COMP90042 Project Report: Automated Fact Checking For Climate Science Claims

Student ID: 1326642

1 Introduction

Misinformation is a pervasive problem in the digital age. Fact-checking is an NLP task that aims to assess the veracity of a claim using evidence. We examined different methods for evidence retrieval (Section 2) and label prediction (Section 3). We encoded the claim and evidence texts, selected the most relevant evidence using a similarity metric, and predicted the final label using a classifier or a dense retriever. In this report, we explain the methods we explored, justify our decisions, and compare their performance. The final approach is discussed in Section 4, where we address the limitations and challenges of the current model and suggest future improvements.

2 Evidence Retrieval

The task of evidence retrieval presents significant challenges due to the enormity of our corpus and the inherent ambiguity within the evidence data. The evidence corpus, boasting more than 1.2 million records, necessitates an efficient encoding strategy. While advanced models such as BERT-large demonstrate high performance, they may not be suitable for this context due to their extensive parameter count. In this section, we delve into a two-fold approach: classical and transformer-based approach, aiming to leverage the strengths of both strategies for effective evidence retrieval.

2.1 Classical Approach

2.1.1 Data Preprocessing and Tokenization

The essential part of our classic approach lies in efficient data preprocessing and tokenization, especially for handling a massive corpus. Our preprocessing begins with eliminating non-digital and non-alphabetic characters, transforming all letters to lowercase, and removing stopwords. These steps were aimed to remove data noise and reduce computation costs.

However, it's worth noting that data preprocessing can significantly impact evidence retrieval performance. Strict preprocessing techniques may accidentally remove critical information, resulting in relevant evidence appearing dissimilar to the claim text. Furthermore, such techniques can overly truncate evidence, resulting in high similarity scores due to their reduced length.

For traditional models, we utilized the word_tokenize function from the Natural Language Toolkit (NLTK) library, chosen for its efficiency in parsing text data.

2.1.2 Encoding

In terms of encoding, we evaluated two widely-used traditional techniques typically employed in text analysis: Term Frequency-Inverse Document Frequency (TF-IDF) from Scikit-learn and Word2Vec from Gensim. These methods aim to convert textual data into numerical vectors which provides convenience for follow-up operations.

TF-IDF is a straightforward method that calculates the importance of a word in a document based on its frequency and inverse document frequency, which makes it effective for keyword-based tasks. However, it only considers the frequency of words and their importance in a specific document, thus ignoring the semantic meaning, and also suffers from unseen words or overlappings.

Word2Vec could capture semantic relationships and similarities between words. It can handle complex linguistic patterns and word associations.

2.2 Transformer-based Approach

Transformer-based models like BERT and its variants offer context-sensitive word embeddings. The model was Pre-trained on large-scale corpora, which allow embeddings to encapsulate semantic and syntactic information, and it can be tuned further on our data for task-specific optimization.

Those models are usually computationally expensive. However, there are still limitations in capturing semantic similarity, which is elaborated in Section 4.

The experimented transformer models are listed below, their detailed performance will be discussed in the next section.

- **BERT:** BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model pre-trained on a large text corpus and can be fine-tuned for various NLP tasks (**?**).
- **RoBERTa:** RoBERTa is an optimized version of BERT which is trained on more data and with larger batch size. It generally outperforms BERT on various NLP tasks and is available in base and large sizes. (**?**)
- **DistilBERT:** DistilBERT is a smaller, more efficient version of BERT, created by distilling the knowledge of BERT into a smaller model. It provides a good trade-off between performance and speed/memory requirements. (**?**)

2.3 Similarity Metrics

The choice of similarity metric is critical when retrieving the most relevant evidence. Different metrics may produce different results and have different levels of computational efficiency. In this section, we evaluate several similarity measures widely used in natural language processing and information retrieval tasks.

- **Cosine Similarity:** Cosine similarity measures the cosine of the angle between two vectors, which in our case are the vector representations of the claim and evidence text. This measure is effective when the magnitude of the vectors may not be indicative of their similarity, which makes it a popular choice for NLP tasks since word embedding does not have informative magnitudes. Cosine similarity is computationally efficient, especially when dealing with sparse vectors such as TF-IDF representations.
- Word Mover's Distance : WMD (WMD) is a novel distance function between text documents that leverages word embeddings to measure the semantic similarity between words(**?**).

  The WMD distance measures the minimum amount of distance that the embedded words of one document need to travel to reach the embedded words of another document.
- **Siamese Network :** Siamese network (SN) is a unique architecture of neural networks designed for tasks that involve finding the similarity or the relationship between two comparable things(**??**). The architecture comprises two identical subnetworks that share the same parameters and weights. Each subnetwork processes one of the two inputs, after which a distance metric, in our case cosine similarity, is applied to the outputs of the two subnetworks to calculate the similarity between the inputs.

While WMD and SN are theoretically more robust and likely to retrieve high-quality evidence, they share common crucial limitations. Firstly, their performance is contingent upon a well-performing word embedding model and can suffer significantly if this requirement is not met. Furthermore, these methods are computationally expensive, which can be a significant hurdle when processing large data sets.

Our proposed solution involves adopting a hybrid approach. Initially, we employ efficient methods like the TF-IDF vectorizer and cosine similarity to retrieve a subset of potentially relevant evidence. Subsequently, we apply more complex and robust methods such as WMD and Siamese Networks on this reduced set. This strategy significantly decreases the corpus size, making these computationally intensive models feasible in practice. Moreover, it combines the efficiency of traditional models with the potential of advanced techniques to yield high-quality evidence, potentially enhancing the overall performance of our evidence retrieval system.

3 Classification

Classification is the heart of machine learning, but it is not the main focus of our project. Label prediction relies on the relevance and accuracy of the evidence obtained by the evidence retrieval model. In this section, we will discuss the criteria for choosing a classification model. The inputs are identical - the concatenated embeddings of the claim text and all the ground truth evidence generated by fine-tuned BERT. These models are considered under

Model	Key Parameter	Accuracy
SVM	Kernel = rbf	0.60
LR	Max Iterations = 500	0.4481
RF	Num of Estimators = 100	0.4416
RNN	Dropout Rate = 0.3	0.4351
LSTM	LSTM Units = 50	0.4512

Table 1: Label prediction accuracy of classification models. Only key hyperparameters are shown.

the presumption that the evidence retrieval model has provided the most relevant and accurate evidence. Thus, the main objective is to discern the model that best classifies the relationship between the claim and the provided evidence.

3.1 Traditional Machine Learning Models

We began by evaluating traditional machine learning models for multi-class classification tasks. Grid search was conducted for popular sentiment classification models, including Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), and Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells (**?**). Results along with parameters for each model are shown in Table 1. The SVM model with a Radial Basis Function (RBF) kernel achieved the highest performance, outperforming the other models.

3.2 Pretrained Transformer Models

For the transformer-based approach, we explored fine-tuned pre-trained models such as BertForSequenceClassification to predict claim labels. The performance of this model is listed separately in Table 2. Our experiments show that these models perform much worse than ideal, and several possible reasons for this will be discussed in Section 4.

3.3 Empirical Weighted Vote

In addition to the classification methods we experimented above, we proposed Empirical Weighted Vote (EWV). For each retrieved evidence, it concatenates the claim and the evidence, then fed it into a classifier to predict the probabilities of the claim being supported or refuted. The weight for each piece of evidence is updated by adding the product of the predicted probability and its similarity with the claim.

After assessing all the pieces of evidence, the algorithm checks if both the support and refute weights are below a predefined threshold t. If they are, it concludes that there is not enough information to determine the veracity of the claim. If both the support and refute weights are above t, it checks the absolute difference between them. If this difference is below t, it indicates a dispute. If the support weight is greater than the refute weight, the algorithm indicates that the claim is supported and vice versa. In this framework, t serves as a hyperparameter that requires tuning to achieve an optimal value. A smaller t value will increase the likelihood of support or refute. Therefore, the choice of t should balance the need for decisive claim verification and the risk of overconfidence when evidence is insufficient or conflicting.

4 Results and Discussion

The results of our exploration across the various techniques on both the development and test datasets are presented in Tables 2 and 3. Our experiments have highlighted the Support Vector Classifier (SVC) as the best-performing classifier, with Word Mover's Distance (WMD) in conjunction with Word2Vec proving to be the most effective approach for evidence retrieval. While transformer-based models appear to lag behind in performance for both evidence retrieval and label prediction tasks, the potential of these models should not be underestimated. With more computational resources dedicated to extensive fine-tuning, it is plausible that these models could exhibit improved performance. Similarly, the Siamese network, despite being theoretically optimal for evidence retrieval, did not perform ideally. It suffered from the quality of the BERT embeddings used as input. Given the high computational demand of the Siamese network, it is not feasible to enhance its performance within our current resource constraints. Future work could focus on improving the quality of the input embeddings or exploring more efficient architectures. In this section, we primarily discussed two key challenges.

- **Semantics completeness:** The model does not detect that a sentence is semantically complete despite being grammatically correct. To better illustrate the challenges of our current approach, we list a statement and two examples of potential evidence:

  1. The science is clear, climate change is making extreme weather events, including tornadoes, worse.

ID	Tokenizer & Encoder	Similarity Metric	Classifier	F-score	Accuracy	HMean
Model 1	word_tokenize & TF-IDF	Top 5 Cosine	SVM	0.0738	0.4351	0.1219
Model 2	word_tokenize & TF-IDF	Top 10 Cosine	SVM	0.0724	0.4481	0.1209
Model 3	word_tokenize & TF-IDF	Cosine	SVM+EWV	0.0647	0.2662	0.1042
Model 4	Keras Tokenizer	Top 5 Cosine	LSTM	0.0923	0.4481	0.1594
Model 5	word2Vec	WMD	SVM	0.0967	0.4221	0.1645
Model 6	Pretrained Bert	Top 10 Cosine	SVM	0.0738	0.4351	0.1219
Model 7	Finetuned Bert	Top 10 Cosine	SVM	0.0546	0.4416	0.0972
Model 8	Finetuned Bert	Top 10 Cosine	Auto_Seq	0.0546	0.5454	0.0984
Model 9	Finetuned Roberta	Top 10 Cosine	Auto_Seq	0.0585	0.4342	0.1048
Model 10	Finetuned DistilBERT	Top 10 Cosine	Auto_Seq	0.0412	0.3636	0.0729
Model 11	Finetuned Bert	Siamese	Auto_Seq	0.0196	0.4285	0.0376

Table 2: Performance comparison of different methods on the dev set. Here $Auto\_Seq$ stands for AutoModelForSequenceClassification from transformers, which is a universal class holding different pre-trained models

ID	F-score	Accuracy	HMean
Model 1	0.0727	0.3947	0.1228
Model 4	0.0679	0.3947	0.1159
Model 5	0.0922	0.4342	0.1521
Model 6	0.0461	0.3289	0.0808
Model 7	0.0808	0.4342	0.1362

Table 3: Performance comparison of different methods on the test set. Table 2 contains details of these models.

2. This is worsened by extreme weather events caused by climate change.
3. The main impact of global warming on the weather is an increase in extreme weather events such as heat waves, droughts, cyclones, blizzards, and rainstorms.

This example illustrates the problem of lexical overlap in evidence retrieval. Sentence 2 has a higher similarity score with sentence 1 because they have more words in common. Frequency-based vectorizers such as TF-IDF rank sentence 2 higher than sentence 3 (0.47 vs 0.28). Although powerful pre-trained models such as BERT can capture contextual representations, and reduce the gap between them (0.88 vs 0.80), it still considers sentence 2 as more relevant. However, sentence 2 is not semantically complete, as it does not specify the subject, therefore shouldn't be treated as evidence.

- **Label Prediction:** The lack of transparency in the process of predicting ground-truth labels poses challenges for model interpretation and performance evaluation. All the experimented models showed stronger performance in predicting "supports" and "refutes" labels compared to others. In our EWV algorithm, we interpret "not enough info" as a state where the probabilities of "supports" and "refutes" are both low. However, this interpretation may still not fully capture the complexity of the "not enough info" label, suggesting the need for further exploration and refinement.

5 Conclusion

In conclusion, our research proposed a variety of approaches to automated fact-checking, blending traditional and advanced transformer models. The result in the final evaluation is F1 = **0.05340**, A = **0.40260**, HMean = **0.09430**. The effectiveness of our model was limited due to challenges discussed in Section 4. Future work aims to overcome existing challenges and further enhance the reliability and accuracy of automated fact-checking systems.