AMERICAN UNIVERSITY OF BEIRUT

FACTFORMER: A TRANSFORMER BASED FACT CHECKER

by

ADEMOLA ABDULGANIYU ADESOKAN

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Computer Science
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
August 2023

AMERICAN UNIVERSITY OF BEIRUT

FACTFORMER: A TRANSFORMER BASED FACT CHECKER

by

ADEMOLA ABDULGANIYU ADESOKAN

Approved by:

Dr. Shady Elbassuoni, Associate Professor — Advisor
Computer Science

Dr. Izzat El Hajj, Assistant Professor — Member of Committee
Computer Science

Dr. Haidar Safa, Professor — Member of Committee
Computer Science

Date of thesis defense: August 29, 2023

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Adesokan Ademola Abdulganiyu

Last	First	Middle
Adesokan	Ademola	Abdulganiyu

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes

___ **As of the date of submission of my thesis**

___ **After 1 year from the date of submission of my thesis .**

✔ **After 2 years from the date of submission of my thesis .**

___ **After 3 years from the date of submission of my thesis .**

September 5, 2023

Signature Date

Acknowledgements

With profound reverence and humility, I extend my heartfelt gratitude to the Almighty Allah. It is through His boundless benevolence that I have been bestowed with the strength and determination necessary to embark upon and see to the successful fruition of this research work.

I extend my profound appreciation to Dr. Shady Elbassuoni, my supervisor, whose support and guidance have been a beacon for me in the undertaking of this research work. His willingness to accommodate my needs during difficult times has been a testament to his understanding and compassion, for which I am immensely grateful. My gratitude also extends to the esteemed members of my thesis committee, Dr. Izzat El Hajj and Dr. Haidar Safa, for their investment of time and for the invaluable feedback and constructive criticism they offered. Their insights have greatly enriched my research.

I cannot put into words the depth of my gratitude towards my beloved wife, Maryam Bashir, who has been an unshakeable pillar of support throughout this journey. Her understanding, the sacrifices she has made, and her unwavering belief in me have been vital to my accomplishments. I am also indebted to my dear friends, Mariam Safieldin, Philip Adebayo, and Assaad El Halabi, among many others, who have bolstered me with their support and encouragement. Their friendship has been an inexhaustible wellspring of motivation. My sincere appreciation goes out to my sponsor, the Mastercard Foundation, whose generosity and support for my Master's degree program have made this journey possible. I am immensely thankful for the opportunities they have afforded me.

The unwavering love and support of my cherished family, both nuclear and extended, have provided the bedrock upon which my academic achievements are built. I am deeply grateful to my family for accompanying me through every step of this journey. My heartfelt gratitude extends to my dear mother, who has been a constant source of inspiration and motivation. Her boundless love, care, and support have been my guiding light, steering me through life's vicissitudes and bolstering my resolve to chase my dreams.

Lastly, I owe an immeasurable debt of gratitude to my late father (May Allah forgive him and have mercy on him and give him strength and pardon him), whose support and motivation have been instrumental in my academic journey. His unwavering belief in the importance of education, coupled with his ceaseless encouragement, even in the face of daunting challenges, has been a driving force in my academic pursuits. I cherish every moment spent with him before his passing, and I will forever treasure the profound impact he has had on my life.

ABSTRACT
OF THE THESIS OF

Ademola Abdulganiyu Adesokan for Master of Science
Major: Computer Science

Title: FactFormer: A Transformer Based Fact Checker

Misinformation can undermine public trust and lead to misguided actions based on unreliable sources and fact-checking efforts. Traditional manual fact-checking systems suffer from several challenges, including issues related to scaling, performance, and complexity. In response to this challenge, we introduce **FactFormer**, an automatic fact-checking system that retrieves evidence from trustworthy sources for a given claim and subsequently classifies the claims into different labels based on the retrieved evidence. Our retrieval model adopts the extractive question-answering technique. This approach treats claims as questions and trusted sources as context from which evidence, construed as answers, is retrieved. We harnessed the capabilities of the Bidirectional Encoder Representations from Transformers (BERT) and Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) architectures, fine-tuning them specifically for the task of evidence extraction. Subsequently, claim verification was accomplished using a multi-headed BERT combined with a fully connected network layer. During the evaluation phase, our retrieval models demonstrated state-of-the-art results: the BERT model yielded an exact match rate of 89.89% and an F1-measure score of 93.93%, while the DistilBERT model achieved an exact match rate of 90.19% and an F1-measure score of 93.98% when evaluated with a maximum evidence length of 100 words. Our claim verification model achieved a high accuracy score of 90% using the existing manually annotated Fact Extraction and VERification (FEVER) dataset with three classes, outperforming other state-of-the-art papers. We further conducted end-to-end system experiments and evaluations using our retrieved evidence to demonstrate its ability to generalize well when compared to the manually annotated FEVER-2 dataset with two labels. Our claim verification model performance on FEVER-2 with DistilBERT achieved 87.14%, outperforming the manual FEVER-2 with an 86.54% accuracy score. In conclusion, our approach significantly enhances fact-checking by improving both evidence retrieval and claim classification.

Table of Contents

ILLUSTRATIONS

TABLES

ABBREVIATIONS

AMD	Advanced Micro Devices
BERT	Bidirectional Encoder Representations from Transformers
BEVERS	Baseline fact Extraction and VERification System
BM25	Best Match 25
CD	Claim-Dissector
CLS	Classification
CPU	Central Processing Unit
CVM	Claim Verification Model
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
ESIM	Enhanced Sequential Inference Model
FEVER	Fact Extraction and VERification
GAT	Graph Attention Network
GB	Gigabyte
GEAR	Graph-based evidence aggregating and reasoning
GHz	gigahertz
GNNs	Graph Neural Networks
GPU	Graphics Processing Unit
ID	Identifier
IR	Information Retrieval
KGAT	Kernel Graph Attention Network
NEI	Not Enough Information
NLP	Natural Language Processing
ProoFVer	Proof System for Fact Verification
PUBHEALTH	Public health
ReLU	Rectified Linear Unit
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RoBERTa	Robustly Optimized BERT Pretraining Approach
RMSProp	Root Mean Square Propagation
S2ORC	Semantic Scholar Open Research Corpus
SciFact	Scientific Fact-checking
SEP	Separator
Seq2seq	Sequence to Sequence
SISER	Semantic-Infused Selective Graph Reasoning
SQuAD	Stanford Question Answering Dataset
TF-IDF	Term Frequency - Inverse Document Frequency

Transformer-XH	Transformer eXtra Hop
V2	Version 2

CHAPTER 1

INTRODUCTION

The abundance of fake news and false claims on the Web and social media poses a very serious threat to society. Misinformation has shown an alarming tendency to spread faster and broader than true information [1]. The repercussions, which are still evident today, have been detrimental in various sectors including academics [2], politics [3], media [4], health [5] and law [6], among others. This has all led to a recent and wide interest in establishing robust and scalable approaches for fact-checking in various communities, including Media, Social Science, and Computing.

Automatic fact-checking can be a challenging task particularly when claims are short in nature and lack any context or evidence, which is often the case in social media and on the Web. A common approach for fact-checking that has been recently adopted by many is to first retrieve textual evidence for a given claim from auxiliary trusted sources, and then use the retrieved evidence to verify or refute the given claim [7]–[16].

In this paper, we propose **FactFormer**, a transformer-based fact-checking approach that verifies a given textual claim by retrieving textual evidence for that claim from a trusted source and then using the retrieved evidence to either verify or refute the given claim or determining that the retrieved evidence is not sufficient enough to judge the validity of the given claim. Our approach thus consists of two main components: 1) an evidence retrieval model, and 2) a claim verification model. Our evidence retrieval model is a transformer-based model that takes in a textual claim and a trusted source and retrieves a span of text from the document that either supports or refutes the input claim (similar to a question-answering model). Our claim verification model is also a transformer-based model that takes in a textual claim and textual evidence and determines whether the claim is true, false, or unverifiable based on the given evidence. An overview of our complete transformer-based fact-checking approach is shown in Figure 1.1.

To the best of our knowledge, our approach is the first to employ a question-answering transformer-based model to retrieve textual evidence for fact-checking and to couple this with another transformer-based model to verify claims. To evaluate our approach, we conducted three different experiments using the Fact Extraction and VERification (FEVER) dataset [17], which is a standard fact-checking benchmark. The FEVER dataset consists of various claims along with manually extracted textual evidence from Wikipedia, and a label for each claim indicating whether the associated textual evidence supports, refutes, or is not sufficient to judge the validity of the claim. Our first experiment evaluated the

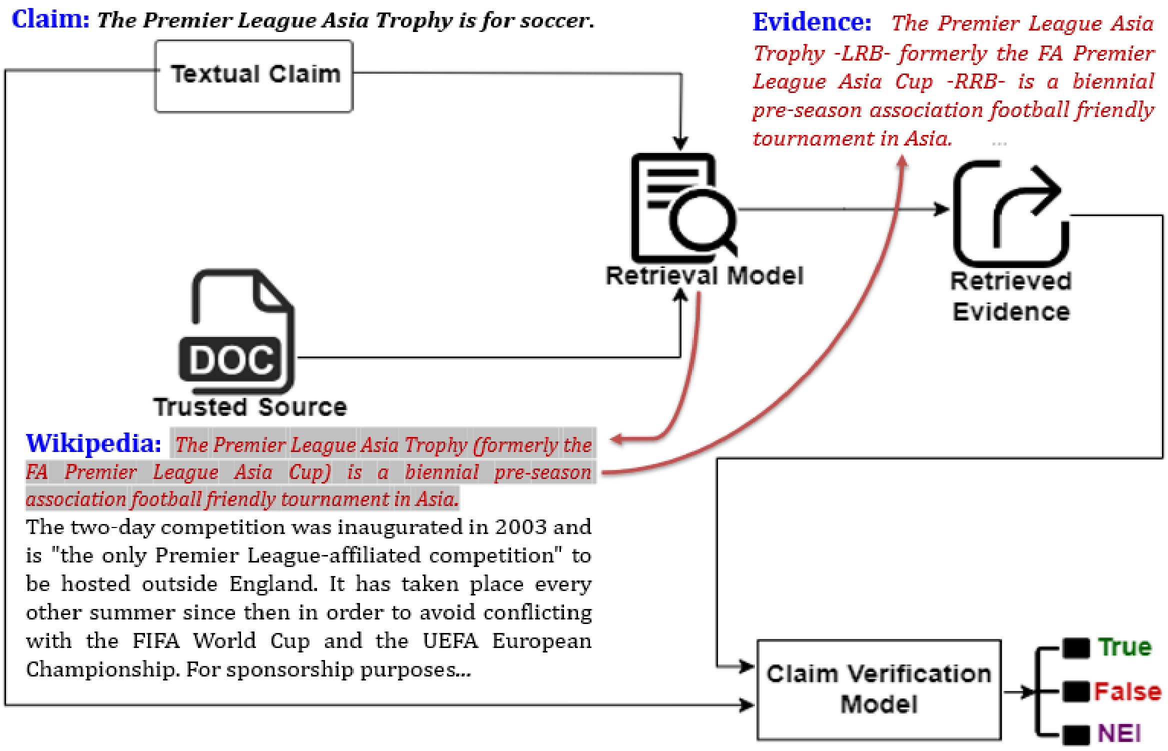


Figure 1.1: An Overview of **FactFormer**

effectiveness of our retrieval model by comparing evidence retrieved from Wikipedia by our model with the manually extracted evidence in the FEVER dataset, and it showed the effectiveness of our approach in extracting relevant textual evidence to given claims. Our second experiment compared our claim verification model to various state-of-the-art claim verification models on the test set of the FEVER dataset, and it demonstrated that our model outperformed all compared-to models, achieving a classification accuracy of more than 90%, with over 12% improvement compared to the best performing state-of-the-art model. Our third experiment was designed to evaluate our end-to-end approach with its two components, and we achieved an overall classification accuracy of over 88% on the test set of the FEVER dataset.

The rest of the paper is organized as follows. In Chapter 2, we review related work on automatic fact-checking. In Chapter 3, we present our proposed fact-checking approach with its two components. the evidence retrieval model and the claim verification model. We then describe our experiments to evaluate our proposed approach and their results in Chapter 4 and we conclude and discuss future directions in Chapter 5.

CHAPTER 2

RELATED WORK

Automatic fact-checking has received significant attention in recent years, leading to the development of various fact-checking approaches. These approaches encompass a range of methods, from simple Information Retrieval (IR) and Natural Language Processing (NLP) approaches all the way to transformer-based deep-learning models. Some of these approaches focused only on retrieving textual evidence for given claims, while others like ours, focused on both textual evidence retrieval and claim verification.

For instance, [18] introduced an evidence retrieval method that uses trusted sources and information retrieval techniques. Their system faces difficulties with analyzing complex languages accurately. Similarly, BEVERS [8] applied document retrieval techniques, combining fuzzy string search with TF-IDF, to better select relevant evidence. They also explored claim verification using Graph Neural Networks (GNNs), transformer models, and a gradient-boosting classifier. The BEVERS pipeline faces issues with phrasing mismatches, synonym detection, unexplored design choices, and untested features. In terms of more defined fact-checking tasks and datasets, [19] provided WIKIFACTCHECK-ENGLISH, sourced from English Wikipedia, that includes claims, their contexts, and annotated evidence. This dataset allows for testing various learning models. Notably, models emphasizing sentiment and opposites performed better than those using artificial data, emphasizing the need for specific techniques for real-world claims. WIKIFACTCHECK-ENGLISH struggles to verify real-world claims because of its feature limitations and the complexity of real data. Meanwhile, [20] developed FacTify, a method for retrieving textual passages based on knowledge graphs. This combines BM25's matching capabilities with the nuance of word embeddings, aiming for accurate and thorough evaluations from various sources. FacTify's limitations encompass non-reliance on primary sources, repetitive facts, varied scoring methods, sourcing from unexpected articles, and potential contextual confusion for users.

In the area of NLP, researchers are keen on automating fact-checking to better understand text meaning. One such method is ProoFVer [14], which focuses on explainability by using seq2seq models to create proofs based on natural logic. These proofs are formed by examining the changes between claims and evidence, with each change being tagged by a specific logic operator to determine if the claim is true. The system is effective in finding similarities between claims and evidence but can struggle when information is limited. While ProoFVer performs well on the FEVER leaderboard, it has challenges, especially in handling certain expressions and the order of its logic steps. Moving to the health sector, [21] explored fact-checking in public health. They developed the "PUBHEALTH" dataset, consisting of 11.8K claims paired with journalist-written explanations. The team introduced two models: one to predict claim truthfulness and another to produce explanations. They ensured the predictions and explanations aligned well, using the ROUGE metric and guidelines from [22], aiming to use this approach in other specialized areas. Their approach's key limitation is its reliance on ROUGE metrics, which may not fully capture the quality or relevance of explanations derived from varied sources. Lastly, Claim-Dissector (CD) [13], a new system for gathering evidence and verifying claims, was presented. This system uses various training methods and hyperlinks, combined with multi-head attention, to provide detailed reasons and assess claims. However, it does face issues like being overly confident and overly dependent on Wikipedia, sometimes neglecting the importance of source reliability. A significant drawback of CD is its propensity to overvalue evidence blocks, which can cause errors, notably in unfamiliar domains, as demonstrated when it misread weapon smuggling situations between Somalia and Estonia in their study.

To better understand the relationship between evidence and claims, researchers are increasingly using graphs to improve fact-checking. One notable method is the Kernel Graph Attention Network (KGAT) [9]. It uses an approach called the Enhanced Sequential Inference Model (ESIM) for sentence retrieval, which compares well with the Graph-based evidence aggregating and reasoning (GEAR) method. Although Bidirectional Encoder Representations from Transformers (BERT) is more effective than ESIM, KGAT proves better than both GEAR and another method called the Graph Attention Network (GAT). KGAT has two specific versions, KGAT-Node for selecting evidence and KGAT-Edge for combining different pieces of evidence. Its focused attention and ability to choose and combine evidence effectively make KGAT stand out. This is evident as KGAT scores better than BERT-base on the FEVER score. KGAT's dot-product attentions are unclear and challenging to interpret, affecting model transparency and performance. In order to solve KGAT problems with certain types of reasoning and issues called over-smoothing, [12] introduced the Semantic-Infused Selective Graph Reasoning (SISER) method. This method focuses on understanding the meaning behind data, using graphs to calculate, and selecting the right pieces of data. It is very good at finding the right evidence and checking if claims are true by comparing them to this evidence. However, SISER has some limitations in its reasoning process, especially when using GNNs. Despite these challenges, test results show that SISER does perform better than KGAT. Nevertheless, SISER did not dwell into an in-depth exploration of node kernels and kernel-based attentions' properties which can improve the performance, and enhance transparency.

In the ongoing efforts to combat misinformation, the introduction of transformers has played a crucial role in improving the performance of fact-checking systems. Leading the way, [10] utilized transformers for fact-checking by adopting methods similar to the UKPAthene approach to gather relevant evidence from Wikipedia. They employed the BERT model to retrieve sentences and used both pointwise and pairwise techniques. To improve their results, they applied hard negative mining. For claim verification, they compared the claim with the top five evidence sentences and combined the outcomes using a three-class classification system. The authors did not study the impact of their work on faster training or using transformers for the whole FEVER pipeline. Subsequently, [23] tackled the complex problem of multi-hop reasoning, essential for verifying political statements and finding supporting evidence. They compared the performance of Transformer-XH multi-hop architecture and BERT using the PolitiHop dataset. The results showed that Transformer-XH generally performed better than BERT. However, when they tested the multi-hop model on larger datasets, its performance slightly decreased. Their method suffers from reduced effectiveness with bigger evidence sets. Expanding on this research, [24] highlighted the importance of verifying claims, especially during critical events like pandemics. They introduced a system named VERISCI, designed for scientific claim verification using the SCIFACT dataset. This system has components like ABSTRACTRETRIEVAL for finding relevant abstracts, RATIONALESELECTION for identifying key rationales, and LABELPREDICTION for deciding claim truthfulness. These components are based on the Robustly Optimized BERT Pretraining Approach (RoBERTa) architecture and aim to provide accurate claim verification by referring to evidence from scientific papers. The SciFact approach risks false negative retrievals in its fact-checking corpus, especially when using broader research datasets like S2ORC, due to potential mismatches between claim sources and newer or overlooked abstracts.

Our approaches (**FactFormer**) stand out from other related works as we utilized a Question and Answer-based model for evidence retrieval, we then used this evidence to classify claims using a multi-headed model that combines transformer and fully-connected layer models. The transformer is utilized for sentence embedding and feature extraction, while the fully-connected layer is used for classification. This combination allows us to capture the semantic and high-level representation between the claim and evidence. Therefore, our method effectively addresses the limitations observed in most of the aforementioned works above, which primarily relate to performance, scalability, and complexity.

CHAPTER 3

APPROACH

Our proposed fact-checking approach, depicted in Figure 1.1, consists of two main components. The first is a transformer-based evidence retrieval model that takes as input a textual claim and a trusted source and retrieves textual evidence that is most relevant to the textual claim from the trusted source. In this paper, we do not address how the trusted source is retrieved given the input textual claim. In principle, any IR model can be used to retrieve such sources from an underlying corpus, which can be Wikipedia, trusted news sources, or even the entire Web. In our experiments and since we rely on the FEVER dataset, we use the Wikipedia article associated with the claim in the dataset. Moreover, we only assume that we extract one textual evidence per claim from one single source, and we assume that the textual evidence is a span of text from the trusted source, so similar in spirit to extractive question answering.

The second component of our fact-checking approach is a transformer-based claim verification model that takes as input the textual claim and the retrieved textual evidence retrieved for that claim using the just-described evidence retrieval model and decides whether the claim is true or false depending on whether the retrieved evidence supports or refutes the claim, respectively. The model is also able to decide that the claim cannot be verified using the retrieved textual evidence, and thus the claim verification model is a three-way classifier with three possible labels: true, false, or not enough information (NEI for short). Our claim verification model relies on one textual evidence to determine whether the claim is true, or false or whether it does not have enough information to determine the validity of the claim.

Our fact-checking approach can thus be generalized by retrieving multiple textual pieces of evidence from multiple trusted sources and using those multiple pieces of evidence together to decide on the validity of a given claim. This is subject to our future work. Next, we describe each component of our fact-checking approach in detail.

3.1 Evidence Retrieval Model

Our evidence retrieval model is inspired by extractive transformer-based question-answering models. However, instead of the model taking in a question and a document that contains the answer to the question, and extracting a span of text from the given document as the answer to the question, our model takes in a claim and a trusted source and extracts a span of text that is most relevant to the claim from the trusted source. Our model architecture is shown in Figure 3.1. As can be seen from the figure, our model is a BERT-based model [25] that uses a pre-trained tokenizer to encode both the claim and trusted source, and where the special tokens [CLS] and [SEP] are used to indicate the start of the claim and the start of the trusted source, respectively. Next, the encoded input sequence is passed through the transformer layers and the dot product is computed between the output embeddings and the start weights. A *softmax* activation is then applied to generate a probability distribution across all words in the trusted source, indicating the likelihood of each word being the start token of the textual evidence for the input claim. Similarly, the end token of the textual evidence is determined by calculating the dot product between the output embeddings and the end weights. The tokens with the highest start and end scores, representing the most probable start and end points of the trusted source are identified and the tokens from the start token to the end token are concatenated, taking into account any sub-word tokenization if necessary. The resulting sequence of tokens is converted back into a string, representing the retrieved textual evidence for the input claim from the underlying trusted source. In our experiment, we model the FEVER dataset to the format of the SQuAD dataset version 2 (v2) [26], the FEVER's claim, trusted source, and evidence are seen as SQuAD's question, context, and answer, respectively. A salient distinction lies in the way each dataset marks the location of the answer or evidence. While SQuAD v2 elucidates the precise character position marking the start of the answer within its context, FEVER uniquely identifies evidence using a sentence ID rather than a character index. To address this distinction, we find the starting position of the evidence in the FEVER dataset's trusted source. We do this by comparing the evidence text to parts of the trusted source. This method allows us to identify the starting position of the evidence as a number, similar to how SQuAD v2 identifies answer start positions. Subsequently, we used this formatted FEVER dataset to fine-tune two transformer models: pre-trained BERT and Bidirectional Encoder Representations from Transformers (DistilBERT) [27]. DistilBERT is a smaller and faster variant of BERT that retains most of its performance. Finally, we compared the performance of these two models.

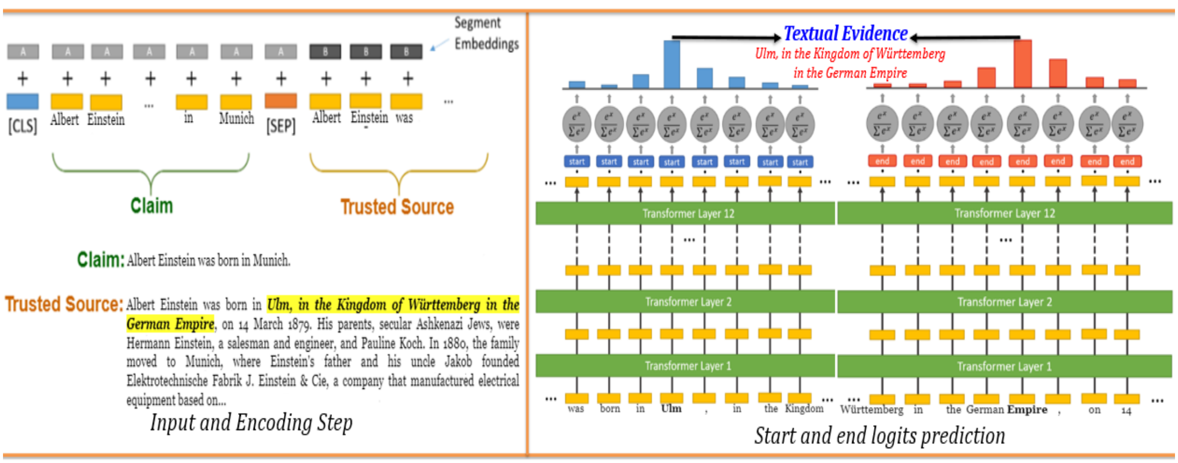


Figure 3.1: Evidence Retrieval Model

3.2 Claim Verification Model

Similar to our evidence retrieval model, our claim verification model is a BERT-based model. Our model takes as input a claim, and the textual evidence retrieved from the trusted source using our evidence retrieval model. The claim and the evidence are then passed through a pre-trained BERT embedding layer.

We utilize a BERT pre-trained model *'bert-base-nli-cls-token'* to generate sentence embeddings for the claim and evidence sentences. BERT has been trained to capture the semantic meaning of individual tokens and the contextual relationships between them. The following are summary of the embedding steps:

✓ **Token Embeddings:** Each token is mapped to its corresponding token embedding, which captures the semantic meaning of the individual tokens.

✓ **Transformer Encoder:** The token embeddings are fed into a transformer encoder, consisting of self-attention and feed-forward layers, which capture contextual relationships and extract high-level representations. This transformer encoder helps to capture the overall sentences' semantic meaning.

✓ **Pooling Strategy:** To obtain a fixed-dimensional representation of the input sequence. The [CLS] token, as well as the output representation from the transformer encoder, is used to create a sentence-level representation. This pooling operation reduces the dimensionality of the [CLS] token representation from *512x768* to *1x768*.

✓ **Output Layer:** The [CLS] token serves as the final embedding for the input sentence pair, encapsulating the overall semantic meaning of the two sentences together. This representation is used for further processing, such as claim-evidence verification and classification.

In our approach, as illustrated in Figure 3.2, we use the final embeddings obtained from the claim and evidence sentences of the output layer. These embeddings serve as the input for performing element-wise addition, subtraction, and multiplication operations between the claim and evidence variables. The resulting addition, subtraction, and multiplication variables are then concatenated with the original claim and evidence variables, forming the features array.

To process the features array, we used a fully-connected layer architecture, which consists of three components: an input layer, a hidden layer, and an output layer. The input layer takes the concatenated features array as input. The hidden layer comprises 300 neurons that perform computations and transformations on the input data (training and test sets), capturing complex patterns and relationships within the features. The output layer (prediction in Figure 3.2) of the fully-connected network varies depending on the type of classification problem. For binary classification, the output layer consists of a single neuron with a sigmoid activation function. This neuron outputs a value between 0 and 1, representing the probability of the positive class. In the case of multiclass classification, the output layer contains neurons equal to the number of classes. This highlights the robustness of our classifier when dealing with various categories of labels. Furthermore, the activation function used in this scenario is *softmax*, which outputs a vector of class probabilities that sum up to 1.

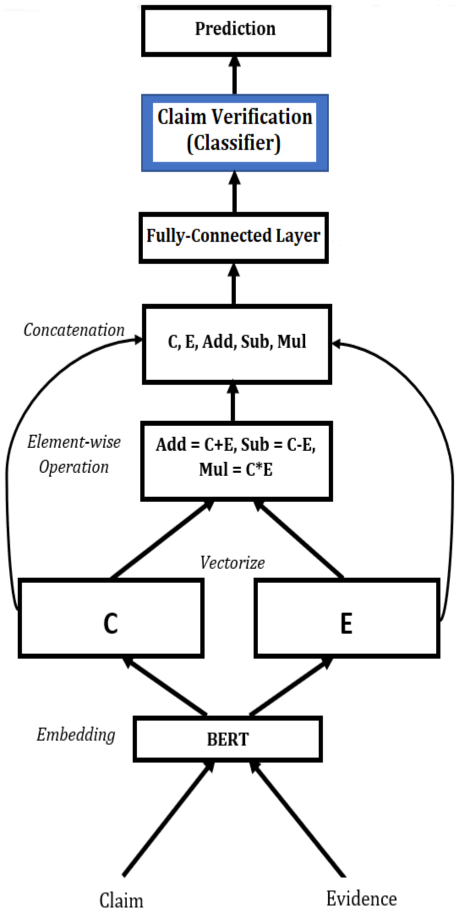


Figure 3.2: Claim Verification Model Architecture

CHAPTER 4

EXPERIMENTS

In this Chapter, we first describe the FEVER dataset we used to train and test our fact-checking approach with its models: the evidence retrieval model and the claim verification model. We next explain how we trained those two models using the FEVER dataset and we provide the test results for each model separately. Next, we explain how we evaluated the whole approach in an end-to-end fashion and we provide some error analysis for the whole approach. All our training and test jobs were conducted on a single Alienware Aurora RyzenTM Edition R14 Gaming Desktop with an NVIDIA®GeForce RTXTM 3090 super GPU, an AMD®RyzenTM 9 5950X up to a 4.90 GHz CPU (16 cores, 32 threads), and 128 GB of RAM.

4.1 Dataset

To train our two models, we used the FEVER dataset, a standard and heavily used fact-checking benchmark. This dataset consists of over 185,000 short textual claims. Each claim in the dataset is associated with textual evidence and one of three labels, indicating whether the textual evidence supports, refutes, or provides insufficient information to draw a conclusion about the validity of the claim. In the first two cases, the claim is also associated with the Wikipedia article from which the textual evidence was extracted. In the third case where the textual evidence does not provide sufficient information to judge the validity of the claim, there is no Wikipedia article associated, as the textual evidence was generated manually to act as evidence that lacks sufficient information to judge the validity of the claims. The dataset is already split into 145,449 training and 19,998 test sets and we used those to train and test our models, respectively.

4.2 Models Training and Testing

4.2.1 *Evidence Retrieval Model*

In practice, our **retrieval models** should retrieve evidence for the SUPPORTS, REFUTES, and NEI labels from the Wikipedia dataset. However, we did not retrieve evidence for claims labeled as NEI because the Wikipedia corpus does not contain sufficient information to support or refute such claims. This is in line with the design of the FEVER dataset, where the NEI label is applied to claims that cannot be verified due to the absence of relevant evidence in the available Wikipedia snapshot. In view of the aforementioned reason, the size of our retrieval models' data was reduced to 121,193 for training, 7,736 for development, and 7,736 for test sets, focusing on the binary classes of SUPPORTS and REFUTES. While converting the dataset from FEVER to SQuAD format, the dataset shrank to 72,455 training and 4,224 development sets because of challenges in determining the evidence start positions for some claims and trusted sources. This refined dataset became the foundation for fine-tuning our models for training and validation. Finally, the trained models were employed to retrieve test data, which was then used for claim verification in our end-to-end system evaluation. As outlined in Chapter 3.1, we fine-tuned two transformer-based retrieval models: BERT and DistilBERT, both in their uncased base configurations. During their fine-tuning with the retrieval data, we used the Adam optimizer with standard beta values (beta 1 = 0.9, beta 2 = 0.999) and an epsilon of 1e-08. We ran the experiment on 3 epochs and the learning rate started at 2e-05 and decayed polynomially over 13,587 steps. We avoided techniques like weight decay, gradient clipping, and AMSGrad. Just-In-Time compilation was enabled for efficiency, and we chose mixed floating-point 16 to achieve a balance between computational speed and model accuracy.

4.2.2 *Claim Verification Model*

Our claim verification model was trained using the training set of the FEVER dataset with the following hyperparameters: RMSprop optimizer, learning rate of 0.001, dropout rate of 0.2, 20 epochs, verbosity level of 1, and a batch size of 1024. The model's fully-connected layer consists of 300 neurons, which are initialized using a uniform kernel initializer and activated using the ReLU activation function. For binary classification, the model employs binary crossentropy as the loss function, has 1 neuron in the output layer with a sigmoid activation function, and a uniform kernel initializer. For multiclass classification, the model utilizes categorical crossentropy as the loss function, the number of neurons in the output layer is determined by the number of classes, and the activation function used is softmax.

4.3 Result

4.3.1 *Retrieval Model Result*

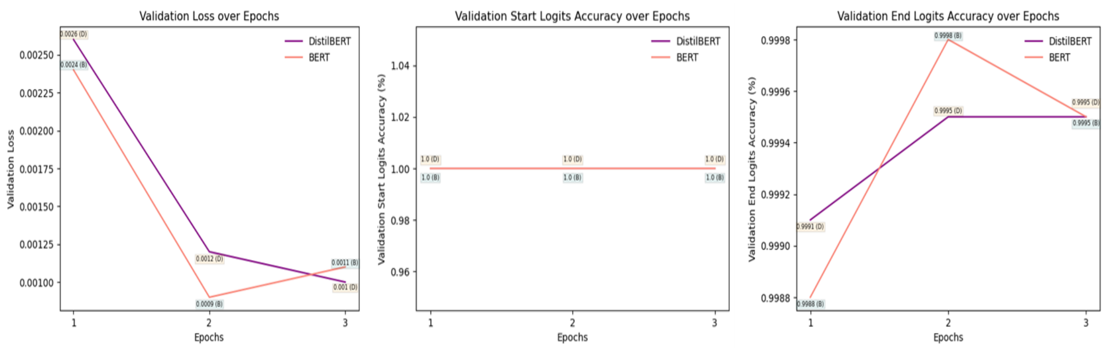
In Chapter 4.2.1, we discussed how we arrived at the validation data and we will further expand on the evaluation of our retrieval task using this data. Our evidence retrieval evaluation employed the exact match and F1-measure, commonly utilized in question-answering tasks. We also present results such as training duration across the three epochs, epoch-wise loss, and start and end logits accuracy to provide a comprehensive view of our models' performance during training and validation. The training duration for our retrieval models was significantly shorter and faster with DistilBERT, completing in 1250 seconds, compared to BERT, which took 3327 seconds. This increased efficiency of DistilBERT can be attributed to its reduced number of parameters, streamlined architecture, optimized computations, and quicker convergence, among other factors.

During training, as showcased in Figure 4.1a, BERT achieved a superior average loss value of 0.0101, surpassing DistilBERT's 0.0132. BERT's intricate design, enriched with more parameters led to a closer fit to the training data. The training accuracies for the start and end logits are presented in Figure 4.1b and 4.1c. BERT with 99.88% marginally outperformed DistilBERT with 99.84% by 0.04% for the start logits. However, for the end logits, DistilBERT with an average of 99.89% slightly performs better than BERT with 99.81% by 0.08%. Subsequently, our validation results further corroborate

(a) (b) (c)

Figure 4.1: (a) Training Loss (b) Training Evidence Start Accuracy (c) Evidence End Accuracy

these performance differentials, Figure 4.2a highlights that BERT, with an average loss of 0.0015, marginally outperformed DistilBERT, which had a loss of 0.0016. Both models demonstrated exemplary validation accuracy for the start and end logits, reaching 100% and 99.94%, respectively, as depicted in Figure 4.2b and 4.2c.



(a) (b) (c)

Figure 4.2: (a) Validation Loss (b) Validation Evidence Start Accuracy (c) Validation Evidence End Accuracy

We also delve into analyzing the exact match scores and F1 measures as functions of varying maximum evidence lengths, ranging from 30 to 130, for both the BERT and DistilBERT models. Our findings indicate that increasing the maximum length of retrieved evidence generally improves both exact match and F1 measure scores. Interestingly, we observed that exact match scores plateau at a maximum evidence length of 90, while F1 measure scores remain relatively stable from a length of 110 onwards. Based on these findings, we chose a maximum evidence length of 100, which yielded an exact match score of 90.19% for DistilBERT and 89.89% for BERT, as well as an F1 measure of 93.98% for DistilBERT and 93.93% for BERT, respectively.

4.3.2 *Claim Verification Model Result*

In our experiments on the FEVER-3 dataset, claim verification model achieved an accuracy of 90.23%, with macro-average and weighted-average scores of 89.69% and 90.03% respectively (Table 4.1).

Model / Labels	Claim Verification		
	Precision	**Recall**	**F1-Score**
SUPPORTS	**0.8960**	0.9393	**0.9171**
REFUTES	0.8272	0.7271	0.7739
NEI	**0.9997**	**0.9997**	**0.9997**
Accuracy	**0.9023**		
Macro Average	**0.9076**	**0.8887**	**0.8969**
Weighted Average	**0.9004**	**0.9023**	**0.9003**

Table 4.1: Claim Verification for FEVER-3

To compare our performance with other baselines, we evaluated the accuracy of our label predictions against two prominent FEVER score accumulators, which incorporate the latest research papers and results from the FEVER leaderboard (Tables 4.2 and 4.3). Based on the results in Table 4.1, our model demonstrated 90% state-of-the-art accuracy score and came first among over 100 submissions to the FEVER CodaLab competition [28] (although we took the top eight and our model in Table 4.2). Furthermore, our model outperformed all other models in the FEVER leaderboard [29] (also we took the top three and our model in Table 4.3).

Our findings highlight the competitive label accuracy of our model, positioning it as a leading performer in the field of fact-checking on the FEVER dataset.

4.4 End-to-End System Evaluation

Our comprehensive evaluation involved assessing the FEVER-2 dataset, which comprises over 7,000 test entries. Each entry contains a claim and corresponding evidence, retrieved as detailed in Chapter 4.2.1. This dataset includes two label categories: SUPPORTS and REFUTES. Our primary objective was to compare and contrast the claims and labels automatically classified by our model, along with the corresponding evidence, against

S/N	Other FEVER Works	Label Accuracy	Year
1.	**Our Model (CVM)**	**90.23%**	2023
2.	BEVERS [8]	80.24%	2023
3.	[15]	80.06%	2022
4.	ProoFVer [14]	79.47%	2022
5.	[13]	79.27%	2023
6.	SISER [12]	77.5%	2022
7.	[7]	76.85%	2019
8.	[9]	74.07%	2019
9.	[10]	71.86%	2020

Table 4.2: Result Comparison for Label Accuracy

S/N	FEVER Leaderboard [29]	Label Accuracy	Year
1.	**Our model (CVM)**	**90.23%**	2023
2.	DREAM [7]	76.85%	2019
3.	RoBERTa-Base Joint MSPP [11]	74.4%	2022
4.	KGAT [9]	74.1%	2019

Table 4.3: FEVER Leaderboard Score

those in the manually annotated original FEVER dataset. To this end, we used the evidence retrieved by our model to classify each claim into one of these binary labels. The evaluation aimed to quantify the system's accuracy in label classification and assess its suitability for real-world applications.

The findings of our experiments, as presented in Table 4.4, reveal that DistilBERT achieved an accuracy of 87.14% for label classification, outperforming BERT's performance with 84.18% on the FEVER-2 dataset. This result aligns with the exact match and F1-measure scores obtained from the evidence retrieval task in Section 4.3.1, where DistilBERT slightly outperformed BERT. However, our evaluation revealed that, when utilizing evidence retrieved via our DistilBERT model, the claim verification model exhibited better performance compared to the manually annotated evidence with a margin of 0.6% in the FEVER-2 classification (as shown in Tables 4.4 and 4.5). Consequently, these results substantiate that our automated evidence retrieval methodology, which employs the DistilBERT model, enhances the accuracy of label classification within the FEVER dataset relative to manually annotated evidence. This evaluation provided valuable insights into the performance of our system and its ability to accurately classify FEVER labels using retrieved evidence. The results highlight the potential of our approach in addressing real-world challenges in fact-checking systems that require effective label classification.

Labels \ Models	BERT			DistilBERT		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
SUPPORTS	**0.9210**	0.8518	**0.8851**	0.9015	0.9208	0.9110
REFUTES	0.6869	**0.8166**	**0.7462**	0.7898	0.7472	0.7679
Accuracy	0.8418			**0.8714**		
Macro Average	0.8040	**0.8342**	**0.8156**	0.8456	0.8340	0.8395
Weighted Average	0.8544	0.8418	**0.8455**	0.8697	0.8714	0.8703

Table 4.4: Claim Verification for our FEVER-2 Automatic Retrieved Evidence

Labels \ Models	FEVER Manual Evidence		
	Precision	Recall	F1-Score
SUPPORTS	0.9114	0.8993	0.9053
REFUTES	0.7553	0.7803	0.7676
Accuracy	**0.8654**		
Macro Average	0.8333	0.8398	0.8364
Weighted Average	0.8669	0.8654	0.8661

Table 4.5: Claim Verification for FEVER-2 Manually Annotated Evidence

4.5 Error Analysis

Figure 4.3 visually presents examples of our models, highlighting the evidence retrieval process using BERT and DistilBERT, as well as the claim classification performed by the claim verification based on the retrieved evidence. Our models demonstrated the ability to retrieve relevant evidence and classify claims accurately into appropriate labels.

However, Figure 4.4 showcases instances of misclassification by our models, with the retrieval task being the most prominent source of errors. We observed that some evidence from Wikipedia is outdated and contains newer updates, as illustrated in Figure 4.4. For example, Liverpool was valued at $1.55 billion according to our retrieved evidence, whereas the updated information on Wikipedia lists the value at $1.944 billion. Additionally, we encountered issues such as retrieved evidence being short and incorrect, leading to misclassification. Other noticeable problems include missing evidence in the trusted source and difficulties in locating evidence start during the dataset conversion from FEVER to SQuAD format.

Claim: *roald dahl is a writer* **True Label:** *SUPPORTS* **Predicted Label:** *SUPPORTS*

Trusted Source: *roald dahl 13 september 1916 23 to november 1990 was a british novelist writer poet screenwriter and wartime fighter ace of norwegian descent. his books have sold more than 250 million copies worldwide. dahl has been called one of the greatest storytellers for children of the 20th century. dahl was born in wales to affluent norwegian immigrant parents and spent most of his life in england. he served in the royal air force raf during the second world war...* *(512 tokens)*

BERT Retrieved Evidence: *british novelist writer poet screenwriter and wartime fighter ace of norwegian descent*

DistilBERT Retrieved Evidence: *roald dahl 13 september 1916 to 23 november 1990 was a british novelist writer poet screenwriter and wartime fighter ace of norwegian descent*

FEVER Manual Evidence: *roald dahl 13 september 1916 23 november 1990 was a british novelist short story writer poet screenwriter and fighter pilot*

Claim: *anderson silva was born on the day of the 15th* **True Label:** *REFUTES* **Predicted Label:** *REFUTES*

Trusted Source: *anderson da silva portuguese pronunciation des siwv born 14 april 1975 is a mixed martial artist and boxer he is a former ufc middleweight champion and holds the record for the longest title reign in ufc history at 2457 days...* *(512 tokens)*

BERT Retrieved Evidence: *14 april 1975*

DistilBERT Retrieved Evidence: *born 14 april 1975 mixed martial artist and boxer*

FEVER Manual Evidence: *anderson da silva born april 14 1975 is a brazilian mixed martial artist and former ufc middleweight champion middleweight mma mixed martial artist mixed martial arts ufc middleweight champion*

Figure 4.3: FactFormer System Samples for Retrieval and Claim Verification Models

Claim: *liverpool fc was valued at $1.55 billion* **True Label:** *SUPPORTS* **Predicted Label:** *REFUTES*

Trusted Source: *according to a 2018 report by deloitte the club had an annual revenue of €424.2 million for the previous year and forbes valued the club at $1.944 billion. in 2018 annual revenue increased to €513.7 million, and forbes valued the club at $2.183 billion. in 2019 revenue increased to €604 million £533 million according to deloitte with the club breaching the half a billion pounds mark in april 2020 the owners of the club came under fire from fans and the media for deciding to furlough all non playing staff during the covid 19 pandemic. in response to this, the club made a u turn on the decision and apologised for their initial decision. in april 2021 forbes valued the club at $4.1 billion, a two year increase of 88%, making it the worlds fifth most valuable football club...* *(512 tokens)*

BERT Retrieved Evidence: *$1.944 billion*

DistilBERT Retrieved Evidence: *forbes valued the club at $1.944 billion in 2018 annual revenue increased to €513.7 million and forbes valued the club at $2.183 billion*

FEVER Manual Evidence: *liverpool was the ninth highest-earning football club in the world for 2014 15 with an annual revenue of $ 391 million and the worlds eighth most valuable football club in 2016 valued at $ 1.55 billion liverpool ninth highest-earning deloitte football money league eighth most valuable forbes list of the most valuable football clubs*

Claim: *marilyn monroe was a part of the war effort* **True Label:** *SUPPORTS* **Predicted Label:** *REFUTES*

Trusted Source: *marilyn monroe born norma jeane mortenson june 1 1926 august 4 1962 was an american actress model and singer known for playing comic blonde bombshell characters she became one of the most popular sex symbols of the 1950s and early 1960s as well as an emblem of the eras sexual revolution she was a top billed actress for a decade and her films grossed $200 million equivalent to $2 billion in 2022 by the time of her death in 1962 long after her death monroe remains a pop culture icon in 1999 the american film institute ranked her as the sixth greatest female screen legend from the golden age of hollywood born and raised in los angeles, monroe spent most of her childhood in a total of 12 foster homes and an orphanage before marrying james dougherty at age sixteen...* *(512 tokens)*

BERT Retrieved Evidence: *an emblem of the eras sexual revolution*

DistilBERT Retrieved Evidence: *actress model and singer known for playing comic blonde bombshell characters she became one of the most popular sex symbols of the 1950s and early 1960s as well as an emblem of the eras sexual revolution*

FEVER Manual Evidence: *while working in a factory in 1944 as part of the war effort she was introduced to a photographer from the first motion picture unit and began a successful pin-up modeling career war effort united states home front during world war ii first motion picture unit*

Figure 4.4: Error Samples from FactFormer

CHAPTER 5

CONCLUSION AND FUTURE WORK

We introduce **FactFormer**, an advanced fact-checking system that leverages finetuned BERT and DistilBERT models for accurate evidence retrieval while our claim verification approach combines BERT for sentence embedding and a fully-connected layer for claim classification, resulting in superior performance compared to other baseline models in terms of label accuracy. Specifically, our claim verification model exhibits notable enhancements over the DistilBERT retrieval model when applied to the FEVER-2 dataset, achieving accuracy levels that outperformed the manually annotated data result. These results highlight the potential of our models for automatic fact-checking.

Our method, designed to address the challenge of combating misinformation, is not limited to specific domains and can be applied to various datasets, including medical, politics, news, health, and more. This adaptability paves the way for enhanced fact-checking systems across diverse domains.

In future research, we plan to expand our approach by incorporating multiple pieces of evidence to determine the final label classification. Additionally, we intend to test our models on different datasets to assess their effectiveness and generalizability.

Bibliography

[1] W.-Y. Sylvia Chou, A. Gaysynsky, and J. N. Cappella, "Where we go from here: Health misinformation on social media," *American Journal of Public Health*, vol. 110, no. S3, Oct. 2020. DOI: 10.2105/ajph.2020.305905.

[2] S. Pappas, *Fighting fake news in the classroom*, Jan. 2022. [Online]. Available: https://www.apa.org/monitor/2022/01/career-fake-news.

[3] M. L. Jenna, *How Fake News Affects U.S. Elections | University of Central Florida News — ucf.edu*, https://www.ucf.edu/news/how-fake-news-affects-u-s-elections/, [Accessed 14-Jul-2023], 2020.

[4] ADL, *The Dangers of Disinformation — adl.org*, https://www.adl.org, [Accessed 12-Jul-2023], 2022.

[5] WHO, *Infodemics and misinformation negatively affect people's health behaviours, new who review finds*, Sep. 2022. [Online]. Available: https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds.

[6] E. Shields, *Deepfakes and Disinformation*, https://mediajustice.org/news, [Accessed 14-Jul-2023], 2020.

[7] W. Zhong, J. Xu, D. Tang, *et al.*, "Reasoning over semantic-level graph for fact checking," *arXiv preprint arXiv:1909.03745*, 2019.

[8] M. DeHaven and S. Scott, "Bevers: A general, simple, and performant framework for automatic fact verification," *preprint arXiv:2303.16974*, 2023.

[9] Z. Liu, C. Xiong, M. Sun, and Z. Liu, "Fine-grained fact verification with kernel graph attention network," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7342–7351. DOI: 10.18653/v1/2020.acl-main.655. [Online]. Available: https://aclanthology.org/2020.acl-main.655.

[10] A. Soleimani, C. Monz, and M. Worring, "Bert for evidence retrieval and claim verification," in *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, Springer, 2020, pp. 359–366.

[11] L. Di Liello, S. Garg, L. Soldaini, and A. Moschitti, "Paragraph-based transformer pre-training for multi-sentence inference," *arXiv preprint arXiv:2205.01228*, 2022.

[12] E. Park, J.-H. Lee, J. D. Hyeon, S. Kim, I. Kang, and S.-H. Na, "Siser: Semantic-infused selective graph reasoning for fact verification," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 1367–1378.

[13] M. Fajcik, P. Motlicek, and P. Smrz, "Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 184–10 205. DOI: 10.18653/v1/2023.findings-acl.647. [Online]. Available: https://aclanthology.org/2023.findings-acl.647.

[14] A. Krishna, S. Riedel, and A. Vlachos, "Proofver: Natural logic theorem proving for fact verification," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1013–1030, 2022.

[15] P. S. H. Lewis, "Improving neural question answering with retrieval and generation," Ph.D. dissertation, UCL (University College London), 2022.

[16] R. Casillas, H. Gómez-Adorno, V. Lomas-Barrie, and O. Ramos-Flores, "Automatic fact checking using an interpretable bert-based architecture on covid-19 claims," *Applied Sciences*, vol. 12, no. 20, p. 10 644, 2022.

[17] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: A large-scale dataset for fact extraction and verification," *arXiv preprint arXiv:1803.05355*, 2018.

[18] I. Touahri and A. Mazroui, "Automatic verification of political claims based on morphological features.," in *CLEF (Working Notes)*, 2019.

[19] A. Sathe, S. Ather, T. M. Le, N. Perry, and J. Park, "Automated fact-checking of claims from Wikipedia," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 6874–6882, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.849.

[20] G. Ercan, S. Elbassuoni, and K. Hose, "Retrieving textual evidence for knowledge graph facts," in *The Semantic Web*, P. Hitzler, M. Fernández, K. Janowicz, *et al.*, Eds., Cham: Springer International Publishing, 2019, pp. 52–67, ISBN: 978-3-030-21348-0.

[21] N. Kotonya and F. 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. [Online]. Available: https://aclanthology.org/2020.emnlp-main.623.

[22] K. Sokol and P. Flach, "Desiderata for interpretability: Explaining decision tree predictions with counterfactuals," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 10 035–10 036, Jul. 2019. DOI: 10.1609/aaai.v33i01.330110035. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5154.

[23] W. Ostrowski, A. Arora, P. Atanasova, and I. Augenstein, "Multi-hop fact checking of political claims," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed., Main Track, International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 3892–3898. DOI: 10.24963/ijcai.2021/536. [Online]. Available: https://doi.org/10.24963/ijcai.2021/536.

[24] D. Wadden, S. Lin, K. Lo, *et al.*, "Fact or fiction: Verifying scientific claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 7534–7550. DOI: 10.18653/v1/2020.emnlp-main.609. [Online]. Available: https://aclanthology.org/2020.emnlp-main.609.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] R. Pranav, Z. Jian, L. Konstantin, and L. Percy, *SQuAD: 100,000+ Questions for Machine Comprehension of Text — arxiv.org*, https://arxiv.org/abs/1606.05250, [Accessed 13-Jul-2023], 2016.

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[28] CodaLab, *CodaLab Competition*, competitions.codalab.org, [Accessed 02-09-2023], 2018.

[29] Leaderboard, *Papers with code - fever benchmark (fact verification) — paperswithcode.com*, https://paperswithcode.com/sota/fact-verification-on-fever, [Accessed 02-09-2023], 2019.