Analysis of Language Embeddings for Classification of Unstructured Pathology Reports

Aishwarya Krishna Allada[1], Yuanxin Wang[1], Veni Jindal[1], Morteza Babee[2], H.R. Tizhoosh[2], Mark Crowley[1]

Abstract— **A pathology report is one of the most significant medical documents providing interpretive insights into the visual appearance of the patient's biopsy sample. In digital pathology, high-resolution images of tissue samples are stored along with pathology reports. Despite the valuable information that pathology reports hold, they are not used in any systematic manner to promote computational pathology. In this work, we focus on analyzing the reports, which are generally unstructured documents written in English with sophisticated and highly specialized medical terminology. We provide a comparative analysis of various embedding models like BioBERT, Clinical BioBERT, BioMed-RoBERTa and Term Frequency-Inverse Document Frequency (TF-IDF), a traditional NLP technique, as well as the combination of embeddings from pre-trained models with TF-IDF. Our results demonstrate the effectiveness of various word embedding techniques for pathology reports.**

I. INTRODUCTION

Cancer is one of the leading causes of mortality in the world. A computer-aided framework for cancer diagnosis requires a pathologist to make a detailed report after analyzing the tissue on glass slides collected from a patient's biopsy sample [1]. Pathology reports are made up of histopathological indicators and detailed analysis of specific cells and tissue types, which are essential for malignancy diagnosis. Most of these reports are written in highly unstructured manner and have no direct connection to the tissue samples. Also, each patient's report is a customized document having high discrepancies in vocabulary, such as misspelled words and lack of punctuation. It is common to find clinical diagnoses intermixed with nuanced explanations, multiple terminologies used to mark the same malignancy and data about various carcinomas in a single report [2]. Also, it may be possible that some of the reports don't have any relevant keywords in them to directly identify the disease. Cancer registries are facing a considerable challenge in the manual analysis of the enormous quantity of pathology reports, with the rise in the number of patients with cancer and the improvement in treatment complexity [2], [3]. When primay diagnosis is not clearly not mentioned, the process of identifying the disease from a pathology report is challenging, time-consuming and requires extensive training, when done manually [2]. This paper demonstrates how to extract meaningful embeddings from written pathology reports to help classify various types of cancer. This may be used for multi-modal learning in conjunction with histopathology images and molecular data. Our primary focus is to evaluate and compare the effectiveness of existing machine learning methods for automatic classification of a given pathology report to its respective primary diagnosis. We demonstrate that contextualized word embeddings combined with Term Frequency-Inverse Document Frequency (TF-IDF) feature vectors, when given as inputs to a Deep Neural Network (DNN), can be an effective method for classification, achieving 93.77% accuracy in our study. Additionally, our experiments with digital pathology reports will allow researchers to develop a versatile way of extracting essential details from free-text pathology reports which could benefit a variety of diagnostic tasks.

II. RELATED WORK

In the field of biomedical research, information extraction using NLP spans from rule-based systems [4] down to domain-specific systems using feature-based classification [3], and to the recent deep networks for end-to-end feature extraction and classification [2].

In case of classification tasks or retrieving specific features from reports, successful studies in NLP for understanding pathology reports have been reported [5]. The Cancer Text Information Extraction System (caTIES) is a framework developed in a caBIG project [6], that focuses on the extraction of key details from Surgical Pathology Reports (SPR) to achieve high precision and recall. On the other hand, a system named Open Registry [7] was able to filter out the pathology reports having cancer specified in them, based on the disease codes.

In 2010, the Automated Retrieval Console (ARC) [8], was introduced, where machine learning models are used to predict the degree of association of a given radiology or pathology report to cancer. The performance of this approach varied from F-measure of 0.75 for lung cancer to 0.94 for colon cancer. However, this approach utilized domain-specific rules, which may be disadvantageous when working with a wide variety of pathology reports. Other works have performed a classification of the pathology reports by extracting the TF-IDF features [9]. The extracted features were given as input to XGBoost, SVM and Logistic Regression, where improved ensemble results were obtained with XGBoost classifier. A large number of algorithms which convert words to fixed-dimensional vectors which can be used to preserve syntactic and semantic relationships in a text corpus were introduced. These include word2vec [10] and GloVe [11] which use co-occurrences of words in the text and produce dense vectors such that words appearing in similar contexts have similar word embeddings. Major

[1] Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada (akallada, yuanxin.wang, v2jindal, mcrowley)@uwaterloo.ca

[2] † KIMIA Lab, University of Waterloo, Waterloo, ON, Canada (morteza.babaie, tizhoosh)@uwaterloo.ca

improvements could be achieved in the NLP field came when unsupervised model architectures were proposed to represent words as a fixed dimensional dense vector [10], [12]. The architectures were Continuous Bag of Words (CBOW) that predicts the target word given the context and the Skip-gram model, which predicts the context based on the target word. Another major advance was BERT (Bidirectional Encoder Representations from Transformers) [13], which improves fine-tuning-based approaches by taking into account both left and right context unlike the traditional algorithms which used the left-to-right direction only.

III. METHODOLOGY

We now elaborate on our pathology reports dataset, data pre-processing steps, pre-trained models and TF-IDF technique.

A. Understanding the Dataset

The data we use are a cleaned and processed subset of the publicly available *TCGA* (The Cancer Genome Atlas) [9]. We obtained a subset of approximately 1,960 pathology reports describing the tissues of organs referred to as the disease type, each report having 0 to 2500 words. The disease type typically consists of totally seven classes, namely, "Kidney Renal Papillary Cell Carcinoma", "Kidney Renal Clear Cell Carcinoma", "Lung Adenocarcinoma", "Lung Squamous Cell Carcinoma", "Testicular Germ Cell Tumors", "Kidney Chromophobe" and "Thymoma".

B. Data Pre-processing

The main challenge in classification with Deep Neural Networks using text data is transforming the data into a clean format, which can be converted into numerical vectors. Before initializing the data pre-processing step, all samples consisting of empty documents were removed. Further, we removed all bullet numbering, stop-words, and special-, numeric-, or null-characters. Occurrences of spatial dimensions of tumor or organ size were also standardized by converting *"$l \times b \times h$" cm* into a single entity with no spaces (i.e. as **lxbxhcm**).

After data pre-processing, we have performed the k-fold cross validation on the data to estimate the performance of the model on unseen data with $k = 5$.

C. Pre-Trained Models and TF-IDF

In this section, we describe several contextualized word embedding models along with TF-IDF and their techniques to convert text into vectors.

1) ***BioBERT****:* BioBERT [14] is an application of the BERT-based model [13], which is popularly used in the biomedical field. This model is obtained upon pre-training the BERT base model on the biomedical corpus. We have used BioBERT-v1.1 for our experiments, which was obtained by pre-training BioBERT on PubMed database for 1M steps. The vocabulary size of the model is 28,996, each having 768 features. The output text from the data pre-processing step is tokenized using the BioBERT tokenizer which uses WordPiece Tokenization [15] that breaks down a word into multiple subwords belonging to the BERT vocabulary. For example, the WordPiece tokenization of the word "penicillin", which will not be present in the vocabulary directly, is split into the subwords, "pen", "##i" and "##cillin", which are available in the BERT vocabulary. The tokenized words are then fed to the classifier model for classification.

2) ***Clinical BioBERT****:* Clinical texts such as physician notes have different linguistic features compared to non-biomedical or general texts. This difference encouraged the necessity for a specifically trained model for clinical domain texts and Clinical BioBERT was introduced. Clinical BioBERT [16] is initialized from BioBERT (BioBERT-Base v1.0 + PubMed 200K + PMC 270K) and is trained on all the MIMIC-III notes (880M words), which is a database containing health reports of the ICU admitted patients at the Beth Israel Hospital in Boston, MA. This data is used to pre-train the model for 150k steps with batch-size set to 32. Like BioBERT, the vocabulary size of Clinical BioBERT is also 28,996 tokens, each having 768 features following WordPiece Tokenization. The processed data is tokenized using the *"Bioclinical_BERT"* tokenizer, which is then used as input data to the classifier model.

3) ***BioMed-RoBERTa****:* BioMed-RoBERTa [17] is a recent model initialized from RoBERTa-base, which is pre-trained for 12.5K steps with a batch size of 2048 using 2.68M scientific papers (7.55B tokens) from Semantic Scholar [18] [19]. The vocabulary size of this model is 50,265 tokens with 768 features, which is acquired using BPE (byte pair encoding [20]) word pieces with *\u0120* as the special signaling character. The *"biomed_roberta_base"* tokenizer from HuggingFace is used for tokenization.

4) ***Term Frequency-Inverse Document Frequency****:* Term Frequency-Inverse Document Frequency(TF-IDF), is a metric that specifies the significance of a given word to a document in a document set. Term frequency is defined as the frequency of a term in a document [21], whereas Inverse Document Frequency (IDF) gives more importance to words frequently found in a set of documents. By multiplying the number of times, a word appears in a document (Term Frequency) and the number of times a word occurs in several documents (Inverse Document Frequency), we obtain a statistical measure which is used to evaluate words based on their value among the rest of the terms. Therefore each sentence should have a representation according to the meaning of each word in the sentence.

IV. EXPERIMENTAL SETUP

In this section, we will discuss the experimental setup of our analysis. Fig.1 depicts the deep neural network topology we have used. The main purpose of this network is to analyze the word embeddings as an initial investigation for NLP in Digital Pathology. The proposed network can be customised with one or two input layers, based on the analysis. Firstly, the data is pre-processed and based on the maximum length of tokens amongst all the pre-processed reports, we chose 300 as the maximum length of each report.

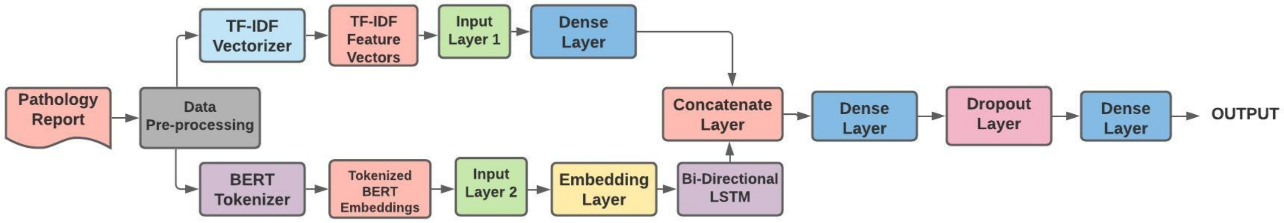

Fig. 1. Deep Neural Network Topology

To understand the effectiveness of each pre-trained models, respective tokenizers of BioBERT, Clinical BioBERT and BioMed-RoBERTa are used to tokenize the data. In the case of TF-IDF, upon tuning, the pre-processed text is vectorized with maximum features of 300 and a minimum threshold value of 5. The tokenized data is then converted into an array of vectors which is given as the input to the DNN.

Now, for the DNN having single input, the respective token embeddings from the pre-trained word embedding model or feature vectors from TF-IDF vectorization along with their labels are passed to the classifier for training. On the other hand, for the DNN having both pre-trained word embeddings and TF-IDF feature vectors, both the token embeddings and the feature vectors along with their labels are given as input to the model having two input layers.

For analysing the contextualized word embedding models, we have extracted the weights from the respective pre-trained model's word embedding layer. The weights are then converted into an embedding matrix, which is initialized as weights to the embedding layer in our DNN classifier.

The word embeddings obtained from the pre-trained model tokenizer are passed through the Embedding Layer, followed by the Bidirectional LSTM layer. On the other hand, the TF-IDF feature vectors are passed through the dense layer with "ReLU" activation function. The respective vectors from both the layers are concatenated and are then passed through the dense layer with "ReLU" activation function, followed by a dropout layer with the dropout rate of 0.3. The vectors are then finally sent to the output dense layer having "softmax" activation function. The model is trained using Adam optimizer with its default learning rate of 0.01 and "categorical cross-entropy" loss function is used. The best vectorization method is decided based on the evaluation metrics such as precision, recall, F1-Score and classification accuracy, which are calculated as an average of all the 5 folds and the results obtained are mentioned in Section V.

V. RESULT ANALYSIS AND DISCUSSIONS

This section describes the quantitative results obtained by our experiments which show that our approach of combining contextualized word embeddings along with TF-IDF feature vectors provides best results than the model with single input. The results of experiments on vectorization techniques using DNN classifier are as shown in the Table I.

TABLE I
EVALUATION OF VECTORIZATION METHODS FOR THE CLASSIFICATION OF DISEASE TYPES WITH A DNN CLASSIFIER

Vectorization Method	**Precision**	**Recall**	**F1-Score**	**Accuracy**
BioBERT alone	0.76	0.78	0.76	79.76%
BioMed-RoBERTa alone	0.83	0.84	0.83	81.03%
Clinical BioBERT alone	0.80	0.81	0.81	81.29%
TF-IDF alone	0.81	0.78	0.79	88.90%
BioMed-RoBERTa plus TF-IDF	0.90	0.93	0.91	90.58%
BioBERT plus TF-IDF	0.88	0.92	0.90	91.65%
Clinical BioBERT plus TF-IDF	**0.91**	**0.92**	**0.91**	**93.77%**

A. RESULTS

Table I compares the performance of various vectorization methods for the disease type classification on the DNN classifier on unseen data. Amongst the vectorization techniques used, *Clinical BioBERT embeddings in combination with TF-IDF feature vectors* yields the best accuracy of 93.77%, precision of 0.91, F1-score of 0.91 and recall of 0.92. The next best were the combinations of *BioBert with TF-IDF* and *BioMed-RoBERTa with TF-IDF*. We believe this is because the Clinical BioBERT embeddings obtained using the BioBERT model were initially pre-trained on a medical corpus and then further trained on clinical texts with similar terminology, whose words are more likely to appear in pathology reports. Also, TF-IDF vectorization on these reports contributes to give important information about the word distributions in a pathology report. Thus the combination of them performed the best on our classification model. A tedious and error prone task that is important in this field is creating an embedding for cancer and tumor detection phrases from any given pathology report.

Retrieving the disease type is one of the most critical aspects of deciphering a pathology report, which will be very useful while combining content-based image retrieval with visual information. The accuracy obtained by these models supports the use of machine learning techniques to extract meaningful and relevant information from pathology reports.

B. ABLATION STUDIES

The basic idea behind ablation is to remove certain components from the experiment to understand the contribution of that component towards the overall model. Our study shows that model performance decreases to almost 80% upon removing the TF-IDF vectorizer from the topology shown in Figure 1. Thus, it is reasonable to say that TF-IDF helps to identify the important medical terms in the pathology report. Upon removing the pre-trained word embeddings from our topology, the classification accuracy decreases to 88.90% as compared to the overall accuracy of 93.77%. Thus, the pre-trained model embeddings helps to improve the model due to its abilities to interpret domain specific terminologies in the biomedical field.

VI. CONCLUSIONS

In this paper, we examined the classification of pathology reports of seven different diseases and reported several experiments by evaluating the word embeddings of the pre-trained models, TF-IDF vectorization technique and the combination of both of them. We found that the combination of TF-IDF with pre-trained model word embeddings was always outperforming contextualised word embeddings and TF-IDF when performed individually. The best performance for pathology report classification was observed for the model that concatenated the embeddings from Clinical BioBERT with the TF-IDF vectors. This seems to form a reasonable baseline and provides valuable insights in the future of digital pathology report analysis.

REFERENCES

[1] E. M. F. El Houby, "Framework of computer aided diagnosis systems for cancer classification based on medical images," *Journal of Medical Systems*, vol. 42, no. 8, p. 157, July 2018.

[2] S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanthan, "Hierarchical attention networks for information extraction from cancer pathology reports," *Journal of the American Medical Informatics Association*, vol. 25 (3), pp. 321–330, 2017.

[3] R. Weegar, J. F. Nygård, and H. Dalianis, "Efficient Encoding of Pathology Reports Using Natural Language Processing," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., 9 2017, pp. 778–783",. [Online]. Available: https://doi.org/10.26615/978-954-452-049-6100

[4] N. Kang, B. Singh, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Using rule-based natural language processing to improve disease normalization in biomedical text," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 876–881, 10 2012.

[5] T. D. Imler, J. Morea, C. Kahi, and T. F. Imperiale, "Natural language processing accurately categorizes findings from colonoscopy and pathology reports," *Clinical Gastroenterology and Hepatology*, vol. 11, no. 6, pp. 689–694, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1542356513000104

[6] R. S. Crowley, M. Castine, K. Mitchell, G. Chavan, T. McSherry, and M. Feldman, "caTIES: A grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 253–264, may 2010. [Online]. Available: https://doi.org/10.11362Fjamia.2009.002295

[7] P. Contiero, A. Tittarelli, A. Maghini, S. Fabiano, E. Frassoldi, E. Costa, D. Gada, T. Codazzi, P. Crosignani, R. Tessandori, and G. Tagliabue, "Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system," *Journal of Biomedical Informatics*, vol. 41, no. 1, p. 24–32, Feb. 2008. [Online]. Available: https://doi.org/10.1016/j.jbi.2007.03.003

[8] L. D'Avolio, T. Nguyen, W. Farwell, Y. Chen, F. Fitzmeyer, O. M. Harris, and L. Fiore, "Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (arc)," *Journal of the American Medical Informatics Association : JAMIA*, vol. 17 4, 2010.

[9] S. Kalra, L. Li, and H. R. Tizhoosh, "Automatic classification of pathology reports using TF-IDF features," 2019. [Online]. Available: http://arxiv.org/abs/1903.07406

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013.

[11] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: http://arxiv.org/abs/1810.04805

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, Sep 2019. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btz682

[15] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[16] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pre-training approach," 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[18] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," 2020.

[19] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, R. Kinney, S. Kohlmeier, K. Lo, T. Murray, H.-H. Ooi, M. Peters, J. Power, S. Skjonsberg, L. Wang, C. Wilhelm, Z. Yuan, M. van Zuylen, and O. Etzioni, "Construction of the literature graph in semantic scholar," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. New Orleans - Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 84–91. [Online]. Available: https://www.aclweb.org/anthology/N18-3011

[20] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," 2016. [Online]. Available: http://arxiv.org/abs/1508.07909

[21] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach," in *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2014, pp. 1–4.