An Ensemble Approach for Automatic Structuring of Radiology Reports

Morteza Pourreza Shahri[1]
CodaMetrix

Amir Tahmasebi[1]
CodaMetrix

Bingyang Ye[1]
CodaMetrix

Henghui Zhu[2]
Amazon

Javed Aslam[1]
CodaMetrix

Timothy Ferris[3]
Mass General Physicians
Organization

[1]{morteza,amir,bingyang,jay}@codametrix.com
[2]henghui@amazon.com
[3]ferris.timothy@mgh.harvard.edu

Abstract

Automatic structuring of electronic medical records is of high demand for clinical workflow solutions to facilitate extraction, storage, and querying of patient care information. However, developing a scalable solution is extremely challenging, specifically for radiology reports, as most healthcare institutes use either no template or department/institute specific templates. Moreover, radiologists' reporting style varies from one to another as sentences are telegraphic and do not follow general English grammar rules. We present an ensemble method that consolidates the predictions of three models, capturing various attributes of textual information for automatic labeling of sentences with section labels. These three models are: 1) Focus Sentence model, capturing context of the target sentence; 2) Surrounding Context model, capturing the neighboring context of the target sentence; and finally, 3) Formatting/Layout model, aimed at learning report formatting cues. We utilize Bi-directional LSTMs, followed by sentence encoders, to acquire the context. Furthermore, we define several features that incorporate the structure of reports. We compare our proposed approach against multiple baselines and state-of-the-art approaches on a proprietary dataset as well as 100 manually annotated radiology notes from the MIMIC-III dataset, which we are making publicly available. Our proposed approach significantly outperforms other approaches by achieving 97.1% accuracy.

1 Introduction

Electronic medical records (EMRs), such as radiology reports, contain patient clinical information and are often in the form of "natural language" written or transcribed by providers (Denny et al., 2008). Gathering and disseminating patient information from such notes is required for patient care management. Natural Language Processing (NLP)-driven solutions have been proposed to augment clinical workflows to facilitate such information extraction and structuring processes. Segmentation of medical reports into topically cohesive sections (Cho et al., 2003) is essential for NLP tasks such as relation extraction, Named Entity Recognition (NER), and Question and Answering.
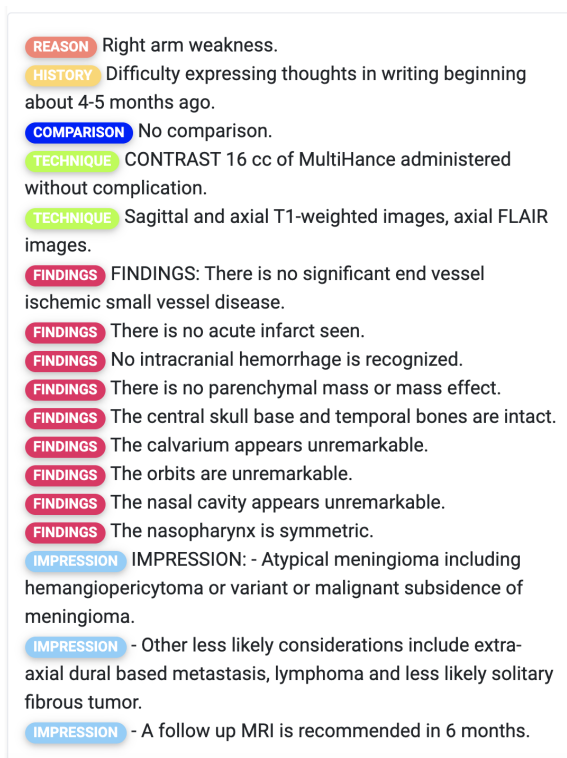

Figure 1: Snapshot of the output of our proposed model on a radiology report. Labels are shown in front of every extracted sentence.

Developing a universal and scalable report segmenting solution is extremely challenging as most healthcare institutes use either no template or institute specific templates. Moreover, providers' style of reporting varies from one to another as sentences are written in a telegraphic format and generally do not follow English grammar rules. Nonethe-

less, in the case of radiology reports, the reports are often composed of similar sections, including the reason for the visit, the performed examination, a summary of observations and findings, and finally, the radiologist's impression and recommendation based on the observations.

To extract and structure patient information from notes, most clinical institutes take the approach of developing their specific set of patterns and rules to extract and label the sections within the clinical reports. This requires a substantial amount of effort for defining rules and maintaining them over time. With advancements in machine learning and NLP, researchers have more recently utilized supervised machine learning methods for automatic structuring of radiology reports (Apostolova et al., 2009; Tepper et al., 2012; Haug et al., 2014; Singh et al., 2015; Rosenthal et al., 2019). These machine learning approaches can be divided into three main themes: 1) Methods that solely rely on extracting features from the format of the text and, therefore, are biased on the specific format of the training data (Tepper et al., 2012); 2) More recent efforts that are focused on learning to label based on the context (Rosenthal et al., 2019); and finally, 3) The hybrid approaches that combine formatting and context-driven features (Apostolova et al., 2009). The two latter methods require a reasonably large amount of annotated reports and yet are not scalable solutions as they do not adequately address inter-institute variability unless model training is fine-tuned using annotated data from the target institute.

In this work, we frame the structuring of the radiology reports as a multi-class sentence classification problem. More specifically, this work presents a novel framework to identify various sections in the radiology reports and to label all sentences within the note with their corresponding section category. We propose an ensemble approach that takes advantage of formatting cues as well as context-driven features. We incorporate Recurrent Neural Networks (RNN) and sentence encoders accompanied by a set of engineered features from the reports for the task of section labeling in radiology reports. The proposed approach considers the context of the current text span and the surrounding context that helps make more accurate predictions.

We were motivated by how a non-expert human self-teaches to perform such a task, paying attention to the context while taking formatting cues into account. We hypothesize that each of the three models learns unique and non-overlapping attributes for solving the problem at hand, and therefore, an ensemble approach seems reasonable.

In order to avoid the requirement of access to a large annotated training corpus, we follow a weak learning approach in which we automatically generate the initial training data using generic rules that are implemented using regular expressions and pattern matching.

We consider seven types of section categories and label each sentence with one of these categories. Our approach is not limited to these specific categories and it can be adapted for any template format and writing style. This is thanks to incorporating a broad set of features that are independent of physicians/institutions. Figure 1 depicts a snapshot of the output of our proposed model for automatic labeling of the sentences within a radiology report. The label in front of each line represents the predicted label for the following sentence.

We train and evaluate our proposed approach on a large multi-site radiology report corpus from Mass General Brigham, referred to as MGB. We demonstrate that our proposed solution significantly outperforms common existing methods for automated structuring of radiology reports (Apostolova et al., 2009; Singh et al., 2015) as well as several baseline models. Moreover, we manually annotated 100 reports from the MIMIC-III radiology reports corpus (Johnson et al., 2016), and we report performances on this dataset as well. We also make this dataset publicly available to other researchers[1].

Our main contributions in this study are as follows:

1. Investigating the importance of different types of features, including formatting and layout, as well as semantics and context in section labeling of radiology notes at the sentence level.

2. Achieving state-of-the-art performance for automatic labeling of radiology notes with predefined section labels through an ensemble approach incorporating models that are capable of learning context and formatting features.

3. Contributing 100 manually-annotated clinical notes with section labels at sentence-level randomly selected from the MIMIC-III corpus.

---

[1] https://doi.org/10.5281/zenodo.4074194

The rest of the paper is organized as follows. In section 2, we briefly review current methods for segmenting and structuring clinical reports. Next, we describe our proposed pipeline in section 3. In section 4, we present and discuss our results on independent test sets, and finally, the conclusions and potential future work are presented in section 5.

2 Related Work

There have been numerous efforts to address the need for automatic structuring of clinical notes via section labeling, including rule-based methods, machine learning-based methods, and hybrid approaches (Pomares-Quimbaya et al., 2019).

Taira et al. (2001) proposed a rule-based system comprising a structural analyzer, lexical analyzer, parser, and a semantic interpreter to identify sections in radiology reports. Denny et al. (2008) developed a hierarchical section header terminology and a statistical model to extract section labels from sentences. RadBank was introduced by Rubin and Desser, which recognizes the structure of radiology reports and extracts the sections for indexing and search, which falls in rule-based methods (Rubin and Desser, 2008). A known shortcoming of rule-based approaches is that they perform well only on reports that follow a specific template and are written following strict structures. As a result, rule-based systems require updating rules/patterns for each new dataset with new formatting and structure. Furthermore, rule-based approaches perform poorly on reports lacking a coherent structure and/or are not written using a predefined template.

Machine learning-based methods solve this problem by training models that can be applied to other datasets without substantial changes as they learn to rely on features beyond formatting and layout. Singh et al. (2015) presented a system based on the Naïve Bayes classifier to identify sections in radiology reports. (Tepper et al., 2012) employ Maximum Entropy to label various sections in discharged summaries and radiology reports. Cho et al. (2003) proposed a hybrid system to extract and label sentences from medical documents. Their proposed system is composed of a rule-based module that detects the sections with labels and a machine learning classifier that detects the unlabeled sections. Apostolova et al. (2009) employed a set of rules for creating a high-confidence training set and applied Support Vector Machines (SVM) trained

on additional formatting and contextual features to label the sentences from radiology and pathology reports. The main challenge in training such machine learning-based approaches is the need for a relatively large annotated training data.

To the best of our knowledge, the most recent work is proposed by Rosenthal et al. (2019) in which they present a system based on an RNN and a BERT (Devlin et al., 2019) model for predicting sections in EHRs. They use sections from the medical literature (e.g., textbooks, journals, web content) with similar content in EHR sections.

Even though the existing methods address the problems mentioned earlier for the complex task of automatic structuring of radiology reports, an ensemble of several models is shown to yield lower generalization error, as opposed to training individual models (Kotu and Deshpande, 2014).

3 Methodology

3.1 Approach

In this work, we formulate the task of automated structuring of radiology reports as a supervised multi-class text classification problem. We define the label set as *Reason for Visit*, *History*, *Comparison*, *Technique*, *Findings*, *Impression*. Any sentence that cannot be categorized as one of the classes above is labeled as *Others*.

Suppose we have the context $C = s_1 s_2 ... s_n$, where s_i is a sentence in the radiology report. We define a mapping function f that for each sentence s_i from the set of sentences in the report, it maps the sentence to its associated label. The context C can be the entire radiology report or a few sentences from the report. The following sections describe the details of our proposed methodology.

3.2 Dataset

Since we do not have access to a publicly-available dataset, we build our own training set using the radiology reports from a multi-institution clinical corpus from Mass General Brigham referred to as the MGB dataset. We randomly selected 856 radiology reports from 12 different clinical sites, i.e., Mass General Brigham. Taking the template and specific formatting/layout of the notes, we develop a weak labeler using regular expressions to detect keywords, including *Findings, Impression, Technique, Comparison, Reason for Visit, History, Indications, Type*, and *Procedure*. Subsequently, we consider all of the sentences between two observed
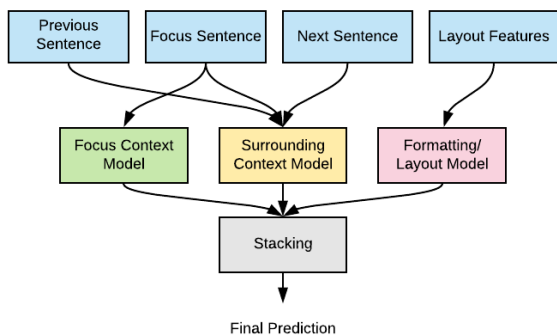
Figure 2: The ensemble model composed of the Focus Context, Surrounding Context, and Formatting/Layout models that combines the three prediction using the *Stacking* method.

keywords as the preceding section. For instance, if the keyword "Findings" appears at position 400 and "Impression" appears at location 700, any sentence in the range of [400, 700) is labeled as *Findings*. One should note that the occurrence of all the keywords in radiology reports is not guaranteed. Therefore, we only pick the sections that appear in the report. Next, we merge *Technique, Procedure*, and *Type* into one category since they convey the same concept. We also combine *History* and *Indications* into one class. Furthermore, we manually correct the automatically assigned labels of sentences using the BRAT annotation tool (Stenetorp et al., 2012). Since only one human annotator corrected the labels, there is no inter-annotator agreement. We split this dataset into three: 686 reports (80%) as the training set, 85 reports (10%) for training the ensemble model, and 85 reports (10%) as the test set.

To comply with data privacy and Health Insurance Portability and Accountability Act (HIPAA), we cannot release this dataset. Nonetheless, we randomly select a separate 100 reports from MIMIC-III corpus with the *CATEGORY* code of Radiology. Subsequently, we manually annotate this dataset, similar to the way we created the MGB dataset, and we employ it as an independent test set. The annotations were performed by two of the co-authors as non-domain experts. No inter-annotator agreement was measured as there was no overlap between labeled reports by two annotators.

3.3 Preprocessing

The preprocessing includes removing special characters while keeping lowercase, uppercase, and digits from the text and replacing all other char-

acters with space. We use Simple Sentence Segment[2] for sentence parsing. Subsequently, all of the sentences are tokenized using the SentencePiece tokenizer (Kudo and Richardson, 2018).

We utilize GloVe (Pennington et al., 2014) word embeddings trained in-house on the entire set of radiology reports from multiple-sites (more than two million radiology reports). The pre-trained word embeddings are 300-dimensional. We also repeated our experiments by utilizing the BERT (Devlin et al., 2019) embeddings, trained in-house on the same corpus of radiology reports, as mentioned above. Overall, the GloVe embeddings yield higher performance for the desired task compared to the BERT embeddings. Therefore, for all of the experiments, we report the performance using the GloVe embeddings.

3.4 Model

Figure 2 demonstrates the proposed ensemble architecture. As can be seen from the figure, the three models aim to capture and encode formatting information, focus sentence context, as well as the context from the surrounding sentences of the focus sentence.

The intuition for having three models is that relying on one source, either context or format alone, is insufficient to capture all necessary text attributes for the labeling task. For example, a sentence such as "Microlithiasis." may occur in *History*, *Findings* or *Impression* sections and only by taking sentence context, the surrounding context, and the formatting cues altogether, one can determine the most appropriate label.

We combine the individual models' predictions using the *Stacking* method (Wolpert, 1992) to derive the final prediction. The architecture of each model is discussed in detail in the following sections.

3.4.1 Focus Context Model

As shown in Figure 3, the proposed architecture for the Focus Context model is composed of a Bi-directional Long Short-Term Memory (LSTM) with 64 units. Subsequently, we encode the sentence using the LSTM's output sequences using max-over-time and mean-over-time pooling and concatenate these two vectors (Blanco et al., 2020). This approach enables us to extract meaningful features from the focus sentence context. The encoded

---

[2]https://github.com/noc-lab/simple_sentence_segment

sentence is next passed to a fully-connected layer with 100 neurons with the *ReLU* activation function and a dropout value of 50%. We stack two more fully-connected layers with sizes of 30 and 16 with the *ReLU* activation functions and dropout values of 50% and 30%, respectively. Finally, the weights are passed to the output layer that employs a *Softmax* activation function to make the final prediction.
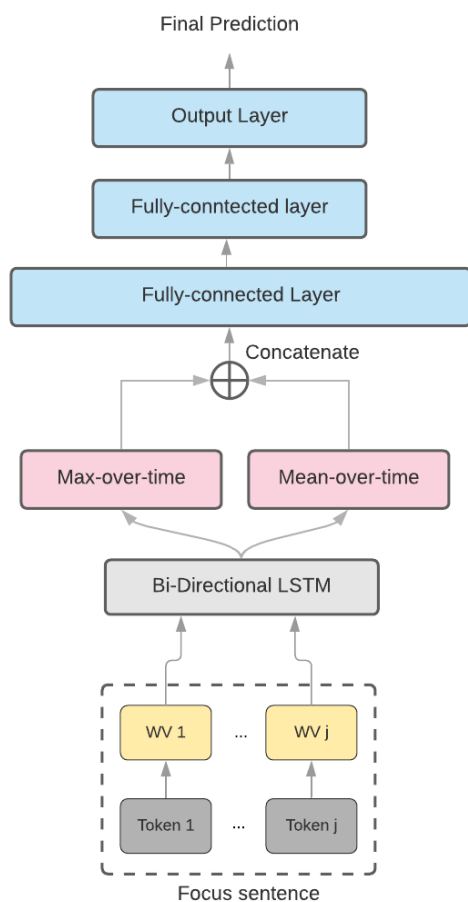

Figure 3: The network architecture of the Focus Context model.

3.4.2 Surrounding Context Model

Figure 4 demonstrates the proposed architecture for the Surrounding Context model. The surrounding context is defined as the sentence immediately before and the sentence immediately after the focus sentence. The most efficient size of the surrounding context can be determined through hyper-parameter tuning, which is beyond the scope of this work and is considered for future work. Each sentence is fed into a Bi-directional LSTM layer. The LSTM layer for the focus sentence comprises 64 units, whereas the LSTM layers of surrounding sentences have 16

units. Next, each Bi-LSTM layer's output sequence is fed into a max-over-time pooling layer to encode the sequence. The three sentence encoders' outputs are concatenated and passed into a fully-connected layer with 50 neurons and *ReLU* activation function. This layer is followed by a Dropout layer with a value of 50%. The weights are passed to a fully-connected layer with ten neurons and a dropout value of 30%. Subsequently, the output is fed into a second fully-connected layer with seven neurons and the *Softmax* activation function to obtain the final prediction. In cases where the focus sentence appears at either the beginning or end of a report, we use an empty string for the sentence before or after.

3.4.3 Formatting/Layout Model

We propose a third model to learn formatting/layout related features using neural networks. Motivated by a prior work (Apostolova et al., 2009), we define 17 features that are described as follows:

1. Number of uppercase, lowercase, and digits in the sentence (three features).

2. Normalized relational position of the focus sentence to each section headers by searching keywords such as reason, history/indications, procedure/technique, comparison, findings, and impression (six features).

3. If the last character of the previous sentence, the current sentence, and the next sentence is either period or colon (six features).

4. Normalized position of the current sentence in the report (one feature).

5. If the first token in the sentence is uppercase or not (one feature).

These features are utilized as input to a neural network with a stack of three fully-connected layers with 100, 16, and seven neurons. We add the *ReLU* activation functions for the first two layers and the *Softmax* function for the last layer. The first two layers are followed by dropout layers with values of 50%.

3.4.4 Ensemble: Stacking

As the last step, we train a Logistic Regression (LR)-based ensemble model using the three models described in the previous sections and using a holdout stacking set. We start making predictions
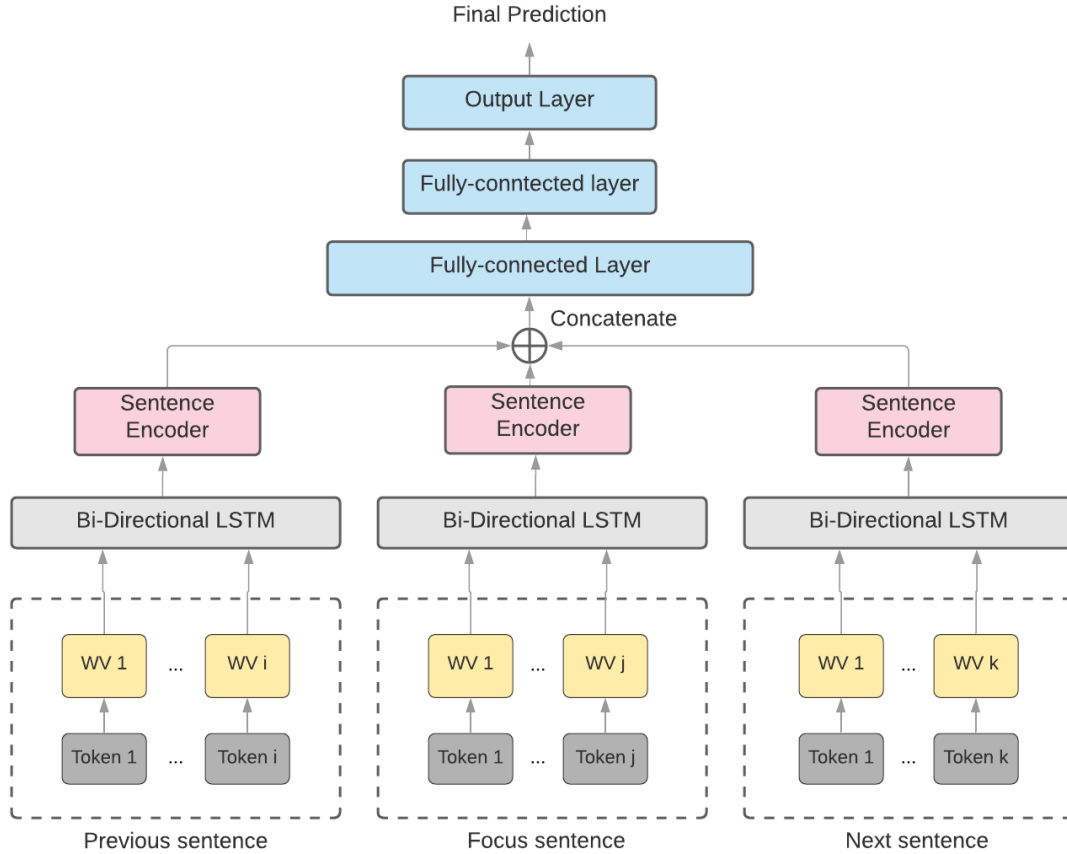
Figure 4: The network architecture of the Surrounding Context model.

using the three models on the holdout set, and we train an LR classifier on their predicted probabilities using Equation 1,

$$p(y = 1) = \sigma(w^T x + b) \quad (1)$$

where w and b are parameters to learn from data, and σ is the Sigmoid function. We perform "one-versus-rest" for multi-class classification. The trained classifier can be utilized for making accurate predictions on the test set.

3.5 Experimental Setup

We implement four baseline models to compare with our proposed model. The first baseline is a rule-based model using the regular expressions specifically assembled based on the format of radiology reports from the MGB dataset. We refer to this model as the *MGB Rule-based* model. The second baseline is also a rule-based model composed of rules designed specifically for the MIMIC-III dataset. We refer to this model as the *MIMIC Rule-based* model. The third baseline model is a neural network consist of similar architecture to ours, but instead of stacking, we concatenate the outputs and

pass it to a fully-connected layer. We refer to this model as the *Merged* model. We also compare our proposed ensemble model with a Linear SVM model with "balanced" class weights, trained on preprocessed sentences in the form of uni-gram TFIDF vectors.

Most prior approaches utilize specific labeling schema that differ from ours and the corresponding labeled datasets are not publicly available (Cho et al., 2003; Rubin and Desser, 2008; Apostolova et al., 2009; Singh et al., 2015). As a result, we cannot provide a fair comparison of our proposed model with such approaches. Moreover, some studies employ external data sources during training, e.g., journals and textbooks (Rosenthal et al., 2019), which is also not compatible with the radiology report labeling schema, and restricts us from comparing our model with their work. Nevertheless, we implement the two existing methods presented by Apostolova et al. (2009) and Singh et al. (2015), which label sections in radiology reports. Since we did not have access to their code, we tried to replicate their methods to the best of our knowledge and understanding.

Model	MGB-test		MIMIC-III	
	Accuracy	m-F1	Accuracy	m-F1
MGB Rule-based	62.7%	47.3%	29.7%	23.6%
MIMIC Rule-based	57.5%	31.0%	33.4%	30.7%
Linear SVM	89.4%	82.2%	66.2%	63.3%
Apostolova et al. (2009)	90.3%	84.3%	72.1%	69.4%
Singh et al. (2015)	85.9%	74.7%	68.8%	63.6%
Formatting/Layout	92.3%	75.2%	42.1%	40.6%
Focus Context	89.4%	74.3%	62.0%	55.3%
Surrounding Context	93.7%	88.8%	71.2%	67.9%
Merged Ensemble	94.3%	89.3%	73.3%	69.2%
Stacking Ensemble	**97.1**%	**93.7**	**77.5**%	**74.0**%

Table 1: Comparison of the results of various models on the MGB-test set and 100 MIMIC-III notes. m-F1 stands for macro F1 score across seven classes.

We implement our proposed model using Keras [3]. We utilize Adam optimizer with a learning rate of 0.001 and Categorical Cross-Entropy loss. We split the training set into two sets: 90% for training and 10% as the validation set. We use early stopping by picking the best validation accuracy value among 30 epochs for the models with the patience value of five. We also set the patience value to 200 among 600 training epochs for the Layout model.

We run our experiments on an *Amazon c5.18xlarge EC2* instance[4]. The average running time for the focus context, surrounding context, Formatting/Layout, and Merged models are roughly 80, 70, 60, and 60 minutes, respectively.

4 Results and Discussion

4.1 Model Comparison

We compare our proposed Stacking Ensemble model with several prior work as described above. We also report the performance of individual models used in our Stacking Ensemble model to investigate the importance of each model independently. Table 1 summarizes the performance of different approaches in terms of accuracy and macro F1 on the MGB-test set as well as 100 MIMIC-III notes.

It can be observed that, overall, our proposed Stacking Ensemble model outperforms all other approaches on both test sets. By comparing the performance of the three models composing our proposed ensemble model, we observe that the

Surrounding Context model achieves the highest performance among three, emphasizing the importance of the surrounding context in such a labeling task. Furthermore, it can be observed that the Formatting/Layout model performs worse on MIMIC-III set than the MGB-test set. This could be because reports from the MGB set are structured more consistently than MIMIC-III notes. In other words, MIMIC-III notes are not prepared using a specific and consistent template.

Another observation is that the rule-based models, i.e., MGB Rule-based and MIMIC Rule-based, perform poorly compared to machine learning-based approaches even though they are tailored specifically based on the corresponding reports' format and structure. Moreover, we observe that the MIMIC Rule-based model yields lower accuracy than MGB Rule-based model on the MGB-test set and vice versa. This confirms that the performance of rule-based approaches significantly varies across different datasets, and overall, rule-based approaches suffer from generalization and scaling.

Finally, the proposed Stacking Ensemble model yields lower performance on MIMIC-III test set compared to the MGB-test set. This could be because there are significant differences between the two sets of radiology reports in terms of content and format: MGB-test set notes are from inpatient and outpatient care and in general, follow a consistent format; however, MIMIC-III reports are discharge notes from the Emergency Department lacking a consistent structure.

To evaluate the sensitivity of our proposed model to a particular split of the data, we perform 10-

fold cross-validation on the training set (i.e., split between 90% training and 10% validation). The $mean \pm std$ of accuracy and macro F1 across 10-folds are $97.0\% \pm 0.2\%$ and $93.0\% \pm 0.2\%$, respectively.
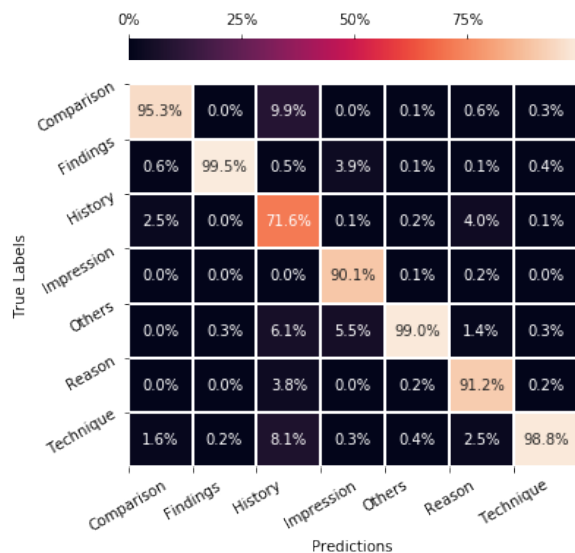

Figure 5: Confusion matrix showing the percentages of true and mislabeled predictions on the MGB-test set.

4.2 Error Analysis

We further investigate the performance of the Stacking Ensemble model for each class label separately. Figures 5 and 6 depict the confusion matrices between the predictions and actual labels from each class for MGB-test set and MIMIC-III notes, respectively. It can be observed that among all classes, "History" is the most challenging, and it is occasionally misclassified as "Comparison" in the case of MGB reports, and with "Others" and "Reason" classes in the case of MIMIC-III reports. We consider two possible reasons for this: 1) the similarity of the context between "History" and the other classes as mentioned earlier; and 2) the adjacency of these sections within the radiology reports.

4.3 Analysis of Stacking Ensemble Input

To further investigate each type of model's importance in the final ensemble decision, we analyze the weights resulting from the ensemble. We observe the different distribution of weights for different label types. For example, weights are equally distributed among three models for "Finding" and "Impression" sections. On the other hand, we observe unbalanced weight distribution
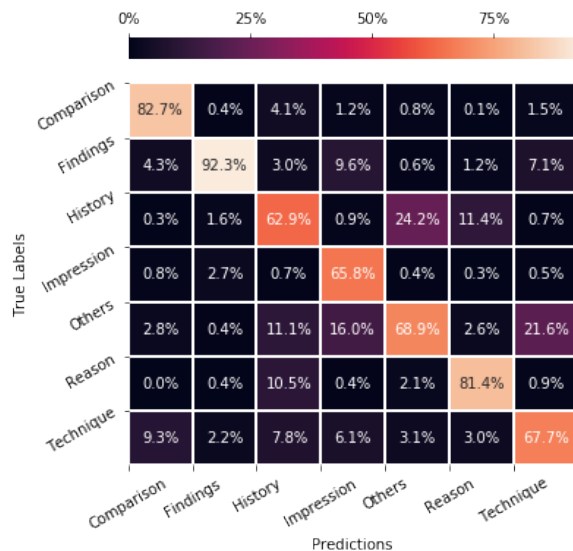

Figure 6: Confusion matrix showing the percentages of correct and incorrect predictions on MIMIC-III set.

for "Technique" and "Comparison" classes. Figure 7 shows the mean of weights for the "Findings" and "Technique" classes on the MGB-test set. It can be seen that all the models are equally important for the "Findings" class, whereas, for the "Technique" class, there is less emphasis on the Formatting/Layout model than the Focus Context and Surrounding Context models.
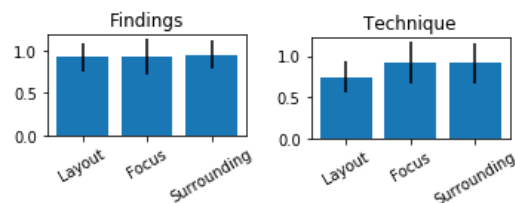

Figure 7: Comparison of the errors of inputs to the Stacking Ensemble model for the Findings and Technique classes on the MGB test set.

4.4 Fine-tuning the Stacking Model

As can be seen from Table 1, the proposed ensemble model trained on MGB data does not perform as well on the MIMIC-III set. We try to improve the performance of the proposed ensemble model on the MIMIC-III set by fine-tuning the ensemble part on a MIMIC-III data subset. We split the MIMIC-III data into 20% for fine-tuning and 80% for testing. Table 2 demonstrates the results of running the Stacking Ensemble model on 80% of the MIMIC-III data with and without fine-tuning. As can be seen from the table, we can obtain a 5.5%

increase in accuracy score and a 6.9% increase in macro F1 score. This is achieved by only fine-tuning the ensemble step using a small subset of the MIMIC-III data, while the individual models are still trained on the MGB data.

Model Type	Accuracy	m-F1
Without fine-tuning	76.3%	73.9%
With fine-tuning	**81.8**%	**80.8**%

Table 2: Comparison between the performance of Stacking Ensemble model with and without fine-tuning on MIMIC-III data.

To show that the results are not sensitive to any specific split of data, we perform five-fold cross-validation on the MIMIC-III reports by utilizing 20% of reports for training the Logistic Regression classifier, and 80% for testing. We obtain a mean value of 81.5% accuracy with a standard deviation of 0.7% and a mean value of 80.4% macro F1 score with a standard deviation of 0.9%, which shows the insensitivity to the specific split of data.

A known shortcoming of our proposed approach is the sensitivity to the accuracy of the sentence segmentation. Poor sentence parsing results in miss-labeling, specifically, if error in sentence parsing results in combining sentences belonging to two different sections. To address this issue, we are currently working on training a clinical note-specific sentence parsing algorithm.

5 Conclusions and Future Work

In this work, we propose an ensemble approach for automatically labeling sentences in radiology reports with section labels. Through the proposed ensemble approach, we achieve the state-of-the-art performance of 97.1% on a relatively sizeable multi-site test set from Mass General Brigham. Our proposed ensemble method is composed of three parallel models that capture various structural and contextual attributes of radiology reports, including formatting/layout, focus context, and the surrounding context. Furthermore, We compared our proposed ensemble model against each of its components and concluded that the combination of all models is more accurate than any individual model.

As future work, we plan to incorporate performance calibration in our ensemble model. It adds the importance of individual models to the ensemble model and enables us to obtain higher performance for unseen data. We also plan to extend this work to other types of reports, i.e., pathology reports and discharge summaries. Another potential future work is to add Conditional Random Field (CRF) to our proposed model.

References

Emilia Apostolova, David S Channin, Dina Demner-Fushman, Jacob Furst, Steven Lytinen, and Daniela Raicu. 2009. Automatic segmentation of clinical texts. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5905–5908. IEEE.

Alberto Blanco, Olatz Perez-de Viñaspre, Alicia Pérez, and Arantza Casillas. 2020. Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. *Computer Methods and Programs in Biomedicine*, 188:105264.

Paul S Cho, Ricky K Taira, and Hooshang Kangarloo. 2003. Automatic section segmentation of medical reports. In *AMIA Annual Symposium Proceedings*, volume 2003, page 155. American Medical Informatics Association.

Joshua C Denny, Randolph A Miller, Kevin B Johnson, and Anderson Spickard III. 2008. Development and evaluation of a clinical note section header terminology. In *AMIA Annual Symposium proceedings*, volume 2008, page 156. American Medical Informatics Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter J Haug, Xinzi Wu, Jeffery P Ferraro, Guergana K Savova, Stanley M Huff, and Christopher G Chute. 2014. Developing a section labeler for clinical documents. In *AMIA Annual Symposium Proceedings*, volume 2014, page 636. American Medical Informatics Association.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Vijay Kotu and Bala Deshpande. 2014. *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alexandra Pomares-Quimbaya, Markus Kreuzthaler, and Stefan Schulz. 2019. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC medical research methodology*, 19(1):155.

Sara Rosenthal, Ken Barker, and Zhicheng Liang. 2019. Leveraging medical literature for section prediction in electronic health records. 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)*, pages 4864–4873, Hong Kong, China. Association for Computational Linguistics.

Daniel L Rubin and Terry S Desser. 2008. A data warehouse for integrating radiologic and pathologic data. *Journal of the American College of Radiology*, 5(3):210–217.

Mark Singh, Akansh Murthy, and Shridhar Singh. 2015. Prioritization of free-text clinical documents: a novel use of a bayesian classifier. *JMIR medical informatics*, 3(2):e17.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Ricky K Taira, Stephen G Soderland, and Rex M Jakobovits. 2001. Automatic structuring of radiology free-text reports. *Radiographics*, 21(1):237–245.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical section segmentation in free-text clinical records. In *Lrec*, pages 2001–2008.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.