Enhancing medical NLI with integrated domain knowledge and sentiment analysis

Oleksandr Chaban[1], Eduard Manziuk[1]

[1] *Khmelnytskyi National University, 11, Institutes str., Khmelnytskyi, 29016, Ukraine*

Abstract

Recent advancements in biomedical embeddings derived from language models, such as BioELMo, have demonstrated superior performance in textual inference tasks within the medical domain. In this study, we aim to enhance medical Natural Language Inference (NLI) by integrating structured domain knowledge through a domain knowledge. We employed a state-of-the-art domain knowledge embedding algorithm, MultE, applied to the Unified Medical Language System (UMLS), and combined these embeddings with the BioELMo model. Additionally, we integrated domain-specific sentiment information using MetaMap to further enhance model performance. In our research, we employed the MedNLI dataset, consisting of 14,049 expert-annotated premise-hypothesis pairs derived from clinical notes. Our methods involved the BioELMo embedding model integrated with domain knowledge embeddings and sentiment vectors, processed through a bidirectional LSTM and attention-based architecture. The results showed that our approach achieved an accuracy of 81.14%, precision of 80.08%, recall of 79.62%, F_1-score of 79.85%, and AUC-ROC of 85.06%, significantly outperforming baseline models. These findings indicate that integrating domain-specific knowledge can tangibly enhance the effectiveness of NLI in the medical field. Overall, this work demonstrates the potential of combining advanced embeddings with structured domain knowledge, providing a robust framework for improving clinical decision support and automated medical record analysis.

Keywords

medical natural language inference, domain knowledge embeddings, smart healthcare systems, artificial intelligence, deep learning, clinical decision support

1. Introduction

Over the past decade, natural language inference (NLI) has emerged as a critical task within the broader fields of artificial intelligence (AI) and natural language understanding (NLU). NLI focuses on identifying the logical relationships between a given premise and a hypothesis, such as determining whether the hypothesis is a consequence, contradiction, or neutral with respect to the premise. Such a task is foundational for a lot of applications, including machine reading comprehension, dialogue systems, and information retrieval [1]. While noteworthy progress has been achieved in general domains like fiction [2] and travel [3], the medical domain remains a relatively unexplored frontier [4]. The inherent complexity and specialized nature of medical language, which often includes jargon, abbreviations, and nuanced context-dependent meanings, poses unique challenges in developing effective NLI models for this field.

The introduction of MedNLI [5], a clinically annotated dataset specifically designed for NLI in the medical domain, represents a pivotal step toward bridging the gap between general NLI models and their application in healthcare. MedNLI enables the evaluation and refinement of embedding methods tailored to medical texts, which is essential for the development of downstream applications such as clinical decision support systems and automated medical record analysis. However, the complex nature of medical texts exacerbates the challenges in inference modeling, particularly when these texts are laden with domain-specific terminology and subtle contextual cues. In this context, the importance of explainable artificial intelligence (XAI) becomes paramount [6]. Transparent and interpretable AI models [7] are crucial for earning the trust of healthcare professionals [8] and ensuring the reliability of inferences made from medical data.

[1]*ICST-2024: Information Control Systems & Technologies, September , 23 – 25, 2024, Odesa, Ukraine*

entee94@gmail.com (O. Chaban); eduard.em.km@gmail.com (E. Manziuk)

0009-0001-4710-3336 (O. Chaban); 0000-0002-7310-2126 (E. Manziuk)

Despite the advancements brought by MedNLI, the medical NLI field continues to struggle with significant challenges. The primary problem addressed in this study is the integration of structured domain knowledge and sentiment analysis into NLI models to improve their performance on complex medical texts. This study aims to enhance the accuracy and reliability of medical NLI by combining advanced contextual embeddings with domain-specific knowledge, thus providing a robust solution for clinical decision-making and medical record analysis.

2. Related works

The development of advanced contextual word embedding techniques, such as ELMo [9] and BERT [10], has significantly transformed the landscape of natural language processing (NLP). These models excel at capturing the intricate nuances of language, thereby enabling state-of-the-art performance across a broad range of tasks. Within the specialized field of biomedical NLP, models like BioBERT [11] and BioELMo [12] have been specifically fine-tuned on vast biomedical corpora, such as PubMed abstracts, to enhance the understanding of medical texts. These specialized models have established new benchmarks in medical NLI, demonstrating the importance of domain-specific pre-training [13].

Despite these advancements, the integration of external domain knowledge into NLI models has emerged as a critical area of research to further elevate their performance. Various methods have been explored to incorporate such knowledge into NLP models. For example, ExBERT [14] by Gajbhiye et al. enhanced the BERT model by integrating external knowledge through knowledge-enriched attention mechanisms, local inference collection, and knowledge-enhanced inference composition. Another valuable resource for integrating domain-specific knowledge is the Unified Medical Language System (UMLS) introduced by Amos et al. [15], a comprehensive biomedical ontology. Leveraging UMLS, Sengupta et al. [16] developed knowledge-directed attention-based techniques and methods that integrated medical concept definitions into pre-trained language models. These approaches, when combined with traditional word embeddings such as GloVe presented by Raymundo-Pereira et al. [17] and FastText introduced by Zeghdaoui et al. [18], demonstrated promising improvements in various NLP tasks.

Incorporating domain-specific sentiment information into NLI models marks a significant advance in improving their performance, particularly in the medical domain. For instance, Sharma et al. [19] tackled the challenge of enhancing medical NLI by integrating embeddings from a UMLS-based knowledge graph with domain-specific sentiment information within the BioELMo framework. Their study demonstrated notable improvements in the MedNLI dataset, underscoring the potential of blending domain knowledge with sentiment analysis. However, a critical issue highlighted in their work, and in similar studies, is the incomplete exploration of clinical domain knowledge features. The problem of effectively harnessing the full scope of clinical domain knowledge and sentiment information remains unresolved, leaving a gap in the development of more robust and interpretable NLI models for the medical field [20, 21]. This unaddressed challenge underscores the need for further research to fully realize the potential of these approaches.

Hence, this study aims to address the critical gap in medical NLI by developing a more comprehensive approach that integrates domain knowledge embeddings and sentiment analysis derived from UMLS. Existing methods have not fully leveraged the intricate relationships and sentiment nuances inherent in medical texts, which are crucial for accurate inference. To overcome these limitations, this study proposes a novel approach that combines contextual word embeddings with domain-specific knowledge and integrates sentiment information associated with medical concepts. This work provides the following scientific contributions:

- A novel approach for embedding domain-specific knowledge from UMLS into advanced medical NLU models like BioELMo that aims to enhance the foundational architecture of NLI models, enabling them to better capture the complexities of medical language.
- An enhanced technique for incorporating sentiment information linked to medical concepts from UMLS that demonstrates substantial improvements in the performance and accuracy of medical NLI tasks.

The structure of this paper is organized as follows. Section III details the datasets used, particularly the MedNLI dataset, and elaborates on the proposed approach, including the integration

of BioELMo embeddings, domain knowledge from UMLS, and sentiment analysis via MetaMap. In Section IV, experimental outcomes are presented, comparing the performance of the proposed model with various baselines, followed by an analysis of key performance metrics. Finally, Section V summarizes the findings, emphasizes the enhancements in medical NLI performance, and discusses potential future research directions.

3. Methods and materials

3.1. The proposed approach

In this study, we set up the classification task as a standard NLI problem. This means deciding if a given *hypothesis* can be inferred from a given premise, and then classifying it as either entailment (true), contradiction (false), or neutral (undetermined). We also refer to the methodology described by [19], using the BioELMo embedding model, which includes contextual information from ELMo embeddings trained on ten million PubMed abstracts, combined with the advanced Enhanced Sequential Inference Model (ESIM) [16] for the NLI task. Figure 1 demonstrates the general scheme of our approach.

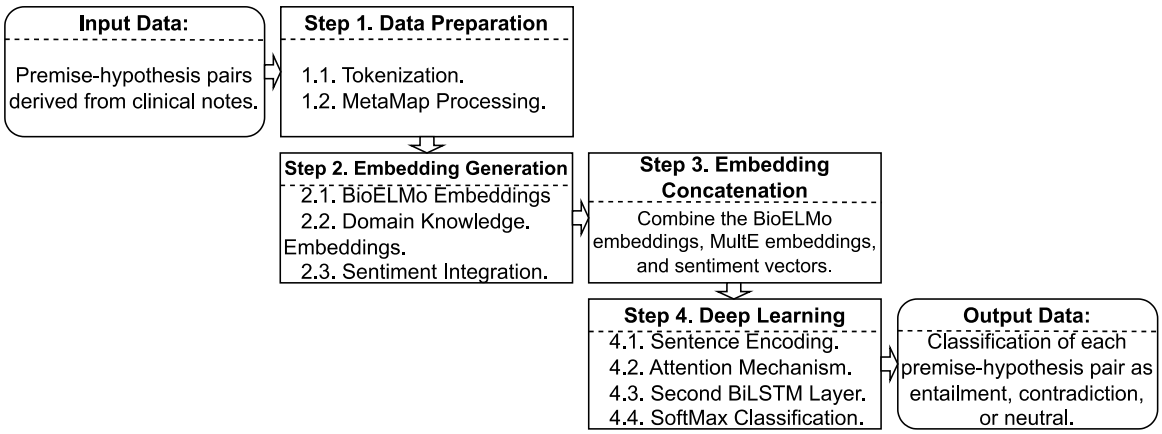

Figure 1: The diagram illustrates the proposed approach for classifying premise-hypothesis pairs derived from clinical notes, detailing the steps from data preparation, and embedding generation to deep learning output and final classification.

Below, we present a step-by-step description of an enhanced approach that we propose in this work.

Input Data: The MedNLI dataset consists of 14,049 expert-annotated premise-hypothesis pairs derived from clinical notes.

Step 1. Data Preparation.

1.1 Tokenization. Each sentence (*premise* and *hypothesis*) is tokenized using the Classical Language Toolkit (CLTK) [23].

1.2 MetaMap Processing. Sentences are processed with MetaMap to extract UMLS concepts and associated sentiment information. Each concept is aligned with constituent words.

Step 2. Embedding Generation:

2.1 BioELMo Embeddings. Contextual word embeddings are generated for each token using the BioELMo model pre-trained on ten million PubMed abstracts.

2.2 Domain Knowledge Embeddings. The MultE model [24] is utilized to generate embeddings for each UMLS concept extracted by MetaMap.

2.3 Sentiment Integration. A sentiment vector is created for each token, where each word has a 1-D vector indicating positive (0) or negative (1) sentiment.

Step 3. Embedding Concatenation.

Next, we combine the BioELMo embeddings, MultE embeddings, and sentiment vectors for each token, similar to our previous work [25], to form a comprehensive embedding (Figure 2).

Step 4. Model Architecture.

4.1 Sentence Encoding. Bidirectional LSTM (BiLSTM) layers are used to encode the embeddings of *premises* and *hypotheses* separately.

4.2 Attention Mechanism. A pairwise attention matrix is computed between the encoded *premise* and *hypothesis*.

4.3 Second BiLSTM Layer. A second BiLSTM layer is applied to the attended representations of *premise* and *hypothesis*. Moreover, we perform max and average pooling on the BiLSTM outputs.

4.4 Softmax Classification. The pooled outputs are fed into a Softmax layer for classification into entailment, contradiction, or neutral categories.

Output Data: The model outputs the classification of each premise-hypothesis pair as entailment, contradiction, or neutral.

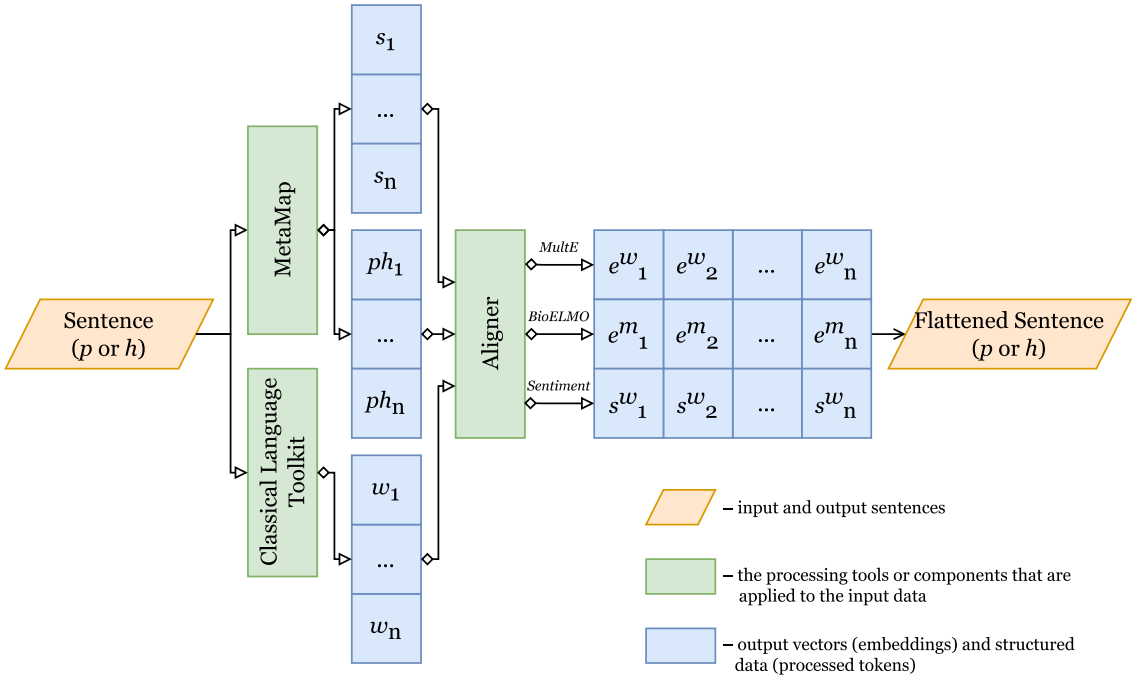

Figure 2: The diagram of the proposed approach based on the BioELMo model to combine contextual word embeddings with domain-specific embeddings; *w* represents the CLTK-tokenized form of the *premise* (*p*) or *hypothesis* (*h*), *ph* signifies the MetaMap-tokenized form of the sentence (*p* or *h*), *s* represents the sentiment vector, and e^w and e^m indicate the aligned word embeddings and MultE embeddings, respectively, with the aligned sentiment vector denoted as s^w.

The architecture of the BiLSTM model features two sentence encoders, each processing the word embeddings of *premise* and *hypothesis* via bidirectional LSTM layers. An attention layer is created by calculating a pairwise attention matrix between the encoded *premise* and *hypothesis*, followed by a second bidirectional LSTM layer applied separately to *premise* and *hypothesis*. Max and average pooling operations are then performed on the LSTM layer outputs, which are subsequently fed into a SoftMax model for classification.

This research uses the Metathesaurus with over 1 million biomedical textual concepts and over 5 million concept labels, all linked by various relationships. Each concept is categorized under one or more Semantic Types, connected through the Semantic Network. The UMLS framework has 127 semantic types and 54 relationships, such as "disease," "symptom," and "laboratory test," with relational types like "is-a," "part-of," and "affects."

Furthermore, MetaMap is a useful tool for linking biomedical text to UMLS Metathesaurus concepts and semantic types. It breaks sentences into phrases based on medical concepts and provides details like each concept's unique ID, sentence position, related semantic types, preferred medical name, and the unique ID for the preferred concept (e.g., "chest pain" is linked to "angina"). It also assigns a Boolean value to indicate if the concept is mentioned negatively (1) or not (0). For example, in "The patient showed no signs of pain," "pain" would be marked negative. Although each phrase can map to multiple medical concepts, this study only considers the mapping with the highest MetaMap Indexing (MMI) score, ensuring each word in a sentence corresponds to at most one medical concept.

To build the domain knowledge, we used MetaMap to process the entire MedNLI dataset, extracting relevant UMLS information into a smaller, more focused set. Medical concepts from both *premises* and *hypotheses* are matched to UMLS's standardized terms, aligning synonymous phrases to the same concept (e.g., "blood clots" matches "thrombus"). This results in 7,496 unique medical concepts within the MedNLI dataset, each represented as a node in the domain knowledge. Relationships are drawn from both the Metathesaurus and the Semantic Network, creating a condensed subgraph of the UMLS. The resulting edge lists in the domain knowledge include 117,467 triples from the Metathesaurus and 23,824,105 triples from the Semantic Network.

For embedding the Domain knowledge, we employed an enhanced version of the MultE model. We hypothesize that a context-aware and regularized MultE model can surpass other domain knowledge embedding models, like in the work [26]. MultE represents entities (nodes) and relationships (edges) as vectors, using non-linear transformations and matrix dot products to evaluate the compatibility between head h (*hypothesis*) and tail t (entities) connected by a relationship r. The improved formalization of MultE embeddings is presented as follows:

$$\sigma_{MultE}^{hrt} = \mathrm{ReLU}(W_r r + b_r)^T \mathrm{ReLU}(W_h h + b_h) \cdot \mathrm{ReLU}(W_t t + b_t), \quad (1)$$

where W_r, W_h, and W_t are weight matrices and b_r, b_h, and b_t are bias terms.

Formulas (1) introduces non-linear transformations to capture more complex interactions.

The enhanced MultE model, as defined above, is employed to integrate domain knowledge embeddings with BioELMo. As depicted in Figure 1, each sentence, whether a *premise* or *hypothesis*, is tokenized using the CLTK and then processed with MetaMap to extract UMLS concepts. These concepts are aligned by associating the UMLS concept for a phrase with all its constituent words.

After aligning tokens using CLTK and MetaMap, we apply BioELMo and the enhanced MultE to generate for each word *w* the embedding vectors $e_{BioELMo}^{w}$ and e_{MultE}^{w}. Instead of simple concatenation, we also employ a weighted sum approach to form the representation of each word, allowing the model to dynamically adjust the importance of each embedding type. The representation is thus formed as follows:

$$e^{w} = \alpha \cdot e_{BioELMo}^{w} + \beta \cdot e_{MultE}^{w}. \quad (2)$$

where α and β are learnable coefficients; this approach provides a flexible and optimized combination of embeddings.

To enhance domain knowledge further, we integrate sentiment information for each concept separately. MetaMap provides a sentiment Boolean for each concept, which we use to create a 1-D vector s^w that contains 0 for positive or nonmedical concepts and 1 for negative concepts. This 1-D vector is aligned with e_{MultE}^{w} in the same manner as previously described. Additionally, we integrate a sentiment weighting mechanism where the sentiment vector's impact is modulated based on its relevance to the word context using an attention mechanism. Consequently, the final embedding for each word is represented as follows:

$$e^{w} = \sum_{i=1}^{n} a_i \left(e_{BioELMo,i}^{w} + e_{MultE,i}^{w} + s_i^{w} \right), \quad (3)$$

where a_i are the attention weights calculated as:

$$a_i = \frac{\exp\left(e_{BioELMo,i}^{w} + e_{MultE,i}^{w} + s_i^{w} \right)}{\sum_{j=1}^{n} \exp\left(e_{BioELMo,j}^{w} + e_{MultE,j}^{w} + s_j^{w} \right)},$$

where *n* is the number of elements in the concatenated vector.

In formula (3), we use the standard ESIM model [5], inputting the weighted embeddings for each word in both the *premise* and *hypothesis* to train the model for the medical NLI task.

By integrating these advanced techniques (1)–(3), we aim to significantly enhance the performance of medical NLI tasks of our model.

3.2. Data collection and processing

This study utilizes the MedNLI dataset [5], a valuable resource in the domain of NLI specifically tailored for clinical applications. MedNLI is derived from clinicians' notes within the MIMIC-III clinical dataset, which is renowned for being the most extensive publicly accessible collection of de-identified patient records. The dataset comprises 14,049 premise-hypothesis pairs, meticulously categorized into three subsets: 11,232 pairs for training, 1,395 pairs for validation, and 1,422 pairs for

testing. Each pair is annotated with one of three labels: entailment, contradiction, or neutral, indicating whether the *hypothesis* logically follows from the *premise*, contradicts it, or is unrelated, respectively.

In terms of linguistic characteristics, the average length of *premises* is 20 words, while *hypotheses* average 5.8 words. The dataset is designed to encompass a wide range of clinical language variability, with the maximum length of *premises* reaching 202 words and *hypotheses* extending up to 20 words. This variability underscores the complexity and richness of the clinical narratives captured within the MedNLI dataset.

The MedNLI dataset was preprocessed to ensure uniformity and consistency across all data points. This involved tokenizing the clinical notes using the CLTK and processing each sentence through MetaMap to extract relevant UMLS concepts and their associated sentiment information.

3.3. Performance criteria

Here, we provide the performance metrics for three-class classification tasks based on MedNLI dataset using standard metrics: accuracy, precision, recall, F_1-score, and area under the curve (AUC-ROC).

First of all, we used the Accuracy metric, which is the ratio of correctly predicted cases to the total number of cases.

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} \text{TP}_k}{\sum_{k=1}^{K} (\text{TP}_k + \text{FP}_k + FN_k)}, \tag{4}$$

where *K* is the number of classes, *k* stands for the index of each class, TP – true positives, FP – false positives, FN – false negatives, and TN – true negatives for each class.

Precision for a specific class *k* is the ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k}. \tag{5}$$

Recall for a specific class *k* is the ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}. \tag{6}$$

The F_1-score is the harmonic mean of Precision and Recall. For a specific class *k*:

$$\text{F}_1 - \text{score}_k = 2 \times \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{7}$$

AUC-ROC is a performance measurement for classification problems at various threshold settings. For multi-class classification, the average AUC can be computed by averaging the AUC of each class against all other classes.

$$\text{AUC} - \text{ROC} = \frac{1}{K} \sum_{k=1}^{K} \text{AUC}_k, \tag{8}$$

where AUC_k is the AUC for class *k* computed as:

$$\text{AUC}_k = \int_0^1 \text{TPR}_k(\text{FPR}_k) d(\text{FPR}_k),$$

where TPR_k (True Positive Rate) and FPR_k (False Positive Rate) are functions of the threshold.

4. Results and discussion

For the MultE model, we configured the word embedding dimensions to 100, which ensures that each word is represented in a 100-dimensional space. This dimensionality was chosen to balance the representation's richness with computational efficiency. We employed Stochastic Gradient Descent (SGD) for optimization, selected due to its effectiveness in handling large-scale and sparse data. The initial learning rate was set to 10^{-3}, a common starting point that allows for significant parameter updates in the early stages of training. The batch size was fixed at 64, ensuring that the model processes a reasonable amount of data per iteration.

For the ESIM, we utilized BiLSTM networks. Each BiLSTM had a hidden state dimension of 500, which allows the model to capture complex dependencies and context from both forward and backward sequences in the input data. Dropout, a regularization technique to prevent overfitting, was applied with a rate of 0.4. This rate was chosen to effectively reduce the chance of overfitting by randomly setting 40% of the hidden units to zero during training. The initial learning rate for the ESIM model was also set to 10^{-3}, which aids in making significant updates to the weights initially, speeding up convergence.

The batch size for ESIM was set to 32. A smaller batch size allows for more updates per epoch, which can lead to better generalization, albeit at the cost of increased training time. We limited the training process to a maximum of 64 epochs, ensuring that the model does not overfit the training data. Additionally, an early stopping mechanism was implemented; training was halted if the development loss did not decrease for 5 consecutive epochs. This criterion prevents the model from overtraining, which could lead to poor performance on unseen data.

Table 1 presents a comparative analysis of various models evaluated on a specific task, with metrics including Accuracy (4), Precision (5), Recall (6), F_1-score (7), and AUC-ROC (8).

Table 1 shows the performance of our proposed models compared to the baselines, demonstrating that integrating domain knowledge embeddings enhances model performance. In fact, the highest accuracy is achieved by the proposed approach (BioELMo + Sentiment) at 81.14%. This model outperforms all others, including ESIM-know (80.22%) and BioELMo + Sentiment (80.60%). Notably, models integrating sentiment or domain knowledge tend to perform better, indicating the effectiveness of these additional features.

Table 1
Performance comparison of various models on MedNLI dataset, evaluating metrics including accuracy, precision, recall, F1-score, and AUC-ROC. All values are presented in %. The highest values are in bold.

Model	Accuracy	Precision	Recall	F_1-score	AUC-ROC
fastText [18]	73.50	72.00	71.05	71.57	75.16
GloVe [17]	76.42	75.12	74.60	74.82	78.20
BioELMo [12]	79.73	78.55	78.15	78.23	82.17
ESIM-know [16]	80.22	79.05	78.67	78.76	83.08
fastText + Sentiment	78.59	77.58	77.80	77.17	80.70
GloVe + Sentiment	79.10	78.08	77.43	77.74	81.35
BioELMo + Sentiment [19]	80.60	79.48	79.30	79.19	84.03
The proposed approach	81.14	80.08	79.62	79.85	85.06

The proposed approach also leads in precision with 80.08%, closely followed by BioELMo + Sentiment (79.48%) and ESIM-know (79.05%). Higher precision values suggest that the proposed model is better at correctly identifying relevant instances, reducing false positives. BioELMo + Sentiment shows a slightly higher recall (79.30%) compared to the proposed approach (79.62%). This implies that the proposed model is more effective in identifying true positives, with fewer false negatives. Additionally, the F_1-score, which balances precision and recall, is highest for the proposed approach (79.85%), reflecting its superior overall performance. ESIM-know (78.76%) and BioELMo + Sentiment (79.19%) also demonstrate strong performance, but the proposed model's integration of sentiment analysis enhances its balanced effectiveness. Moreover, the proposed approach achieves the highest AUC-ROC score of 85.06%, indicating excellent performance in distinguishing between classes. This score suggests that the model has a strong ability to rank positive instances higher than negative ones.

It is also worth noticing that the fastText model shows the lowest performance across all metrics, which is expected given its relatively simpler architecture compared to more advanced embeddings and the use of domain knowledges or sentiment analysis. Integrating sentiment features consistently boosts performance, as seen in the fastText + Sentiment, GloVe + Sentiment, and the proposed approach. Similarly, leveraging domain-specific embeddings like BioELMo and integrating domain knowledges result in significant performance improvements.

Figure 3 presents the ROC curves and AUC values for investigated deep learning on the MedNLI dataset through 5-fold cross-validation.

The baseline performance, represented by the dashed red line, serves as a reference for random guessing with an AUC of 0.5. The models compared include fastText, GloVe, BioELMo, and ESIM-know, both with and without sentiment analysis integration. Among the individual models, BioELMo (green line) shows a notable performance with an AUC of 82.17, while ESIM-know (yellow line) achieves a slightly higher AUC of 83.08.

When sentiment analysis is integrated, the models' performance generally improves. The BioELMo + Sentiment model (gray line) achieves an AUC of 84.03, while the GloVe + Sentiment model (brown line) reaches an AUC of 81.35. The proposed approach outperforms all individual and sentiment-enhanced models with an AUC of 85.06, indicating its superior discriminative capability in the medical NLI task.

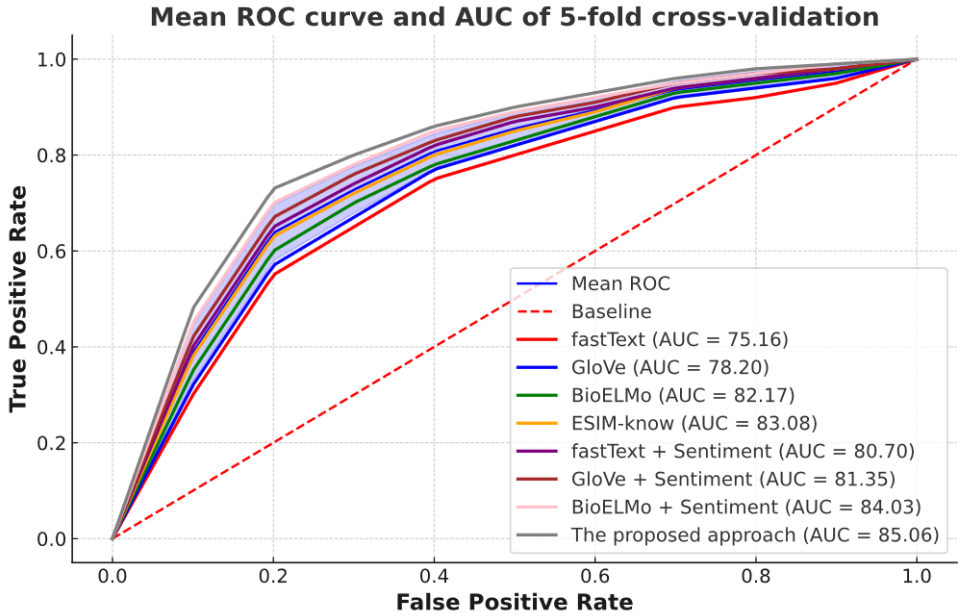

Figure 3: The ROC curves and corresponding AUC values for various deep learning models on the MedNLI dataset, illustrating the performance of each model under 5-fold cross-validation.

Overall, Figure 3 clearly demonstrates that integrating domain-specific enhancements and sentiment analysis can significantly boost model performance in medical NLI tasks, with the proposed approach leading the pack in terms of accuracy and reliability.

The quality performance of our approach is demonstrated through several instances. For example, in the sentence pair *premise*: "Diagnosis of COPD" and *hypothesis*: "Patient has chronic obstructive pulmonary disease," the term 'COPD' is correctly identified as synonymous with "chronic obstructive pulmonary disease," leading to the correct classification as entailment.

Our model also effectively captures negative sentiment, such as in the pair *premise*: "Patient presents with abdominal pain, no signs of infection for the past 6 months," and *hypothesis*: "Patient has no current infection," where BioELMo misclassifies it as contradiction while our model correctly identifies it as entailment.

However, there are still incorrect cases. For instance, the pair *premise*: "He was walking steadily at that moment" and *hypothesis*: "The patient never had a steady walk" is incorrectly classified as entailment.

The absence of negative sentiment detection for "walking steadily" by MetaMap contributes to this mistake. Another challenging example is the *premise*: "There were no changes in blood pressure, and the initial blood tests were normal," and *hypothesis*: "The patient has normal blood tests." Our model classifies it as entailment despite the gold label being neutral, indicating the need for better temporal context handling.

Overall, the proposed approach demonstrates superior performance across all metrics. This suggests that combining advanced embeddings with sentiment analysis effectively enhances the model's capability in the given task. The results highlight the importance of integrating domain-specific knowledge and additional contextual features to enhance model performance.

5. Conclusion

This study demonstrates that integrating domain knowledge embeddings from models like MultE, derived from UMLS, with BioELMo and sentiment analysis using MetaMap, significantly enhances the performance of medical NLI tasks. The proposed approach achieved an accuracy of 81.14%, precision of 80.08%, recall of 79.62%, F1-score of 79.85%, and AUC-ROC of 85.06%, outperforming baseline models such as BioELMo alone and ESIM with integrated knowledge. These improvements highlight the effectiveness of combining domain-specific knowledge and sentiment analysis to capture the complex nuances of medical texts. However, limitations persist, particularly in accurately detecting subtle negative sentiment and effectively handling temporal information within clinical data. The model occasionally misclassifies sentiments in nuanced contexts, indicating a need for more refined sentiment analysis techniques.

Future research should focus on developing more sophisticated methods for temporal context processing and refining sentiment analysis to better capture subtle cues. Expanding this approach to other medical datasets and exploring its applicability to broader clinical decision support tasks will also be critical in advancing the field.

6. References

[1] A. Turkmen, O. Can, Natural language processing for ontology development in IoT-enabled smart healthcare, in: Studies on the Semantic Web, IOS Press, Amsterdam, The Netherlands, 2024, pp. 88–108. doi:10.3233/ssw230027.

[2] I. K. Raharjana, D. Siahaan, C. Fatichah, User stories and natural language processing: A systematic literature review, IEEE Access 9 (2021) 53811–53826. doi:10.1109/access.2021.3070606.

[3] P. Eleftheriadis, I. Perikos, I. Hatzilygeroudis, Evaluating deep learning techniques for natural language inference, Appl. Sci. 13.4 (2023) 2577. doi:10.3390/app13042577.

[4] C. Herlihy, R. Rudinger, MedNLI Is not immune: Natural language inference artifacts in the clinical domain, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2021, pp. 1020–1027. doi:10.18653/v1/2021.acl-short.129.

[5] A. Romanov, C. Shivade, Lessons from natural language inference in the clinical domain, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2018, pp. 1586–1596. doi:10.18653/v1/d18-1187.

[6] P. Radiuk, O. Barmak, E. Manziuk, I. Krak, Explainable deep learning: A visual analytics approach with transition matrices, Mathematics 12.7 (2024) 1024. doi:10.3390/math12071024.

[7] D. W. Joyce, A. Kormilitzin, K. A. Smith, A. Cipriani, Explainable artificial intelligence for mental health through transparency and interpretability for understandability, NPJ Digit. Med. 6.1 (2023) 6. doi:10.1038/s41746-023-00751-9.

[8] O. Wysocki, J. K. Davies, M. Vigo, A. C. Armstrong, D. Landers, R. Lee, A. Freitas, Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making, Artif. Intell. 316 (2023) 103839. doi:10.1016/j.artint.2022.103839.

[9] M. Ulčar, M. Robnik-Šikonja, Cross-lingual alignments of ELMo contextual embeddings, Neural Comput. Appl. 34 (2022) 13043–13061. doi:10.1007/s00521-022-07164-x.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36.4 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.
[12] Q. Jin, B. Dhingra, W. Cohen, X. Lu, Probing biomedical embeddings from language models, in: Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 82–89. doi:10.18653/v1/w19-2011..
[13] P. Radiuk, O. Kovalchuk, V. Slobodzian, E. Manziuk, O. Barmak, Human-in-the-loop approach based on MRI and ECG for healthcare diagnosis, in: Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine, CEUR-WS.org, Aachen, Germany, 2022, pp. 9–20.
[14] A. Gajbhiye, N. A. Moubayed, S. Bradley, ExBERT: An external knowledge enhanced BERT for natural language inference, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 460–472. doi:10.1007/978-3-030-86383-8_37.
[15] L. Amos, D. Anderson, S. Brody, A. Ripple, B. L. Humphreys, UMLS users and uses: A current overview, J. Am. Med. Inform. Assoc. 27.10 (2020) 1606–1611. doi:10.1093/jamia/ocaa084.
[16] S. Sengupta, C. Heaton, P. Mitra, and S. Sarkar, Leveraging external knowledge resources to enable domain-specific comprehension. arXiv, Jan. 15, 2024. doi: 10.48550/arXiv.2401.07977.
[17] P. A. Raymundo-Pereira, N. O. Gomes, S. A. S. Machado, O. N. Oliveira, Wearable glove-embedded sensors for therapeutic drug monitoring in sweat for personalized medicine, Chem. Eng. J. 435.2 (2022) 135047. doi:10.1016/j.cej.2022.135047.
[18] M. W. Zeghdaoui, O. Boussaid, F. Bentayeb, F. Joly, Medical-Based text classification using FastText features and CNN-LSTM model, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 155–167. doi:10.1007/978-3-030-86472-9_15.
[19] S. Sharma, B. Santra, A. Jana, S. Tokala, N. Ganguly, P. Goyal, Incorporating domain knowledge into medical NLI using knowledge graphs, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 6092–6097. doi:10.18653/v1/d19-1631.
[20] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, S. Yu, A survey on incorporating domain knowledge into deep learning for medical image analysis, Med. Image Anal. 69 (2021) 101985. doi:10.1016/j.media.2021.101985.
[21] X. Xu, L. Zhao, J. Li, L. Li, Incorporating medical domain knowledge into data-driven method: A vessel attention guided multi-granularity network for automatic cataract classification, Expert Syst. With Appl. 241 (2024) 122671. doi:10.1016/j.eswa.2023.122671.
[22] A. R. Aronson, F.-M. Lang, An overview of MetaMap: Historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17.3 (2010) 229–236. doi:10.1136/jamia.2009.002733.
[23] K. P. Johnson, P. J. Burns, J. Stewart, T. Cook, C. Besnier, W. J. B. Mattingly, The classical language toolkit: An NLP framework for pre-modern languages, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Stroudsburg, PA, USA, 2021, pp. 20–29. doi:10.18653/v1/2021.acl-demo.3.
[24] Z. Huang, B. Li, J. Yin, Knowledge graph embedding via multiplicative interaction, in: Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence, Association for Computing Machinery, New York, NY, USA, 2018, pp. 138–142. doi:10.1145/3194206.3194227.
[25] E. Manziuk, W. Wojcik, O. Barmak, I. Krak, A. Kulias, V. Drabovska, V. Puhach, S. Sundetov, A. Mussabekova, Approach to creating an ensemble on a hierarchy of clusters using model decisions correlation, Przeglad Elektrotechniczny 96.9 (2020) 108–113. doi:10.15199/48.2020.09.23.

[26] M. Moradi, N. Ghadiri, Quantifying the informativeness for biomedical literature summarization: An itemset mining method, Comput. Methods Programs Biomed. 146 (2017) 77–89. doi:10.1016/j.cmpb.2017.05.011.