Text Classification for infection detection from Radiology report

Student ID : 31334679

Name: Jirarote Jirasirikul

Masters of Data Science

FIT5126 - Minor Thesis 1

Supervisors:

Gholamreza (Reza) Haffari,

Ehsan Shareghi

1. Introduction

With the rapidly growing amount of data, digitalization of documents is no longer a barrier for Data Scientists. Every report is now provided in an electronics form or physical paper with electronic copy, this includes our personal medical document and treatment history. These information are stored somewhere within databases and we could retrieve them by queries when necessary. Our problem became a paradox as we are now overflowing with these data. Humans have limited capacity of concentration, which they commonly use for information that they are interested in. Regrettably, lots of information that is potentially valuable is often overlooked. Natural Language Processing has become important because we could use computation to help manage these data and assist in detecting interest and unusual patterns. These bring missout beneficial data into our attention and lessen precious information unattended.

In health services, these problems are no less vulnerable. As a matter of fact, we could say they are more vital to not leave any data unnoticed because this could lead to a life-death situation. We see these as a gap and opportunity to assist these devoted medical staff in identifying potential life-threatening diseases in patients and bring them into attention to get earlier treatment. One of the reports that medical staff commonly use is a radiology report. This report is a summarization of the assessment performed by radiologist in order to address general practitioner concern on the patient. In order to conclude evaluation, multiple medical procedures were performed to detect and diagnose, some relevant and some not. However, radiologist tasks are only to address general practitioner focus. Out of those assessments some would be picked and pointed to diagnose immediate patient illness. The rest remain in the report as supporting documents. However as mentioned earlier, general practitioners have their hands full on concentrating treatment for immediate illness. Unfortunately, some information might have been overlooked.

Our goal is to use Machine Learning to figure out some common patterns using supervised and unsupervised learning methods on these radiology reports. There are two main tasks that would be required in order to address this problem: (1) Understanding medical reports which are embedded with specialized domain knowledge. (2) Build prediction models to calculate probability of patients infected with fungal disease. These tasks are connected and contribute to quality prediction performance. We will face some foresee challenges in this thesis as we are processing medical information: (1) data privacy of medical information (2) medical domain specialize corpora (3) precision and accuracy of the prediction on disease.

This thesis focuses on Text classification using the state-of-the-art of Natural Language Processing to handle radiology reports and detect unique patterns that could show signs of infection. The prediction results will assist General Practitioners to be aware of additional complications that may occur in their patients. Conceivably, this will reduce beneficial data misout to the bare minimum.

2. Substantive Literature Review

Natural language processing (NLP) is a branch of study in Machine Learning that helps computers understand, interpret and process human language. This involves a range of computational techniques for analyzing and representing naturally occurring texts (Liddy, n.d.; Louis, 2020). Conventional natural languages are typical unstructured information and NLP aims to extract useful information which represents those raw data by feature engineering (Z. Liu et al., 2020).

In this literature review, our goal is to look back into the history of text representation methods in natural language processing, understand the benefits and limitations of each improved method, in order to see the possibility to adapt these techniques into medical use cases. By doing so, we hope to find a better way to extract valuable information from medical reports that may be left unnoticed and could be extremely beneficial to people's life.

A typical machine learning system, that builds to assist in given tasks such as classification or forecasting, are usually follow consists of three components:

Machine Learning = Data representation + Objective + Optimization

Data representation is one of the core components to build a good model. In NLP, our data is usually in a form of Text, Word or Sentence representation. The Figure 1 diagram below highlights some of the significant algorithms created along the NLP journey (*Evolution of Natural Language Processing*, 2020). We will explore, summarize and argue throughout the following subsection.

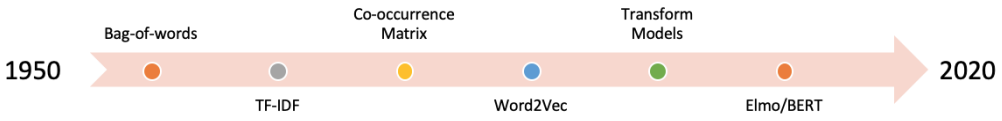


Figure 1: Evolution of NLP's Text representation

2.1 General Text Representation

Computer ("Computer," 2021) is an electronic device for storing and processing data in binary form. This means that in order to make computers understand the meaning of human language. They need to transform language text 'words' into a form of 'numbers'. Generally, computers automatically convert each character of the alphabet symbol into numbers called ASCII numbers (Hieronymus & Laboratories, n.d.). Although, this does not give a representation of the words rather than an alphabet, using the same methodology could still be applied. By representing words as numbers, we could count occurrences of each word in passage or documents (*Evolution of Natural Language Processing*, 2020; Zhang et al., 2010) and perform statistical inference. This method is called the "**Bag-of-Words**" (BoW) representation which is a successful and popular approach for document categorization where word distribution could determine a document's topic. However, BoW representation

has disadvantages as it does not consider type words such as common words or stopwords that appear frequently but has less valuable meaning. Term frequency-Inverse document frequency (TF-IDF) has been introduced to address this problem by putting weight to important words that might occur less frequently shown in Figure 2. Nevertheless, these approaches disregard valuable information on the position of words in a context.

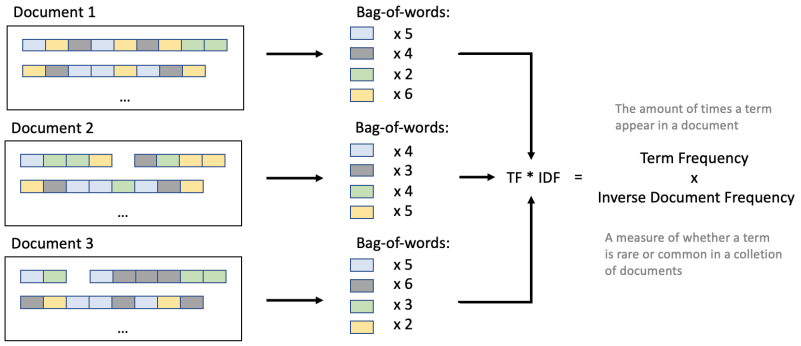


Figure 2: Illustrate a construction of Bag-of-words and TF-IDF from documents

Alternatively, words can also be represented in a form of vector instead of number. This text representation is called **one-hot-encoding** which will give an index of words in the dictionary (*Andre Ye*, n.d.)

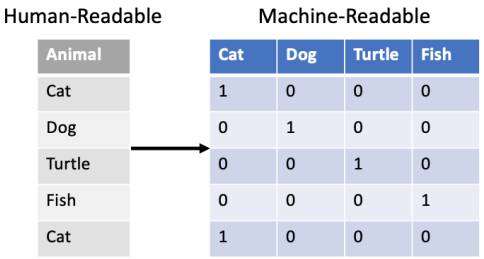


Figure 3: One-hot-encoding from word to vector of index in dictionary

A **Co-occurrence Matrix** (Manning et al., 2008; Vargas, 2017) was created by analyzing the context in which a word is used and taking the neighboring words of each word into account. This technique generated word embeddings that track context of the word by using large matrices, Figure 4. As a consequence, this method requires a large amount of memories which are not preferable.

Sentence : I love Programming. I love Math. I tolerate Biology.

	I	love	Programming	Math	Tolerate	Biology	.
I	0	2	0	0	1	0	2
love	2	0	1	1	0	0	0
Programming	0	1	0	0	0	0	1
Math	0	1	0	0	0	0	1
Tolerate	1	0	0	0	0	1	0
Biology	0	0	0	0	1	0	1
.	1	0	1	1	0	1	0

Figure 4: Co-occurrence matrix build from sentence by counting neighbor of each word

As NLP progressed, word embedding became a frequent norm to achieve state-of-the-art results in machine learning. Researchers seek to find an improvement of embedding context that is relevant into the word so that semantic meaning of the word could be extracted. **Word2Vec** was introduced in late 2013 (Mikolov et al., 2013) by generating vectors that could represent the meaning of words after going through a large scale corpus. The simplest part of this idea is by looking for previous n-1 words from the focus word and calculating the probability of the possible word that fit the focus word, Figure 5. There are 2 algorithms that infer from this idea which is CBoW and skip-gram. While CBoW is looking at n-1 words to try identify the focus words. Skip-gram, on the other hand, is trying to find the highest possibility of words that would come before or follow the focus word. Hence after massive learning of corpus, it became possible to form a vector that represents each word.

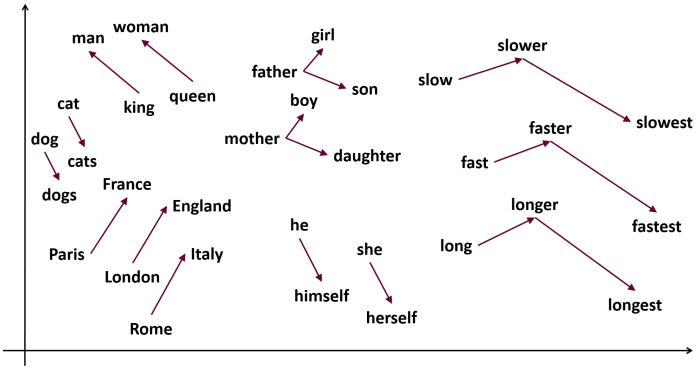


Figure 5: Example of words in 2D dimensions of Word2Vec's vectors (Pal, 2019)

Classic example that shows how Word2Vec works and how it is embedded with semantic meaning are:

King - Man + Woman = Queen

However, Word2Vec accuracy is not as high as it claims (Church, 2017, p. 2). There are words that have similar vector with comparable probability but represent totally different meanings for instance in sample context example, i.e. Monarch, Princess, Prince, could lead to different meanings of the sentence. Nevertheless, Word2Vec is the most basic and standard text representation that is the most well known up-to-date. This is because the Word2Vec approach is simple and easily accessible which overrules its inaccuracy and incorrectness. The Word2Vec contributes a large impact to the NLP fields and fundamentals to text representation.

Meanwhile, the Neural network based model has started to become an increasingly popular and dominant approach to tackle NLP tasks, as they produce state-of-art results (Conneau et al., 2017; Joulin et al., 2016). They tend to be relatively slow when performing in both train and test datasets. Facebook AI researcher has developed the **FASTTEXT** approach (Mikolov et al., 2013), which was inspired by enhanced efficiency in word representation learning, using a linear model with rank constraint and loss approximation on word representation. FASTTEXT allows models to train billions of words within ten minutes on word representation, yet, achieved performance on par with state-of-the-art evaluated in 2 NLP tasks.

The concept of pre-train word representations became key components in natural language processing nowadays modelling (M. E. Peters et al., 2018; Vaswani et al., 2017). But because of ambiguity, word tokens could have various meanings. Pre-train word representation should not only learn the meaning of the word itself but also need to learn its context, which drives pre-train contextualized word representation. Ideally, this kind of word representations should be able to contain (1) characteristic of words (e.g. syntax and semantics) (2) application in linguistic context. Therefore, embedding these attributes into word representation is challenging. Embeddings from Language Model or **ELMo** (M. Peters et al., 2018) are presented as methods to derive word representation vectors from the context using Long-short term memory (LSTM) recurrent neural networks. Moreover, ELMo is trained bi-directional to be able to get word representation that covers both prior and after meaning of the word on a large text corpus. ELMo representations are deep because their outputs are vectors stacked that come from all internal layers of the Bi-LSTM. ELMo have outperformed previous representations and work extremely well in practice.

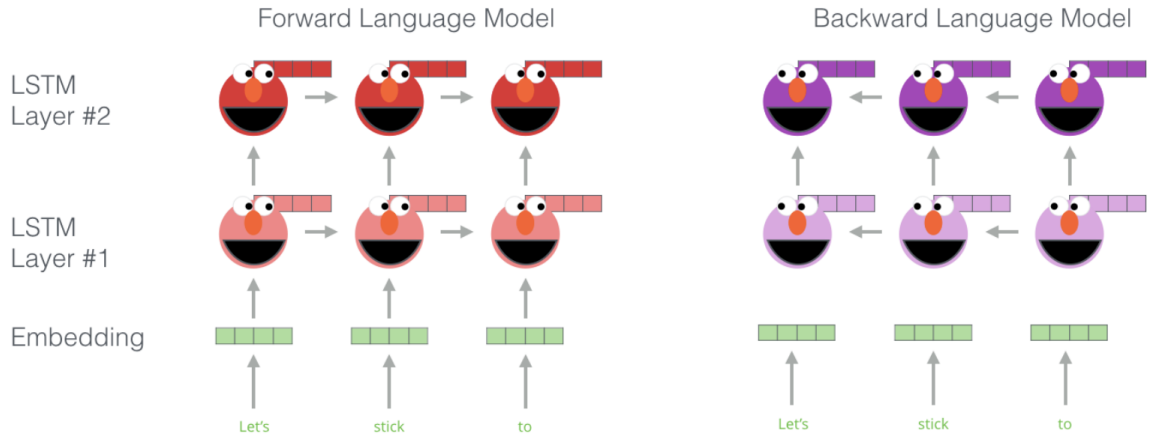


Figure 6: Illustrate on ELMo's LSTM hidden layers within Forward and Backward Language Model (Alammar, n.d.)

$$\sum_{k=1}^{N} \Big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big)$$

Figure 7: Jointly maximize likelihood equations of ELMo's bi-LSTM forward and backward combined (M. Peters et al., 2018)

In 2017, **The transformer** paper about "attention is all you need" was published (Vaswani et al., 2017; Radford et al., n.d.). This comes with the fact that the attention mechanism deals with long-term dependencies better than LSTM, which could only recurrent back with limited window-size. Additionally, the transformer is using an encoding-decoding technique that is better with handling machine translation, Figure 8. By assigning weight for all its 'relevance' or 'attention' to each token and having each 'attention head' focus on different things, make transformers be able to outperform previous word representation. Moreover, the transformer could be trained significantly faster than LSTM architectures.

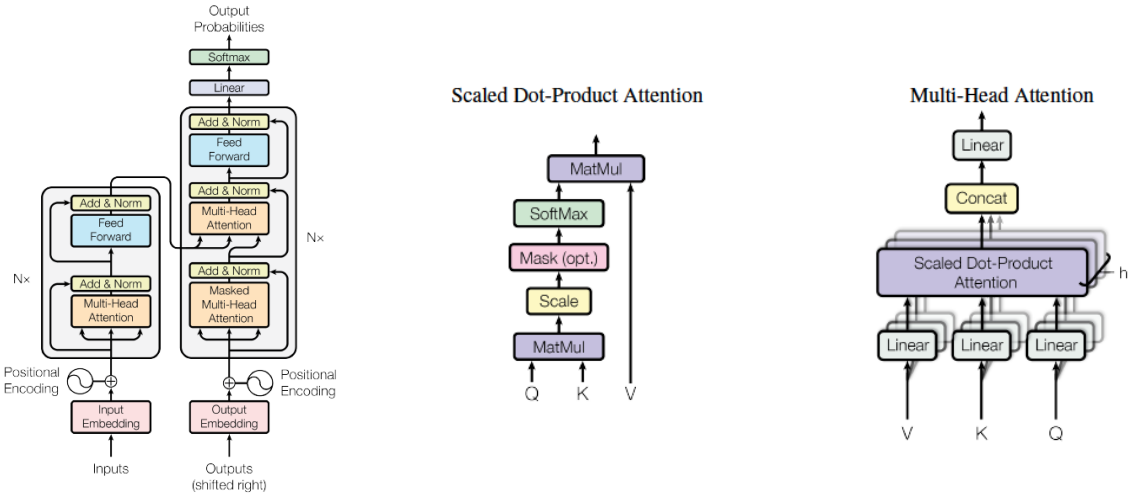


Figure 8: (left) High-level Transformer model architecture
(middle) Scaled Dot-Product Attention module
(right) Multi-Head Attention consists of several attention layers running in parallel.
(Vaswani et al., 2017)

Recently in 2019, based on the work of ELMo and The transformers, researchers have taken these two method advantages, combined them together and created **BERT** (Bidirectional Encoder Representations from Transformers), which is our current state-of-the-art (Devlin et al., 2019). Given that ELMo has the advantage of being able to capture bidirectional meaning while the transformer performs better on capturing dependencies of each token using attention mechanism. Similarly to all other pre-train models, BERT is trained on a large corpus with provided 2 pre-train models with different size $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ for easy appliance.

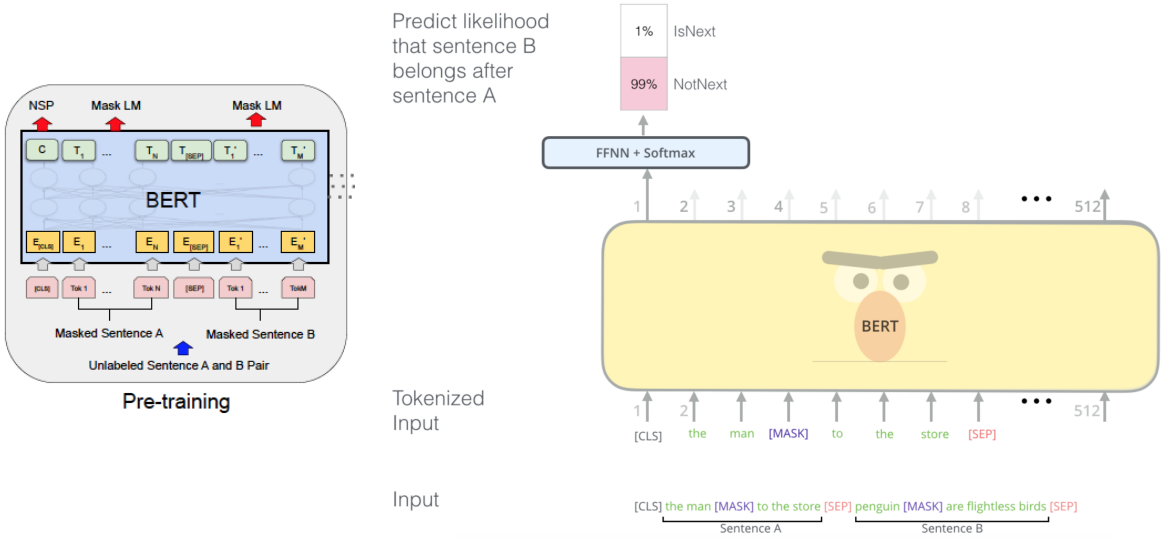


Figure 9: (left) BERT unsupervised Mask LM pretraining, (right) Illustrate BERT model in classification.
(Alammar, n.d.; Devlin et al., 2019)

BERT is the most recent state-of-the-art model. We will look into more detail of its architecture. Original BERT pre-trained using 2 unsupervised tasks called Masked Language Model (MLM) and Next sentence prediction (NSP). However, later research proves that NSP actually hurts its performance because the model is not able to learn long-range dependencies (Y. Liu et al., 2019). MLM was done by masking the token with [MASK] token in each sequence at random and making the model try to predict the correct word. This approach allows machines to capture important meaning using attention mechanisms. NSP, on the other hand, was simply adding a [CLS] token at the start of the paragraph and using [SEP] to separate between different sentences. This allows BERT to understand the relationship between 2 sentences. RoBERTa (Y. Liu et al., 2019), which carefully studies the effects of BERT hyper parameters, shows that removing NSP during pre-train matches or slightly improves downstream task performance. Nevertheless, BERT is a Language Model that encodes sentences and words into its representation which later on can be used in other models. For example: Text classification using Feed Forward Neural Network (Figure 9 - right).

Compared to the start of NLP journey, we have come so far on text representation that we have proven that contextualized word representation is better than non-contextualize word representation. Nowadays, we most likely use Language Model to represent our information before performing any modeling on-top, even though it is hardly recognized in its original form after transformation. All previous research has proven that somehow these word representations preserve characteristics of word's semantics and context because the result enhances prediction accuracy in various NLP tasks, i.e. Named Entity Recognition, Question answering, Sentiment analysis, and Text classification (*Tracking Progress in Natural Language Processing*, n.d.). This leads further on our topic to getting state-of-art performance in a more specifically medical domain NLP task.

2.2 Biomedical Text representation

In this section, we will look specifically toward text representation in the biomedical domain. Biomedical is a domain specific field that has its word distribution shifted from general domain corpora (Lee et al., 2019). Although NLP advancements yield a promising result in the general domain, applying to the biomedical text domain still remains as a challenge because of its own linguistic characteristics that are unique from the general text.

Now that we understand the history of NLP's text representations, we will focus on the recent state-of-the-art language model BERT. Our goal is to find the possibility of customization of BERT into a domain specific task. Regardless of our literature review that focuses on BioMedical Data, transferring knowledge of models from general corpora to a specific domain have a record of proven results. For example: SciBERT (Beltagy et al., 2019) and FinBERT (Yang et al., 2020) extend their corpora into the science and the finance domains. They share similar motivation that the general corpora lack quality to represent word distribution within each domain (Beltagy et al., 2019; Yang et al., 2020). Although using general corpora BERT could achieve decent results, compared with BERT that perform additional training on specific domains still show significant difference. Additional training on the Pre-train BERT model shows an improvement when handling domain specific tasks.

In the biomedical domain, **BioBERT** (Lee et al., 2019)**, ClinicalBERT** (Alsentzer et al., 2019) **and BlueBERT** (Peng et al., 2019) were driven by similar reasons. Their motivations are to improve BERT performance in handling BioMedical NLP tasks. They all perform additional training on pre-train BERT to learn large-scale BioMedical corpora. **BioBERT** approach initializes setting from general BERT and training additionally on medical corpora of PubMed abstracts (*PubMed*, n.d.) and PMC full-text articles (*Home - PMC - NCBI*, n.d.).

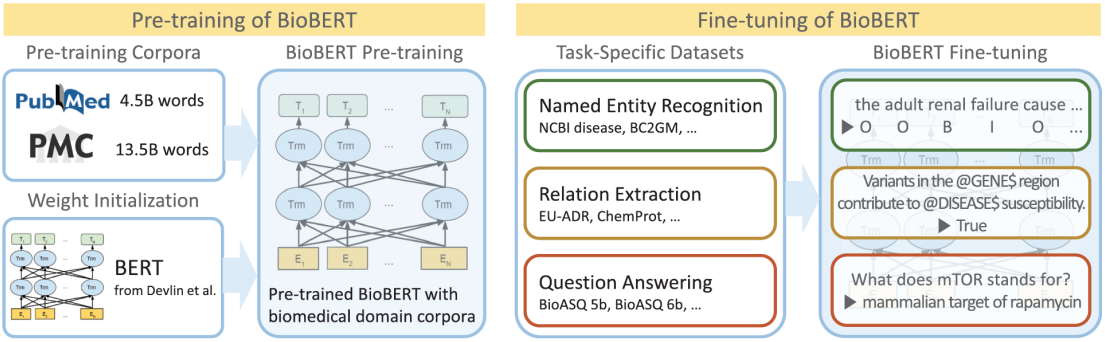


Figure 10: Overview of the pre-training and fine-tuning of BioBERT (Lee et al., 2019)

However, they were trained in 3 different ways on corpus for comparison, PubMed, PMC and PubMed+PMC, without changing the default architecture of general BERT. Moreover, BioBERT fine-tunes the model on-top to perform 3 specific NLP Tasks i.e Named Entity Recognition, Relation Extraction and Question Answering. All of these test results surpass original BERT on the same task. BioBERT claims that pre-training BERT on biomedical corpora is crucial and they provided their extension of BERT pre-trained models as open-source (https://github.com/google-research/bert). The interesting part within BioBERT paper is that training on more dataset doesn't guarantee a better performance. More specifically, training on only 'PubMed' can also perform better in some tasks, which we could see from Table 1.

Table 1: Biomedical question answering test results

			BERT	BioBERT v1.0			BioBERT v1.1
Datasets	Metrics	SOTA	(Wiki + Books)	(+ PubMed)	(+ PMC)	(+ PubMed + PMC)	(+ PubMed)
BioASQ 4b	S	20.01	27.33	25.47	26.09	**28.57**	27.95
	L	28.81	44.72	44.72	42.24	**47.82**	44.10
	M	23.52	33.77	33.28	32.42	**35.17**	34.72
BioASQ 5b	S	41.33	39.33	41.33	42.00	44.00	**46.00**
	L	56.67	52.67	55.33	54.67	56.67	**60.00**
	M	47.24	44.27	46.73	46.93	49.38	**51.64**
BioASQ 6b	S	24.22	33.54	**43.48**	41.61	40.37	42.86
	L	37.89	51.55	55.90	55.28	**57.77**	**57.77**
	M	27.84	40.88	48.11	47.02	47.48	**48.43**

Notes: BioASQ 4b/5b/6b datasets are designed for NLP question answering tasks from BioASQ leaderboard (http://participants-area.bioasq.org/). Strict Accuracy (S), Lenient Accuracy (L) and Mean Reciprocal Rank (M) scores on each dataset are reported. The best scores are in bold, and the second best scores are underlined. The test results are compared between three types of BERT model that learn from five different corpora. (Lee et al., 2019)

However, even within the same bioMedical domain **ClinicalBERT** believes that clinical narratives have differences in linguistic characteristics from both general text and non-clinical biomedical text used above (Alsentzer et al., 2019). They list out nearest neighbors for 3 sentinel words for each of 3 categories within clinical and bioBERT domains. ClinicalBERT appears to show greater cohesion within its domain than BioBERT, Table 2.

Table 2: Nearest neighbors for 3 sentinel words for each 3 categories (Alsentzer et al., 2019)

Model	Disease			Operations			Generic		
	Glucose	Seizure	Pneumonia	Transfer	Admitted	Discharge	Beach	Newspaper	Table
BioBERT	insulin	episode	vaccine	drainage	admission	admission	coast	news	tables
	exhaustion	appetite	infection	division	sinking	wave	rock	official	row
	dioxide	attack	plague	transplant	hospital	sight	reef	industry	dinner
Clinical	potassium	headache	consolidation	transferred	admission	disposition	shore	publication	scenario
	sodium	stroke	tuberculosis	admitted	transferred	transfer	ocean	organization	compilation
	sugar	agitation	infection	arrival	admit	transferred	land	publicity	technology

Their research focuses on specialized clinicalBERT models by training on two different specialized corpora of clinical notes and discharges summaries using two BERT models of $BERT_{BASE}$ and BioBERT for comparison. Their results show that additionally specialized BioBERT received better performance. Moreover, out of five given tasks they performed, the researchers argue that pure BioBERT only performs better on de-identification (de-ID) tasks because of its fundamental facets.

Table 3: Accuracy (MedNLI) and Extract F1 score (i2b2) across various clinical NLP tasks

Model	MedNLI	i2b2 2006	i2b2 2010	i2b2 2012	i2b2 2014
BERT	77.6%	93.9	83.5	75.9	92.8
BioBERT	80.8%	**94.8**	86.5	78.9	**93.0**
Clinical BERT	80.8%	91.5	86.4	78.5	92.6
Discharge Summary BERT	80.6%	91.9	86.4	78.4	92.8
Bio+Clinical BERT	**82.7%**	94.7	87.2	**78.9**	92.5
Bio+Discharge Summary BERT	**82.7%**	94.8	**87.8**	78.9	92.7

Notes: MedNLI is a natural language inference task (Romanov & Shivade, 2018) and i2b2 are named entity recognition (NER) tasks (*I2b2: Informatics for Integrating Biology & the Bedside*, n.d., p. 2; Sun et al., 2013, p. 2; Uzuner et al., 2011, p. 2)

Similarly, **BlueBERT** additionally trains the pre-trained BERT with their own version of corpora that includes PubMed (*PubMed*, n.d.) and MIMIC-III clinical documents (*MIMIC*, n.d.). However, their researcher came up with a benchmark that is similar to the General Language Understanding Evaluation benchmark or GLUE (*GLUE Benchmark*, n.d.) but for the biomedical domain. Biomedical Language Understanding Evaluation benchmark or BLUE (*Ncbi-Nlp/BLUE_Benchmark*, 2019/2021) consists of 5 tasks and 10 datasets that are specifically NLP tasks for the biomedical domain.

Table 4: BLUE benchmark datasets and tasks (Peng et al., 2019)

Corpus	Train	Dev	Test	Task	Metrics	Domain	Avg sent len
MedSTS, sentence pairs	675	75	318	Sentence similarity	Pearson	Clinical	25.8
BIOSSES, sentence pairs	64	16	20	Sentence similarity	Pearson	Biomedical	22.9
BC5CDR-disease, mentions	4182	4244	4424	NER	F1	Biomedical	22.3
BC5CDR-chemical, mentions	5203	5347	5385	NER	F1	Biomedical	22.3
ShARe/CLEFE, mentions	4628	1075	5195	NER	F1	Clinical	10.6
DDI, relations	2937	1004	979	Relation extraction	micro F1	Biomedical	41.7
ChemProt, relations	4154	2416	3458	Relation extraction	micro F1	Biomedical	34.3
i2b2 2010, relations	3110	11	6293	Relation extraction	F1	Clinical	24.8
HoC, documents	1108	157	315	Document classification	F1	Biomedical	25.3
MedNLI, pairs	11232	1395	1422	Inference	accuracy	Clinical	11.9

With their defined benchmark, they have performed a comparison between multiple state-of-the-art models such as ELMo, BioBERT and their own pretrain BERT. They show that their results are the best with slight improvement from BioBERT. However as they are the one who define this benchmark, this could possibly be due to some bias of their training dataset. Regardless, this Benchmark is useful for the development and experiment of this thesis.

Table 5: Baseline performance on BLUE benchmark task test sets (Peng et al., 2019)

Task	Metrics	SOTA*	ELMo	BioBERT	Our BERT			
					Base (P)	Base (P+M)	Large (P)	Large (P+M)
MedSTS	Pearson	83.6	68.6	84.5	84.5	**84.8**	84.6	83.2
BIOSSES	Pearson	84.8	60.2	82.7	89.3	**91.6**	86.3	75.1
BC5CDR-disease	F	84.1	83.9	85.9	**86.6**	85.4	82.9	83.8
BC5CDR-chemical	F	93.3	91.5	93.0	**93.5**	92.4	91.7	91.1
ShARe/CLEFE	F	70.0	75.6	72.8	75.4	**77.1**	72.7	74.4
DDI	F	72.9	78.9	78.8	78.1	79.4	**79.9**	76.3
ChemProt	F	64.1	66.6	71.3	72.5	69.2	**74.4**	65.1
i2b2	F	73.7	71.2	72.2	74.4	**76.4**	73.3	73.9
HoC	F	81.5	80.0	82.9	85.3	83.1	**87.3**	85.3
MedNLI	acc	73.5	71.4	80.5	82.2	**84.0**	81.5	83.8
Total			78.8	80.5	82.2	**82.3**	81.5	79.2

Notes: SOTA, state-of-the-art as of April 2019. BlueBERT has performed 4 different sizes and corpora of models for comparison.

However, these papers did not deeply study into the details of factors affecting each domain language. **BioMegatron** paper performs a more detailed analysis on subword vocabulary, model size, and domain transfer to evaluate impact in NLP tasks (Shin et al., 2020). They conclude that a language model performs best when it's targeted on domain and application. This also aligns with the "no free lunch" theorem in machine learning that there are no master models that can perform well for all tasks but good enough if we only targeted one.

2.3 Conclusion on Text representation

The process of performing additional training for fine-tune downstream tasks on pre-trained data-rich models is called **Transfer learning**. This technique has emerged as a powerful method in NLP because it gave rise to diversity of approaches, methodology and practice (Raffel et al., 2020). Transfer learning is an optimal way to build a model because it saves time by avoiding training from scratch and gets better performance as extension training converges better (CITED). Transfer learning enables us to develop skillful models with less data. This fit our scenarios of our BERT Language Model that was pre-trained on general corpora then transferred to biomedical corpora.

From an overview standpoint, our approach to build models for downstream NLP tasks has been changed. Instead of directly using input data such as sentences or documents to make classification, we use Language Model to capture semantic meaning and characteristics of those input and embedded into its data representation. Using this representation, we build another model such as a classifier on top to fine-tune for downstream NLP tasks, Figure 11. In order to maximize LM performance, transfer learning technique is used to learn from general corpora then shifted to domain specific corpora.

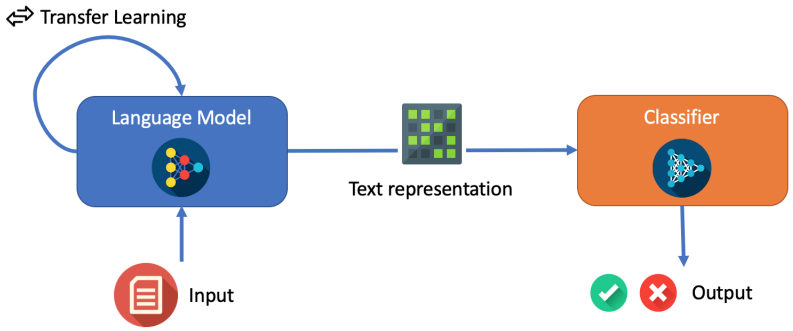


Figure 11: Overview of Modeling with Language model and Text representation

Identical to our thesis, the goal is a classification task given a radiology report to predict the presence of a fungal disease. We could simply use a Feed-forward Neural Network (FNN) with softmax to get our results, Figure 12. However, one of our constraints that we foresee is that we might not have a lot of labeled data for tasks. Semi-supervised (Zhu & Goldberg, 2009) and Active learning (Settles, 2012) might be our alternative solutions, allowing us to leverage the unlabelled data, for the next step of our research

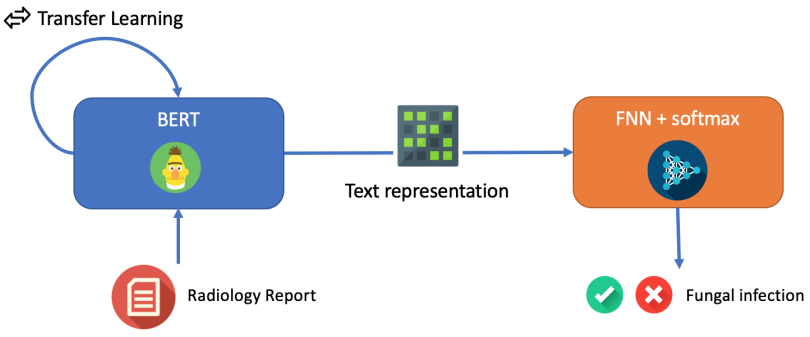


Figure 12: Overview of Modeling of Fungal infection disease detection

3. Summary of the State of the Art

The study of Natural Language Processing has been developed and improved over the past decade. Researchers in the field have built on top or come up with methodology to understand the semantic meaning of words. However, natural language is inherently complicated. There could be countless hidden meanings behind a single word. Humans learn to understand those by experience. Machines, unlike humans, could only learn based on the input we have given them directly, or external knowledge sources they have access to.

From our literature review, we have seen various attempts to extract those features. At every stage, researchers came up with a good text representation that could improve the state-of-the-art. However from those literatures, we can conclude that non-contextualize word representations such as Bag-of-words, Co-occurrence Matrix, and Word2Vec are no longer a good approach because of the sensitivity of meanings on each word. Contextualized word representation became more popular because it could capture those hidden meanings and dynamically adjust depending on our focus. The latest state-of-the-art BERT follows the same manner.

BERT or Bidirectional Encoder Representations from Transformers is a Language Model that takes input and converts it into text representation. It is a pre-trained Language Model that was learned from the large general corpora of BookCorpus and English Wikipedia. It looks at the whole input context that was given and performs transformations to embedded semantic meaning of each word in its given context. BERT's output is later used on downstream models to predict and classify NLP tasks.

However, as we stress the importance of words sensitivity towards its contexts, general BERT is not customized for a specific domain task that has its own corpora with distribution of words shifted from general ones. Transfer learning became an important element to guide the BERT model into a better performance. This was done by additional training the pre-trained BERT. There is multiple evidence of improvements when pre-trained BERT on biomedical corpora such as BioBERT, ClinicalBERT and BlueBERT. But because it is only as good as what it was trained for, our Radiology report customized version doesn't yet exist. Nevertheless, BERT has surpassed multiple state-of-the-art in both general and biomedical NLP domains. It has generated various promising outcomes throughout several NLP benchmark tasks. Considering all the elements mentioned above, we propose building a system that performs Text Classification using BERT with pre-trained radiology reports data. Our system should achieve a satisfactory outcome of detecting the presence of fungal disease. We hope to use our system to assist General Practice with their diagnosis to help every patient and save their lives.

4. Plan for your research project

4.1 Introduction & Background

The amount of biomedical data is rapidly growing. Reports are generated as electronics documents instead of physical. All information recorded is kept and stored in databases waiting for someone to reveal. Regrettably, reports such as radiology summaries, which contain lots of information, are usually generated for the purpose of addressing general practitioner immediate focuses of patients. The rest of assessment results that were performed but not relevant to GP concerns are usually overlooked. These reports sometimes contain valuable information that may identify or indicate additional disease that patients are possibly infected with. Consequently, Natural language text processing for extracting these information could be crucial to patients' benefit. Machines could process data in parallel with GP investigation and will help reduce information being overlooked. The result of machine learning will help assisting, analyse and diagnostic to ensure that unusual patterns are being detected and lesson precious information unattended.

Recent progress of biomedical text processing was made possible with the improvement of state-of-the-arts in Natural Language Processing (NLP). Text representation modeling has brought a revolution in understanding specialized corpora in various domains including biomedical. For instance, Embeddings from Language Model (ELMo) and Bidirectional encoder representation from transformer (BERT) are machine learning algorithms that could learn semantic meaning representation from given context. These approaches allow word representation to dynamically embed contextualized word corpora and greatly improve understanding on sentences structure.

4.2 Purpose Statement & Research questions

Machine learning capabilities have been demonstrated in various real-life applications. Tackling challenges in the biomedical domain could be challenging, yet, rewarding. Our goal is to assist General Practitioners by proving that machine learning can assist them to identify unusual patterns and detect disease infection from radiology reports with natural language processing state-of-the-art modeling.

We will be answering the following research questions:

- **RQ1:** Can we specialize existing state-of-the-art word representation systems such as BERT, BioBERT or ClinicalBERT onto radiology reports?
- **RQ2:** Using specialized word representations on radiology reports, can we achieve satisfactory performance on text classification tasks of detecting disease infection? We will evaluate performance both statistical quantitative and domain experts feedback qualitative.
- **RQ3:** How can we leverage labeled data classification from other biomedical tasks onto unlabel radiology reports that we have accesses to improve our detection system further?

4.3 Research Methodology

In this section, we will describe our approach to address each individual research question. Going through details of how to set up the experiment, evaluation and expected outcome, we hope to see promising results from each research question to pass onto one another.

Research Question 1

Can we specialize existing state-of-the-art word representation systems such as BERT, BioBERT or ClinicalBERT onto radiology reports?

Generally, to answer this question we need to understand radiology reports' characteristics. Our assumption is that each domain has its own corporas. When approaching tasks within different domains, model's corporas need to be adjusted and fine tuned to fit their shifted distribution. However, our task of radiology reports is still within the biomedical domain. Previous BERT has also been built based on some existing medical data such as PubMed, which is biomedical magazine and Clinical Notes by General Practitioners. We will set up our experiment to see distribution of corporas by perform word frequency analysis on 4 different datasets in our focuses:

- English Wikipedia - used by General BERT
- PubMed - used by bioBERT
- Clinical Notes - used by clinicalBERT
- Radiology reports

Hopefully, this will explain the distribution shift of words and stress the importance of specialized word representation systems. Nevertheless, existing BERT have been trained over these datasets. This would be a great measurement when we conclude the importance of selecting the BERT pre-train checkpoint that we will be extending our models from.

Our main task to prove this research question would be by additional train three BERT models on radiology report:

1. radiology BERT, initialize from $\text{BERT}_{\text{BASE}}$
2. radiology BERT, initialize from BioBERT
3. radiology BERT, initialize from clinicalBERT

With these newly built models, we will evaluate their performance of learning radiology report's word distribution by conducting an experiment of masked words accuracy on models before and after learning about radiology's corporas. Our hypothesis is that we should see some improvement in radiology BERT over models before training. Moreover, if radiology reports word distribution is really shifted from other biomedicals. These prediction accuracy should be noticeably significant.

Research Question 2

Using specialized word representations on radiology reports, can we achieve satisfactory performance on text classification tasks of detecting disease infection?

From the above results from Research Question 1, we will select those Language Models to perform classification models on-top. We will use simple Feed-forward neural network model (FNN) to perform our analysis on 3 benchmarks:

- BLURB Benchmark - Microsoft's Biomedical Language Understanding and Reasoning benchmark
- BLUE Benchmark - Biomedical Language Understanding Evaluation benchmark
- Radiology report Benchmark - Our given label datasets provided by subject domain experts supporting this project from Monash University - Medicine Alfred Hospital

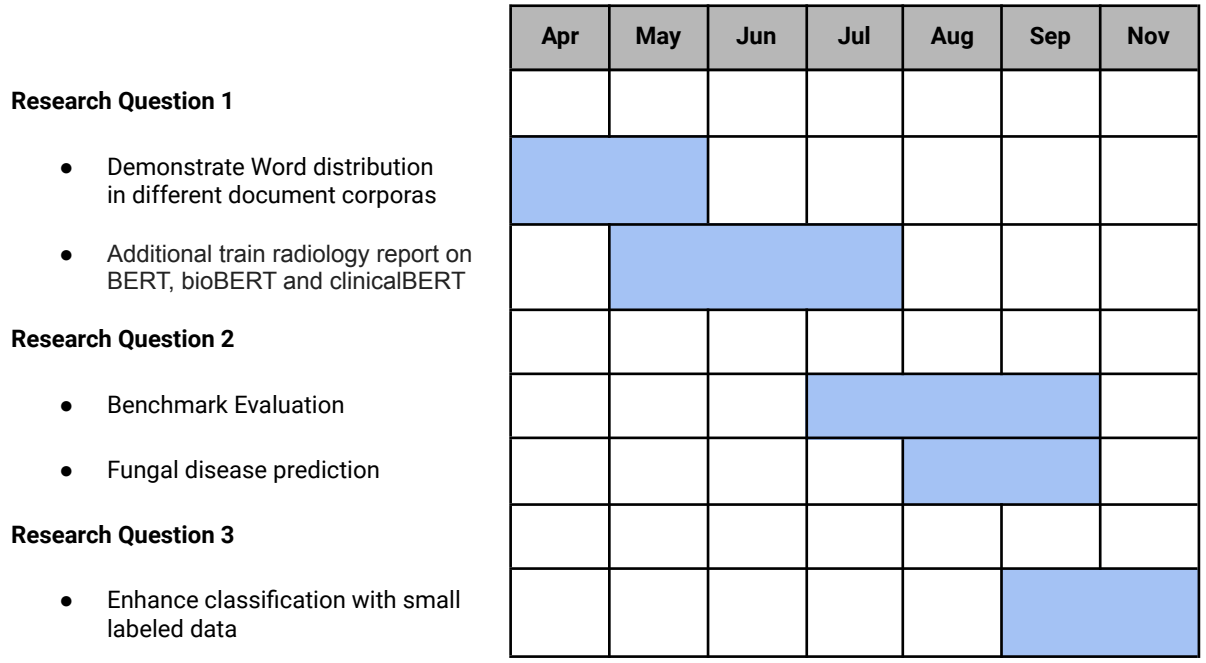
These benchmarks will be our evaluation of our model succession. We will conduct quantitative and qualitative analysis on model performance. For quantitative, we will measure our models by bringing out statistical results onto these benchmarks datasets. Our hypothesis is that we will see our radiology report trained BERT can retain on-par in BLURB and BLUE benchmarks tasks and show enhancement when performing disease classification on a Radiology report benchmark given by subject domain experts. Once we achieve satisfactory results on quantitative analysis. We will perform qualitative analysis by setting up a system to run in parallel with General Practitioner diagnostic on recent radiology reports as assistive tools that will flag potential infection. We evaluate our qualitative analysis by survey those General Practitioners if the displayed possibility infection does help assist with their diagnostic decisions. At this stage, our research should have produced a minimum viable product that could flag potential possibility towards fungal infection disease.

Research Question 3

How can we leverage labeled data classification from other biomedical tasks onto unlabel radiology reports that we have accesses to improve our detection system further?

As we proceed to this question, we have already built a promising word representation system that could adapt to radiology report words distribution. However, we may or may not achieve satisfactory results in predicting infection disease. Part of it may be due to the fact that building word distribution can be trained using an unsupervised method which does not require data to be labeled. On the contrary, classification tasks need label data for supervised tasks to identify which cases could be infected by fungal disease. Unfortunately, the cost of labeling these datas are oftenly high especially in medical domains that need subject matter experts. Our access to labeled data is limited. In theory, this data could be collected overtime during GP documentation of patients, but in reality, we may only have a few thousand samples to generate this prediction model. This might not be enough for the model to converge. To address this Research Question 3, we would further seek to try on transfer learning approaches that build models on similar other biomedical tasks then pass on to radiology reports. Our set up for evaluation still remains the same with those 3 benchmarks above. We might explore further on active learning and semi-supervised learning that could minimize learning on labeled data while maximizing the possible outcome.

4.4 Timeline



5. Conclusion

In this paper, we have addressed the general problem of valuable data being overlooked. This problem has occurred and raised concerns in the biomedical domain as well, where various reports of physical examination on the patient have been documented. This information could be crucial as it could determine patients' health. We acknowledge this problem and remodel this challenge into NLP tasks of identifying the fungal infections based on radiology reports. In particular, we will be performing text classification using radiology reports as predictors to classify the possibility of fungal infections.

Our studies in literature have shown us that with recent NLP approaches it is possible to solve the challenge. We have gone through the history of text representation in NLP and improvement of the state-of-the-arts. The latest paper exhibits contextualizing word representation using bidirectional encoder representation from transformer (BERT). This method of Language Modeling allows us to embed semantic meaning of sentences into vectors dynamically. Thus, BERT that was pre-trained on general corporas require adjustment to demonstrate maximum performance when handling specialized domain corporas such as biomedical. Specifically, extended BERT needs to be trained on corporas of the focus because even within the same domain word distribution could be shifted. We acknowledge that limitation and will be focusing our thesis onto radiology report distribution of words. We believe that if our Language model could correctly be embedded with radiology report characteristics. Our downstream classification accuracy to predict the possibility of patients infected by fungal infection disease would be enhanced.

6. Reference List

Alammar, J. (n.d.). *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)*. Retrieved April 26, 2021, from http://jalammar.github.io/illustrated-bert/

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. *ArXiv:1904.03323 [Cs]*. http://arxiv.org/abs/1904.03323

Baker, S., Korhonen, A., & Pyysalo, S. (n.d.). *Cancer Hallmark Text Classification Using Convolutional Neural Networks*. 9.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *ArXiv:1903.10676 [Cs]*. http://arxiv.org/abs/1903.10676

Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162. https://doi.org/10.1017/S1351324916000334

Computer. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Computer&oldid=1015399018

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2017). Very Deep Convolutional Networks for Text Classification. *ArXiv:1606.01781 [Cs]*. http://arxiv.org/abs/1606.01781

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Evolution of Natural Language Processing. (2020, September 18). Fresh Gravity. http://www.freshgravity.com/evolution-of-natural-language-processing/

GLUE Benchmark. (n.d.). Retrieved April 27, 2021, from https://gluebenchmark.com/

Hieronymus, J. L., & Laboratories, B. (n.d.). *ASCII Phonetic Symbols for the World's Languages: Worldbet*. 48.

Home—PMC - NCBI. (n.d.). Retrieved April 26, 2021, from https://www.ncbi.nlm.nih.gov/pmc/

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical Text Classification using Convolutional Neural Networks. *ArXiv:1704.06841 [Cs]*. http://arxiv.org/abs/1704.06841

i2b2: Informatics for Integrating Biology & the Bedside. (n.d.). Retrieved April 27, 2021, from https://www.i2b2.org/NLP/HeartDisease/

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *ArXiv:1607.01759 [Cs]*. http://arxiv.org/abs/1607.01759

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, btz682. https://doi.org/10.1093/bioinformatics/btz682

Liddy, E. D. (n.d.). *Natural Language Processing*. 15.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv:1907.11692 [Cs]*. http://arxiv.org/abs/1907.11692

Liu, Z., Lin, Y., & Sun, M. (2020). Representation Learning and NLP. In Z. Liu, Y. Lin, & M. Sun, *Representation Learning for Natural Language Processing* (pp. 1–11). Springer Singapore. https://doi.org/10.1007/978-981-15-5573-2_1

Louis, A. (2020, July 7). *A Brief History of Natural Language Processing—Part 2*. Medium. https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Marafino, B. J., Davies, J. M., Bardach, N. S., Dean, M. L., & Dudley, R. A. (2014). N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the*

American Medical Informatics Association, *21*(5), 871–875. https://doi.org/10.1136/amiajnl-2014-002694

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

MIMIC. (n.d.). Retrieved April 27, 2021, from https://mimic.physionet.org/about/mimic/

Ncbi-nlp/BLUE_Benchmark. (2021). [Python]. NLM/NCBI BioNLP Research Group (PI: Zhiyong Lu). https://github.com/ncbi-nlp/BLUE_Benchmark (Original work published 2019)

Pal, S. (2019, June 22). *Implementing Word2Vec in Tensorflow*. Medium. https://medium.com/analytics-vidhya/implementing-word2vec-in-tensorflow-44f93cf2665f

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *ArXiv:1906.05474 [Cs]*. http://arxiv.org/abs/1906.05474

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv:1802.05365 [Cs]*. http://arxiv.org/abs/1802.05365

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

PubMed. (n.d.). PubMed. Retrieved April 26, 2021, from https://pubmed.ncbi.nlm.nih.gov/

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). *Improving Language Understanding by Generative Pre-Training*. 12.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J.

(2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv:1910.10683 [Cs, Stat]*. http://arxiv.org/abs/1910.10683

Romanov, A., & Shivade, C. (2018). Lessons from Natural Language Inference in the Clinical Domain. *ArXiv:1808.06752 [Cs]*. http://arxiv.org/abs/1808.06752

Settles, B. (2012). Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

Shin, H.-C., Zhang, Y., Bakhturina, E., Puri, R., Patwary, M., Shoeybi, M., & Mani, R. (2020). BioMegatron: Larger Biomedical Domain Language Model. *ArXiv:2010.06060 [Cs]*. http://arxiv.org/abs/2010.06060

Stop One-Hot Encoding Your Categorical Variables. | by Andre Ye | Towards Data Science. (n.d.). Retrieved April 5, 2021, from https://towardsdatascience.com/stop-one-hot-encoding-your-categorical-variables-bbb0fba89809

Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, *20*(5), 806–813. https://doi.org/10.1136/amiajnl-2013-001628

Tracking Progress in Natural Language Processing. (n.d.). NLP-Progress. Retrieved April 28, 2021, from http://nlpprogress.com/

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556. https://doi.org/10.1136/amiajnl-2011-000203

Vargas, E. (2017, February 9). *A Comprehensive Introduction to Word Vector Representations*. Medium. https://medium.com/ai-society/jkljlj-7d6e699895c4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Yang, Y., UY, M. C. S., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. *ArXiv:2006.08097 [Cs]*. http://arxiv.org/abs/2006.08097

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0

Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*(1), 1–130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006