Automated Fact Checking

1000447
University of Melbourne

1 Introduction

In the era of information overload, the pervasive influence of misinformation and disinformation threatens the integrity of public discourse. This report introduces our research on an automated fact-checking system designed to counter misinformation in today's information landscape. Our system performs two critical functions: Evidence Retrieval, the extraction of relevant evidence for a particular claim, and Claim Classification, determining the claim's validity based on the evidence, classified as Supports, Refutes, Not Enough Info, or Disputed. Addressing this complex task extends beyond basic text retrieval and requires understanding intricate claim-evidence relationships and evaluating source reliability. Our approach leverages recent advancements in information retrieval and natural language inference, comparing various pre-trained Bert models including Roberta Large MNLI, Distil Roberta Base, and Facebook Bart MNLI.

2 Methodology

The pipeline for the automated fact checking is split into two steps - evidence retrieval and claim classification. This section describes the data processing, augmentation and techniques used to design the automated fact check.

2.1 Data Processing

We leverage SpaCy, a Python library for Natural Language Processing, to preprocess the text data. The pre-processing steps includes lemmatization, case normalization, stop word removal, punctuation removal, and alphabetic word filtering. These steps are designed to reduce the complexity of the data, decrease noise, and enhance the performance of subsequent analytical tasks. The entire corpus of documents is processed in batches for efficiency, with progress tracked to ensure the smooth execution of the task.

2.2 Data Augmentation

We mitigated class imbalance in our dataset, specifically in the 'Refutes' and 'Disputed' categories, through a data augmentation strategy. Utilizing the BM25 algorithm, we created new instances by retrieving the top 15 similar evidence passages (hard positives) for randomly selected unique claims from these underrepresented classes. The top seven evidence passages per claim were used to create new instances, resulting in a more balanced dataset and increased potential for effective model learning.

Class	Before Augmentation	After Augmentation
SUPPORTS	1343	1343
NOT_ENOUGH_INFO	1930	1930
REFUTES	457	1227
DISPUTED	392	1162

Table 1: Class Counts Before and After Data Augmentation

2.3 Evidence Retrieval

We implement three evidence retrieval techniques – Term Frequency Inverse Document Frequency (TF-IDF), Best Match (BM25), and BM25 along with different Bert models to re-rank evidence. The first two methods are producing sparse matrices whereas BERT models produce a dense vector.

2.3.1 TF-IDF

TF-IDF is an algorithm used in information retrieval that measures how important a word is to a document in a collection or corpus. It achieves this by multiplying two metrics: term frequency (how often a word appears in a document) and inverse document frequency (which diminishes the weight of words that occur very frequently in the corpus and increases the weight of words that occur

rarely). The claim is first pre-processed, transforming it into a format suitable for analysis, a TF-IDF vector is then generated for the claim. This vector is compared with all evidence vectors using cosine similarity to find the top n evidence that are most like the claim.

2.3.2 BM25

BM25 is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It's an improvement over TF-IDF because it considers the document length and the term frequency saturation, meaning that after a certain frequency, additional occurrences of a term do not contribute as much to the relevance. The claim is similarly pre-processed at the outset. BM25 scores are then calculated for the claim against all evidence passages, indicating the relevance of each passage to the claim. The top n evidence with the highest BM25 scores is selected as the most relevant evidence.

2.3.3 Finetuned Sentence Transformers

We fine-tune multiple pre-trained Sentence Transformer models for the task of evidence retrieval. Specifically, we utilized three models: 'distilroberta-base', 'roberta-large-mnli', and 'facebook/bart-large-mnli'. The former is computational efficienct, while the latter two are large-scale models trained on the Multi-Genre Natural Language Inference (MNLI) dataset, known for their robust understanding of semantic relationships (Reimers and Gurevych, 2019). The rationale behind this methodology is to leverage the capabilities of a pre-trained Sentence Transformer model to differentiate between related and unrelated pairs of claims and evidence, which is central to the task of evidence retrieval. To prepare the model for training, we generated a set of InputExample instances from our training data, each composed of a claim and corresponding evidence. We confined our training set to entries labelled as 'Supports', 'Refutes', or 'Disputed'. We also experimented using pairing instances with 'Supports' and 'Refutes' with a label of 1 and vice versa with the other labels with a label of 0. Our Sentence Transformer model consisted of two primary components: a **word_embedding_model** for generating word embeddings, and a **pooling_model** for aggregating these into sentence-level embeddings. We adopted the Multiple Negatives Ranking Loss function for fine-tuning the model, a strategy designed to position semantically related sentence (sentences with a relationship) pairs closer in the embedding space than unrelated pairs. For example, any claim and evidence pair with 'Supports', 'Refutes', or 'Disputed' have a clear relationship between them while there is no clear relationship for a claim evidence pair with 'Not Enough Info'.

Cat.	Input Example	Classes
1	Claim, Evidence	Supported, Refuted
2	Claim, Evidence (Label 1)	Supported, Refuted
2	Claim, Evidence (Label 0)	Not Enough Info, Disputed

Table 2: Input Example Categories and Corresponding Classes

The model was trained over four to six epochs. This comprehensive approach harnessed the strengths of pre-trained Sentence Transformers, adapting them to our specific task, and optimizing their ability to discern between related and unrelated claim-evidence pairs.

2.4 Claim Classification

The methodology for fact classification begins with the preparation of the dataset, a crucial step for effective machine learning models as emphasized by (Webersinke et al., 2022). Each claim-evidence pair is tokenized with special tokens added, specifically the [CLS] (classification) token at the beginning and the [SEP] (separator) token between the claim and evidence. This tokenization helps the model differentiate between the two distinct parts of the input and comprehend the relationship between them. Encoding also includes attention masks, which help the model focus on the relevant tokens in sequences that have been padded to a maximum length of 512 tokens for uniformity. The model is then trained over multiple epochs, during which it learns to classify the encoded claim-evidence pairs into one of the four classes.

3 Results

The results section constitutes of results for each of the above techniques described in section 2.

3.1 Evidence Retrieval: BM25 and TF-IDF

The evidence retrieval systems were evaluated based on an F-score (F) that compares the retrieved evidences to the ground truth evidences and computes the mean over the precision and recall (Table 2).

Figure 1: Automated Fact Checking Pipeline

Model	3 Evidences	4 Evidences	5 Evidences
BM25	0.1193	0.1182	0.1183
TFIDF	0.0902	0.0874	0.093

Table 3: Evidence Retrieval F Scores for Different Numbers of Evidences Retrieved

3.2 Evidence Retrieval: BM25 + BERT Model

The results on the dev set for only three and four number of evidences retrieved are presented below while they attain the best performance compared to retrieving one, two or five evidences (Table 3).

Model	Batch Size	F-Score	Evidences
Roberta MNLI	16	0.178	3
Facebook Bart MNLI	16	0.2064	3
Distilroberta	32	0.1704	3
Roberta MNLI	16	0.1733	4
Facebook Bart MNLI	16	0.1974	4
Distilroberta	16	0.17926	4

Table 4: F-Score of Different Models for Different Numbers of Evidences Retrieved Input Example - Cat 1

The Cat 2 Input Example set up was discarded due to GPU limitations and no significant improvement over Cat 1.

3.3 Claim Classification Accuracy

This metric assesses solely how well the system classifies the claim and evaluates the classification component. Distilroberta-base is the model used to classify claims and different parameters used to train this model along with the data it's trained on is presented in the table below.

Model	Batch Size	F-Score	Evidences
Facebook Bart MNLI	3	0.156	3
Distilroberta	8	0.1711	3

Table 5: F-Score of Different Models for Different Numbers of Evidences Retrieved Input Example - Cat 2

Model	Batch	Acc.	Epochs	Data
Distilroberta	16	0.4118	4	Train
Distilroberta	16	0.47	4	Aug. Train
Distilroberta	32	0.446	2	Train
Distilroberta	32	0.51	2	Aug. Train
Distilroberta	37	0.55	1	Aug. Train

Table 6: Claim Classification Performance

3.4 Evidence Retrieval and Claim Classification (Harmonic mean of F and A)

Finally, we evaluated the overall performance of our system by calculating the harmonic mean (H) of the F-score (F) and Accuracy (A) on the dev set. This measure gives us a comprehensive view of our system's capability in both retrieving evidence and classifying claims accurately. The best performing claim classification model was used for classification and BM25 was first used to retrieve top 50 evidence matches for a claim and then re-ranked with the models below.

Model	F	A	H	Evidences
Roberta MNLI	0.1741	0.4805	0.2556	3
Facebook Bart MNLI	0.2064	0.474	0.2876	3
Distilroberta base	0.1757	0.4545	0.2535	3
Roberta MNLI	0.1733	0.4805	0.2547	4
Facebook Bart MNLI	0.1974	0.4805	0.2798	4
Distilroberta base	0.1717	0.4765	0.253	4

Table 7: Performance Metrics for Different Models when 3 and 4 Evidences Retrieved.

4 Evaluation

4.1 Evidence Retrieval BM25 and TF-IDF

The BM25 model consistently outperforms the TFIDF model across all tested evidence counts, achieving the highest F-score of 0.1193 with three evidence. The performance of both models appears relatively stable across different numbers of evidence, suggesting limited impact of additional evidences on these models.

4.2 Evidence Retrieval BM25 + Sentence Transformers

BM25 serves as a baseline in our evaluation of evidence retrieval improvements with more advanced models. All Sentence Transformer models surpass BM25's baseline F score, with the Facebook Bart MNLI model, used in conjunction with BM25, outperforming Roberta MNLI and Distilroberta across all evidence counts, demonstrating its superior retrieval capability. A slight decline in F-score with increasing evidence count suggests the possibility of additional false positives, but the minimal decrease suggests that performance is still consistent. Critically, the Sentence Transformer models act as re-rankers, depending on BM25's initial candidate sentences. Without the inclusion of ground truth evidence in BM25's top-ranked sentences, models miss the chance to improve retrieval, highlighting the importance of precise initial filtering. Despite Facebook Bart MNLI's superior F-scores, the performance of all models depends on the quality of BM25's initial filtering, thus, emphasizing the need for optimizing the initial retrieval stage.

4.3 Claim Classification Accuracy

The Distilroberta model's performance improved with larger batch sizes and augmented training data, with accuracy rising from 0.4118 to 0.446 with increased batch size, and up to 0.55 when using augmented data and a batch size of 37. This suggests that data augmentation and larger batch sizes accelerate learning, leading to higher accuracy in fewer epochs. Improvements can also be attributed to the addition of evidence samples from BM25 (data augmentation), indicating that adding more evidence data for all classes could further enhance model accuracy.

However, training on disputed labels could have affected performance due to potential inaccuracies in labelling claims with clear 'Supported' or 'Refuted' relationships as 'Disputed'.

While these findings underscore the benefits of data augmentation and batch size adjustments in enhancing model performance, it should be noted that optimal configurations may vary across different models, tasks, or datasets.

4.4 Evidence Retrieval and Claim Classification

The performance of sentence transformer models in evidence retrieval varies, and the interdependence of retrieval and classification tasks impacts overall performance. Facebook Bart Large MNLI outperformed the others for both 3 and 4 evidences retrieved in terms of F-score, but Roberta Large MNLI had higher accuracy when retrieving 4 evidences. DistilRoberta base had a lower F-score, but claim classification accuracy on its retrieved evidences improved significantly when retrieving 4 evidences compared to 3. This suggests that even if a model's retrieval performance is not the highest, its classification accuracy can still improve with more evidence. The harmonic mean decreases when moving from 3 to 4 evidences for all models, indicating that the optimal number of evidences for these models in this task might be around 3. In summary, achieving optimal model performance necessitates balancing interdependent tasks of evidence retrieval and claim classification, which is influenced by the optimal number of evidences retrieved, a principle potentially applicable across various models and tasks.

4.5 Final Evaluation on Codalab

For final evaluation on Codalab, we picked the best performing pipeline: **BM25 + FB Bart Sentence Transformer for evidence retrieval and Distilroberta Base** from Table 6 the results are in the table below.

Model	F	A	H	Eval Set
FB Bart MNLI	0.1144	0.3896	0.1769	Final Eval Set

Table 8: Performance Metrics for Facebook Bart Large MNLI Model

The Facebook Bart Large MNLI model displayed a drop in performance from the development set to the test set, with the F-score falling from 0.2064 to 0.1144, accuracy dropping from 0.474 to 0.3896, and the harmonic mean decreasing from 0.2876 to 0.1769. This decline suggests possible overfitting to the development set or it could also be due to the test set being more complex or different in nature compared to the development set. Despite this, the model ranked 148th out of 367, placing it in the top 40%. Future improvement could stem from further fine-tuning, diversified training data, or ensemble methods.

References

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text. *SSRN Electronic Journal*.