AN EXPERIMENTAL EVALUATION OF DEEP NEURAL NETWORK MODEL PERFORMANCE FOR THE RECOGNITION OF CONTRADICTORY MEDICAL RESEARCH CLAIMS USING SMALL AND MEDIUM-SIZED CORPORA

***Fatin Syafiqah Yazi*[1*], *Wan-Tze Vong*[2], *Valliappan Raman*[3], *Patrick Hang Hui Then*[4], *Mukulraj J Lunia*[5]**

[1,2,3,4]Faculty of Engineering, Computing and Science, Swinburne University of Technology,
93350 Kuching, Malaysia

[5]Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India

Email: fyazi@swinburne.edu.my[1*](corresponding author), wvong@swinburne.edu.my[2],
vraman@swinburne.edu.my[3], pthen@swinburne.edu.my[4], mukul_lunia99@outlook.com[5]

DOI: https://doi.org/10.22452/mjcs.sp2021no2.5

ABSTRACT

Corpora come in various shapes and sizes and play an essential role in facilitating Natural Language Processing (NLP) tasks. However, the availability of corpora specialized for Evidence-Based Medicine (EBM) related tasks is limited. The study is aimed to discover how the size of a corpus influence the performance of our Deep Neural Network (DNN) model developed for contradiction detection in medical literature. We explored the potential of the EBM Summarizer corpus by Mollá and Santiago-Martínez, a medium-sized corpus to be used with our contradiction detection model. The dataset preparation involves the filtering of open-ended questions, duplicates of claims, and vague claims. As a result, two datasets were created with the claim input represented by sniptext in one dataset and longtext in the other. Experiments were conducted with varying numbers of hidden layers and units of the model using different datasets. The performance of the DNN model was recorded and compared with the result of using a small-sized corpus. It was found that the DNN model performance did not improve even after it was trained with a larger dataset derived from the medium-sized corpus. The factors may include the limitation of the DNN model itself and the quality of the datasets.

Keywords: Evidence-based medicine, contradiction detection, medical literature, deep neural network, deep learning

1.0 BACKGROUND

The Evidence-Based Medicine (EBM) practice requires clinicians and researchers to be au courant with the current state of knowledge in the biomedical field. The best current evidence is gathered, appraised, and evaluated to make an informed clinical decision for their patients with their good clinical judgment, which also has to be aligned with the patients' preferences and values [1]. The evidence in the form of research claims can be found in review articles, journal articles, practice guidelines, editorials, and many other forms of medical literature [2].

The gathering of the evidence will involve the formulation of clinical questions as part of the process. A good clinical question must be as detailed as possible and should accept only "Yes" or "No" as the answer. When designing such questions, the PICO framework is used as a guideline which specified the elements to be included: Participants/Problem (P), Intervention (I), Comparison (C) and Outcome (O) [3]. The evidence or research claims that answer the clinical question may assert either a positive answer ("Yes") or a negative answer ("No") to the question. These assertion values were found to be the key in detecting whether the research claims that answer the same clinical question are contradictory or not [4], [5]. If two research claims that answer the same question has assertion values that differ from each other, they can be considered contradictory [4], as illustrated in Fig. 1.

The automated detection of contradictory research claims can become a strong support for EBM practitioners to carry out their tasks. One of the challenges faced by EBM practitioners is the number of medical literature to be reviewed and appraised for their quality. In 2020, the MEDLINE archive held more than 27 million cumulative citations, recording daily new additions of 2,600 citations. Amongst the vast amount of literature that is continuously growing in number [6], the presence of contradictory research claims proves to be another hurdle that may complicate the EBM practice [7], [8].

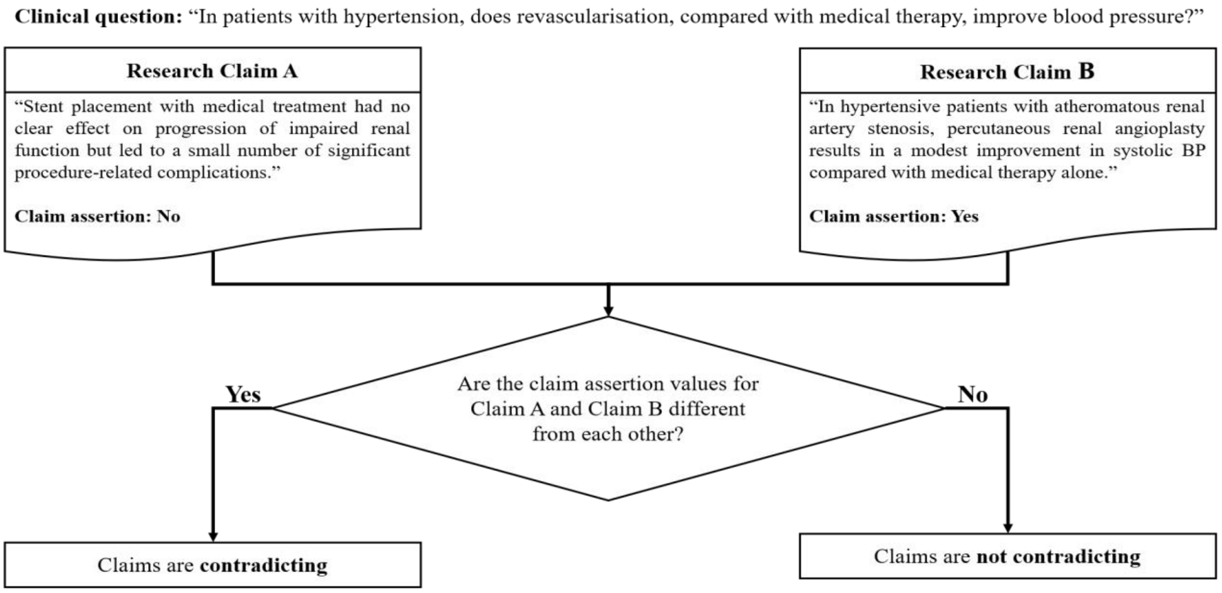

Clinical question: "In patients with hypertension, does revascularisation, compared with medical therapy, improve blood pressure?"
Research Claim A
"Stent placement with medical treatment had no clear effect on progression of impaired renal function but led to a small number of significant procedure-related complications."
Claim assertion: No
Research Claim B
"In hypertensive patients with atheromatous renal artery stenosis, percutaneous renal angioplasty results in a modest improvement in systolic BP compared with medical therapy alone."
Claim assertion: Yes
Are the claim assertion values for Claim A and Claim B different from each other?
Yes
No
Claims are contradicting
Claims are not contradicting

Fig. 1: Example of contradictory research claims found in the ManConCorpus

In our previous study [9], we developed two deep neural network (DNN) models using several techniques such as the bidirectional Long-Short-Term Memory (LSTM), Global Vectors (GloVe), and Bidirectional Encoder Representations from Transformers (BERT) for the automatic detection of contradictory research claims through the classification of claim assertion value. The Manual Contradiction Corpus (ManConCorpus) [10], a collection of contradictory research claims from systematic reviews of the cardiovascular topic, was used to evaluate the two DNN models. The corpus is considered to be small in size, with 259 claims included in it. The outcome of the study revealed an encouraging performance of the model built using BERT, which surpassed the results from earlier studies done on contradiction detection in medical literature using the same corpus as the dataset in [4], [5].

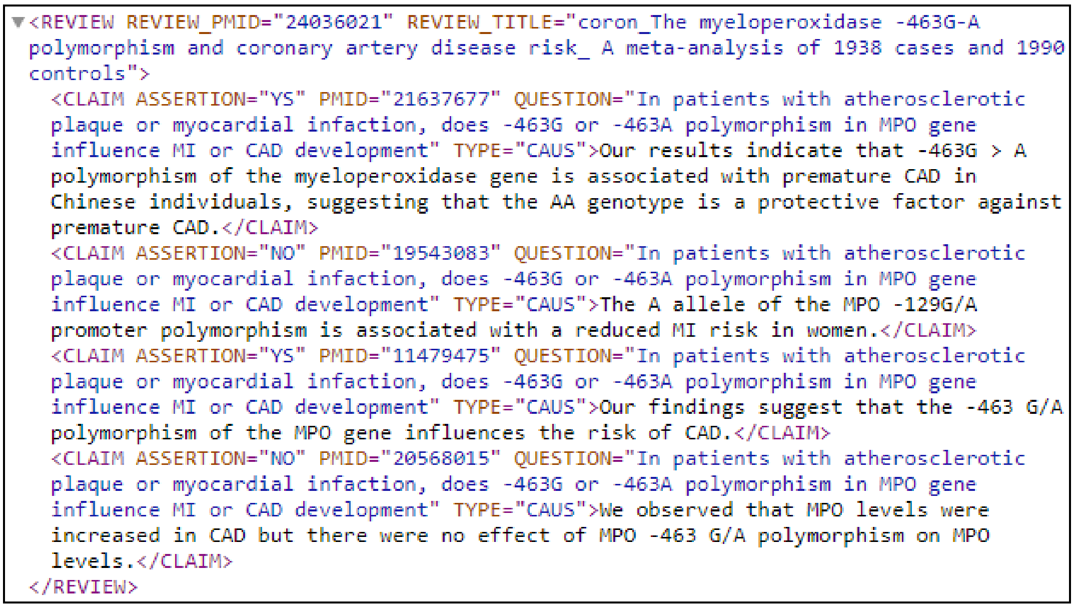
The ManConCorpus is regarded as a small corpus. Its size has been hypothesized to be the obstacle in achieving a better result for contradiction detection in medical literature [5], [9], [11]. We extend our preceding work in [9] by evaluating the performance of the previously developed BERT-based model using a medium-sized corpus [12]. This study focuses on investigating the influence of corpus size on the performance of the DNN model. The experimental results are compared with the results from the previous study, which utilized a small-sized corpus.

2.0 RELATED WORKS

The existence of corpora which compiles biomedical texts and contradictory research claims found in medical literature, can be a great catalyst in enabling more research conducted in this area. The GENIA corpus [13], [14] was one the earliest corpus developed as a resource for natural language processing (NLP) tasks in bio-text mining. It consists of 2000 abstracts extracted from MEDLINE articles concerning the terms "human", "blood cells", and "transcription factor". The corpus has molecular events annotated with biological and linguistic information, including the presence of negation and uncertainties. Selected event types from the corpus have been included in the BioNLP'09 corpus, which was used by Sarafraz [15] in 2012 to recognize the contradiction in medical literature through the usage of machine learning and rule-based methods.

The BioScope corpus [16] is another derivative of the GENIA corpus, which was constructed and made available for the study on negation and uncertainties in biomedical literature. The corpus is made up of mostly clinical free-texts (radiology reports), biological full papers and abstracts from GENIA corpus. The texts were annotated for negations, speculations and linguistic scopes, and more than 20,000 sentences have been included in the corpus, with more than 10% of them annotated for negation or uncertainty.

In 2016, the Manual Contradiction Corpus (ManConCorpus), containing contradictory research claims mainly on the cardiovascular topic, was manually constructed by Alamri and Stevenson [10]. The ManConCorpus is created out of 24 systematic reviews and has a total of 259 abstracts, out of which 180 introduces positive claims (`Yes'), and 79 introduces negative claims (`No'). The process of annotating the corpus includes the formulation of clinical questions using the PICO framework based on the review objectives. The question has to be close-ended and can be answered with either “Yes” or “No”. Then, the annotator has to identify the research claim by choosing the best sentence from the review abstract that answers the PICO question created earlier. The chosen claim will be labelled with the assertion values “YS” for claims that assert a “Yes” answer and “NO” for claims that assert a “No” answer to the question. Lastly, the claims will be annotated with their claim type, which can be either causal or evaluative claims.



```
▼<REVIEW REVIEW_PMID="24036021" REVIEW_TITLE="coron_The myeloperoxidase -463G-A
 polymorphism and coronary artery disease risk_ A meta-analysis of 1938 cases and 1990
 controls">
   <CLAIM ASSERTION="YS" PMID="21637677" QUESTION="In patients with atherosclerotic
   plaque or myocardial infaction, does -463G or -463A polymorphism in MPO gene
   influence MI or CAD development" TYPE="CAUS">Our results indicate that -463G > A
   polymorphism of the myeloperoxidase gene is associated with premature CAD in
   Chinese individuals, suggesting that the AA genotype is a protective factor against
   premature CAD.</CLAIM>
   <CLAIM ASSERTION="NO" PMID="19543083" QUESTION="In patients with atherosclerotic
   plaque or myocardial infaction, does -463G or -463A polymorphism in MPO gene
   influence MI or CAD development" TYPE="CAUS">The A allele of the MPO -129G/A
   promoter polymorphism is associated with a reduced MI risk in women.</CLAIM>
   <CLAIM ASSERTION="YS" PMID="11479475" QUESTION="In patients with atherosclerotic
   plaque or myocardial infaction, does -463G or -463A polymorphism in MPO gene
   influence MI or CAD development" TYPE="CAUS">Our findings suggest that the -463 G/A
   polymorphism of the MPO gene influences the risk of CAD.</CLAIM>
   <CLAIM ASSERTION="NO" PMID="20568015" QUESTION="In patients with atherosclerotic
   plaque or myocardial infaction, does -463G or -463A polymorphism in MPO gene
   influence MI or CAD development" TYPE="CAUS">We observed that MPO levels were
   increased in CAD but there were no effect of MPO -463 G/A polymorphism on MPO
   levels.</CLAIM>
 </REVIEW>
```

Fig. 2: A snippet of the ManConCorpus

The availability of ManConCorpus has encouraged more research on contradiction detection in the biomedical domain. Tawfik and Spruit [5], [11] used the ManConCorpus to evaluate their automated two-phase model that outperformed the original systems developed by Alamri [4] in the extraction of research claims and detection of conflicts and contradiction. In 2019, Tawfik and Spruit again utilized the ManConCorpus in investigating the performance of models built using machine learning, deep learning and hybrid of both. Despite the promising findings from the studies, previous studies suggested that further study is needed due to the small size of the ManConCorpus corpus and the limited types of contradiction instances available in the corpus since they may influence the performance of an automated contradiction detection model.

Alamri and Tawfik proposed possible enhancement to the existing model by annotating more details about the claims aside from their assertion value, highlighting claims in abstracts with colour codes signifying their assertion value, and incorporating general sentences in ManConCorpus to test the claim extraction performance of the proposed methods. As revealed by these limitations and possible future extensions of existing models, further research on the detection and recognition of contradictory claims in medical literature with a training corpus of a larger size and refined definition of contradictions is called for.

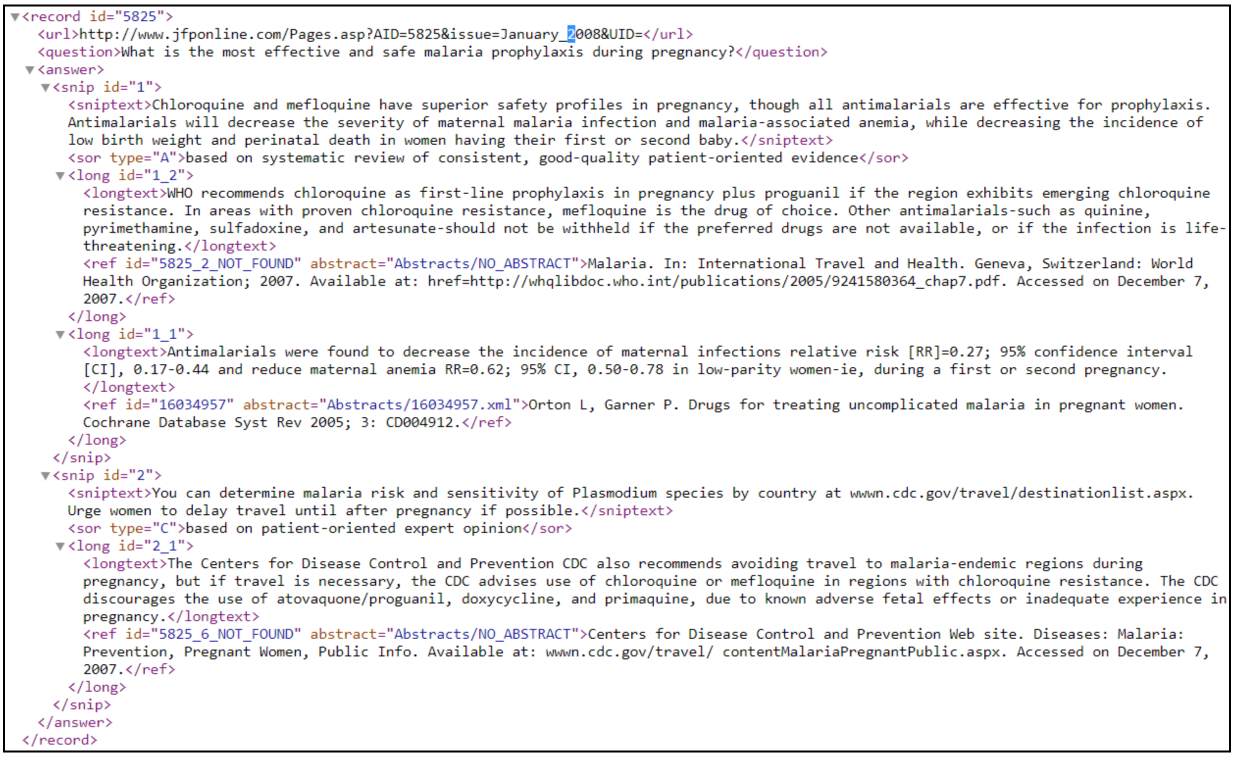
In the same year, Mollá and Santiago-Martínez built a corpus of clinical questions and summarised findings from medical literature to support the development and testing of text processing tools that may be helpful in the EBM practice. The EBM Summarizer (EBMSum) corpus [12] consists of 456 questions, 1,396 answer components with 1,225 of them are grade specified, 3,036 detailed answer justifications, and 2,908 unique PMID of referenced PubMed articles.

In contrast with ManConCorpus, the EBMSum corpus was not constructed from systematic reviews but from the articles in the Clinical Inquiries section of the Journal of Family Practice (JFP). Their earlier study reported that using the Clinical Inquiries section of the JFP is more convenient in terms of building a corpus compared to systematic reviews [17]. The information extracted from articles in the JFP Clinical Inquiries section includes:

1. The question, directly obtained from the article’s title.
2. The answer parts or “snip” from the evidence-based answer section of the article. A question may have more than one snip.

3. The Strength of Recommendation (SOR) grading of each answer part. The evidence grading follows the Strength of Recommendation Taxonomy (SORT) grading scale that is used by JFP.
4. The justifications of the evidence-based answers or "long" which are summarised information from each referenced medical literature.
5. The reference ID or the PMID of referenced medical research publications from PubMed.

Mollá and Santiago-Martínez's corpus has been used in multiple studies, mainly on biomedical text summarization [18], [19]. Bavani, Ebrahimi, Wong and Chen [18] reported that the corpus is considered middle-sized, and a larger corpus can help in generating a more reliable result.



```
▼<record id="5825">
   <url>http://www.jfponline.com/Pages.asp?AID=5825&issue=January_2008&UID=</url>
   <question>What is the most effective and safe malaria prophylaxis during pregnancy?</question>
 ▼<answer>
   ▼<snip id="1">
       <sniptext>Chloroquine and mefloquine have superior safety profiles in pregnancy, though all antimalarials are effective for prophylaxis.
       Antimalarials will decrease the severity of maternal malaria infection and malaria-associated anemia, while decreasing the incidence of
       low birth weight and perinatal death in women having their first or second baby.</sniptext>
       <sor type="A">based on systematic review of consistent, good-quality patient-oriented evidence</sor>
     ▼<long id="1_2">
         <longtext>WHO recommends chloroquine as first-line prophylaxis in pregnancy plus proguanil if the region exhibits emerging chloroquine
         resistance. In areas with proven chloroquine resistance, mefloquine is the drug of choice. Other antimalarials-such as quinine,
         pyrimethamine, sulfadoxine, and artesunate-should not be withheld if the preferred drugs are not available, or if the infection is life-
         threatening.</longtext>
         <ref id="5825_2_NOT_FOUND" abstract="Abstracts/NO_ABSTRACT">Malaria. In: International Travel and Health. Geneva, Switzerland: World
         Health Organization; 2007. Available at: href=http://whqlibdoc.who.int/publications/2005/9241580364_chap7.pdf. Accessed on December 7,
         2007.</ref>
       </long>
     ▼<long id="1_1">
         <longtext>Antimalarials were found to decrease the incidence of maternal infections relative risk [RR]=0.27; 95% confidence interval
         [CI], 0.17-0.44 and reduce maternal anemia RR=0.62; 95% CI, 0.50-0.78 in low-parity women-ie, during a first or second pregnancy.
         </longtext>
         <ref id="16034957" abstract="Abstracts/16034957.xml">Orton L, Garner P. Drugs for treating uncomplicated malaria in pregnant women.
         Cochrane Database Syst Rev 2005; 3: CD004912.</ref>
       </long>
     </snip>
   ▼<snip id="2">
       <sniptext>You can determine malaria risk and sensitivity of Plasmodium species by country at wwwn.cdc.gov/travel/destinationlist.aspx.
       Urge women to delay travel until after pregnancy if possible.</sniptext>
       <sor type="C">based on patient-oriented expert opinion</sor>
     ▼<long id="2_1">
         <longtext>The Centers for Disease Control and Prevention CDC also recommends avoiding travel to malaria-endemic regions during
         pregnancy, but if travel is necessary, the CDC advises use of chloroquine or mefloquine in regions with chloroquine resistance. The CDC
         discourages the use of atovaquone/proguanil, doxycycline, and primaquine, due to known adverse fetal effects or inadequate experience in
         pregnancy.</longtext>
         <ref id="5825_6_NOT_FOUND" abstract="Abstracts/NO_ABSTRACT">Centers for Disease Control and Prevention Web site. Diseases: Malaria:
         Prevention, Pregnant Women, Public Info. Available at: wwwn.cdc.gov/travel/ contentMalariaPregnantPublic.aspx. Accessed on December 7,
         2007.</ref>
       </long>
     </snip>
   </answer>
</record>
```

Fig. 3: A snippet of the EBMSum corpus

In 2018, EBM-NLP corpus [20] containing 5,000 annotated abstracts extracted from medical articles on clinical RCT was developed by Nye et al. in 2018. They retrieved 5,000 abstracts from MEDLINE articles about RCTs related to cardiovascular diseases, cancer, and autism via PubMed. The texts were annotated with the P (Population), I (Intervention), and O (Outcome) elements based on EBM's PICO framework, with the I and C components being combined into a single element I. The corpus was created in acknowledgement of the need for larger corpora to support research on NLP application in the EBM practice, especially for biomedical evidence synthesis.

To date, the ManConCorpus is the only existing corpus that is specialized for the task of contradiction detection in medical literature. Although other existing biomedical corpus was constructed mainly to facilitate information extraction and text summarization in medical literature, they still possess the potential to be useful for contradiction detection tasks depending on the way they are utilized.

3.0 METHODS

The EBM Summarizer (EBMSum) corpus by Mollá and Santiago-Martínez [12] is used as the main dataset in this study. This corpus was chosen as it is a corpus specialized for EBM-related NLP tasks. Like the ManConCorpus, it contains information such as questions, answer parts, and justifications, making it suitable to be used with the current DNN model with minimal load and time spent in annotating any missing information. Compared to the ManConCorpus, the EBMSum corpus contains a larger amount of questions and research claims with a total of 456 questions, 1,396 answer components with 1,225 of them are grade specified with 3,036 detailed answer justifications.

To the best of our knowledge, most of the extant research relied on the ManConCorpus as it is the only available corpus specialized for the task of contradiction detection in medical literature to date [4], [5], [9], [11], [21]. The small size of the ManConCorpus is one of the limitations highlighted in previous studies. In this study, we propose the usage of an alternative medium-sized [18] corpus, the EBMSum corpus [12], in aiding the task of recognizing contradictory medical research claims. The performance of the DNN model developed in our previous study is evaluated using the datasets extracted from the EBMSum corpus.

3.1 Dataset Preparation

The existing contradiction detection model requires inputs in the form of questions, research claims, and their claim assertion values as the input. While all of the inputs are readily available in the ManConCorpus, which was a corpus specialized for contradictory claims detection, it is a different case with the EBMSum corpus. The EBMSum corpus contains the information extracted from the articles in the Clinical Inquiries section of the Journal of Family Practice (JFP). The questions are directly obtained by extracting the title of the article. The corpus also consists of the answer parts called sniptext and answer justifications called longtext, which can be regarded as the research claim input component for the DNN model. Although the input components for question and claim can be obtained directly from the corpus, the corpus lacks the annotation of the claim assertion value. To enable the usage of the EBMSum corpus with the current DNN model, it needs to be annotated with the assertion value.

In obtaining the claim assertion value, it is essential for the question to be close-ended and can only be answered with “Yes” or “No” answers. However, some of the questions in the corpus are open-ended and may not be formulated using the PICO framework. In this study, only close-ended questions that can be answered with “Yes” or “No” were selected as part of the dataset. With the elimination of questions that did not fulfil the criteria, the dataset holds 1123 entries.

The next step is annotating the dataset entries with their claim assertion values. The annotation was done by two annotators with experience in conducting research on computational linguistics related to medicine. Both annotators possess advanced level of English proficiency with educational backgrounds in the biomedical and computer science field. The annotation process was done by following the same steps used by Alamri and Stevenson in annotating the ManConCorpus. The annotation process should follow the guideline set by the creator of ManConCorpus:

1. The claim should be annotated with “YS” when the claim asserts a positive answer to the question.
2. The claim should be annotated with “NO” when the claim asserts a negative answer to the question. “NO” should also be used if the claim neither asserts nor negates the question.

In addition to that, any claims that do not clearly answer the question or are totally unrelated to the question are considered low-quality entries and are removed from the dataset. Should there be any disagreement between the annotation of different annotators, discussions should be done to resolve the differences and decide on the best annotation before the corpus annotation is finalized.

There are two candidates of data from the EBMSum corpus that can be used as the research claim input component for the DNN model: the longtext and the sniptext. We created two separate datasets for the model performance evaluation; one uses sniptext as the claim, while the other contains longtext. Duplicates exist for both sniptext and longtext as one answer part can have multiple justifications, and likewise, one justification being associated with multiple answer parts. The numbers of unique claims in the sniptext and longtext datasets are 340 and 746, respectively. It is noted that the size of the dataset has become considerably smaller after the elimination of open-ended questions, duplicates of sniptext and longtext, as well as low-quality claims. Both of the datasets contain imbalanced class distribution by having claims with assertion value “YS” with a higher number than “NO”. The class distribution of each of the datasets is shown in Table 1.

Table 1: The class distribution of the sniptext and longtext dataset extracted from the EBMSum corpus

Dataset	Sniptext	Longtext
Claim assertion value - YS	189	419
Claim assertion value - NO	151	327
Total claims	340	746

After all the preparation tasks have been carried out, the dataset is complete with all the necessary input components: the selected questions and research claims from the EBMSum corpus with the claim assertion values annotated. The dataset is now ready to be used with the DNN model to classify the claim assertion value of question-claim pairs in detecting contradictory claims.

3.2 Feature Extraction

In this study, the BERT pre-trained model was utilized to extract features from the texts in the EBMSum corpus. BERT is a Transformer-based pre-trained model developed by Google and is said to be the first model to overcome the limitations of existing unidirectional language models and achieve true bidirectionality compared to pseudo-bidirectionality by previous models [22], [23]. The model has been reported to achieve state-of-the-art for various datasets and NLP tasks, including scientific and medical literature [22]–[24]. It is publicly available and can be fine-tuned or used as a feature extractor to suit the objective of the task.

In our experiments, BERT, specifically the BERT-base-uncased model, was used as a sentence encoder. The representations produced by BERT are contextualized embeddings in which one word may have different representations depending on its context in the sequence. The BERT embedding scheme yields two outputs; the sequence output and the pooled output. Sequence output is the representation of each token in context, while the pooled output is the contextual representation of the input sequence as a whole. The pooled output, which captured the context of the questions and claims were retrieved to be processed by the subsequent layers of the DNN model for training and classification.

3.3 Deep Neural Network Model

Contradiction detection in medical texts by recognizing the claim assertion value is a text classification task where we are trying to identify the claim assertion value of each question-claim pair. In this study, the deep learning approach was applied to carry out the classification task. Fig. 4 illustrates the design of the DNN model.

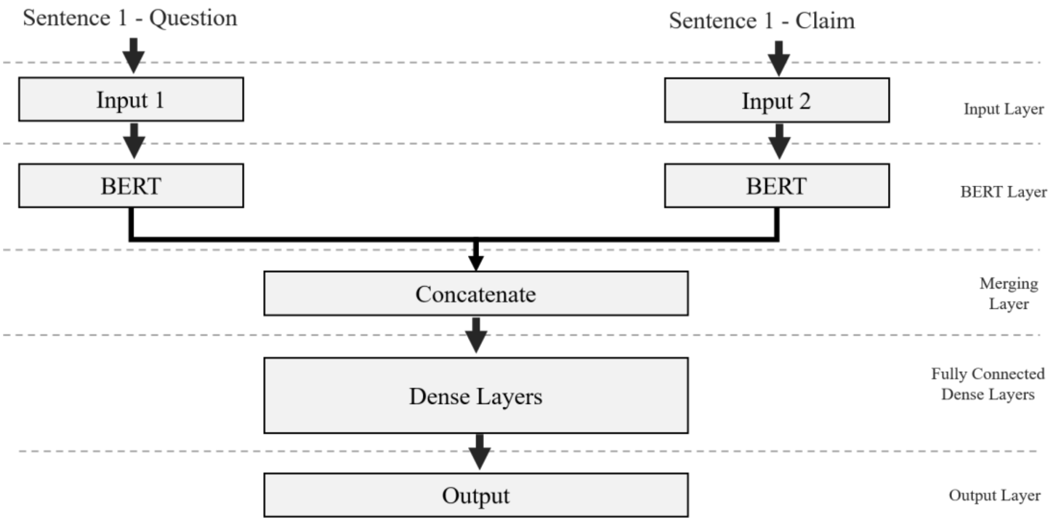


Fig. 4: The architecture of the DNN model

The DNN model used in this study is the model developed in our previous work [9], which utilized BERT pre-trained model [23]. Before being fed to the model, the inputs are tokenized using the BERT tokenizer. The BERT tokenizer utilizes the WordPiece algorithm in building a vocabulary of 30,000 tokens. The WordPiece tokenizer works by breaking each word in the input sentences into subwords called ‘wordpieces’ that minimizes the out-of-vocabulary (OOV) occurrences [25]. Two special tokens, which are the “[CLS]” token at the beginning of each input sequence and “[SEP]” tokens as sentence separators, are introduced by the BERT tokenizer. The pooled output, which is the contextual representation of the whole input sentence, is derived from the “[CLS]” token, where its embedding is considered to sufficiently capture the holistic information of the input sentence [26].

The Siamese-like architecture was applied in designing the model, considering that the inputs come in pairs. It allows the model to accept more than one input via multiple input channels. Both inputs will undergo BERT embedding with the same weights. The outputs from both input channels are then merged together. Following that, the output of the merging layer will pass through hidden layers assigned with the rectified linear (ReLU) activation function. The number of hidden layers and units are the variables in the experiments conducted in this study.

The output layer with a single node using the Sigmoid activation function will produce a binary classification outcome that predicts the claim assertion values based on the input questions and claims. The interpretation of the prediction of class labels is subject to a threshold of 0.5. In combating overfitting, the learning rate for the Adam optimizer is specified to be relatively low. Dropout layers of rate 0.3 are also introduced, and early stopping is applied to stop training when the loss value is not improving.

4.0 EXPERIMENTAL RESULTS AND DISCUSSION

The claim assertion value is considered an important key in the detection of the claim assertion value. Hence, the performance of the model in recognizing the claim assertion value of a claim to its question reflects the performance of the detection of contradictory research claims. To evaluate the model performance, two experiments were done using the sniptext and the longtext dataset extracted from the EBMSum corpus. The dataset with sniptext consists of 340 unique claims, while the dataset with longtext contains 746 entries.

The results are recorded as the average precision, recall and F1 score across 10-fold stratified cross-validation. Precision measures the ratio of correctly predicted "YS" class (true positives) to the total of both correct and wrong predictions of assertion values as "YS" (true positive and false positive). Recall or sensitivity measures the ratio of correctly predicted "YS" class (true positives) to the actual number of the "YS" label (true positives and false negatives). F1 is an accuracy measure that takes both precision and recall into account. Unlike accuracy, F1 focuses more on false negatives and false positives than true negatives and is considered a great accuracy measure for imbalanced datasets. Stratification ensures that both "Yes" and "No" classes are represented in all dataset splits. The sizes of the datasets are not large; hence, the cross-validation evaluation method is chosen to optimize the usage of small to medium-sized corpora. Also, in order to compare the result with the previous study in [9], similar performance metrics were chosen.

The results are recorded as the average precision, recall, and F1 score across the 10-fold cross-validation. Table 2 describes the results of the experiments using the EBM Summarizer corpus. The result from the preceding study [9], which uses the smaller-sized ManConCorpus, are also included for comparison purposes.

Table 2: The results of the experiments using the EBMSum corpus compared with the results from the previous study using the ManConCorpus

Hidden layers	Hidden units	Average Precision			Average Recall			Average F1		
		ManCon Corpus	EBMSum (sniptext)	EBMSum (longtext)	ManCon Corpus	EBMSum (sniptext)	EBMSum (longtext)	ManCon Corpus	EBMSum (sniptext)	EBMSum (longtext)
1	1024	0.84	0.814	0.696	0.973	0.921	0.712	0.901	0.857	0.683
2	256, 128	0.867	0.87	0.755	0.987	0.932	0.787	0.922	0.893	0.757
2	1024, 512	0.809	0.863	0.70	0.987	0.928	0.688	0.889	0.887	0.671
3	256, 128,6 4	0.815	0.642	0.711	0.987	0.85	0.836	0.892	0.706	0.746
3	512, 256, 128	0.852	0.86	0.666	0.947	0.94	0.828	0.921	0.891	0.72

The model achieved precision values of 0.642 to 0.87 when using the sniptext dataset and 0.666 to 0.755 when using the longtext dataset. The result indicates that the model has a lower false-positive rate when using sniptext than longtext. For recall, the model recorded values between 0.85 and 0.94 using the sniptext dataset. Meanwhile, the recall value when using the longtext dataset ranged from 0.688 to 0.836. The model achieved a higher sensitivity when used with the sniptext dataset compared to the longtext dataset. On the other hand, the model scored from 0.706 to 0.893 for F1 when using the sniptext dataset while for the longtext dataset, the F1 score is between 0.671 to 0.757.

The overall result shows that the sniptext dataset contributed to better precision, recall and F1 of the DNN model in comparison to the longtext dataset. The results obtained from our preceding study using the small-sized

ManConCorpus [9] show that deep learning approach can improve the performance of contradiction detection model through the classification of the claim assertion value, which also outperform the models developed in the earlier studies [4], [5] that uses the machine learning approach evaluated on the ManConCorpus as well. However, the results achieved by both of the datasets derived from the EBMSum corpus did not surpass the outcome outlined by our previous work. In general, it can be said that training with a larger corpus alone does not improve the performance of the DNN model.

An assumption was made where a better result is anticipated with the usage of a DNN model having trained with more data. Although the longtext dataset is of a bigger size than the other datasets, the performance of the DNN model is not better than when using the small-sized datasets. One of the most significant differences between the sniptext and longtext is the length of the sentences. Sniptexts are shorter and typically contain one or two sentences with a maximum length of 74 words. Longtexts are considerably longer and can contain up to 446 words. As a comparison, the claims in the ManConCorpus used in the previous study [9] is no longer than 55 words. It was found that one of the major issues with Transformer-based models like BERT is their limitation in capturing long-term dependency when used with longer sequences due to the predefined fixed-length context [27]. The existing contradiction detection model will have to be modified for the sake of better processing of longer inputs which may result in a better outcome when using the longtext dataset.

The features of the corpus itself may also influence the model performance. In the dataset preparation phase, due to the presence of open-ended questions, duplicate entries of sniptext and longtext, and the existence of vague claims that either assert unclear or having an unrelated answer to its question. The elimination process excluded so much data from the EBMSum, leaving less than 1000 claims out of over 3000 claims that are usable as part of the dataset. Data augmentation can be done to transform the excluded data to suit the dataset in order to include more entries in the dataset, for instance, by converting open-ended questions into close-ended questions based on the claims. However, the existence of duplicates and vague claims may still cause many of the data to be eliminated. Nevertheless, securing good-quality data that satisfy the criteria to be included in the dataset may allow a better generalization of the effect of using a larger corpus on the DNN model performance result.

5.0 CONCLUSION

In this study, we conducted experiments to investigate how corpus size influence the performance of the contradiction detection model. The DNN model used in this study is the model developed in our previous work [9] using BERT. Two datasets were derived from the medium-sized EBMSum corpus to be used with the existing model. Based on the result, it was discovered that training with a bigger dataset does not necessarily produce a better performance. The model did not achieve a better result using the dataset derived from the medium-sized EBMSum corpus compared to the smaller ManConCorpus. This may be caused by BERT having difficulty capturing long-term dependency due to the fixed-length context [27] as well as the quality of the corpus itself, which prompted many of the data to be excluded from the dataset. Further enhancement can be done by improving the current model to overcome the issue of long-term dependency and acquiring better-quality data to be used as part of the dataset.

6.0 ACKNOWLEDGMENT

This research is funded by the Fundamental Research Grant Scheme (FRGS/1/2019/ICT02/SWIN/03/01) under the Ministry of Higher Education, Malaysia

REFERENCES

[1] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson, “Evidence based medicine: what it is and what it isn’t,” *BMJ*, vol. 312, no. 7023, pp. 71–72, 1996.

[2] G. H. Guyatt and D. Rennie, “Users’ Guides to the Medical Literature,” *JAMA J. Am. Med. Assoc.*, vol. 270, no. 17, pp. 2096–2097, 1993.

[3] W. S. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward, “The well-built clinical question: a key to evidence-based decisions.,” *ACP journal club,* vol. 123(3), no. 3. pp. A12–A13, 1995.

[4] A. D. Alamri and M. Stevenson, “The Detection of Contradictory Claims in Biomedical Abstracts,” University of Sheffield, 2016.

[5] N. S. Tawfik and M. R. Spruit, “Automated Contradiction Detection in Biomedical Literature,” in *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2018, pp. 138–148.

[6] G. Rosemblat, M. Fiszman, D. Shin, and H. Kilicoglu, “Towards a characterization of apparent contradictions in the biomedical literature using context analysis,” *J. Biomed. Inform.*, vol. 98, p. 103275, 2019.

[7] R. Gauch, *It’s Great! Oops, No It Isn’t*. Springer Science & Business Media., 2008.

[8] V. Prasad *et al.*, “A decade of reversal: An analysis of 146 contradicted medical practices,” *Mayo Clin. Proc.*, vol. 88, no. 8, pp. 790–798, 2013.

[9] F. S. Yazi, W.-T. Vong, V. Raman, P. H. H. Then, and M. J. Lunia, “Towards Automated Detection of Contradictory Research Claims in Medical Literature Using Deep Learning Approach,” in *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2021, pp. 116–121.

[10] A. Alamri and M. Stevenson, “A corpus of potentially contradictory research claims from cardiovascular research abstracts,” *J. Biomed. Semantics*, vol. 7, no. 1, pp. 1–9, 2016.

[11] N. S. Tawfik and M. R. Spruit, “Towards Recognition of Textual Entailment in the Biomedical Domain,” in *International Conference on Applications of Natural Language to Information Systems*, vol. 11608 LNCS, 2019, pp. 368–375.

[12] D. Mollá, M. E. Santiago-Martínez, A. Sarker, and C. Paris, “A corpus for research in text processing for evidence based medicine,” *Lang. Resour. Eval.*, vol. 50, no. 4, pp. 705–727, 2016.

[13] T. Ohta, Y. Tateisi, and J. D. Kim, “The GENIA corpus: An annotated research abstract corpus in molecular biology domain,” *Proc. Second Int. Conf. Hum. Lang. Technol. Res.*, pp. 82–86, 2002.

[14] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, “GENIA corpus - A semantically annotated corpus for bio-textmining,” in *Bioinformatics*, 2003, vol. 19, no. SUPPL. 1.

[15] F. Sarafraz, “Finding conflicting statements in the biomedical literature,” University of Manchester, 2012.

[16] V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik, “The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes,” in *BMC Bioinformatics*, 2008, vol. 9, no. SUPPL. 11.

[17] D. Mollá, M. E. Santiago-Martínez, “Development of a Corpus for Evidence Based Medicine Summarisation,” *Proc. Australas. Lang. Technol. Assoc. Work.*, p. 86–94, 2011.

[18] E. S. Bavani, M. Ebrahimi, R. Wong, and F. Chen, “Appraising UMLS coverage for summarizing medical evidence,” in *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2016, pp. 513–524.

[19] A. Sarker, Y. C. Yang, and M. A. Al-Garadi, “A light-weight text summarizer for fast access to medical evidence,” *medRxiv*. 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.05.22.20110742v1.

[20] B. Nye *et al.*, “A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature,” *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2018, pp. 197.

[21] S. Usha, D. P. Apoorva, S. Aishwarya, E. Amrutha, and H. Chaithali, “Data Analysis and Detection of Contradiction in Biomedical Literature,” *J. Crit. Rev.*, vol. 7, no. 18, pp. 1047–1056, 2020.

[22] C. Cohn, “BERT efficacy on scientific and medical datasets: a systematic literature review,” *Coll. Comput. Digit. Media Diss.*, Nov. 2020. [Online]. Available: https://via.library.depaul.edu/cdm_etd/24.

[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4171–4186.

[24] A. Ezen-Can, "A Comparison of LSTM and BERT for Small Corpus," *arXiv Prepr.*, 2020. [Online]. Available: http://arxiv.org/abs/2009.05451.

[25] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Linear-Time WordPiece Tokenization," *arXiv Prepr.*, 2020. [Online]. Available: http://arxiv.org/abs/2012.15524.

[26] T. Kim, K. M. Yoo, and S. Lee, "Self-Guided Contrastive Learning for BERT Sentence Representations," *arXiv Prepr.*, 2021. [Online]. Available: https://arxiv.org/pdf/2106.07345.pdf.

[27] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 2978–2988, 2019.