevancy scores are used. These scores are then combined for each sentence to produce a final relevancy score, as demonstrated by:

$$\textbf{\textit{combined-score}} = \frac{\textit{avg-token-score} + \textit{sentence-score}}{2}$$

This combined score assigns equal importance to both token-level and sentence-level relevancy, thus offering a more comprehensive measure of the relevance of each sentence to the claim.

3.3.1 Ranking: After calculating the relevancy and similarity scores, the corresponding evidence sentences for each claim are sorted in descending order based on their scores, as depicted in Figure 3. The top-ranked sentence (top-1) is regarded as the primary evidence. Subsequent top-ranked sentences (from top-2 to top-n) serve as multiple pieces of evidence determined by the number of sentences in the documents. This sorting process is carried out independently for token-level, sentence-level, and combined scores. Subsequently, the top-ranked sentences are considered primary or secondary evidence based on their ranking.

3.4 Claim Verification

In order to classify claims as either supports or refutes based on the retrieved evidence in Section 3.3, our claim verification approach (DisC), as depicted in Figure 4, employs pre-trained models similar to the one used in our evidence retrieval task and efficient feature transformations. Through our method, pairs of claims and evidence are processed to classify the claims as binary tasks "SUPPORTS" and "REFUTES" for our DisFact dataset, and we also made it to classify multiclass labels such as "SUPPORTS," "REFUTES," and "NOT ENOUGH INFO," ensuring robustness and generalization on other publicly available fact-checking datasets. As shown in Figure 4, we explain the different components of our model as follows:

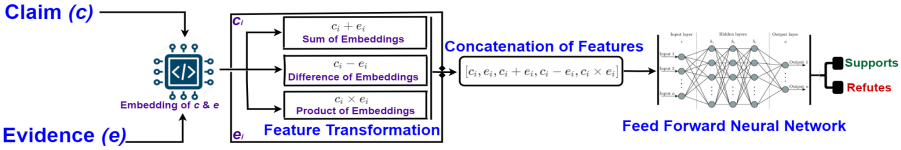

Fig. 4: ***Claim Verification System (DisC).***

a. Preprocessing: In the preprocessing phase, unnecessary characters are removed from the text, and text normalization is applied. Loading the data entails extracting claims and evidence, as well as assigning labels for training purposes. The emphasis on clean, structured data is imperative for improving model performance and ensuring generalization. Subsequently, the careful preprocessing of the data sets the stage for optimal model performance and generalizability.

b. Embeddings: In our approach, we employed the Sentence Transformer model to produce high-quality sentence embeddings. Specifically, we opted for the 'bert-base-nli-cls-token' due to its pre-training on natural language inference tasks. This enables a comprehensive comprehension of the semantic connections between claims and evidence. Consequently, this choice facilitated a deeper understanding of the relationships between different elements in the context of our study.

c. Feature Transformation: In feature transformation, textual data is converted into numerical features to be utilized in model training. This process encompasses the generation of embeddings for both claims and evidence and the creation of combined features, including sums, differences, and products. By combining these features, the model's capacity to comprehend intricate relationships between claims and evidence is significantly enhanced. Moreover, our feature transformation approach mitigates the necessity for intricate feature engineering. The generation of embeddings is carried out utilizing the Sentence Transformer model to derive features based on the provided claims and evidence. Denoting c_i as the embedding for claim i and e_i as the embedding for the corresponding evidence, our feature combinations are structured as follows:

- ***Sum of Embeddings:*** In exploring the sum of embeddings, we aim to capture the combined semantic information of both the claim and the evidence. This feature facilitates an understanding of the overall semantic similarity by combining the magnitudes of the embeddings, which enhances comprehension of how the combined meaning of both texts relates to each other. Equation 1 is used to determine the sum of embeddings is given by:

$$\|c_i + e_i\| = \sqrt{\sum_{j=1}^{d} (c_{ij} + e_{ij})^2} \tag{1}$$

- ***Difference of Embeddings:*** In our feature extraction, we examined the variance in embeddings to emphasize the disparities between the claim and the evidence. This method effectively captures dissimilarity by quantifying the spatial separation between the embeddings within the semantic domain. Our focus on these distinctions is essential for determining whether the evidence substantiates or contradicts the claim. This approach is represented by Equation 2:

$$\|c_i - e_i\| = \sqrt{\sum_{j=1}^{d} (c_{ij} - e_{ij})^2} \tag{2}$$

- ***Element-wise Product of Embeddings:*** The third feature in our approach encompasses the element-wise product of embeddings, allowing us to capture the interactions between each dimension of the embeddings and their corresponding dimensions. This method highlights specific feature-level interactions, providing insights into the specific relationships between the claim and the evidence. Mathematically, the element-wise product between the embeddings c_i and e_i is represented in Equation 3 as:

$$(c_i \times e_i)_j = c_{ij} \cdot e_{ij} \tag{3}$$

d. Concatenation of Features: The amalgamation of features generates the ultimate feature vector for each claim-evidence pair. This vector is a composition

of the separate embeddings and the trio of combined features. By joining these components, the resulting vector encapsulates a comprehensive array of information regarding both the claim and the evidence. Specifically, it encompasses their implications, combined implications, divergences, and synergies. This is mathematically formulated in Equation 4 as:

$$\boldsymbol{features}_i = [c_i, e_i, c_i + e_i, c_i - e_i, c_i \times e_i] \quad (4)$$

e. FeedForward Neural Network (FFNN) Model: The combined features were fed into sequential feedforward neural networks comprising dense and dropout layers, which are known for their capability to capture intricate patterns and interconnections within data through multiple layers. The incorporation of dense layers and dropout mechanisms enhances generalization and mitigates overfitting. This distinctive architecture, coupled with the utilization of combined features such as sum, difference, and product, confers enhanced efficiency upon the model in terms of learning from the provided data.

4 Experimental Setup and Results

This section outlines the experimental setup of our models. The performance of our models was rigorously evaluated using an 80% train and 20% test split, as shown in Table 1, with metrics such as accuracy, precision, recall, F1-score, and mean reciprocal rank (MRR). These experiments were conducted on a robust computational system featuring dual NVIDIA® Tesla V100 GPUs and an Intel® Xeon® Gold CPU, providing computing power to process our models.

Table 2: ***Claim Verification and Claim Generation Hyperparameters***

Claim Verification				Claim Generation	
Hyperparameter	**Value**	**Hyperparameter**	**Value**	**Hyperparameter**	**Value**
Model Name	bert-base and FFNN	Hidden Layer	relu	Random Seed	42
Model Max Length	512	Output Layer (binary)	sigmoid	Epoch	50
Max Length	50	Output Layer (multiclass)	softmax	Tokenizer Model	t5-base
Dropout Rate	0.2	Kernel Initializer	uniform	Max Length	512
Epochs	15	Optimizer	RMSprop	Max Length	50
Verbose	1	Loss (binary)	CrossEntropyLoss/BCE	Batch Size	64
Batch Size	1024	Input Layer (Dense) Size	300 Units	Label	Supports/Refutes

Table 2 details the hyperparameters and their candidate values used in our claim generation and claim verification models. Additional information on our code and the hyperparameters used for all models can be found on the GitHub link provided in the abstract.

4.1 Results

The experiments are conducted to study the performance of our retrieved evidence and claim classification; our model possesses advantages in different reasoning scenarios, such as single or multiple evidence(s), which shows the effectiveness of our approach.

4.1.1 DisFact Evaluation

a. Retrieval Task - Table 3 presents the results of evidence retrieval and claim verification models of the fact verification system of our disaster domain dataset. As our dataset is new to the fact-checking field, the results will establish a baseline for future work utilizing it. Multiple testing scenarios were carried out to compare the effectiveness of single (top-1) and multiple evidence (top-2 or more).

A comparative analysis presented in the retrival task of Table 3 illustrates that sentence-level ranking for top-1 yields the highest performance across all

metrics. In contrast, token-level scores are notably lower, indicating that focusing solely on token-level information is less effective for evidence retrieval. Precision decreases as more top-N sentences are analyzed (from top-1 to top-5) across all ranking types as more sentences are included. Conversely, recall increases, demonstrating that more considered sentences retrieve more relevant evidence. Notably, sentence-level ranking consistently achieves the highest MRR values, indicating that the most pertinent evidence is often found in the top-ranked positions.

Table 3: ***DisFact Retrieval and Classification Results***

Relevancy Type	**Top-N**	**Retrieval Task**				**Classification Task**			
		P	**R**	**F1**	**MRR**	**P**	**R**	**F1**	**LA**
Token-Level	Top-1	29.17	28.98	26.33	0.1002	81.92	80.43	80.20	80.43
Token-Level	Top-2	25.32	41.95	29.13	0.1455	81.83	80.94	80.81	80.94
Token-Level	Top-3	22.88	51.42	29.63	0.1737	81.80	80.98	80.86	80.98
Token-Level	Top-4	21.24	58.03	29.21	0.1933	82.29	81.87	81.82	81.88
Token-Level	Top-5	19.84	64.24	28.59	0.2095	81.51	81.26	81.23	81.26
Sentence-Level	Top-1	82.05	82.05	81.11	0.7493	83.70	79.18	78.39	79.10
Sentence-Level	Top-2	52.23	91.16	63.81	0.8050	82.49	77.38	76.39	77.30
Sentence-Level	Top-3	37.87	94.87	51.74	0.8205	83.13	82.70	82.63	82.68
Sentence-Level	Top-4	30.50	96.45	44.18	0.8261	82.76	82.63	82.62	82.64
Sentence-Level	Top-5	25.71	97.46	38.86	0.8294	82.07	78.76	78.14	78.69
Combined-Level	Top-1	77.14	77.02	75.78	0.6806	83.67	78.35	77.43	78.31
Combined-Level	Top-2	50.62	86.70	61.31	0.7377	83.47	83.08	83.04	83.09
Combined-Level	Top-3	37.56	91.49	50.91	0.7590	83.45	83.41	83.41	83.41
Combined-Level	Top-4	30.68	93.71	44.11	0.7666	83.22	82.64	82.55	82.63
Combined-Level	Top-5	26.06	95.25	39.09	0.7711	83.01	83.00	82.99	83.00

The superior performance of sentence-level and combined-level rankings emphasizes the significance of considering broader contextual information rather than concentrating solely on individual tokens. This approach ensures that the retrieved evidence is not only relevant but also contextually coherent.

The implications for model design suggest that future models should prioritize sentence-level and combined-level approaches for evidence retrieval tasks. Combining sentence-level context with token-level details can lead to more accurate and reliable fact-checking systems. Moreover, the results indicate that optimizing for sentence-level relevance may yield the best immediate improvements in precision and recall, while combined-level optimization can enhance the balance between these metrics, resulting in robust overall performance.

b. Claim Verification - The findings presented in the classification task of Table 3 indicate excellent performance across various metrics, particularly at the top-2 and top-3 ranks, with the highest accuracy and F1-Score observed at the top-3 level for our claim verification system. Furthermore, the combined-level ranking consistently outperforms both token-level and sentence-level rankings, emphasizing the significance of integrating multilevel information for evidence retrieval. Additionally, the results imply that leveraging the top-ranked pieces of evidence yields optimal overall performance by effectively balancing precision, recall, F1-Score, and label accuracy (LA).

4.1.2 Beyond DisFact: *FEVER and SciFact*

In expanding the assessment of our automatic fact verification pipeline, we extended our evaluation to encompass the FEVER and SciFact datasets. The FEVER dataset, a publicly available dataset with 185,455 claims, is manually annotated and accompanied by 5,416,537 supporting Wikipedia documents. On the other hand, SciFact, which shares similarities in structure with FEVER, comprises scientific articles and encompasses 1,409 claims along with 5,183 article abstracts. This expansion broadens the scope and generalizability of our study and provides a more comprehensive understanding of the performance of our methods across different datasets. It is worth noting that our evaluation was on the training and development sets of both datasets.

a. FEVER - For the FEVER evidence retrieval task, our DisRev model demonstrates superior precision and F1-score compared to other models, as illustrated in Table 4, thereby showcasing its effectiveness in identifying relevant evidence. The DisRev model exhibits a significant improvement of 12% in precision and approximately 9% in F1-score over its closest counterpart, KGAT. However, our recall lags slightly behind that of KGAT. Analysis of Table 5 reveals that our DisC model surpasses all other models in the FEVER classification task with an accuracy of 85.93%, signifying its exceptional performance in validating claims based on retrieved evidence. Furthermore, our model outperforms the nearest competitor, ProoFVer-SB, by 5% in label accuracy.

Table 4: *FEVER Retrieval*

Model	P	R	F1
Stammbach [21]	25.49	90.79	39.81
KGAT [20]	27.29	**94.37**	42.34
DREAM [1]	26.67	87.64	40.90
Soleimani [18]	24.97	88.32	38.93
Ours (DisRev)	**39.46**	87.71	**50.89**

Table 5: *FEVER LA Score*

Model	Label Acc
Stammbach [21]	80.59
ProoFVer-SB [3]	80.74
KGAT [20]	78.29
GEAR [2]	74.84
DREAM [1]	79.16
Soleimani [18]	74.59
Ours (DisC)	**85.93**

Table 6: *Evaluation on SciFact*

Models	SS Only	SS+Label		
	R	P	R	F1
Zhang [9]	60.7	67.0	52.1	58.7
Wadden [17]	43.4	48.5	38.8	43.1
Pradeep [6]	57.4	60.8	53.8	57.1
Li [5]	57.4	63.8	48.9	55.2
Zhang [4]	58.5	66.5	51.1	57.8
Ours (Both)	**87.7**	**75.7**	**74.3**	**75.0**

b. SciFact - Our DisRev model demonstrates exceptional performance within the SciFact dataset, achieving an 87.7% recall and 75.0% F1 score, thus clearly outperforming other models documented in Table 6. This underscores the proficiency of our model in effectively generalizing scientific claims, surpassing the scope of the disaster dataset. Given that recall holds greater significance in retrieval tasks, our model notably surpasses the subsequent best model in recall, achieving 27% for sentence selection exclusively. Furthermore, for sentence selection combined with labeling, our model excels across all metrics compared to the closest models.

Our model's ideal performance can be attributed to incorporating straightforward and precise relevance and similarity methods. These techniques guarantee the prioritization of the most pertinent evidence, enhancing precision and recall metrics. Additionally, our approach harnesses the power of context-aware pre-trained embedding models, benefiting from comprehensive pre-training across diverse datasets, consequently bolstering its generalization prowess.

5 Limitation

a. Domain Shift/Mismatch: In claim generation, we train T5 on FEVER data but used it for DisFact generated data resulted in a few hallucinated claim due to domain shift and data mismatch, which were manually corrected. T5

sometimes struggles to differentiate between generating claims based on input data and relying on pre-learned knowledge, leading to errors with unfamiliar concepts, vocabulary, or styles in the claims.

b. Impact of Noisy Data: Despite using preprocessing to eliminate noise, generating claims from the FEMA dataset may still require these noises, such as special characters like the dollar ($) sign (for example, if $450 used in a claim was preprocessed to be 450). However, this noise could actually enhance the DisRev model's ability to support or refute claims. By analyzing the impact of this noise, we could uncover opportunities to improve the model's robustness.

c. Challenges with Long Documents: DisFact tackles long-document problems using BigBird, a model that can handle sequences of up to 4096 tokens via sparse attention. However, the model might still need help with sequences beyond this limit or when crucial information spans multiple segments. Investigating these cases is crucial, as doing so could suggest better segmentation techniques or models that capture cross-segment relationships more effectively.

6 Conclusion and Future Work

DisFact's use of embeddings (without additional training) for evidence retrieval and automatic dataset generation from the FEMA website has significantly advanced fact-checking for disaster-related claims. By integrating contextual relevancy and similarity information through pre-trained models, DisFact has outperformed existing methods. Its capability to handle long documents and diverse datasets underscores its robustness in real-time fact-checking scenarios. These findings validate DisFact's potential to greatly improve the accuracy and reliability of automated fact-checking systems in disaster management contexts. In the future, the focus will be on expanding the dataset to encompass a wider array of disaster-related and social media data sources. This expansion aims to enhance DisFact's capabilities and position it as a premier automated fact-checking solution for disaster management and other fields.

Acknowledgement

This research project received support from the NSF - USA CNS-2219615 and the Kummer Institute for Student Success, Research, and Economic Development at the Missouri University of Science and Technology through the Kummer Innovation and Entrepreneurship Doctoral Fellowship.

References

1. W. Zhong, J. Xu, D. Tang, Z. Xu, N. Duan, M. Zhou, J. Wang, and J. Yin, "Reasoning Over Semantic-Level Graph for Fact Checking," in *Proceedings of the 58th Annual Meeting of the ACL*, Association for Computational Linguistics, 2020.
2. J. Zhou, X. Han, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "GEAR: Graph-based evidence aggregating and reasoning for fact verification," in *Proc. of ACL*, 2019, pp. 892–901.
3. A. Krishna, S. Riedel, and A. Vlachos, "Proofver: Natural logic theorem proving for fact verification," *Transactions of the ACL*, vol. 10, pp. 1013-1030, 2022.
4. Z. Zhang, J. Li, F. Fukumoto, and Y. Ye, "Abstract, rationale, stance: a joint model for scientific claim verification," in *Proc. of the 2021 Conf. on EMNLP*, Punta Cana, Dom. Republic, pp. 3580–3586, ACL, 2021. doi: 10.18653/v1/2021.emnlp-main.290.

5. X. Li, G. Burns, and N. Peng, "A paragraph-level multi-task learning model for scientific fact-verification," arXiv preprint arXiv:2012.14500, 2020.
6. R. Pradeep, X. Ma, R. Nogueira, and J. Lin, "Scientific claim verification with VerT5erini," in *Proc. of the 12th LOUHI Workshop*, 2021, pp. 94–103, ACL.
7. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.
8. M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big Bird: Transformers for Longer Sequences," in *NeurIPS*, vol. 33, 2020, pp. 17283–17297, Curran Assoc., Inc.
9. Z. Zhang, J. Li, and F. Fukumoto, "An Efficient Approach for Improving the Recall of Rough Abstract Retrieval in Scientific Claim Verification," in *Intl. Conference on Artificial Neu. Net.*, Springer Nature Switzerland, Cham, 2023, pp. 63-74.
10. M. Shahbazi and D. Bunker, "Social media trust: Fighting misinformation in the time of crisis," *Int. J. of Info. Mgt*, vol. 77, 2024, doi: 10.1016/j.ijinfomgt.2023.102780.
11. N. B. Tiller, J. P. Sullivan, and P. Ekkekakis, "Baseless Claims and Pseudo science in Health and Wellness: A Call to Action for the Sports, Exercise, and Nutrition-Science Community," *Sports Medicine*, vol. 53, no. 1, pp. 1-5, 2023.
12. S. Ahmed, D. Madrid-Morales, and M. Tully, "Social media, misinformation, & age inequality in online political engagement," *J. of IT & Pol.*, vol. 20, no. 3, pp. 269-285, 2023.
13. N. van Antwerpen, D. Turnbull, and R. A. Searston, "Perspectives from journalism professionals on the application and benefits of constructive reporting for addressing misinformation," *The Intl. J. of Press/Politics*, vol. 28, no. 4, pp. 1037-1058, 2023.
14. E. S. Apostol, C. O. Truică, and A. Paschke, "ContCommRTD: A distributed content-based misinformation-aware community detection system for real-time disaster reporting," *IEEE Transactions on Knowledge and Data Engineering*, 2024.
15. Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the ACL*, vol. 10, pp. 178-206, 2022.
16. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," in *Proc. of the Conf. of the NAACL: Human Language Technologies*, vol. 1 (Long Papers), 2018, pp. 809-819.
17. D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or Fiction: Verifying Scientific Claims" in *Proc. of EMNLP* 2020, pp. 7534-7550
18. A. Soleimani, C. Monz, and M. Worring, "BERT for evidence retrieval and claim verification," in *Advances in Information Retrieval: 42nd ECIR 2020*, Lisbon, Portugal, Proc., Part II 42, Springer Intl. Publishing, 2020, pp. 359-366.
19. M. DeHaven and S. Scott, "BEVERS: A General, Simple, and Performant Framework for Automatic Fact Verification," in *Proc. of the 6th FEVER Work.*, 2023, pp. 58-65.
20. Z. Liu, C. Xiong, M. Sun, and Z. Liu, "Fine-grained Fact Verification with Kernel Graph Attention Net.," in *Proc. of the 58th the ACL Conf.*, 2020, pp. 7342-7351.
21. D. Stammbach, "Evidence selection as a token-level prediction task," in *Proc. of 4th Workshop on Fact Extraction and VERification (FEVER)*, 2021, pp. 14-20, ACL.
22. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conf. of the NAACL: Human Language Technologies*, vol. 1, 2019, pp. 4171-4186.
23. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.