Development of a Language Model and Opinion Extraction for Text Analysis of Online Platforms

by

Mohiuddin Md Abdul Qudar

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science

in

Computer Science

in the

Faculty of Science and Environmental Studies

of

Lakehead University, Thunder Bay

Committee in charge:

Winter 2021

The thesis of Mohiuddin Md Abdul Qudar, titled Development of a Language Model and Opinion Extraction for Text Analysis of Online Platforms, is approved:

_____ Date _____

_____ Date _____

_____ Date _____

Lakehead University, Thunder Bay

Development of a Language Model and Opinion Extraction for Text Analysis of Online Platforms

Mohiuddin Md Abdul Qudar

Abstract

DEVELOPMENT OF A LANGUAGE MODEL AND OPINION EXTRACTION FOR TEXT ANALYSIS OF ONLINE PLATFORMS

Language models are one of the fundamental components in a wide variety of natural language processing tasks. The proliferation of text data over the last two decades and the developments in the field of deep learning have encouraged researchers to explore ways to build language models that have achieved results at par with human intelligence. An extensive survey is presented in Chapter 2 exploring the types of language models, with a focus on transformer-based language models owing to the state-of-the-art results achieved and the popularity gained by these models. This survey helped to identify existing shortcomings and research needs. With the advancements of deep learning in the domain of natural language processing, extracting meaningful information from social media platforms, especially Twitter, has become a growing interest among natural language researchers. However, applying existing language representation models to extract information from Twitter does not often produce good results. To address this issue, Chapter 3 introduces two TweetBERT models which are domain specific language presentation models pre-trained on millions of tweets. TweetBERT models significantly outperform the traditional BERT models in Twitter text mining tasks. Moreover, a comprehensive analysis is presented by evaluating 12 BERT models on 31 different datasets. The results validate our hypothesis that continuously training language models on Twitter corpus helps to achieve better performance on Twitter datasets. Finally, in Chapter 4, a novel opinion mining system called ONSET is presented. ONSET is mainly proposed to address the need for large amounts of quality data to fine-tune state-of-

the-art pre-trained language models. Fine-tuning language models can only produce good results if trained with a large amount of relevant data. ONSET is a technique that can fine-tune language models for opinion extractions using unlabelled training data. This system is developed through a fine-tuned language model using an unsupervised learning approach to label aspects using topic modeling and then using semi-supervised learning with data augmentation. With extensive experiments performed during this research, the proposed model can achieve similar results as some state-of-the-art models produce with a high quantity of labelled training data.

Dedication

This is dedicated to my family, friends and everyone who believed and supported me throughout my academic journey.

Contents

List of Figures

List of Tables

Acronyms

ABAE Attention- Based Aspect Extraction. 74

ABSA Aspect-Based Sentiment Analysis. vi, 67, 76, 77

AlBERT A lite BERT. 16, 21, 34, 37, 40, 46, 49, 54, 56–60, 63

AM Aspect Mining. 67, 68, 70

ASC Aspect Sentiment Classfication. 67, 70

BC5CDR BioCreative V CDR task corpus: a resource for chemical disease relation extraction. 26–28, 52, 61

BERT Bidirectional Encoder Representations from Transformers. v, 2, 11, 13, 15, 17–19, 34, 37–45, 47–50, 55, 56, 59–64, 68, 71, 74, 78–80

BioBERT Bidirectional Encoder Representations from Transformers for Biomedical Text Mining. 15, 17, 34, 37–39, 60, 62–64

CoLA The Corpus of Linguistic Acceptability. 22, 51, 56–59

CVT Cross View Training. 68, 73, 74, 76, 78, 80, 81

DA Data Augmentation. 68, 70–72, 74, 75, 78–80, 82

DeBERTa Decoding-enhanced BERT with Disentangled Attention. 42, 43, 56–59

EDA Easy Data Augmentation. 71, 79

ELMO Embeddings from Language Models. 10, 12, 13, 71

ERNIE Enhanced Language Representation with Informative Entities. 43, 44

GLUE General Language Understanding Evaluation. 17, 22, 34, 41, 43, 44, 55, 56

GRU Gated recurrent units. 70

KGs Knowledge graphs. 43, 44

LDA Latent Dirichlet Allocation. vi, 67, 70, 71, 74, 78, 79

MRPC Microsoft Research Paraphrase Corpus. 51, 56–59

NCBI National Center for Biotechnology Information. 26, 27, 52, 60, 61

NLP Natural language processing. iv, 5–7, 11–14, 16–19, 27, 37, 41, 44, 74, 75

QQP Quora Question Pairs. 23, 51, 57

RACE Large-scale ReAding Comprehension Dataset From Examinations. 52, 55

RoBERTa A Robustly Optimized BERT Pretraining Approach. 17, 19, 21, 41, 42, 56–59, 64

RTE Recognizing Textual Entailment. 24, 51, 57

SciBERT A Pretrained Language Model for Scientific Text. 17, 19, 21, 34, 37, 39, 41, 47–49, 57–60, 62, 63

SciREX A Challenge Dataset for Document-Level Information Extraction. 53

SQuAD Stanford Question Answering Dataset. 24, 43, 51, 55–59

SST The Stanford Sentiment Treebank. 22, 57–59

StrucBERT Incorporating Language Structures into Pre-training for Deep Language Understanding. 43, 56–59

STS Semantic Textual Similarity Benchmark. 23, 51, 57–59

SWAG A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. 26, 52, 55, 57–59

TAC Text Analysis Conference. 29

XLNet Generalized Autoregressive Pretraining for Language Understanding. 18, 19

Acknowledgments

I am immensely thankful to my supervisor, Dr. Vijay Mago and all colleagues at the DaTALab in the CASES building at Lakehead University, for their invaluable insights and expertise. I would also like thank Dr. Abhi Rao for helping proof-read some of the work in Chapter 2. I would like to extend thanks to Andrew Heppner and Maegen Lavallee for their help with scholarship applications.

I am incredibly grateful for NSERC Discovery Grant held by my supervisor, for providing support and fund throughout my degree. Additionally, I would also like to express my gratitude to Lakehead University's HPC for providing high-end GPUs and CASES building for providing the infrastructure.

Chapter 1

Introduction

This study is mainly comprised of articles written during this degree. The main focus of the thesis was to conduct text analysis for online platforms. In Chapter 2, the thesis provides an insight on important topics in the field of language models, emphasising on transformer-based language models, that are essential to understand for the Chapters to follow. It includes the properties of various state-of-art-language models and their application on different types of natural language processing tasks. Chapter 2 serves as a point of reference for researchers to gain an understanding of the recent developments and breakthroughs in the field of language models.

In Chapter 3, we present two TweetBERT models, which are domain specific language representation models, pre-trained on millions of tweets. Twitter is a well-known microblogging social site where users express their views and opinions in real-time, as a result, tweets tend to contain valuable information. Due to this reason, mining useful information from tweets has become a growing interest among natural language researchers. Implementing the existing language model on Twitter text analysis tasks seldomly yields good results. Moreover, no language representation models exist for text analysis that is unique to the

social media domain. Hence, to conduct text analysis for Twitter, TweetBERT was developed and to demonstrate the effectiveness of the approach TweetBERTs were fine tuned on two major Twitter text mining task: sentiment analysis and classification. For Twitter datasets, we show that the TweetBERT models outperform the conventional Bidirectional Encoder Representations from Transformers (BERT) models by more than 7% in Twitter text mining tasks. A thorough and detailed analysis is presented by comparing the results of 12 different BERT models, including TweetBERTs, on 31 different datasets. The results indicates that continuously training language models on Twitter corpus over time improves output on Twitter datasets.

Lastly, in Chapter 4, the thesis provides a a novel opinion mining system called ONSET. Online businesses are highly interested in finding practical solutions to opinion mining, but it is challenging to extract aspects and sentiments from the text. One way to solve this problem is to fine-tune good quality extractions from reviews using state-of-the-art pre-trained language models. However, such fine-tuned language models can produce good results if trained with a large amount of relevant data. In this thesis, we present a technique that can fine-tune language models for opinion extractions using unlabelled training data. The framework is built using a fine-tuned language model that takes into account unsupervised learning to extract aspects with the aid of topic modeling, followed by semi-supervised learning with data augmentation. Based on comprehensive experiments conducted during this research, it was observed that the proposed method can achieve competitive results as some of the recent robust models that are trained with a large amount of labeled data. F1-scores of 87.30% and 88.35% are achieved on SemEval Aspect-Based Sentiment Analysis and Twitter datasets, respectively.

To summarize, this research is primarily a collection of articles written over the course of this degree. The thesis' key emphasis was on developing efficient language models for

conducting text analysis for online platforms.

Chapter 2

Background

All of this chapter will be submitted in a peer-reviewed journal as the following:

- Qudar, M. M. A., & Mago, V. (2020). A Survey of Language Models.

To broaden my expertise in the field of text analysis, I conducted research on topics related to language models during my degree. As a result, I've compiled summaries of a number of recent publications for the background chapter of my thesis. We plan to publish a comprehensive survey article based on the content presented here.

2.1 Introduction

The field of Natural language processing (NLP) has received significant attention from researchers in the past decade with the advancements in the field of deep learning. Neural networks based components have been used by researchers to replace traditional statistical or symbolic methods, which in turn has yielded increasing performance. Language models are one of the basic components of natural language processing. Language modeling is defined as determining a probability of a text component e.g. word or sentence occurring in a given context and a language model is a function that captures the probability distribution of all possible text components in a natural language. For example, let's consider a partial sentence "Please submit your" It is more likely that the next word would be "homework" or "paper" than the next word being "professor". Language models play an important role in various NLP applications such as machine translation [185], grammatical error correction [132], speech recognition [27], information retrieval [32], text summarization [56], question answering [184], and sentiment analysis [151] [146].

Statistical language models are based on Markov's assumption which states that the distribution of a word depends on some fixed number of words that immediately comes before it. The most popular traditional language model is the n-gram model. N-grams can be defined as a group of words that occur continuously in a given text corpus. Based on the number of words used to predict the probability of the occurrence of the given word n-gram models are classified as uni-gram, bi-gram, tri-gram, and so on. Every model uses $n-1$ previous words to determine the probability of the word in question, for example, the trigram model uses two previous words to determine the probability of the word, the bigram model uses the previous word and the unigram model simply indicates the probability of the given word being present in the document. Natural languages are versatile and there are

frequent addition and deletion of words and phrases, which makes it almost impossible to build training data with all possible combinations of words and sentences. This exponential increase in the requirement of training samples with the increase of input sequences is called the curse of dimensionality. In order to overcome this hurdle, neural network-based language models were proposed which calculate the probability distribution based on the feature-vectors of words instead of discrete units like words or sentences. The feature-vectors may be defined as real value vectors of text data that capture the semantic properties of the words thus removing the mutual exclusiveness of words. For example, consider the two sentences "A student is studying in the school" and "A student is learning in the classroom", the feature-vectors of these two sentences are closely aligned because they contain different words, the semantic properties of the words "studying, learning" and "school, classroom" are similar.

In recent years transformer-based language models have shown promising results in a wide range of NLP tasks, which had led to numerous research works with a focus on language models. However, a comprehensive survey to analyze and compare, various attributes of these models has not been published yet. In this survey article, we briefly provide an introduction to the statistical foundations of language models and discuss in detail various neural networks based language models classifying them as static and dynamic language models. In Section 2.2 of this survey, we discuss the procedure followed to extract and select articles for the survey; in Section 2.3 we discuss in detail, various neural network-based language models. In Section 2.4 we provide a comparison between the state-of-the-art transformer based language models; and lastly in Section 2.5 we present some of the benchmark datasets for fine tuning language models. Figure. 2.1 shows the overall structure of the survey.

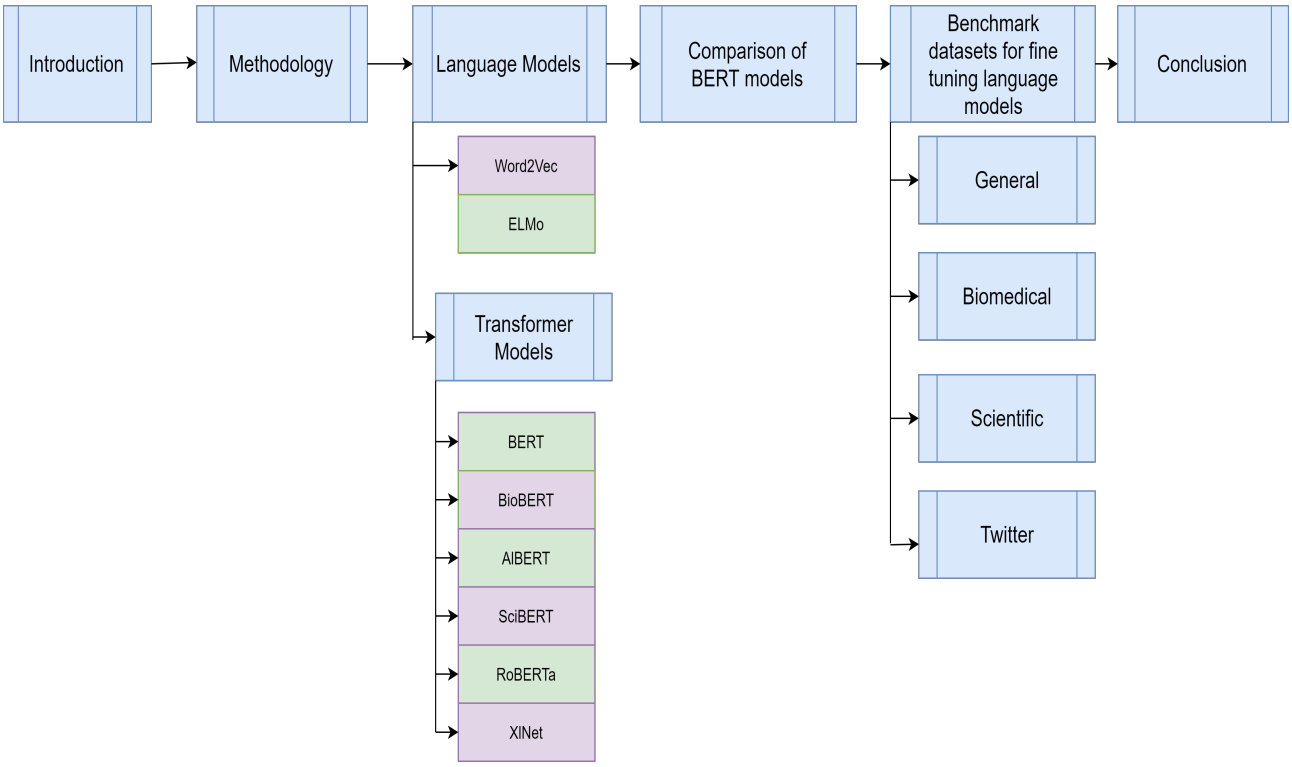

Figure 2.1: Overview of the structure of Chapter 2

2.2 Survey Methodology

With the recent advancements in deep learning models the importance of NLP has significantly increased and thus a vast amount of research has been conducted. To study the impact of the researches done the total number of citations of the selected articles, h-index of the venue where the articles were published and their year of publications were extensively analyzed to study the effect of the researches performed. The articles surveyed in this study were selected searching with keywords such as NLP, text classification, sentiment analysis. Futhermore, article chosen were from 2015 onwards, enabling for a more in-depth analysis of the techniques used in recent papers. Some arxiv papers were chosen because they received a large number of citations in a short period of time. This represents the arxiv paper's

significant influence on the topic. Table A shows the articles selected for the survey, including the total number of citations, h-index on Google scholar of the venue, and the year of publication. Figure. 2.2 shows that no articles were selected from venues that had h-Google index lower than 25 to ensure high quality of articles and shows the number of citations each articles has as of April'20. It can be observed that instead of a linear drop there is a increase in number of articles over 3000+ citations. This is because most the papers in this citation range are written by authors who made a high contribution, such as Tom Mikolov and Christopher Manning. Figure. 2.3 gives a visualization of the year of publication of the articles selected for carrying out this survey. It shows most the articles were published recently.

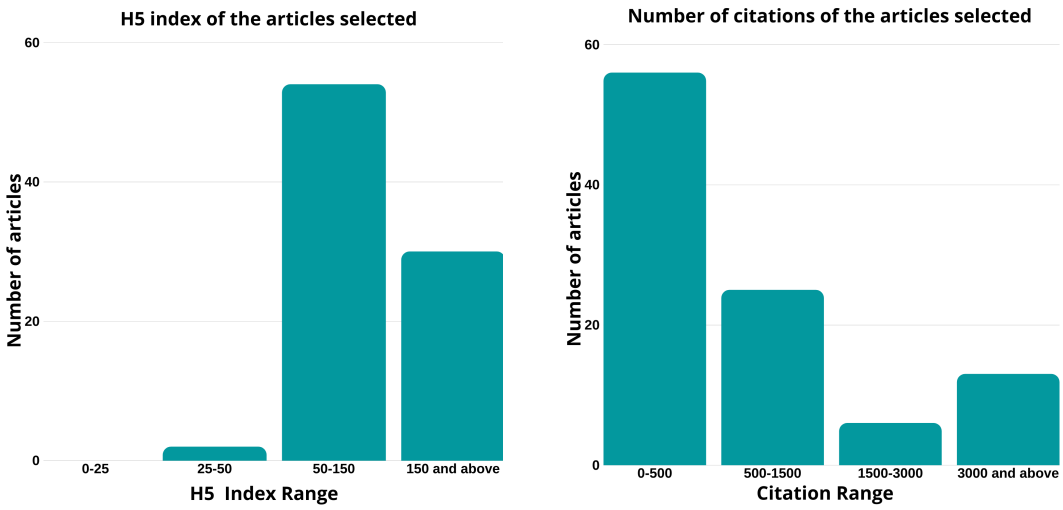

Figure 2.2: The h-Google index of the venues from where the articles were selected and the total number of citations each articles has as of April'20.

For this survey, the name of the authors from the 106 articles that had high citations and high h-index were extracted to form a dataset. A word cloud was created to illustrate a visualization of authors who have made a significant contribution in neural language models.

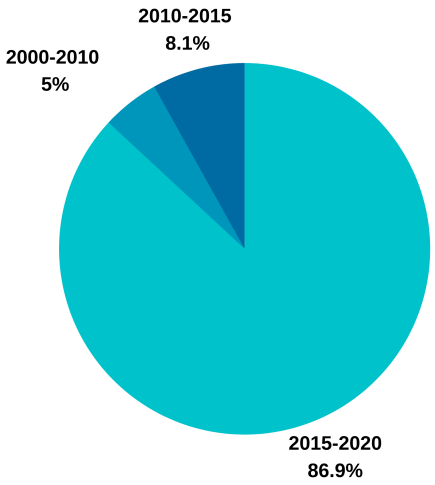

Figure 2.3: The year of publications of the articles that were selected.

As the articles in the survey are from 2015 onward, the word cloud represents authors currently working in the field of language models. The dataset of authors was pre-processed by attaching the author's first and the last name together. This prevents the repetition of the same author having different names, in the word cloud. Figure. 2.4 shows the name of the authors. The size of the author's name corresponds to the frequency of that name appearing in the dataset, thus representing a higher contribution of that particular author.

2.3 Language Models

Capturing the semantic properties of text data using numerical representations is a challenging task and language models exploit the principles of probability to predict the occurrence of a text component based on the previous content. Statistical language models calculate the conditional probability of the occurrence of a word given a set of previous words using

Figure 2.4: Names of the authors in the form of word cloud

equation 2.1

$$\hat{P}(w_1^T) = \prod_{t=1}^{T} \hat{P}(w_t|w_1^{t-1}) \tag{2.1}$$

A neural networks model was proposed in building distributed representations of text data using parallel distributed processing [112]. With significant advancements in the field of neural networks, proposed the first neural probabilistic language model, where the model optimizes to derive at a function that obtains the highest probability of an "out-of-sample" word was proposed [7]. Given a vocabulary of words $V = w_1, w_2, ...w_T$ the model optimizes the function, $\hat{P}$ represents the probability in equation 2.2.

$$f(w_t, ...w_{t-n+1}) = \hat{P}(w_t|w_1^{t-1}) \tag{2.2}$$

In this section, the embeddings of different types of language models are discussed that include word2vec, ELMO and transformer based language models. Word embeddings de-

veloped by various language models are commonly classified as static and contextual word embeddings. Static word embeddings are constructed using a vocabulary and the embedding of a word is constant irrespective of the occurrence of the word whereas the contextual embeddings produce the embeddings based on the context it appears in. Language models like word2vec produce static embeddings, while the more recent transformer-based language models generate dynamic or contextual embeddings. The difference between static and contextual embeddings are tabulated in Table 2.1. Recent language models have primarily focused on transformer-based language models, especially with the introduction of Bidirectional Encoder Representations from Transformers (BERT), since these models have shown to perform well in a variety of NLP tasks such as sentimental analysis and classification problems. Some of the properties of different types of BERT baseline models are illustrated in Table 2.2.

Word2Vec

A language model that used a simple neural network with one hidden layer is called word2vec [114]. Given a large text corpus as input *word2vec* builds distributed representations of words, that when applied in simple mathematical operations produced results that were closely in consensus with human understanding. For example, the difference between the embeddings of the words "king" and "man" when added with the embedding of the word "woman" produces an embedding in close proximity to the embedding of the word "queen". There are two different types of word2vec: the Skip-gram and Continuous Bag of Words (CBOW). The skip-gram model is optimized to predict a target word when given the neighboring or context words and the CBOW model is optimized to predict the context words when given a target word. The dimension of the embeddings depends on the number of neurons in the hidden

Parameters	**Static Word Embeddings**	**Contextualized Word Embeddings**
SEMANTICS	Does not consider the polysemy of a word [81], which is the the ability of words to have multiple meaning. Same embeddings are generated from same word in different contexts [11].	Considers the word semantics in different context taking into account the context of a word [24].
OUTPUT	The output of the training models, for example word2vec, are only vectors [178].	The output of the training is a trained model and vectors [24]. As a result, this trained model can be used to fine-tune different NLP tasks, such as SQuad [143].
NLP TASKS	Word vectors have very shallow word representations [74]. In other words, it only has a single layer for training and each time the network has to be trained from scratch to fine-tune on a NLP task [11].	Weights from the trained model generated can be used to fine-tune the models for a specific natural language task [90] [44]. This process is called transfer learning where instead of training a model from scratch existing neural network models can be modified to train on a small data and give high performance [155].

Table 2.1: Differences between static and contextual word embeddings

layer of the model, and *word2vec* produces static word embeddings with a dimensionality of 300. The advent of *word2vec* was a major breakthrough owing to the simplicity of the model, which enabled researchers to focus on exploiting the advancements in neural networks to build efficient language models.

ELMO

In an attempt to incorporate the concept of polysemy into the embeddings a deep contextualized word embedding model Embeddings from Language Models (ELMO) was proposed

[131]. ELMO uses layers of LSTMs and traverses through a given sentence from both directions trying to predict a given word thus building a more information-rich embedding. Instead of assigning a single embedding to any given word ELMO calculates the embedding of a word based on the sentence it appears in. The embeddings are considered 'deep' because they are formed using the features from all the underlying of the model in contrast to other models which use only the final layer to provide the values [34]. Hence, ELMO generates context-rich embeddings that capture a wide range of syntactic and semantic properties of the words in consideration. The model was able to generate embeddings for words not in the vocabulary or training dataset by taking into consideration the characters in the given word. The model when added with architectures to perform specific NLP tasks both at the input and the output layer achieved state-of-the-art results in six major NLP tasks [131].

Transformer Models

BERT:

Bidirectional Encoder Representations from Transformers (BERT) is similar to ELMO, but uses a pre-trained neural network instead of feature based approach for word representations [36]. BERT's key component is that it applies bidirectional transformer language model while training a corpus [36]. A transformer is a machine learning model that takes into account the ordered sequence of the data, even though it is not necessary that the sequence of words are processed in that order [171]. As a result, it can start to process the end of a sentence without starting to process the beginning. If a language model is trained using a bidirectional transformer it can have a sense of the linguistic context [126]. BERT used two training techniques Masked language model and Next Sentence Prediction [36].

Masked Language Model

Although it is logical to think that bidirectional model performs better than unidirectional models but bidirectional model has its own disadvantage. When using a corpus to train a bidirectional model, it can "allow each word to see itself" [36], since it is bidirectional in nature. To solve this, some percentage of words are randomly masked and the model is asked to predict random words from the input rather than the next word from the sequence [171]. Masking is carried out in three different ways. For example if the sentence to be trained is "My dog is hairy" [36] and the word "hairy" is chosen to be the token, then masking is done either by replacing it with a $< Mask >$ token i.e., "My dog is $< Mask >$" or with a random token e.g. "My dog is apple" or keeping it as it is i.e., "My dog is hairy" [36]. Using these three ways together masking is done to capture the contextual meaning of a word. If only the first method was used, that is only using $< Mask >$ tokens, then the performance of the model would be low as it was never trained on anything other than a masked object. Also sometimes keeping the sentence intact, the model is forced to train on the original representation of the sentence to introduce biasness [171]. This biasness helps the language model to stick to the context [172].

Next Sentence Prediction

The second part for pre-training BERT is done by a method called Next Sentence Prediction [108]. This method requires giving the model a pair of sentence and then testing if the model can predict whether the second sentence comes after the first sentence or not in the corpus. 50% of the time the second sentence is actually related to the first sentence [36]. Next Sentence Prediction is mainly carried out so that the model can understand and relate how two sentences are connected [171], and this helps the model to perform better in various NLP tasks such as Language Inference [32] or Question Answering [184].

BioBERT:

Taking into account the considerable increase in documents generated in the biomedical domain a domain-specific version of BERT called the Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) was proposed [87]. BioBERT uses the architecture and pretraining techniques of BERT while training on a domain-specific corpus. The domain-specific corpus includes abstracts from PubMed - a search engine that contains medical literature and biomedical information and full articles from PubMed Central - a full-text archive of biomedical and life sciences journal literature [1]. The corpora contain 4.5B and 13.5B tokens respectively. The model is initialized with weights from BERT trained on a general English corpus and further trained with the BioMedical corpus for computational efficiency thus using transfer learning. The BioBERT outperformed the existing language models on three biomedical text mining analysis which includes biomedical named entity recognition (0.62% F1 score improvement), biomedical relation extraction (2.80% F1 score improvement), and biomedical question answering (12.24% MRR improvement).

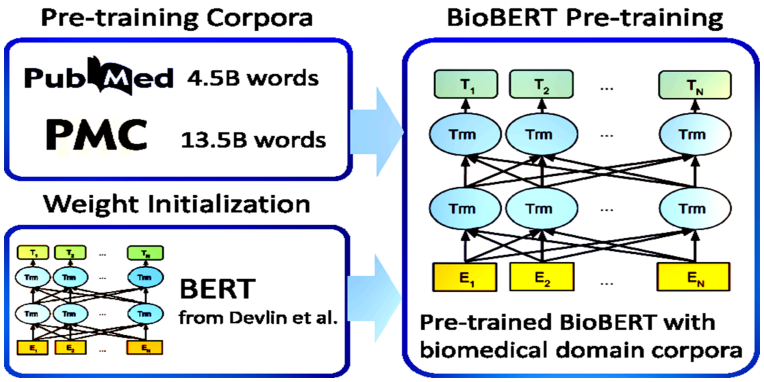

Figure 2.5: Pre-training of biobert with words from PubMed and PMC [87].

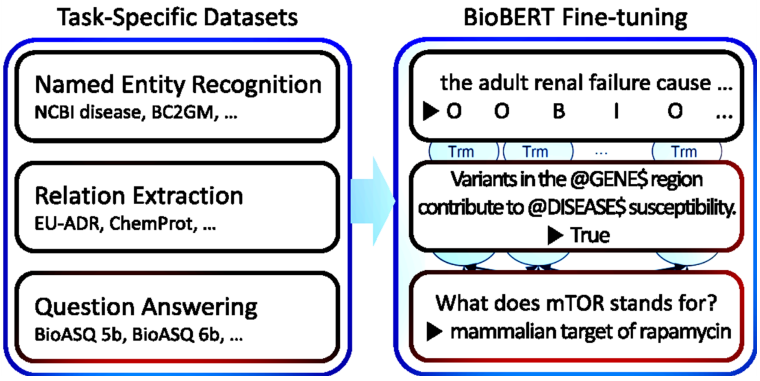

Figure 2.6: Finetuning of BioBERT on specific NLP tasks [87].

AlBERT:

Generally, increasing the number of training corpus and the model size increases the performance of the NLP tasks [84]. However, as the model size increases it becomes difficult to pre-train the model because of the "GPU/TPU memory limitations and longer training times" [84]. To solve this issue A lite BERT (AlBERT) was introduced. AlBERT has the same architecture as BERT. AlBERT uses two parameter-reduction techniques to significantly reduces the number of training parameters of BERT. They are:

- Factorized embedding parameterization, it breaks down the large word matrix into smaller matrices [84]. As a result the size of the word representations is separated from the size of vocabulary embedding [6].
- Cross-layer parameter sharing, which stops the parameters from increasing as the depth of the neural network increases [84].

Both the techniques significantly decrease the training time and increase the training speed of the model [84].

SciBERT:

Similar to BioBERT, a modified version of the BERT model trained on a scientific corpus called the A Pretrained Language Model for Scientific Text (SciBERT) was proposed [6]. The corpus is built using 1.14M articles obtained from the semantic scholar containing predominantly articles from the biomedical domain and approximately 18% from the computer science domain. Unlike BioBERT, SciBERT uses a vocabulary built specifically for scientific text. Using the Sentencepiece[1] library, the authors build a 30K size vocabulary that overlaps with the BERT vocabulary by 42%. SciBERT is evaluated across 5 different NLP tasks on domain-specific datasets. The model achieves state-of-the-art results in 3 out of 7 biomedical datasets, in all 3 datasets in the computer science domain, and in 2 multidomain datasets.

RoBERTa:

A Robustly Optimized BERT Pretraining Approach (RoBERTa) was build upon BERT for pretraining natural language understanding systems. RoBERTa mainly improves on the hyperparameters of BERT and trains on greater quantities of mini-batches and learning rates [104]. Moreover, in RoBERTa BERT's next-sentence pretraining approach is removed which enables RoBERTa to perform better than BERT on language masking approach since many hyperparameters of BERT are not used as next-sentence pretraining task is removed thus enables RoBERTa to perform better for downstreaming tasks. RoBERTa was also trained with a larger quantity of data and also for a longer time thus improving the memory of RoBERTa. When the modifications were applied in the proposed RoBERTa model, there was a significant performance improvement in GLUE benchmark dataset and thus beating the performance of the XlNet-Large model [104]. After introducing the RoBERTa model,

[1]https://github.com/google/sentencepiece

researchers have deduced that by hyperparameter tuning the training approach, the performance on multiple types of NLP task can be improved significantly. RoBERTa is part of Facebook's reasearch that is still in progress for enhancing the self-supervised machanisms that can be able to perform with less amount of data labeling and training time.

XlNet:

Generalized Autoregressive Pretraining for Language Understanding (XLNet) that uses the TransformerXL architecture and it has shown to outperform BERT in 20 different NLP tasks including document ranking, question answering, natural language inference, and sentiment analysis [193] [35]. The model claims to address the limitations of BERT such as,

- Independence Assumption: Given the unmasked tokens, the BERT model assumes that the tokens that are predicted are independent of each other which is an oversimplified assumption since high-range dependency contexts are quite common in natural language [193].

- Noise Input: Artificial symbols such as *MASK* used in the BERT model tends to create noise as such symbols do not exist in the downstream tasks. Hence, these symbols lead to inconsistencies in the pre-training and finetuning phase. The masked tokens can be replaced with original tokens, but the issue will still not be solved as the original tokens can only be used by a small probability [193] [35].

The XLNet method uses the permutation modeling approach by training an autoregressive model on all possible permutations of words in a given sentence. It maximizes its performance on the expected log-likelihood by computing all possible permutations instead of traversing a fixed right-left or left-right modeling. Each position of the context learns to use the contextual information from all possible positions thus capturing information bidirectionally

[142]. Therefore, in this approach, no masking is required and so the input data is not contaminated with masked tokens. Autoregressive factorization is performed on T orders, for a given sequence of x of length T. The model will be able to learn information for all possible position from both sides: left to right and right to left. In order to formulate the XLNet method, let Z_T be a set for all the possible permutations length T index sequence of *1,2,...T*. $z_{<t}$ represents the t-th element and the initial $t-1$ elements of a permutation such that $z \in Z_T$. Thus, the permutation language modeling can be defined as [193]:

$$max_\theta \mathbb{E}_{\mathrm{z}\ \mathrm{Z_T}} \left[\sum_{\mathrm{t=1}}^{\mathrm{T}} \log \mathrm{p}\theta(\mathrm{x_{z_t}} | \mathrm{x_{z_{<t}}}) \right] \tag{2.3}$$

For a given text sequence x, a factorization order z at a given time t was sampled and decomposed the likelihood of $p\theta(x)$, according to the order of factorization. When the model is being trained, x_t has come across every possible element in the text sequence, thus enabling the model to learn the bidirectional content.

2.4 Comparison of BERT models

BERT models have earned considerable attention in the Machine Learning community by providing cutting-edge findings in several NLP tasks, such as Question Answering [143], Natural Language Inference [184], and others. The properties of BERT models are often discussed due to their high performance in text analysis tasks. Table 2.2 shows the properties of various baseline BERT models. In comparison to the BERT model, a greater number of corpora were used to pre-train RoBERTa, SciBERT, and BioBERT models, as shown in the table, and they were able to perform substantially better than BERT. SciBERT, for example, outperformed BERT models when pre-trained with scientific datasets in scientific

text analysis tasks. Table 2.2, also, provides the different types of techniques, vocabulary used to pre-train various BERT models.

Model	Size	Vocabulary	Pre-training		Techniques	Transformer layers	Hidden layers	Parameters (M)
			Corpus	Size				
BERT	Tiny	BaseVocab	Book Corpus and English Wiki	16 GB	MLM and NS	2	128	4
	Mini					4	256	11
	Small					4	512	29
	Medium					8	512	41
	Base					12	768	110
	Large					24	1024	334
AlBERT	Base	BaseVocab	Book Corpus and English Wiki	16GB	MLM and NS with 2 parameter reduction technique	12	768	12
	Large					24	1024	18
	xlarge					48	2048	60
	xxlarge					96	4096	235
RoBERTa	Base	BaseVocab	Book Corpus + English Wiki + CC-News + OpenWebText and Stories	160GB	MLM	12	768	125
	Large					24	1024	255
SciBERT	Base	BaseVocab	Scientific article from computer science and biomedical domain	1.14M papers	MLM and NS	12	768	110
	Base	SciVocab				12	768	110
BioBERT	Pubmed	BaseVocab	Book Corpus + Wiki + Pubmed	7.8B words	MLM and NS	12	768	110
	PMC		Book Corpus + Wiki + PMC	16.8B words		12	768	110
	PMC+ Pubmed		Book Corpus + Wiki + Pubmed + PMC	21.3B words		12	768	110

Table 2.2: Properties of various BERT models

2.5 Benchmark Datasets for Fine Tuning Language Models

General

This section discuss some of the benchmark datasets that are often used to fine-tune language models. The general domain contains datasets such as GLUE [173], SQuAD [143], SWAG [199] and RACE datasets. These datasets have contents that covers a wide range of general knowledge in basic English.

GLUE

General Language Understanding Evaluation (GLUE) consists of datasets used for "training, evaluating, and analyzing" language models [173]. GLUE consist of nine different datasets designed in such a way so that it can evaluate a model's understanding of general language [72][181].

- The Corpus of Linguistic Acceptability (CoLA) is a single-sentence task consisting of more than 10,000 English sentences. Each sentence is given a label indicating if its grammatical or ungrammatical English sentence. The language model's task is to predict the label.

- The Stanford Sentiment Treebank (SST) is also a binary single-sentence classification task containing sentences from movie reviews, along with their sentiment, labeled by humans [157]. The task of language model is to predict the sentiment of a given sentence only.

- The Microsoft Research Paraphrase Corpus (MRPC) is a sentence pair corpus generated from online news sources, with human annotations for whether both the sentences are semantically equivalent or not. Thus, the task is to predict if a given sentence-pair has semantic similarity or not [173].

- Quora Question Pairs (QQP) is similar to MRPC; the task is to predict how similar a given pair of questions are in terms of semantic meaning [173]. However, unlike MRPC, QQP dataset is a collection of questions from the question-answering website Quora[2].

- Semantic Textual Similarity Benchmark (STS) is a collections of sentence pairs extracted from news headlines, video and image captions, and similar sources, where semantic similarity score from one to five is assigned to the sentence pairs. The task is to predict the scores [190].

- The Multi-Genre Natural Language Inference Corpus (MNLI) is a crowd sourced dataset, consisting of sentence pairs with a human annotated premise and a hypothesis sentence. The task is to predict whether the premise sentence "entails" the hypothesis, contradicts the hypothesis sentence or stays neutral [173].

- Question Natural Language Inference (QNLI) is a simplified version of SQuAD dataset which has been converted into a binary classification task by forming a pair between each question and each sentence in the corresponding context. A language model's task would be to determine if the sentence contains the answer to the question. A positive value is assigned if pairs contain the correct answer, similarly a negative value is assigned if the pairs do not contain the answer [173].

[2]https://www.quora.com/

- Recognizing Textual Entailment (RTE) is similar to MNLI, where the language model predicts if a given sentence is similar to the hypothesis, contradicts or stays neutral. RTE dataset is very small compared to MNLI [157].
- The Winograd Schema Challenge (WNLI) is a reading comprehension task, in which a model takes a sentence with a pronoun as an input, and selects an answer from a list of choices that references to the given pronoun [181].

SQuAD

Stanford Question Answering Dataset (SQuAD) is a collection of more than 100,000 questions answered by crowdworkers [143]. It contains 107,785 question-answer pairs on 536 articles. Each question and its following answer is from Wikipedia. SQuAD, unlike previous datasets like MCTest dataset [147], does not provide a list of choices. The dataset has been created in such a way so that a language model can select the answer from the context of the passage and the question. In the beginning when releasing this dataset, logistic regression was performed to evaluate the level of difficultly [130]. It was seen that the performance of the model decreases as the diversity of the model increases. The dataset helps a model to predict the context of a language [147].

RACE

Large-scale ReAding Comprehension Dataset From Examinations is a collection of approximately 28,000 English passages and 100,000 questions [82]. This dataset was developed by English language professionals in a such a way so that a language model can gain an ability to read a passage or paragraph. The dataset is a multiple question answering task, where the model tries to predict the correct answer [174][50]. Other existing question answering

dataset have two significant limitations. First, answer from any dataset can be found by simply a word-based search from the passage, which shows that the model is not able to consider the reasoning factor; this restricts the various types of questions that can be asked. Secondly, most datasets are crowd sourced, which introduces unwanted noise and bias in the dataset. Moreover, RACE is the largest dataset that support neural network training and needs logical reasoning to answer. It also contains option for an answer that might not be present in the training passage, which diversifies the questions that can be asked [108]. RACE also contains content from various fields, allowing the language models to be more generic.

Passage: Apollo ran from 1961 to 1972, and was supported by the two-man Gemini program which ran concurrently with it from 1962 to 1966. Gemini missions developed some of the space travel techniques that were necessary for the success of the Apollo missions. Apollo used Saturn family rockets as launch vehicles. Apollo/Saturn vehicles were also used for an Apollo Applications Program, which consisted of Skylab, a space station that supported three manned missions in 1973–74, and the Apollo–Soyuz Test Project, a joint Earth orbit mission with the Soviet Union in 1975.

Question:

What space station supported three manned missions in 1973-1974

Answer:

Skylab

Figure. 2.5 is a sample from SQuAD dataset [143].

SWAG

A Large-Scale Adversarial Dataset for Grounded Commonsense Inference (SWAG) is composed of approximately 113,000 multiple choice questions, including 73,000 instances for training, 20,000 instances for validating, and 20,000 instances for testing, respectively [199]. The multiple choice questions are derived from video caption, that are taken from ActivityNet Captions and the Large Scale Movie Description Challenge (LSMDC) [154]. The ActivityNet Captions consists of around 20,000 YouTube clips, in which each clip contains one of 203 activity types such as doing gymnastics or playing guitar [13]. LSMDC dataset has approximately 128,000 movie captions including both audio descriptions and scripts. For every captions pairs, constituency parsers have been used for splitting the second sentence of each pair into nouns and verb phrases [199]. Each question from the multiple choice questions was annotated by workers from Amazon Mechanical Turk. In order to improve the quality of the dataset, annotation artifacts were minimized. Annotation artifacts are the stylistic patterns that unintentionally provide suggestions for the target labels.

Biomedical

The biomedical domain contain datasets, such as National Center for Biotechnology Information (NCBI), BioCreative V CDR task corpus: a resource for chemical disease relation extraction (BC5CDR) and MedNLI dataset. These datasets only contain texts related to biomedical domain.

NCBI

The national center for biotechnology information disease corpus is a collection of 793 PubMed abstracts in which abstracts are manually labelled by annotators, where the name

of each disease and their corresponding concepts can be found in Medical Subject Headings [50] or in Online Mendelian Inheritance in Man [28]. Name entity recognition is considered to be an important and challenging task of NLP. For example, *adenomatous polyposis coli* [194] and *Friedrich ataxia* [194] can be both a gene or a disease name. Also, abbreviated disease names are commonly used in biomedical texts, such as AS can stand for *Angelanguage modelan syndrome, ankylosing spondylitis,aortic stenosis, Asperger syndrome* or even *autism spectrum* [184]. Also, doctors have their own way of describing a disease and as a result, it more difficult for any language model to achieve good performance. Evaluating a model on this NCBI dataset would show how the model performs in terms of remembering names, especially in biomedical domains [20].

BC5CDR

BioCreative V CDR task corpus: a resource for chemical disease relation extraction (BC5CDR) dataset consists of chemical induced disease (CID) relation extractions [129]. The corpus is composed of 1,500 PubMed articles with approximately 4,400 annotated chemicals, 5,818 diseases and 3,116 chemical-disease interactions. To study the chemical interactions within diseases in depth, it is also not only important for the corpus to have the annotations of the chemical/diseases, but also their interactions with one another [87]. Moreover, the corpus consists of disease/chemical annotations and relation annotations from the corresponding series of articles. Medical Subject Headings (MeSH) indexers were used for annotating the chemical/disease entities. Comparative Toxicogenomics Database (CTD) was used for annotating the CID relations. In order to attain a rich quality of annotation, comprehensive guidelines along with automatic annotation tools were given. For evaluating the inter-annotator agreement (IAA) score between each of the annotators, Jaccard similarity coefficient was calculated separately for the diseases and chemicals. This dataset has been

used in multiple BioCreative V challenge assignments of biomedical text mining.

Chemical Disease Relations (CDR) are usually physically curated with the aid of CTD. However, this approach of curating manually is expensive. Thus, multiple alternative approaches have been proposed of guiding curation with text-mining mechanisms, which consist of the automatic extraction of CDRs. However, these proposed approaches have not been significantly successful since there are shortages of large training corpora. Moreover, to study the chemical interactions within diseases in depth, it is also not only important for the corpus to have the annotations of the chemicals diseases, but also their interactions with one another. However, there are multiple biomedical corpora that consist of only a few selected diseases and chemicals. In addition, none of the previous corpora have the instances of chemical-disease relation annotations, which includes abstracts having the entire chemical disease, relation annotation, and controlled vocabulary. In the case of the BC5CDR dataset, MeSH vocabulary was used as a controlled vocabulary similar to the existing biomedical information extraction datasets, BC5CDR that includes protein-protein interaction and drug-drug interactions. In contrast to the existing biomedical corpora, BC5CDR dataset is crucially different in terms of annotations (CID relations) from the 1,500 PubMed abstracts.

MedNLI Dataset

MedNLI dataset is a dataset that consists of medical history of the patients which is annotated by doctors. MIMIC-III have been used as the source of sentences. In order to avoid in annotating the data, only the medical prescriptions of deceased patients were used. The doctors performed a natural language inference task (NLI) task on the clinical notes that were provided. The MedNLI dataset has shown to be very handy as it is extremely challenging in having constructive, knowledge specific domains, where there is a shortage of training data. The clinical domain has a shortage of massive-scale annotated datasets for training machine

learning models for natural language tasks, such as question answering, or paraphrasing. This makes the MedNLI a suitable resource in the open-medical field, since it is publicly available. Moreover, designing such a knowledge intensive medical domain dataset is expensive as well, since common approaches such as crowdsourcing platforms cannot be used for annotating the dataset. This is because annotating the dataset requires medical domain experts and thus curating such a dataset is very costly. Previously existing datasets have small sizes, and they target general fundamental natural language tasks such as co-reference resolution or information extraction tasks (e.g. named entity extraction).

BIOSSES

Biosses is one of the benchmark dataset for sentence similarity in the biomedical domain. The dataset is composed of 100 pairs of sentences. The sentences are selected from the Text Analysis Conference (TAC) containing Biomedical Summarization Track Training dataset. The TAC dataset consists of 20 reference articles and for each of the reference articles [158]. The sentence pairs are mainly selected from the citing articles in which the sentence has a citation from any one of the reference articles. The data in TAC dataset is both semantically related. At the same time there are dissimilar sentence pairs that also occur in the annotated texts. Sentences that are citing articles from the same reference article will tend to be somewhat semantically similar [23]. In addition, there are other sentences in which the citing sentence referring to an article is written about different ranges of topics or domains. Such sentence pairs will tend to have less or no similarity at all. Thus, sentence pairs covering different rates of similarity were obtained from the TAC dataset. In order to obtain a higher quality of dataset, only the pairs which gave strong alliance between the scores of the annotators were taken into account [158]. Table 2.3 shows a sample from the original biosses dataset.

Sentence1	Sentence2	Comment	Score
Membrane proteins are proteins that interact with biological membranes.	Previous studies have demonstrated that membrane proteins are implicated in many diseases because they are positioned at the apex of signaling pathways that regulate cellular processes	The two sentences are not equivalent, but are on the same topic	1
This article discusses the current data on using anti-HER2 therapies to treat CNS metastasis as well as the newer anti-HER2 agents	Breast cancers with HER2 amplification have a higher risk of CNS metastasis and poorer prognosis	The two sentences are not equivalent, but share some details	2
We were able to confirm that the cancer tissues had reduced expression of miR-126 and miR-424, and increased expression of miR-15b, miR-16, miR-146a, miR-155 and miR-223	A recent study showed that the expression of miR-126 and miR-424 had reduced by the cancer tissues	The two sentences are roughly equivalent, but some important information differs/ missing	3
Hydrolysis of b-lactam antibiotics by b-lactamases is the most common mechanism of resistance for this class of antibacterial agents in clinically important Gram-negative bacteria	In Gram-negative organisms, the most common b-lactam resistance mechanism involves b-lactamase-mediated hydrolysis resulting in subsequent inactivation of the antibiotic	The two sentences are completely or mostly equivalent, as they mean the same thing	4

Table 2.3: A sample from Biosses dataset showing example annotations[158]

JNLPBA

Joint Workshop on Natural language Processing in Biomedicine and its Application is a corpus of Pubmed abstracts specialized for NER tasks [59]. The types of entities that are selected from the biomedical domain include DNA, RNA, protein of cells and its types. However, few of the entities did not turn out to be prominently significant. For instance, entities for genes include the DNA as well as other gene entities like the protein and RNA [59].

Chemprot

Chemprot is a chemical protein interaction corpus generated from PubMed abstracts [163]. The dataset consists of annotations within protein and chemical entities for identifying chemical protein interactions. The dataset is organized in a hierarchical structure with a total of 23 interactions. The author of the dataset has emphasized on mainly five high level interactions that includes: *upregulator, downregulator, agonist, antagonist, and substrate* [163].

GAD

Genetic Association Database is a dataset that was generated from the Genetic Association Archive [68]. The archive mainly contains gene-disease interactions from the sentences of PubMed abstracts. NER tool was also used in this dataset to detect gene-disease interactions and create artificial positive instances from the labeled archive sentences. On the other hand, negative instances from the dataset that were labeled as negative gene-disease interactions.

HOC

Hallmarks of Cancer dataset is generated from cancer hallmarks annotated on 1,499 PubMed abstracts. Afterwards, the dataset was broadened to 1,852 abstracts. The dataset has binary labels which focuses on labelling the cancer discussions on the abstracts as positive samples. However, the samples which had no mention of cancer were filtered out [78].

Scientific

The scientific domain contain datasets such as SciCite and SCIERC that contain texts related to scientific domain.

SCICITE

SciCite is a dataset composed of citation intents that are extracted from various scientific fields [6]. SciCite has been very recently released [31]. The dataset was extracted from Semantic Scholar corpus of medical and computer science domains, and was annotated by giving label to citation content in four categories the are: *method*, *result*, *comparison*, *background*, and *other*. Language models are used to evaluate how well it performs in classification and question answering tasks on scientific domain.

SCIERC

SCIERC dataset is a publicly available dataset that consists of annotations of around 5,000 scientific abstracts [105]. The abstracts are collected from 12 AI conference/workshop proceedings from the Semantic Scholar Corpus. SCIERC is an extended version of previous existing similar datasets that are also collected from scientific articles, which include SemEval 2017 Task 10 [20] and SemEval 2018 Task 7 [20]. SCIERC dataset is broadened in

terms of summing up the cross-sentences related to one another by using conference links, named entity and relation types.

Twitter

The Twitter domain contain datasets such as gender classification and tweets for sentiment analysis. These datasets only contain tweets.

Twitter US airline dataset

Twitter airline dataset[3] is a collection of 14,640 tweets from six US airlines that includes: United, US Airways, Southwest, Delta and Virgin America. The tweets represent the reviews from each of the customers. The tweets are either labeled as positive, negative or neutral, based on the sentiment expressed. The airline company usually checks the feedback of their quality through traditional approaches such as the customer satisfaction questionnaires and surveys that are filled by customers. However, this approach is time consuming and inaccurate as customers might fill up the surveys in a hurry. Hence, designing an airline sentiment dataset as the Twitter airline dataset is very helpful since users in social media give genuine feedback and reviews about the airlines.

Twitter User Gender Classification

In this dataset, [4] annotators were asked to predict and label if the user of a certain Twitter account is male, female or a brand by only viewing the account. The dataset contains about 20,000 instances with user name, user id, account profile, account image and location.

[3]https://www.kaggle.com/crowdflower/twitter-airline-sentiment
[4]https://www.kaggle.com/crowdflower/twitter-user-gender-classification

2.6 Conclusion

In this Chapter, a survey was presented that discusses state-of-art language models. It can be seen that with the recent developments in neural network models, numerous types of neural language models have been proposed. Each model comes with its own advantages and disadvantages. This survey is mainly about the different types of language models, their architecture, why a model was proposed, and the datasets that were used to pre-train and fine-tune the language models. The survey addresses how each article was selected considering the number of citations of the article, the year of publication, and the h-index of the venue. The name of the authors, from the selected articles, was used to form a word cloud to analyse the authors who were currently working in the field of language model. The survey focuses on the recent transformer based language models and BERT and AlBERT. BERT uses a neural network approach for word representations. The advantage of using BERT is that it applies bidirectional transformer language model and this helps BERT to stick to the context of a text. BERT was pre-trained on a general domain and to develop domain specific language models, BERT models were pre-trained on different domain corpus. For example, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) has the same structure as BERT but it was pre-trained on biomedical corpus for biomedical text analysis. Similarly, to extract information from scientific text SciBERT was released.

This survey article also present different datasets commonly used in the field of language models for pre-training or fine tuning a model. When BERT was introduced it was fine-tuned on a number of datasets, such as RACE, SQuAD, and GLUE, to compare the accuracy of BERT with the existing language models. Both RACE and SQuAD are question answering datasets. SQuAD contains more than 100,000 questions answered by crowdworkers and it

was constructed in such a way that the model has to predict the answer from the context of the passage. Likewise, RACE dataset has several questions and each question has a set of four answers. It was designed in such a way that to answer the questions critical thinking is necessary. Thus, accessing a model's capability to understand a text.

Overall this Chapter can serve as a resource, enabling natural language researchers to comprehend and become aware of recent developments of language models

Chapter 3

TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis

All of this chapter is submitted in a peer-reviewed journal and is currently under revision:

- Qudar, M. M. A., & Mago, V. (2020). "TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis"

This chapter captures my contribution to a larger research initiative that applies artificial intelligence techniques to develop a language model for the Twitter platform. Twitter was chosen because it is open for discussion, unlike other social media like Facebook. Twitter does not have any group option, allowing users to express their feelings and thoughts publicly. One of the main objectives of Twitter is that opinions are heard all over the world. All of these make it easier to extract tweets that contain useful information from users in real-time. We submitted this section of my thesis in to a Journal, where it is currently under review.

3.1 Introduction

Twitter is a popular social networking platform where users tend to express themselves and share their information in real-time [19], as a result, text from Twitter is widely studied by natural language researchers and social scientists. The users tend to write texts that are very colloquial and casual, usually following very little or no grammatical rules [19] [44] [178]. The text is written in a completely different way than traditional writings, primarily due to a restriction in their length. However, their contents are so powerful that they can start a movement or impact the economy of a nation or help the public health authorities to plan in the early stages of epidemic, or in current scenario pandemic [46]. Hence, usage and style of language need to be studied extensively. Using existing language representation models, such as BERT [36] or AlBERT [84], to evaluate such texts is a challenge. As a result, a need for a language model specific to social media domain arises. Deep neural network models have contributed significantly to many recent advancements in NLP, especially with the introduction of BERT. BERT and BioBERT [1] have considerably improved performance on datasets, in the general domain and biomedical domain, respectively [128]. State-of-art research indicates that when unsupervised models are pre-trained on large corpora, they perform significantly better in the NLP tasks [87]. However, language models, such as BioBERT, cannot achieve high performance on domains like social media corpora. This is mainly due to the fact that these models are trained on other domain corpora and the language in social media is irregular and mostly informal. To address this need, in this article, TweetBERT is introduced, which is a language representation model that has been pre-trained on a large number of English tweets, for conducting Twitter text analysis. Experimental results show that TweetBERT outperformed previous language models such as SciBERT [6], BioBERT [87] and AlBERT [84] when analyzing twitter texts [167].

In order to study and extract information from social media texts it is neccessary to have a language model specific to social media domain. Futhermore, the TweetBERT models have been evaluated on 31 different datasets, including datasets from general, biomedical, scientific and Twitter domains. These state-of-the-art language representation models have shown promising results in the datasets for conducting text analysis. To show the effectiveness of our approach in Twitter text analysis, TweetBERTs were fine-tuned on two main Twitter text mining tasks: sentiment analysis and classification. In this chapter, the authors made the following contribution:

- TweetBERT, a domain specific language representation model trained on Twitter corpora for general Twitter text mining, is introduced.
- TweetBERT is evaluated on various Twitter datasets and is shown that both TweetBERTv1 and TweetBERTv2 outperform other traditional BERT models, such as BioBERT, SciBERT and BERT itself in Twitter text analysis.
- A comprehensive and elaborate analysis is provided by evaluating 12 different BERT models including TweetBERTs on 31 different datasets, and their results are compared.
- Pre-trained weights of TweetBERT are released and source code is made available to the public[1].

The structure of the chapter is as follows: the existing work in the field of language models is discussed in Section 3.2. Section 3.3 presents the methodology, where it is described how the data has been collected for pre-training the model, and includes the approaches that were taken for implementing the TweetBERT models. There is also a brief description about datasets that were selected for evaluating all the BERT models. A detailed description of the

[1]https://github.com/mohiuddin02/TweetBERT

datasets was provided in Chapter 2.5. Section 3.4 provides a discussion of the experimental results of the benchmark datasets with the various BERT and TweetBERT models. Finally, the conclusion is presented in Section 3.5.

3.2 Related Works

Recently a vast amount of work has been done, in the field of NLP, using bidirectional language models especially by modifying BERT [139]. BERT is a pre-trained neural network word representation model. It uses bidirectional Transformer, which considers the sequence of data and, therefore, can understand the context of a text. It was pre-trained using texts from BookCorpus [185] and English Wiki [36]. BERT uses two techniques for pre-training: masked language model, and next sentence prediction. Masking is carried out in three different ways in a sentence: by replacing a word with a token, or by replacing the word with a random word, or keeping the sentence as it is. These three ways help a bidirectional model to maintain and learn the context of a text. On the other hand, the next sentence prediction helps BERT to relate and connect two sentences together [140, 198]. This is useful when evaluating sentiment analysis or question answering datasets. However, as BERT has been pre-trained on general corpora, it performs poorly in domain specific tasks. As a result, language models like BioBERT and SciBERT have been introduced. Recent language models have been broken down into two categories: contiual pre-training and pre-training from scratch.

Continual Pre-training

Continual models are those which use weights from another model and modify themselves for a specific task [194]. BioBERT is a continual pre-trained model because it was first initialized

with the weights of BERT, and then pre-trained on various biomedical corpora, such as PubMed abstracts and PMC full-text articles, to make it domain specific [87]. BioBERT was released as Biomedical documents were increasing and biomedical text analysis was becoming popular [117]. For example, more than 2,000 articles are published in biomedical peer-reviewed journals every day [39]. Directly using BERT to evaluate biomedical tasks did not give satisfactory results, thus BioBERT was created [87]. BioBERT has the same architecture as BERT, but it has shown to perform better than BERT on biomedical text analysis [172]. BioBERT was mainly evaluated in three biomedical tasks: biomedical named entity recognition, biomedical relation extraction, and biomedical question answering [194]. Likewise, more models were introduced for specific domains. Lately, Covid-Twitter BERT model (CT-BERT) has been released to analyze tweets related to Covid [122]. CT-BERT has been pre-trained on around 160 million coronavirus tweets collected from the Crowdbreaks platform [122]. CT-BERT is a continual BERT model and has shown an improvement of more than 10% on classification datasets compared to the original BERT [134] model. This model has shown the most improvement in the target coronavirus related tweets. Furthermore, other extensions of BERT models, such as AlBERT [84], were also released. Generally, increasing the training corpus increases the performance of the NLP tasks. Moreover, the model size is directly proportional to the size of the training corpus. However, as the model size increases, it becomes increasely difficult to pre-train the model because there are GPU limitations. To address this factor AlBERT was introduced. It uses two parameter-reduction techniques to significantly reduces the number of training parameters in BERT: factorized embedding parameterization [84], which breaks a large matrix into smaller matrices [139], and performing cross-layer parameter sharing, which cuts down the number of parameters as the neural network size increases. These methods have helped BERT to increase its training speed [139].

Pre-training from Scratch

There are other domains where both BERT and BioBERT provide unsatisfactory results. For example, when extracting information from general scientific texts, BERT performed poorly because it was only pre-trained on general domain corpora. As a result, A Pretrained Language Model for Scientific Text (SciBERT) was released to evaluate scientific datasets [6]. SciBERT also has the same architecture as BERT, but it is not a continual model. SciBERT is pre-trained from scratch and it uses a different WordPiece vocabulary called SentencePiece [92] [193]. SentencePiece vocabulary consists of words that are commonly used in scientific domains [2]. When WordPiece and SentencePiece are compared, it is found that there is a similarity of only about 40%. This shows that there is a huge difference between the words regularly used in general and scientific articles. SciBERT was pre-trained on a corpus from semantic scholar, containing 1.14 million papers from the computer science and biomedical domain [43]. Each paper produced around 3,000 tokens making it similar to the number of tokens used to pre-train BERT [43]. Additionally, an optimized robust language model was build based on BERT for pretraining natural language understanding systems. RoBERTa mainly optimizes the hyperparameters of BERT and trains on large number of mini-batches and learning rates [104]. In RoBERTa next-sentence pretraining approach is removed which enables RoBERTa to perform better than BERT on downstreaming tasks since many hyperparameters of BERT are not used. RoBERTa was trained with a larger quantity of data and also for a longer time thus improving the memory of RoBERTa [21]. These modifications allowed RoBERTa model to significantly improve performance in GLUE benchmark dataset. It has been deduced that by hyperparameter tuning during the training approach, the performance on multiple types of NLP task can be improved significantly. Additionally, RoBERTa is not a continual model. It has been pre-trained on an extremely

large, five different types of corpora: BookCorpus, English Wikipedia, CC-News (collected from CommonCrawl News) dataset, OpenWebText, a WebText corpus [141], and Stories, a dataset containing story-like content [141]. The overall size of the datasets was more than 160GB [104]. Moreover, RoBERTa uses 4 different techniques, unlike BERT, to pre-train. They are:

- Segment-pair with next sentence prediction hyperparameter, which is the same as next sentence prediction as BERT [104].
- Sentence-pair next sentence prediction hyperparameter, where back to back sentences from only one document are connected [104].
- Full-sentences hyperparameter, where sentences within a document are connected.
- Doc-sentences hyperparameter, which is similar to full-sentences but two or more documents are connected [141].

Recently, Decoding-enhanced BERT with Disentangled Attention (DeBERTa) was introduced to improve the performance of RoBERTa and BERT models, by using two novels methods: disentangled attention and enhanced masked decoder [51]. In the disentangled attention technique, each of the words is represented with two vectors that can encode the word's position and content. Disentangled matrices are used to calculate the attention weights of the words with respect to their contents and positions [127]. When designing the DeBERTa, it was observed that the attention weight of words is not only dependent on their contents but also on their relative positions. In DeBERTa's architecture, an enhanced masked decoder is used to interpolate the positions in the decoding the masked tokens in the model's pre-training phase. DeBERTa was trained with 48 Transform layers which had around 1.5B parameters. The model was pre-trained with Wikipedia, BookCorpus [204],

OPENWEBTEXT (public Reddit content) [47], and STORIES (a subset of CommonCrawl) [170]. Due to these changes, the DeBERTa model has outperformed the SuperGLUE benchmark over the human baseline.

Incorporating Language Structures into Pre-training for Deep Language Understanding (StrucBERT) was also introduced to improve the performance of BERT in terms of the model's contextual representation [175]. It mainly tries to keep the structure of an input sentence in proper order. StrucBERT improves the BERT's masked language model task by shuffling the tokens after the masked language model task and then predicting the right order of the context. To understand the correlation of each of the sentences, StrucBERT randomly swaps the order of the sentences and predicts both the next sentence and the previous sentence as a new sentence prediction task [175]. Performing the pre-training tasks in these ways, the new model can efficiently capture the word structures in more detail [49]. The StrucBERT was also pre-trained with English Wikipedia, BookCorpus, and WordPiece vocabulary [175]. The model has outperformed BERT and achieved competitive scores in many downstream tasks such as on GLUE and SQuAD datasets.

Enhanced Language Representation with Informative Entities (ERNIE) was built upon BERT to incorporate Knowledge graphs (KGs), which can present rich structured facts for better language understanding [202]. ERNIE was introduced to overcome two main challenges: (1) For a given text, it is a challenge to effectively extract and encode the relevant information into KGs for language representation (2) The approach of BERT's pre-training for language representation is very different from the knowledge representation procedure that leads to two individual vector spaces [202]. Therefore, a special pre-training objective is required to compose the lexical, syntactic, and knowledge information. To extract and encode the knowledge information, the named entity mentions were first identified in the text and then aligned to their respective entities in KGs [202]. The graph structure of the

KGs was encoded with knowledge embedding algorithms such as TransE [12]. Finally, the informative entity embeddings were fed as input into the ERNIE model. Masked language model and next sentence prediction techniques were applied for pre-training ERNIE. For merging the textual and knowledge features, a new pre-training technique was proposed that randomly masked some of the named entity alignments in the input sentence and simultaneously questioned the model to select proper entities from KGs for completing the alignments [202]. The ERNIE model is unique because of its objective to merge both the textual and knowledge features from the KGs for predicting the tokens and entities. Thus, ERNIE was pre-trained on both large-scale textual corpora and KGs [202]. The knowledge embeddings were trained on Wikidata by TransE as the input embeddings for entities. The experimental results have proved that ERNIE succeeded to obtain comparable results with the recent BERT models on many common NLP tasks.

To address the issue of time taken during the pre-training phase of BERT over large corpora, SqueezeBERT was released [61]. SqueezeBERT was designed with grouped convolutions in order to evaluate whether it could conduct NLP tasks at a faster rate than BERT or not, since grouped convolutions have significantly increased the speed of image processing for computer vision networks [61]. In the SqueezeBERT model, several attention layers were replaced with grouped convolutions, with this novel architecture it was able to run four times faster than the BERT-base model and has also achieved comparable accuracy scores with the BERT model on the GLUE dataset [61]. BERT is also highly dependent on global self-attention blocks that cause the language model to suffer from a large memory footprint because it has around 110M parameters [65]. ConvBERT was introduced to solve this issue, which replaces some of the self-attention heads to model local dependencies by a span-based dynamic convolution mechanism [65]. The mixtures of novel convolution and self-attention heads make it extremely effective to learn the context of texts both at the local

and global level. Therefore, this mixed attention mechanism was able to outperform BERT in multiple downstream tasks with reduced training costs and model parameters. The model was per-trained with the open-sourced dataset: OpenWebText which has a similar size to the mixture of English Wikipedia and BooksCorpus [47].

Non-English BERT models for Twitter analysis

There are BERT models which have been pre-trained in other languages such as Italian, and Spanish language for Twitter analysis [136, 71]. According to BERT documentation, *the Multilingual model is somewhat worse than a single-language model. However, it is not feasible for us to train and maintain dozens of single-language models* [36]. To address the limitations of non-English language models and the size of the vocabulary, ALBERTo language model was developed. AlBERTo was the first Italian language model pre-trained on the writing style of social networking sites [136]. The architecture of AlBERTo is similar to BERT and the model was trained on Google TPU-V2 on 200M tweets in the Italian language. The model was later finetuned on SENTIPOLC (SENTIment POLarity Classification) Dataset and showed state-of-the-art results. SentencePiece segmentation algorithm was used for generating an extensive vocabulary for the ALBERTo model. Similarly, TWilBERT is another pre-trained language model trained with Spanish tweets [48]. TWilBERT was pre-trained on 47M Spanish Tweets. SentencePiece algorithm was used to generate the vocabulary for the TWilBERT model with a size of 30,000 subwords. However, in ALBERTo, the model was unable to learn coherence among the tweets as the flow of tweets cannot be identified directly on a sequence of tweets from the same user [48]. This issue was resolved when designing the TWilBERT model. The authors of TWilBERT pointed out that inter-sentence coherence is an essential perspective of language understanding that could boost up the performance

on downstream tasks that require reasoning on pairs of tweets. For this reason, the authors have proposed a coherence sign as in Twitter conversations, where a flow of tweets can be easily identified as (tweet, reply) pairs [48]. Furthermore, there is a Reply Order Prediction signal mechanism which has boosted up the performance of TWilBERT. This mechanism specializes in learning the coherences of each of the sentences of Twitter conversations internally. To learn the coherences between each of the sentences, Sentence order prediction was used which is an alternative to the next sentence prediction approach. The sentence order prediction approach was first used for the AlBERT model to improve its performance [84]. The sentence order prediction signal is a reformulation of next sentence prediction in which the pairs of the unordered sentences are taken into account as negative samples. Unlike sentence order prediction, next sentence prediction is only better at capturing topic coherence rather than sentence coherence. As a result, next sentence prediction does not provide additional information to the masked language modeling task[193]. Sentence order prediction is a means of pre-training signal for learning the coherences in each of the sentences more effectively. By using this approach TWilBERT models have outperformed multilingual BERT on 14 different text classification tasks which include irony detection, sentiment analysis, emotion detection, hate speech detection, stance detection, and topic detection.

Although there are different types of language models pre-trained on various corpora, but no language model yet exists specific to the social media domain in the English language. Thus, to evaluate datasets from the social media domain, a need for such a language model arises. As a result, the authors developed TweetBERT models. The next section discusses the approach taken to create the model.

3.3 Methodology

This section discusses in detail the source of data collecting, tweet extracting, and corpora used for pre-training TweetBERT. An overview of the pre-training approach is shown in Figure. 3.1. There are two TweetBERT models: TweetBERTv1 and TweetBERTv2. Each of these models are pre-trained using different approaches, but have the same architecture as BERT because it is continual pre-training model. Moreover, Table 3.1 shows the different variation of corpora and vocabulary used to pre-train each BERT model. For example, SciBERT uses SciVocab vocabulary which contain words popular in the scientific domain. Further details are provided in the following subsections.

Pre-training Corpus

For domain specific text mining tasks, language models like BioBERT were pre-trained on PubMed and PMC [87]. Likewise, TweetBERT was pre-trained on English tweets. TweetBERTv1 was pre-trained on a corpus that consists of 140 million tweets. The corpus contains tweets from top 100 personalities[2] and top 100 hashtags of Twitter [159]. Top personalities are the group of people who have the highest number of followers, Twitter platform. TweetBERTv2 was pre-trained on a similar but larger corpus containing 540 million English tweets. Table 3.1 shows the different combination of corpora and WordPiece vocabulary involved in training of BERT models.

To create the training datasets, tweets were collected and pre-processed from *big data analytics platform*[3] developed in DaTALab at Lakehead University, Canada [110]. This platform allows users to extract millions of tweets by simply providing keywords as inputs. The tweets are pre-processed by converting all the texts to their lowercase form and all characters (emo-

[2]https://www.kaggle.com/parulpandey/100-mostfollowed-twitter-accounts-as-of-dec2019
[3]https://twitter.datalab.science/

jis, URLs, hashtags, mentions, punctuation) except for *full stop* and *question mark* have been removed. Moreover, it was also ensured that each tweet post had more than 1 sentence so that the next sentence prediction task could be performed efficiently during the pre-training phase. The authors generated two corpora: Corpus140 and Corpus540 which indicate corpora with 140 and 540 million tweets, respectively. Corpus140 contain 2.3 billion word tokens and Corpus540 contain 10.1 billion word tokens. Each corpus consists of tweets from top trending hashtags and top personalities [159]. The reason behind generating the corpora with the top personalities, followed by millions of Twitter users, was to ensure that the tweets were taken from authentic profile, since Twitter contains many fake accounts and their tweets have no real meaning. Moreover, tweets from top hashtags were used to analyze the pattern and style of informal language used in the Twitter platform by the general users.

Model	Corpora Used	WordPiece Vocab
BERT	English Wiki + BookCorpus	BaseVocab
SciBERT	Scientific articles	SciVocab
TweetBERTv1	English Wiki + BookCorpus + Corpus140	BaseVocab
TweetBERTv2	English Wiki + BookCorpus + Corpus540	BaseVocab + SciVocab

Table 3.1: Shows the different variation of corpora and WordPiece vocabulary involved in BERT models

TweetBERT

TweetBERTs are continual pre-trained models since they were initialized with the weights of uncased BERT-base and AlBERT-base models. Uncased models does not differentiate between lower and upper case words. Since there are no significant differences between upper and lower case words in tweets uncase BERT models were selected. Futhermore, to

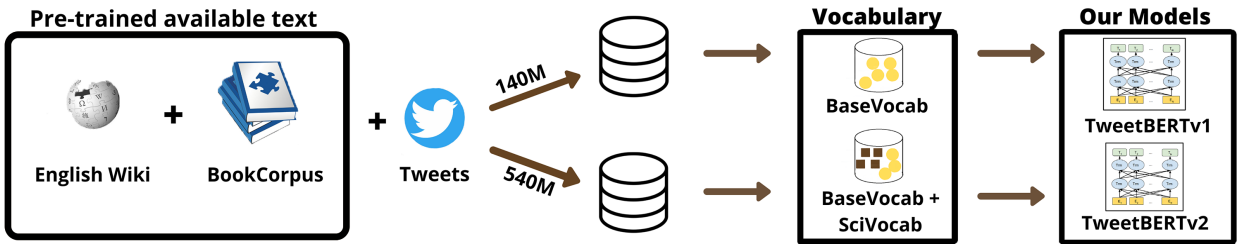

Figure 3.1: Overview of the pre-training TweetBERTs

reduce the risk of overfitting and complexity specifically for tweets, the weights of BERT-base and AlBERT-base was used since the base models have less parameters than the large models [36]. As a result, TweetBERTv1 has the same architecture as BERT. TweetBERTv1 is pre-trained on Corpus140. On the other hand, TweetBERTv2 is pre-trained on Corpus540 using both BaseVocab and SciVocab. Moreover, as TweetBERTv2 was initialized with the weight of AlBERT, it also has the same architecture like BERT, except that is uses two parameter-reduction techniques to reduce the number of training parameters in BERT, which increases the training speed of the model [139] [6]. BaseVocab and SciVocab are the WordPiece and SentencePiece vocabularies of BERT and SciBERT, respectively. TweetBERTs use the same type of vocabulary as BERT so that the initial pre-trained weights of BERT and ALBERT are compatible with TweetBERT models [84]. The vocabulary of SciBERT is used, in TweetBERTv2, so that scientific analysis can be carried out, for example detecting an epidemic or pandemic as opposed to simple sentiment analysis from tweets. TweetBERT can be also be used to evaluate other datasets in different domains, rather than just analyzing tweets.

Moreover, both TweetBERT models use Transformer networks, like BERT. A Transformer is a neural network model that is designed to work with sequential data. Each TweetBERT model contains 24 Transformers that consist of multiple attention layers [30]. An attention

layer contains set of matrices called an attention head, which work with the word tokens that is most relevant with the head. Then, the attention head provides an embedding for the token that contains information about the token itself and a weight relative to how relevant the token is in that context. For example, in the case of TweetBERT, input sequence of a tweet in vector is $t = [t_1, \ldots, t_n]$, where n represents the number of tokens in the input sequence. Each vector t_i is broken down by attention layers into key k, query q, and value v [171]. Hence, for t_i vector it is k_i, q_i and v_i [171]. The attention head then calculates the weight of the token using softmax dot product.

$$\alpha_{ij} = \frac{e^{(q_i^T k_j)}}{\sum_{l=1}^{n} e^{(q_i^T k_l)}} \tag{3.1}$$

where, α is the attention weights between all the query and key vectors of all the pairs of words [171]. The sum of the weights is the output of the attention head, which is denoted by [30]:

$$Output = \sum_{j=1}^{n} \alpha_{ij} v_j \tag{3.2}$$

Futhermore, TweetBERT transformer has a multi-head attention layer, which allows model to link and simultaneously work with words from different representations. Figure 3.1 gives a detailed overview of the approach in making TweetBERT models.

Datasets for fine-tuning

The authors evaluated 12 BERT models on 31 different datasets. The datasets are divided into four domains: general, biomedical, scientific and Twitter. Datasets that have been used for evaluation are discussed in detail in Chapter 2.5. In addition, Table 3.2 and Figure 3.2 provide a brief description and visualization of the datasets used for the evaluation.

Table 3.2: Some of the datasets selected for the evaluation, the table contains the name of the dataset, the task, the number of data point and and year of publication

Name	Task	No. of data points	Year of publication
Multi-Genre Natural Language Inference (MNLI)	Natural language inference	433k	2018
Quora Question Pairs (QQP)	Paraphrase	537k	2018
The Stanford Natural Language Inference (SNLI)	Classification	570k	2018
The Corpus of Linguistic Acceptability (CoLA)	Sentiment Analysis	106k	2018
Sea surface temperature (SST)	Sentiment Analysis	67k	2018
Microsoft Research Paraphrase Corpus (MRPC)	Semantic textual similarity	3.7k	2018
Semantic Textual Similarity Benchmark (STS)	Textual Similarity	7k	2018
Recognizing Textual Entailment (RTE)	Natural Language Inferences	2.5k	2018
Stanford Question Answering Dataset (SQuAD)	Question Answering	100k	2018

Continued on next page

Table 3.2 – *Continued from previous page*

Name	Task	No. of data points	Year of publication
A Large-Scale Adversarial Dataset for Grounded Commonsense Inference (SWAG)	Question Answering	113k	2018
Large-scale ReAding Comprehension Dataset From Examinations (RACE)	Natural Language Inferences	100k	2018
National Center for Biotechnology Information (NCBI)	Natural Language Inferences	705k	2016
BioCreative V CDR task corpus: a resource for chemical disease relation extraction (BC5CDR)	Natural Language Inferences	4.4k	2016
Gene-Disease Associations (GDA)	Sentiment Analysis	12k	2015
A semantic sentence similarity estimation system for the biomedical domain (BIOSSES)	Textual Similarity	100k	2017
A Natural Language Inference Dataset For The Clinical Domain (NLIC)	Sentiment Analysis	14k	2019

Continued on next page

Table 3.2 – *Continued from previous page*

Name	Task	No. of data points	Year of publication
Structural Scaffolds for Citation Intent Classification in Scientific Publications (SciCite)	Classification	11k	2019
A Challenge Dataset for Document-Level Information Extraction (SciREX)	Question Answering	10k	2020
Twitter US Airline Sentiment	Sentiment Analysis	14k	2015
Twitter User Gender Classification	Sentiment Analysis	20k	2020
Sentiment140	Sentiment Analysis	1.6M	2013

3.4 Results

In this section, the parameter settings and training details of pre-training and fine-tuning results on 31 distinct datasets are presented.

Experimental Setup

The total amount of parameters of TweetBERT is around 12M which is the same as the AlBERT-base model. The maximum sequence length is intialized to 256 for speeding up the

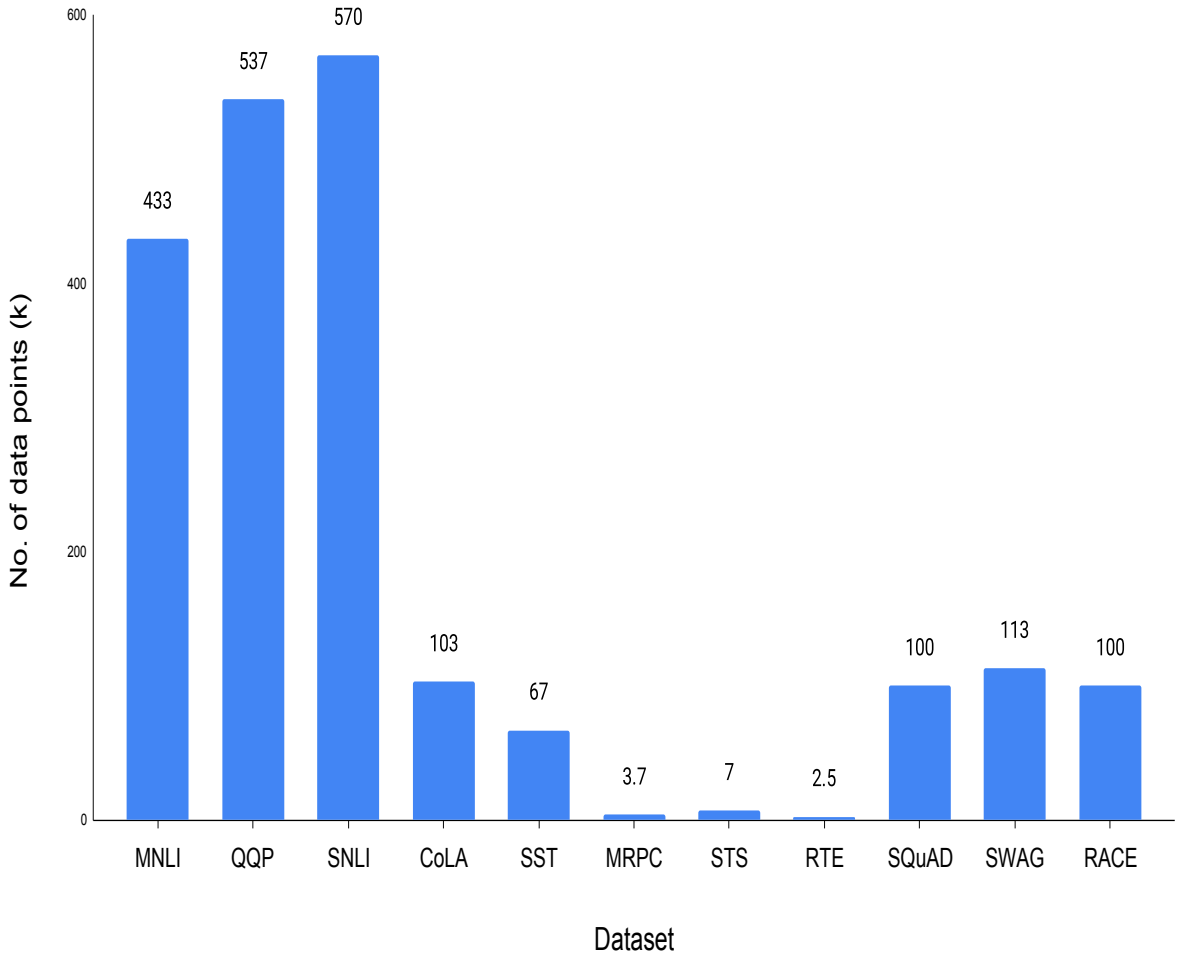

Figure 3.2: Shows the number of data points for the general domain datasets

training process [202]. To have the number of tokens of TweetBERT similar to the AlBERT-base model, the training batch size has been kept as 4,096. The pre-training hyperparameters of TweetBERT are mostly kept the same as the hyperparameters of the AlBERT model except for the batch size for evaluation, the learning rate, and the number of training epochs. In the TweetBERT model, it was observed that the following ranges work best for pretraining the model, i.e, the batch size for evaluation: 32, learning rate: $3e^{-5}$, maximum sequence

length: 256, and the number of epochs: 4. The configurations of TweetBERT are shown in Table 3.3.

To understand the relationship between each sentence of a tweet post, the model was pre-trained with the binarized next sentence prediction task which was inherited from the BERT architecture. In the pre-training phase, when sentences A and B of a tweet post was selected, 50% of the time B was given as the actual next sentence that is followed by A labeled as IsNext (since it is the next sentence after A) and in the remaining 50% of the time, a random sentence from the corpus was given followed by A and was labeled as NotNext.

Hyperparameters	Values
Drop out ratio for attention probability	0.1
Non-linear activation function	gelu
Hidden drop out probability	0.1
Embedding size	128
Intermediate size	3072
Maximum sequence length	256
Attention heads	12
Hidden layers	768
Hidden groups	1
Inner group number	1
Vocab size of token ids	2
Vocab size	33050
Batch size	32
Learning rate	$3e^{-5}$
Number of epochs	4

Table 3.3: Configurations for TweetBERT models

Experimental Results

The 31 datasets used can be divided into four different domains. The general domain, includes eight datasets from GLUE [173], SQuAD [143], SWAG [199] and RACE datasets. Table 3.4 and Table 3.5 shows the performance of the BERT models on the GLUE and

question answering datasets, respectively. It is observed that AlBERT [36] and RoBERTa [104] achieve a higher score than other BERT models. AlBERT performs better in almost all of the GLUE datasets whereas RoBERTa outperforms in general question answering datasets. TweetBERT models are also compared with other state-of-art BERT models: DeBERTa, StrucBERT, ERNIE, ConvBert, and SqueezeBERT. From the results it can be observed, only in the case of MRPC dataset of GLUE, DeBERTa has outperformed the rest of the BERT models. The results of TweetBERT are fairly or sometimes extremely close to that of the highest accuracy. For example, on CoLA dataset AlBERT and TweetBERT achieves an accuracy of 71.42% and 71% respectively. Moreover, to understand the improvement and effectiveness of each TweetBERT models the marginal performance on each dataset is calculated using equation 3.3 [122]. Table 3.6 and Table 3.7 shows the marginal performance between existing BERT models and TweetBERTv1, and Table 3.8 and Table 3.9 shows the mariginal performance of TweetBERTv2, on general domain datasets. Positive value represents by how much the TweetBERT outperformed a BERT model. For example, from Table 3.9 TweetBERT outperformed BioBERT by 12.81% in SQuAD dataset. On the other hand, negative value represents by how much an existing BERT model outperformed the TweetBERT model. To find the most suitable model overall on all the datasets the total of all the marginal performance of each BERT model was calculated. In the *Total* column positive and negative number indicates the value by which TweetBERT performs better or worst than that BERT model. Both Tables 3.6, 3.7, 3.8, and 3.9 show that overall RoBERTa performs the best.

$$\Delta MP = \frac{Accuracy_{BERTmodel} - Accuracy_{TweetBERTs}}{100 - Accuracy_{BERTmodel}} \times 100 \tag{3.3}$$

Secondly, the evaluation of the BERT models on 12 different biomedical domain datasets

Domain	General							
Datasets	GLUE							
	MNLI	QQP	QNLI	SST	CoLA	STS	MRPC	RTE
Metrics	A	A	A	A	A	PC	A	A
BERT	84.43	72.1	90.51	93.58	60.61	86.51	89.3	70.11
BioBERT	86.27	85.65	90.28	93.86	65.83	87.31	85.04	75.72
SciBERT	84.51	73.47	88.34	94.25	61.72	87.14	90.78	66.26
RoBERTa	90.28	92.21	94.72	96.4	68	92.41	90.9	86.65
AlBERT	90.83	92.25	95.37	96.99	71.42	96.94	90.9	89.21
TweetBERTv1	90.91	86.37	91.25	92.43	68.42	90.2	88.64	75.23
TweetBERTv2	90.51	88.83	91.21	94.38	71	94.41	91.79	91.3
DeBERTa	90.8	92.15	95	95.1	69.4	92.3	92.8	90.9
StrucBERT	85.2	88.4	91.8	94.1	57	87.9	89.5	76.6
Ernie	84.9	70.4	91.1	93.8	52.1	82.7	89.3	68.9
ConvBERT	88.1	89.8	93	95.5	66.9	95.8	88.1	77.3
SqueezeBERT	81.4	80.3	90	91.1	45.4	86.6	86.9	71.5

Table 3.4: Shows the performance of different BERT models on GLUE datasets. Highest accuracies are underlined

Domain	General		
Datasets	QA		
	SQuAD	SWAG	RACE
Metrics	A	A	A
BERT	81.66	86.23	69.23
BioBERT	72.22	82.71	80.9
SciBERT	84.69	84.44	78.58
RoBERTa	94.63	90.16	81.31
AlBERT	85.3	88.57	82.37
TweetBERTv1	69.84	85.47	81.96
TweetBERTv2	75.78	88.86	81.74

Table 3.5: Shows the performance of different BERT models on question answering datasets

is shown in Table 3.10. Precision (P), recall (R), and F1-score (F) are used as metrics for measuring performance. It shows that, although BioBERT was pre-trained on millions of biomedical corpus, RoBERTa and TweetBERT outperforms BioBERT in all dataset types

Domain	General								Total
Datasets	GLUE								
	MNLI	QQP	QNLI	SST	CoLA	STS	MRPC	RTE	
BERT	41.61	51.14	7.79	-17.91	19.82	27.35	-6.16	17.12	**140.76**
BioBERT	33.79	5.017	9.97	-23.28	7.57	22.77	24.06	-2.01	**77.887**
SciBERT	41.31	48.62	24.95	-31.65	17.50	23.79	-23.21	26.58	**127.89**
RoBERTa	6.48	-74.96	-65.71	-110.27	1.31	-29.11	-24.83	-85.54	**-382.63**
AlBERT	0.87	-75.87	-88.98	-151.49	-10.49	-220.26	-24.83	-129.56	**-700.61**
DeBERTa	1.19	-73.63	-75	-54.48	-3.20	-27.27	-57.77	-172.19	**-462.35**
StrucBERT	38.58	-17.5	-6.70	-28.30	26.55	19.00	-8.19	-5.85	**17.59**
Ernie	39.80	53.95	1.68	-22.09	34.07	43.35	-6.16	20.35	**164.95**
ConvBERT	23.61	-33.62	-25	-68.22	4.59	-133.33	4.53	-9.11	**-236.55**
SqueezeBERT	51.12	30.81	12.5	14.94	42.16	26.86	13.28	13.08	**204.75**

Table 3.6: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv1 on GLUE datasets

Domain	General			Total
Datasets	QA			
	SQuAD	SWAG	RACE	
BERT	-64.44	-5.51	33.74	**-36.21**
BioBERT	-8.57	15.96	3.66	**11.05**
SciBERT	-97.00	6.62	14.10	**-76.28**
RoBERTa	-461.64	-47.66	14.10	**-495.20**
AlBERT	-105.17	-27.12	-4.37	**-136.66**

Table 3.7: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv1 on question answering datasets

including NER and relation extraction. TweetBERTs performed best or very close to the best in many of the biomedical datasets. The the marginal performance of all the biomedical datasets between existing BERT models and TweetBERTs were calculated and reported in Table 3.11 and 3.12 respectively. Results in both the table indicates that TweetBERT outperforms BERT, BioBERT and SciBERT, on the other hand, RoBERTa and AlBERT performed better.

Domain	General								Total
Type	GLUE								
Datasets	MNLI	QQP	QNLI	SST	CoLA	STS	MRPC	RTE	
BERT	39.05	59.96	7.37	12.46	26.37	58.56	23.27	70.89	**297.93**
BioBERT	30.88	22.16	9.56	8.46	15.13	55.94	45.12	64.16	**251.41**
SciBERT	38.73	57.89	24.61	2.26	24.24	56.53	10.95	74.21	**289.42**
RoBERTa	2.36	-43.38	-66.47	-56.11	9.37	26.35	9.78	34.83	**-98.84**
AlBERT	-3.48	-44.12	-89.84	-86.71	-1.46	-82.67	9.78	19.36	**-279.14**
DeBERTa	-3.15	-42.29	-75.8	-14.69	5.22	27.40	-14.03	4.39	**-112.95**
StrucBERT	35.87	3.70	-7.20	4.75	32.55	53.80	21.80	62.82	**208.09**
Ernie	39.80	53.95	1.68	-22.09	34.07	43.35	-6.16	20.35	**164.95**
ConvBERT	23.61	-33.62	-25.00	-68.22	4.59	-133.33	4.54	-9.11	**-236.54**
SqueezeBERT	48.97	43.29	12.10	36.85	46.88	58.28	37.32	69.47	**353.16**

Table 3.8: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv2 on GLUE datasets

Domain				Total
Datasets	QA			
	SQuAD	SWAG	RACE	
BERT	-32.06	19.09	19.8	**6.83**
BioBERT	12.81	35.56	4.39	**52.76**
SciBERT	-58.19	45.1	14.75	**1.66**
RoBERTa	-351	-13.21	2.3	**-361.91**
AlBERT	-64.76	2.53	-3.57	**-65.80**

Table 3.9: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv2 on question answering datasets

Thirdly, BERT models on four scientific datasets were evaluated. Previously, with the introduction of SciBERT there was statistical evidence that it performed remarkably better on scientific tasks. Although, Table 3.13 show that TweetBERT performed best in only two datasets, Table 3.15 shows that TweetBERTv2 outperformed SciBERT and it is more suitable to use TweetBERTv2 to evaluate scientific tasks rather than using SciBERT. In the TweetBERTv2 model, the vocabulary was composed with both AlBERT and SciBERTs'

Domain	Type	Datasets	Metrics	BERT	BioBERT	SciBERT	RoBERTa	AlBERT	TweetBERTv1	TweetBERTv2
Biomedical	NER	NCBI	P	88.30	88.22	88.57	90.97	90.43	87.62	90.38
			R	89.00	91.25		91.15	91.22	91.33	91.62
			F	88.60	89.71		90.58	89.83	89.70	89.69
		BC5CDR	P	91.3	86.47	90.01	90.28	90.69	89.61	89.22
			R	80.1	87.84		89.12	89.03	86.09	88.86
			F	85.9	87.15		90.64	89.51	87.83	90.41
		Species	P	69.35	72.80	70.89	84.25	83.77	85.18	85.17
			R	74.05	75.36	75.82	87.16	85.90	87.45	88.31
			F	71.63	74.06	73.68	84.76	84.06	84.89	83.53
		BC5CDR	A	91.5	93	93.46	93.73	93.14	92.4	92.83
		JNLPBA	A	74.23	77.54	75.63	78.23	78.33	81.63	81.61
	RE	GAD	P	79.21	77.32	80.18	83.82	83.41	78.18	78.11
			R	89.25	82.68	88.51	90.14	89.73	91.81	91.92
			F	83.25	79.83	80.28	82.78	82.01	84.45	85.57
		EUADR	P	75.45	84.83	74.91	85.84	85.76	77.73	75.95
			R	96.55	90.81	96.64	89.5	90.48	92.31	92.1
			F	84.62	80.92	85.41	85.24	84.11	81.36	79.39
		CHEMPROT	P	76.02	77.02	71.3	80.17	85.32	86.10	85.77
			R	71.60	75.90		78.97	87.55	84.35	87.69
			F	73.74	76.46		79.32	83.29	85.63	85.04
	Sentence	MedSTS	A	78.6	84.5	78.6	89.06	91.06	86.78	90.89
		Biosses	A	71.2	82.7	74.23	88.77	91.25	80.27	83.96
	Inference	MedLNI	A	75.4	80.5	75.36	86.39	90.13	82.16	88.41
	Doc classif	HoC	A	80	82.9	80.12	87.83	91.48	82.71	86

Table 3.10: Shows the performance of different BERT models on biomedical domain dataset. Highest accuracies are underlined

vocabularies. Due to this reason, TweetBERTv2 has outperformed or performed fairly close to SciBERT's accuracy on scientific domain datasets. Moreover, it is advantageous to use the vocabulary of SciBERT when fine-tuning scientific domain datasets since Scivocab has been generated from scientific corpora. Furthermore, BioBERT is pre-trained on bio-medical corpora which is similar to the scientific domain. As a result, the TweetBERT, and SciBERT models have outperformed the BioBERT model on the Biomedical benchmark datasets.

Finally, all the BERT models were evaluated on tweets sentiment and classification datasets. As the TweetBERTs were pre-trained on millions of tweets it outperformed all existing BERT models, as expected. Table 3.16 records the performance of the BERT models including our TweetBERT. Table 3.17 shows that the highest total marginal performance

Domain	Type	Datasets	BERT	BioBERT	SciBERT	RoBERTa	AlBERT
Biomedical	NER	NCBI disease	9.64	0.91	9.88	-9.34	-1.27
		BC5CDR disease	13.68	-0.14	-21.82	-30.02	-16.01
		Species	46.7	49.06	42.59	0.85	5.2
		BC5CDR chemical	10.58	-8.57	-16.20	-21.21	-10.78
		JNLPBA	28.71	18.21	24.62	15.61	15.22
	RE	GAD	7.16	52.71	21.14	9.69	13.56
		EUADR	-21.19	16.32	-27.75	-26.54	-17.3
		CHEMPROT	45.27	38.95	49.93	30.51	14
	Sent sim	MedSTS	38.22	14.70	38.22	-20.84	-47.87
		Biosses	31.49	-14.047	23.43	-75.69	-125.48
	Inference	MedLNI	27.47	8.51	27.59	-31.08	-80.79
	Doc Classifi	HoC	13.55	-1.11	13.02	-42.07	-102.93
Total			**237.63**	**171.62**	**184.67**	**-200.12**	**-354.42**

Table 3.11: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv1 on different Biomedical datasets

Domain	Type	Datasets	BERT	BioBERT	SciBERT	RoBERTa	AlBERT
Biomedical	NER	NCBI disease	9.56	-0.19	9.79	-9.44	-1.37
		BC5CDR disease	31.9	25.36	4	-2.45	8.57
		Species	41.9	36.5	37.4	-8.07	-3.32
		BC5CDR chemical	15.64	-2.42	-9.63	-14.35	-4.51
		JNLPBA	28.63	18.12	24.53	15.52	15.13
	RE	GAD	13.85	28.45	29.67	16.2	19.78
		EUADR	-30.49	-5.18	-37.55	-35.98	-26.3
		CHEMPROT	43.03	36.44	47.87	27.65	10.47
	Sen sim	MedSTS	57.42	41.22	57.42	16.72	-1.90
		Biosses	44.30	7.28	37.75	-42.83	-83.31
	Inference	MedLNI	52.88	40.56	52.96	14.84	-17.42
	Doc Classifi	HoC	30	18.12	29.57	-15.03	-64.31
Total			**306.75**	**244.27**	**283.81**	**-37.21**	**-148.51**

Table 3.12: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv2 on different Biomedical datasets

is 159.13% when SciBERT and TweetBERTv1 are compared. Table 3.18, on the other hand, shows that the lowest marginal performance, 167.15%, is greater than the highest marginal performance from Table 3.17. As a result, it can concluded that TweetBERTv2 performs significantly better than TweetBERTv1 in Twitter domain tasks.

Domain	Scientific			
Datasets	Text Classification			parsing
	paper feild	sci-cite	scie-relation-extraction	genia
Metrics	A	A	A	A
BERT	55.06	84.33	63.55	64.81
BioBERT	56.22	85.11	65.42	67.71
SciBert	65.71	85.42	65.77	72.3
RoBerta	63.48	87.16	66.79	76.95
Albert	62.85	86.68	68.46	<u>78.45</u>
TweetBERTv1	58.12	88.5	<u>68.85</u>	67.98
TweetBERTv2	<u>66.49</u>	<u>88.56</u>	66.82	70

Table 3.13: Shows the performance of different BERT models on scientific domain dataset. Highest accuracies are underlined

Domain	Scientific				Total
Datasets	Text Classification			Parsing	
	paper feild	sci-cite	sci-RE	Genia	
BERT	6.8	26.61	14.54	9	**56.95**
BioBERT	4.33	22.76	9.91	0.83	**37.83**
SciBERT	-22.13	21.12	8.99	-15.59	**-7.61**
RoBERTa	-14.67	10.43	6.2	-38.91	**-36.95**
AlBERT	-12.73	13.66	1.23	-48.58	**-46.42**

Table 3.14: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv1 on different scientific datasets

3.5 Discussion and Conclusion

Twitter is a popular social networking site, which contain valuable data, where analyzing the content is particularly challenging. Tweets are usually written in an informal structure, and as a consequence, using language models trained on general domain corpora like BERT or other domains such as BioBERT often gives unsatisfactory results. Hence, two versions of TweetBERT are introduced, which are pre-trained language representation models used

Domain	Scientific				Total
Datasets	Text Classification			Parsing	
	paper feild	sci-cite	sci-RE	Genia	
BERT	25.43	26.99	8.97	14.74	**76.13**
BioBERT	23.45	23.16	4.04	7.09	**57.74**
SciBERT	2.27	21.53	3.06	-8.3	**18.56**
RoBERTa	8.24	10.9	0.09	-30.15	**-10.92**
AlBERT	9.79	14.11	-5.19	-39.21	**-20.50**

Table 3.15: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv2 on different scientific datasets

Domain	Twitter			
Datasets	Sentiment			
	Airline Sentiment	Gender Classification	Sentiment140	Political Tweets
Metrics	A	A	A	A
BERT	85.2	80.65	85.63	69.99
BioBERT	84.17	80.22	87.84	69.34
SciBERT	82.73	72.23	82.29	64.66
RoBERTa	88.68	80.74	86.71	72.01
AlBERT	87.08	82.22	90.59	69.57
TweetBERTv1	89	85.02	92.74	75.13
TweetBERTv2	92.99	89.75	95.18	78.79

Table 3.16: Shows the performance of BERT models in different Twitter datasets. Highest accuracies are underlined

for Twitter text mining. This chapter also discusses how the data was collected from the *big data analytics platform* for pre-training TweetBERT. Millions of tweets were extracted and cleaned from this platform. Moreover, detailed discussion of pre-training TweetBERT models are included. TweetBERTv1 was initialized using weights from BERT and then pre-trained on a tweet corpus. In the case of TweetBERTv2, first the model is initialized with weights from AlBERT and used vocabularies from both BERT and SciBERT. Two main advantages of using BaseVocab and SciVocab are scientific analysis can be carried out by

Domain	Twitter				Total
Datasets	Sentiment				
	Airline Sentiment	Gender Classification	Sentiment140	Political Tweets	
BERT	25.67	22.58	49.47	17.12	**114.84**
BioBERT	30.51	24.26	40.29	18.88	**113.94**
SciBERT	36.30	34.21	59.0	29.62	**159.13**
RoBERTa	2.82	22.22	45.37	11.14	**81.55**
AlBERT	14.86	15.74	22.84	18.27	**71.71**

Table 3.17: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv1 on different Twitter datasets

Domain	Twitter				Total
Datasets	Sentiment				
	Airline Sentiment	Gender Classification	Sentiment140	Political Tweets	
BERT	52.63	47.02	66.45	29.32	**195.42**
BioBERT	55.72	48.17	60.36	30.82	**195.07**
SciBERT	59.40	54.98	72.78	39.98	**227.14**
RoBERTa	38.07	46.78	63.73	24.22	**172.80**
AlBERT	45.74	42.35	48.77	30.29	**167.15**

Table 3.18: Shows the marginal percentage of existing BERT models in comparison to TweetBERTv2 on different Twitter datasets

studying tweets, and ALBERT is compatible with TweetBERTs and can be used in other evaluating other datasets in different domains rather than just analyzing tweets.

Moreover, this chapter focuses on the datasets used to evaluate BERT models. Evaluation of TweetBERT models and five other BERT models on 31 different datasets from general, biomedical, scientific and Twitter domains and provide a comparison between them. Chapter 2.5 gives a detail description of most of the datasets used. Finally, the results for the evaluation are released. It is shown that TweetBERT significantly outperforms other BERT models on Twitter datasets, and even on some other domain datasets, like BioBERT. The

marginal performance that shows the amount by which a BERT model outperforms another BERT model is calculate. It shows that, especially in the case of Twitter datasets, TweetBERTs has the best performance. TweetBERTv2 outperforms AlBERT by a total of 167.17% when evaluating Twitter datasets. Overall, an extensive discussion is provided about the necessity of language model specific to social media. We introduce TweetBERTs and give comprehensive discussion about the methods, approaches and data used to pre-train TweetBERTs.

Chapter 4

ONSET: Opinion and Aspect Extraction System from Unlabelled Data

All of this chapter is submitted as the following in a peer-reviewed conference:

- Qudar, M. M. A., Bhatia, P., & Mago, V. (2020). "ONSET: Opinion and Aspect Extraction System from Unlabelled Data"

This chapter is about extracting aspects and opinion from unlabelled data from an online platform. For this case only benchmark datasets from SemEval and Twitter were use to compare the results. The system develop during my thesis research will serve as a base architecture for extracting aspects from unlabelled data. We have submitted this section of thesis in a Conference

4.1 Introduction

Online platforms like Amazon, Yelp, and Booking actively extract aspects and opinions from user-generated information feedback and other online sources. These data extractions help gain insight into services, customers reviews, products and also in addressing questions from the customer. An overall opinion about a review or sentence can be extracted from a document-level or sentence-level sentiment analysis. However, more fine-grained information can be extracted from Aspect-Based Sentiment Analysis (ABSA) by mining aspects and examining aspect-level opinions for a discussed entity [100, 17]. For example, a user posts a review for a laptop: "I love the operating system but not the preloaded software" which contains two aspects, (a) a positive sentiment for the "operating system" and (b) a negative sentiment for the "preloaded software".

An ABSA can be divided into two sub-tasks of Aspect Mining (AM) and Aspect Sentiment Classfication (ASC) [100]. The AM sub-task extracts the aspect words from each sentence of reviews, which has been thoroughly investigated by applying unsupervised models [69, 203, 53], supervised models [63, 169, 96, 138, 177, 176, 97], or semi-supervised techniques [25, 26, 121, 94, 188, 93]. The ASC sub-task attempts to determine the sentiment polarities on aspects. These subtasks are performed using a supervised learning approach and required a large set of labelled reviews [180, 165, 166, 98, 99]. The results from these approaches achieve high accuracy. However, manually training a large dataset is very expensive, especially for domain-dependent aspects, i.e. different domains may have various aspect domains. As a result, researchers are encouraged to explore more efficient semi-supervised models for ABSA [60].

In recent years, Latent Dirichlet Allocation (LDA) [10] and its variants have become a major unsupervised approach for aspect extraction [168, 15, 121]. However, using pre-trained

language models, such as BERT [37] or XLNet [192], it is very easy to extract opinions. These language models can be fine tuned to attain high-quality extractions from the data. Fine-tuning such language models needs a huge number of high-quality labelled training data as they have a large number of parameters and training pre-trained language models on small datasets will cause overfitting [111]. Therefore, some systems obtain labelled training data through crowdsourcing [91]. Collecting data using the crowdsourcing technique requires additional tasks such as preparing the questionnaires, launching and managing the project and processing the results. These tasks are time-consuming, complicated and expensive. Additional steps are required to exclude responses from dishonest crowd workers to maintain the quality of data. Also, to eliminate potential mistakes, the labels for a sentence have to be obtained many times, and the results have to be cleaned until they are consumable for downstream tasks. Consequently, there has been a rising interest in collecting quality labelled data using less expensive and in a more effective way [152].

This chapter introduces ONSET, an architecture to reduce the labelled training data required for fine-tuning language models for AM and ASE. It is a novel system in which unsupervised learning, Data Augmentation (DA) and semi-supervised learning are performed to extract opinion and aspect from unlabelled data. The architecture uses Cross View Training (CVT), a semi-supervised learning algorithm. The CVT helps improve the representations of a Bi-LSTM sentence encoder using a mix of labelled and unlabelled data. CVT uses standard supervised learning for labelled examples. On unlabelled examples, CVT acts as both a teacher that makes predictions about the examples and a student that is trained on those predictions [29]. ONSET can mine three main types of information from reviews: aspects, opinions, and sentiments. An example of restaurant review with its aspect, opinion and sentiment is illustrated in Table 4.1.

From Table 4.1 the triplet (service, good, +1) consists of two spans of tokens extracted

Sentence	Aspect	Opinion	Sentiment
The service varies from good to mediocre	service	good	+1
depending upon waiter, seating is always prompt	seating	prompt	+1
though, and the restaurant gets busy in evening.	restaurant	busy	-1

Table 4.1: Shows an example of a restaurant review with its aspect, opinion and sentiment

from the review, where "good" is an opinion about the aspect "service". Positive sentiment is derived based on the sentence containing the aspect and opinion terms indicating a positive sentiment in this example (1 indicates positive, -1 is negative, and 0 is neutral). In this chapter, the authors made the following contributions:

- Introduction of ONSET, an architecture that eliminates the need of using huge amounts of labelled data which is very expensive and time consuming to collect and label.
- Augmentation of data to automatically generate more labelled training data from the existing data in which the aspects are labelled via topic modeling.
- Fine-tuning a language model by a semi-supervised approach for opinion extraction.
- Extensive experimentation carried out on large review datasets of Yelp, Amazon and Twitter; and the source code is made publicly available[1].

4.2 Related Works

Sequence labelling is a challenging task in natural language processing, and it is intended to assign a label to each input token in sequence. The aspect mining problem can be identified as a sequence labelling problem that requires a label sequence $(y_1 \ldots y_n)$ to be predicted

[1]https://github.com/mohiuddin02/ONSET

for a given word sequence ($x_1 \ldots x_n$). These labels can be simplified and written in (B, I, O) scheme, where B identifies the beginning of an aspect, I for the continuation of the aspect, and O for other words [67]. The (B, I, O) schemes can effectively deal with aspects expressing in phrases [96, 188] and aspect opinion mining term extraction [177, 176]. This section discusses the prior works related to the topic modeling using deep learning models, DA, and semi-supervised approaches.

Topic Modeling using Deep Learning Models

Many deep learning and pre-trained language models have been utilised to perform review and mining related tasks. For example a multi-task supervised model with two coupled Gated recurrent units (GRU) layers is proposed to co-extract aspects and sentiment words for aspect-based sentiment analysis [176]. On the other hand, uses a deep learning model with three Long Short-Term Memory (LSTM) layers to execute multi-task learning for AM achieving state of the art results [176, 96]. A two-step attention-based LSTM along with an interactive deep learning network (IMN) has been proposed to learn the model from the token-level AM and ASC [52, 107]. Topic modeling is gaining vast popularity in various text-mining communities. LDA has become the standard unsupervised approach for topic modeling in recent years [55]. Several extensions to LDAs have been proposed for social networks and social media. A novel probabilistic topic model was introduced using LDA to analyze text corpora and infer descriptions of the entities and relationships between those entities using Wikipedia [22]. Moreover, to apply LDA to tweets a TwitterRank system was developed using authors pooling [183]. To discover groups among the entities and topics among the corresponding text both simultaneously a scalable implementation of a semi-supervised learning model (labelled LDA) was developed [179, 144]. Futhermore, a

new model was introduced to incorporate LDA into a BERT community detection process [201]. LDA was also expanded to a managed form, and its implementation was studied in a micro-blogging context [145, 144].

Data Augmentation

Automatic DA techniques are used regularly in the computer vision and speech domain to enhance model's robustness to perform better by increasing the training data [33, 80]. DA techniques are mostly used when working with smaller datasets [156, 160, 76]. Due to the large dataset requirement, it is challenging to develop deep learning models with novel text augmentation methods with generalized rules for transforming language. As a result, fewer comprehensive research has been done that is devoted to novel text augmentation techniques.

Researches have been focused on synonym replacements by using predictive language models and data noising methods for smoothing augmented text [77, 187]. Augmenting sentences by replacing tokens with their respective synonyms shows efficient results for training sentence classifiers [182]. Even though these synonyms replacement techniques are entirely valid, they are often not used because it is computationally expensive to use these methods for text augmentation [196].

A DA technique, Easy Data Augmentation (EDA), was proposed that augments text by using four simple operations: synonym replacement (SYR), random insertion (INS), random deletion (DEL), and random swap operation (SPR) [182]. The EDA method is beneficial, especially for smaller datasets, since it uses these operators to increase the training dataset. The EDA technique has been shown to increase the performance of text classification tasks. However, EDA does not perform well when used with pre-trained models such as ULMFit, and ELMO, BERT [182]. Furthermore, EDA has also shown signs of overfitting due to

having a similar type of data and meaning semantic information also can get deleted when using the random deletion technique. Also, when random tokens are generated using the random insertion operator, the tokens can cause the text to have more noise. Another DA method, "MixDA" was proposed by that allows text to be partially altered so that the augmented data is not distorted [111]. It conducts a convex interpolation on the augmented data and original data, and the result is used as the training data [111]. The interpolation step between the actual existing example and the augmented example would yield reduced inconsistency.

Semi supervised approaches

Most current semi-supervised approaches use labelled data to guide an unsupervised topic model. Expectation Maximization (EM) uses both labelled and unlabelled data to determine generative classification parameters, such as naive Bayes, is a common technique for Semi-Supervised Learning (SSL) [124]. Another approach for semi-supervised learning is to use labelled reviews from the same domain to optimize the supervised model. For example, manually selecting seed words for the topic modeling [25, 26, 121, 94].

However, this approach requires manually defined domain knowledge and does not solely rely on labelled reviews. In the aspect mining model [188], the concept of pre-training was used to learn in advance domain-specific word embedding from unlabelled reviews.

Other researches have used external linguistic tools to obtain adequate word information. It may be considered as a special case in semi-monitored approaches for solving the sentiment classification problem [86, 107, 18] .

MixMatch is a semi-supervised learning paradigm proposed recently. It enhances the previous self-training method by using labelled and unlabelled data interpolations [9, 8]. A new

technique called MixMatchNL has been adapted from MixMatch technique [9]. MixMatch was used in the computer vision domain for training image classifiers [9]. MixMatch achieved higher accuracy in classifying images compared to earlier SSL algorithms with small number of labelled images. MixMatchNL uses huge amount of unlabelled data by guessing the labels and interpolation. For an unlabeled instance, MixMatchNL produces a "soft" guessed label. The guessed labelled is later used as training data.

A deep learning model typically works best when trained on a large set of data with appropriate labels. However, generating a large dataset of manual labels could be a considerable investment for domain-dependent aspects. One solution is to use semi-supervised learning to take advantage of unlabelled reviews. Current semi-supervised learning methods split the training process into two stages: pre-training and supervised learning [188]. A significant disadvantage of these methods is that the first stage of representation learning has no advantage from any labelled reviews. Another semi-supervised learning method is Cross View Training (CVT), which performs semi-supervised learning by rotating the training process with labelled and unlabelled data [29]. CVT algorithm improves the representations of a Bi-LSTM sentence encoder using a mix of labelled and unlabelled data. On labelled examples, standard supervised learning is used. On unlabelled examples, the model acts as both a teacher who makes predictions about the examples and a student trained on those predictions.

4.3 Proposed Model

This section explains the strategy of solving the aspect mining problem using unlabelled datasets. The proposed system targets removing the need to use large amounts of labelled data which is very costly and time-consuming to collect and label. However, there is a need

for ground truth when solving problems with unlabelled datasets. As a result, the proposed system uses labelled datasets, of which the labels have been removed and held as ground truth for comparing the predicted values. The dataset is divided into five-folds, where one fold has been labelled through an unsupervised approach using LDA and BERT models. The remaining four folds have been kept intact for implementing a semi-supervised approach at a later stage. Semi-supervised learning is mainly defined as an approach that requires a small amount of labelled and with a large amount of unlabelled data during training [8]. As Semi-supervised require a large dataset for training, therefore, to increase the training dataset, various DA techniques have been used. Figure. 4.1 shows the overview of the proposed model. First, an unsupervised approach is applied, followed by DA, then a semi-supervised approach using the CVT method to fine-tune a language model.

Unsupervised Learning

For performing aspect extraction, Attention- Based Aspect Extraction (ABAE) was utilised [53]. The ABAE learns a series of aspect embeddings by searching the nearest or representative words in the embedding space. This learning involves four steps. First, identifying a neural word that co-exist within a similar context nearby its embedding space [115]. Second, the word is filtered from the sentence using the attention mechanism [4]. Third, the filtered words are used to create aspect embeddings. Fourth, the common factors are extracted from the embedded sentences using dimension reduction.

MixDA

DA is a way to automatically increase the size of the training data without using human experts. DA has been found helpful in tasks related to computer vision and NLP tasks. DA

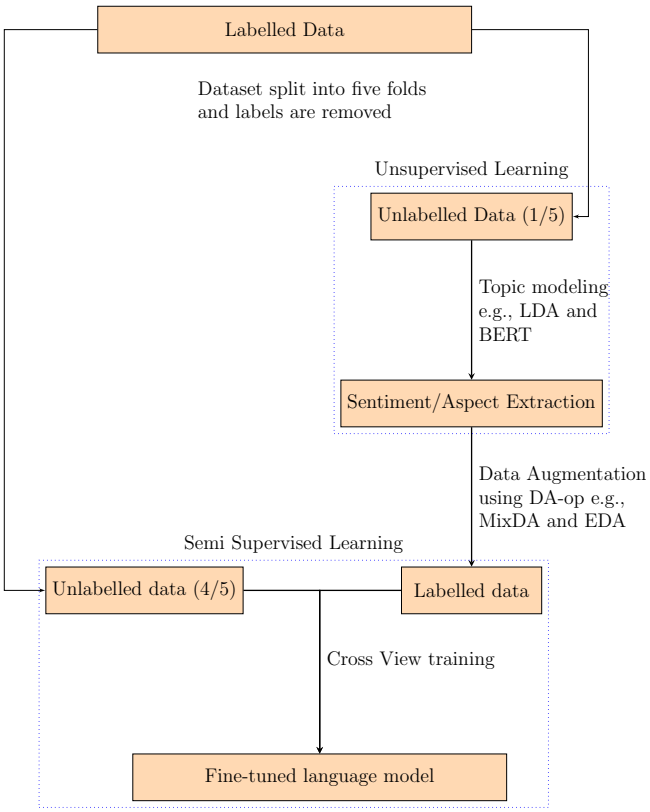

Figure 4.1: Overall architecture of the proposed model

has operators useful in NLP related tasks, such as token substitution with a synonym, token addition, token removal and token swap. A DA trained model can help in identifying the unchanged properties in the data. However, DA has limitations when used for NLP tasks, as DA operators may hamper the semantics of the generated sentence. To counter these issues, MixDA is used for augmenting data [111]. In MixDA, x is considered as a text sequence and y as the one-hot label vector, MixDA trains the model by first applying a data augmentation operator to obtain (x_{aug}, y_{aug}). MixDA then performs interpolation on the original input pair (x, y) and the augmented pair (x_{aug}, y_{aug}) to get $(BERT(x'), y')$. $Bert(x')$ represents the encoding of the sequence lying between the actual and augmented sequence x and x_{aug}

respectively. Later, the resulting $BERT(x')$ is fed to the remaining layers, calculating the loss over y' and back-propagation to minimize the loss [111].

Cross View Training

CVT is a semi-supervised learning algorithm that enhances the representations of a Bi-LSTM sentence encoder using labelled and unlabelled data [29]. The core concept of CVT is to use labelled and unlabelled reviews from the same domain. CVT helps in restoring the model's representation learning by using auxiliary prediction modules from the primary model's predictions since the primary model in CVT has a more robust and complete view of the input [95].

4.4 Experimental Results

Datasets

SemEval ABSA datasets from three domains (restaurant, beer and laptop) were considered for validating the opinion mining tasks. The SemEval datasets include laptop reviews from Amazon Review [54] and restaurant reviews from Yelp Review Dataset [3]. Moreover, two Twitter datasets were also used to extract opinions from Tweets [88]. They are Stanford Twitter Sentiment (STS) and Sanders Twitter Corpus (STC). STS contains 1.6 million tweets with equal number of positive and negative tweets. STC, however, only contains 5K manually classified tweets [88]. Table 4.2 shows the properties of the SemEval ABSA datasets and Figure 4.2 provides a visualization of the datasets. All the datasets contain annotated aspects, but in the proposed model, the annotations were removed, and the model was trained using an unsupervised approach.

SubTask	Sentence	Aspect
Aspect term extraction	I liked the service and the staff, but not the food	Service, Staff and food
Aspect term polarity	The fajitas are their first plate	neutral
Aspect category detection	The restaurant was too expensive	price
Aspect category polarity	The restaurant was too expensive	negative/-1

Table 4.2: Shows the task description for ABSA [137]

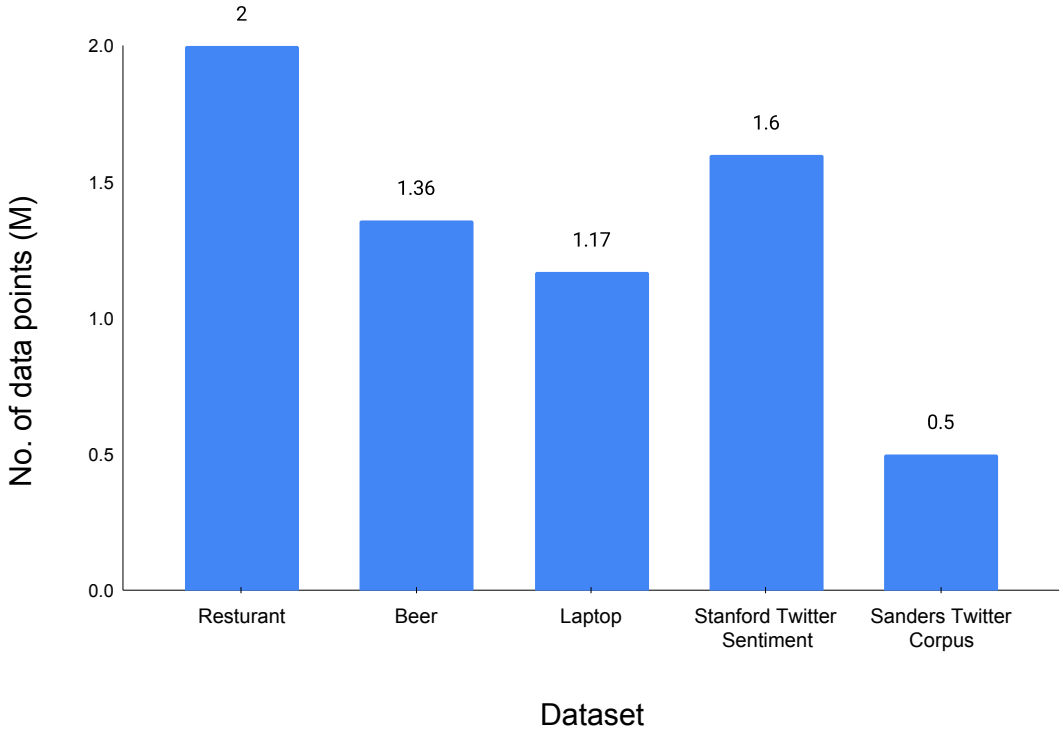

Figure 4.2: Shows the number of data points of the SemEval ABSA datasets used for opinion mining tasks

Training details

The dataset was trained with 20 epochs at a learning rate of $5e^{-5}$ and batch size of 32 for unsupervised learning. For data augmentation, the MixDA technique replaced 5% of words with synonyms and deleted 10% of words. The MixDA technique also generated data by

inserting 5% of the words and swapping 5% of words. For training on CVT a learning rate of 0.005 was set with batch size of 64 having a maximum of 200 words per sentence. The hyperparameters used are listed in detail in Table 4.3.

Topic Modeling	Data Augmentation	Cross View Training
No. of epochs: 20	No. of epochs: 5	No. of epochs: 50
Learning rate: $5e^{-5}$	Learning rate: $5e^{-5}$	Learning rate: 0.005
Batch size: 32	Batch size: 64	Batch size: 64
Embeddings dimension: 200	Mixup parameter: 0.2	Maximum sentence length: 200
Vocab size: 9000	SYR: 5%	Maximum word length: 20
Optimizer: adamax	INS: 5%	Dropout probability: 0.5
Regularizaiton: 0.1	DEL: 10%	
	SPR: 5%	

Table 4.3: Lists all the hyperparameters.

Results

Each dataset was split into five folds, and topic modeling was applied to one-fifth of the dataset to mine each sentence's aspect. Table 4.4 shows the F1-scores obtained from LDA and BERT models. The results show that the BERT model outperformed LDA model in predicting the appropriate topic for each sentence. The BERT is a transformer-based model which is already pre-trained with a large corpus of Wikipedia (2,500M words) and a book corpus (800M words) [37]. Based on the results, it can be observed that the BERT model tries to learn high-level features from the textual data. Furthermore, during the training phase, BERT learns the feature representations bidirectionally making its memory much stronger than LDA, whereas LDA was unable to capture correlations between the topic words of each of the sentences.

After extracting the aspects using the topic modeling approach, the DA techniques were

Model	Restaurant	Beers	Laptops	Stanford Twitter Sentiment	Sanders Twitter Corpus
LDA	75.7	72.12	74.0	79.59	77.79
BERT	**81.57**	**74.25**	**78.85**	**81.46**	**83.29**

Table 4.4: F1-scores of topic modeling using LDA and BERT models

used to increase training data. The two techniques used for data augmentation were EDA and MixDA. DA methods were used in such a way that it produces only augmented instance for each original instance. The number of folds increased from five to six. Now one-third of the total dataset has aspects extracted and two-third is unlabelled. These techniques helped the model to increase performance on the benchmark datasets, shown in Table 4.5. The increase in performance is because the input was more generalized and similar data due to data augmentation. Table 4.5 shows the LDA and BERT model's result with EDA and MixDA. The results show that the model obtained better results with MixDA augmented data than with EDA. The reason is that the EDA only performs SYR, INS, DEL, SPR. However, during the data augmentation phase, the target aspects could get swapped or deleted. These operations would change the overall meaning of a sentence and lead the model to perform poorly. For example: In the sentence, "Everybody was very nice. (+1)", if the DA operators replace "nice" with a negative/neutral word (e.g., "poor","okay") then the statement label would no longer be +1. Similarly, If DEL drops "nice", INS insert "sometimes" after "was", or SPR replaces "Everybody" with "Nobody" then the sentiment label will be wrong. Compared to the EDA approach, the MixDA also performs interpolating on the data with MixUp interpolation. As a result, the encoding of a sequence is within the augmented sequence x_{aug} in the original sequence x. Moreover, the MixDA uses a backpropagation technique, which adds the interpolated encoding sequence to the remaining layers to calculate the predicted aspect's loss and updates the model to reduce the loss.

The DA operators can change the length of the sequence if the tokens get deleted with the Random deletion operator, thus MixDA aligns the label of the sequence y_{aug} with the original y in case the target aspect gets deleted.

Model	Restaurant	Beers	Laptops	Stanford Twitter Sentiment	Sanders Twitter Corpus
LDA+EDA	83.69	77.02	80.68	83.27	80.43
BERT+EDA	84.06	81.25	85.39	88.19	84.72
LDA+MixDA	83.72	78.51	86.11	86.82	83.35
BERT+MixDA	**89.93**	**84.65**	**90.20**	**90.77**	**88.55**

Table 4.5: F1-scores of data augmentation using EDA and MixDA with LDA and BERT models

After performing MixDA, the dataset includes more similar instances, which may cause data overfitting. The semi-supervised approaches are used to reduce the overfitting of data by generalizing and distributing data. After using semi-supervised approaches the overall F1-score have decreased illustrated in Table 4.6 and Table 4.7. For training the proposed model, first data augmentation was applied followed by two semi-supervised approaches: MixMatchNL and CVT. The overall F1-score decreased for all the datasets when the data augmentation was used with the BERT and LDA methods. Table 4.6, shows that the MixMatchNL approach with the BERT model performed better over the MixMatchNL approaches with the LDA method. Also, when the CVT semi-supervised approach was implemented with the BERT, the model performed significantly better than the LDA method's approaches, shown in Table 4.7.

Comparing Table 4.6 and Table 4.7, it can be concluded that the CVT method has performed better than the MixMatchNL method. The CVT model's performance is because the CVT trains the auxiliary modules to observe partial sentences to match the model predictions. In other words, the auxiliary prediction modules are implemented on unlabelled data with different types of views of the input. Training is done by masking the input partially

Model	Restaurant	Beers	Laptops	Stanford Twitter Sentiment	Sanders Twitter Corpus
LDA+EDA+MixMatchNL	79.37	72.16	78.5	80.53	79.00
BERT+EDA+MixMatchNL	81.63	78.27	84.11	82.86	80.75
LDA+MixDA+MixMatchNL	82.87	77.24	80.49	81.41	83.89
BERT+MixDA+MixMatchNL	**84.52**	**82.96**	**85.76**	**86.60**	**83.12**

Table 4.6: F1-scores of semi-supervised approach: MixMatchNL with a combination of EDA, MixDA, LDA, and BERT models

and is trained with the primary prediction module. By this type of training, the auxiliary modules enhance the contextual representations produced by the model. Moreover, each of the auxiliary models is composed of two layers of CNN-BiLSTM sentence encoder. The CNN-BiLSTM process inputs in two directions enabling model's memory to be very strong by learning the representations of a sentence bidirectionally. Whereas the MixMatchNL method generates a "soft" label for each unlabelled sentence estimated by the model. However, a major problem with the MixMatch method is that the labels generated can be noisy depending on proposed model's quality. This issue can be partially resolved using the interpolated label rather than using the "soft" label to reduce noise. However, the noise still exists which can reduce the overall performance of the model [111].

Based on the above observations, it can be concluded that the aspect mining task is best performed with the CVT semi-supervised approach combined with MixDA and BERT methods to extract the linguistically rich semantic information of the input sentences.

Model	Restaurant	Beers	Laptops	Stanford Twitter Sentiment	Sanders Twitter Corpus
LDA+EDA+CVT	80.62	73.38	80.93	82.73	82.60
BERT+EDA+CVT	84.06	80.76	86.69	85.83	86.28
LDA+MixDA+CVT	81.37	80.18	82.17	86.52	84.43
BERT+MixDA+CVT	**88.14**	**85.00**	**87.30**	**88.14**	**88.35**

Table 4.7: F1-scores of semi-supervised approach: CVT with a combination of EDA, MixDA, LDA, and BERT models

4.5 Conclusion

This chapter proposes a novel opinion mining system that extracts aspects and opinions from the text. In the proposed model, an unsupervised approach is used to extract the aspects from the text using topic modeling methods. Afterwards, different DA augmentation techniques are used to generated training data. The augmented data generated helps in improving the performance of the proposed model. However, using the data augmentation technique caused the model to overfit due to having more alike instances. To reduce the overfitting problem, the semi-supervised method was used to generalize the distribution of data. The CVT semi-supervised method helped in further increasing the performance of the proposed model. The proposed model can achieve state-of-art results in multiple opinion mining tasks with a very small amount of training data comparatively. In future, the model can be further optimized for multitask learning and to reduce the requirement of labelled data.

Chapter 5

Conclusion

In this thesis, at first, a survey is presented in Chapters 2 focusing on the state-of-art language models, specifically on the transformer based models because of their significant contributions in the field of NLP. This survey has enabled us to pinpoint the research gap and drawbacks of language models when conducting text mining analysis tasks for the social media domain. It served as a point of reference for researchers to gain an understanding of the recent developments and breakthroughs in the field of language models. Expanding on what has already been developed, in Chapter 3, we introduced two TweetBERT models that have been pre-trained on millions of tweets and are domain specific language presentation models. These two models were evaluated and compared with a number of BERT models on numerous datasets. The results have ensured that TweetBERT models have performed significantly better than the traditional transformer based BERT models when performing Twitter text mining tasks. The outcomes of this research have also demonstrated that continuously training language models over time improves the performance of these models on Twitter datasets.

Later in Chapter 4, we have proposed a unique opinion mining system from unlabelled

data, ONSET. The primary aim of ONSET is to resolve the need for vast volumes of high-quality labeled data to fine-tune state-of-the-art pre-trained language models. The model is designed with a language model via an unsupervised approach in which the labels for each of the texts are extracted by topic modeling methods. Finally, the model is improved by using various data augmentation techniques to increase the size of training data so that the model can perform more efficiently.

The work presented in this thesis was to encourage more extensive work that is highly needed to development models for text analysis for online platforms. Users on online platforms prefer to write texts in an informal manner without following any grammatical rules. The text is written in a radically unstructured manner than conventional writings mainly due to a limit in the length of the post. As a result, it extremely challenging for traditional language models to conduct text mining tasks on such texts that are hardly grammatically correct and highly unstructured. The concerns raised in each chapter can impact various NLP tasks, so finding solutions to mitigate or fix those issues is extremely important.

Bibliography

[1] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *NAACL HLT 2019* (2019), p. 72.

[2] Waleed Ammar et al. "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)* (2018), pp. 84–91.

[3] Nabiha Asghar. *Yelp Dataset Challenge: Review Rating Prediction*. 2016. arXiv: 1605.05362 [cs.CL].

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[5] Yonatan Belinkov and James Glass. "Analysis methods in neural language processing: A survey". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 49–72.

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". 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)* (2019), pp. 3606–3611.

[7] Yoshua Bengio et al. "A neural probabilistic language model". In: *Journal of machine learning research* 3.Feb (2003), pp. 1137–1155.

[8] David Berthelot et al. "Mixmatch: A holistic approach to semi-supervised learning". In: *arXiv preprint arXiv:1905.02249* (2019).

[9] David Berthelot et al. "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring". In: *arXiv preprint arXiv:1911.09785* (2019).

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *The Journal of machine Learning research* 3 (2003), pp. 993–1022.

[11] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in neural information processing systems* (2016), pp. 4349–4357.

[12] Antoine Bordes et al. "Translating embeddings for modeling multi-relational data". In: *Neural Information Processing Systems (NIPS)* (2013), pp. 1–9.

[13] Samuel R Bowman et al. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015), pp. 632–642.

[14] Samuel R Bowman et al. "A large annotated corpus for learning natural language inference". In: (2015), pp. 632–642.

[15] Samuel Brody and Noemie Elhadad. "An unsupervised aspect-sentiment model for online reviews". In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. 2010, pp. 804–812.

[16] Zhai C. and Lafferty J. "A study of smoothing methods for language models applied to ad hoc information retrieval". In: *ACM SIGIR* (2017).

[17] E. Cambria. "Affective Computing and Sentiment Analysis". In: vol. 31. 2. 2016, pp. 102–107. DOI: 10.1109/MIS.2016.31.

[18] Erik Cambria et al. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives". In: *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*. 2016, pp. 2666–2677.

[19] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter". In: *Proceedings of the 20th international conference on World wide web* (2011), pp. 675–684.

[20] Daniel Cer et al. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (Aug. 2017), pp. 1–14. DOI: 10.18653/v1/S17-2001. URL: https://www.aclweb.org/anthology/S17-2001.

[21] Dhivya Chandrasekaran and Vijay Mago. "Evolution of Semantic Similarity—A Survey". In: *ACM Computing Surveys* 54.2 (Feb. 2021). ISSN: 0360-0300. DOI: 10.1145/3440755. URL: https://doi.org/10.1145/3440755.

[22] Jonathan Chang, Jordan Boyd-Graber, and David M Blei. "Connections between the lines: augmenting social networks with text". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 169–178.

[23] Qingyu Chen, Yifan Peng, and Zhiyong Lu. "BioSentVec: creating sentence embeddings for biomedical texts". In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)* (2019), pp. 1–5.

[24] Xinxiong Chen et al. "Joint learning of character and word embeddings". In: *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015).

[25] Zhiyuan Chen and Bing Liu. "Topic modeling using topics from many domains, lifelong learning and big data". In: *International conference on machine learning*. PMLR. 2014, pp. 703–711.

[26] Zhiyuan Chen et al. "Exploiting domain knowledge in aspect extraction". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1655–1667.

[27] Chung-Cheng Chiu et al. "State-of-the-art speech recognition with sequence-to-sequence models". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 4774–4778.

[28] Jason PC Chiu and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 357–370.

[29] Kevin Clark et al. "Semi-supervised sequence modeling with cross-view training". In: *arXiv preprint arXiv:1809.08370* (2018).

[30] Kevin Clark et al. *What Does BERT Look At? An Analysis of BERT's Attention*. 2019. arXiv: 1906.04341 [cs.CL].

[31] Arman Cohan et al. *Structural Scaffolds for Citation Intent Classification in Scientific Publications*. 2019. arXiv: 1904.01608 [cs.CL].

[32] Alexis Conneau et al. "Supervised learning of universal sentence representations from natural language inference data". In: *arXiv preprint arXiv:1705.02364* (2017).

[33] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. "Data augmentation for deep neural network acoustic modeling". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.9 (2015), pp. 1469–1477.

[34] Andrew M Dai and Quoc V Le. "Semi-supervised sequence learning". In: *Advances in neural information processing systems*. 2015, pp. 3079–3087.

[35] Zihang Dai et al. "Transformer-xl: Attentive language models beyond a fixed-length context". In: *arXiv preprint arXiv:1901.02860* (2019).

[36] Jacob Devlin et al. "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)* (June 2019), pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[37] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[38] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. "NCBI disease corpus: a resource for disease name recognition and concept normalization". In: *Journal of biomedical informatics* 47 (2014), pp. 1–10.

[39] Rezarta Islamaj Doğan, Aurélie Névéol, and Zhiyong Lu. "A context-blocks model for identifying clinical relationships in patient records". In: *BMC bioinformatics* 12.S3 (2011), S3.

[40] Timothy Dozat and Christopher D Manning. "Deep biaffine attention for neural dependency parsing". In: *arXiv preprint arXiv:1611.01734* (2016).

[41] E.and Matias Y. Finkelstein L.and Gabrilovich et al. "Placing search in context: The concept revisited." In: *Proceedings of the 10th international conference on World Wide Web* (2015).

[42] Yarin Gal and Zoubin Ghahramani. "A theoretically grounded application of dropout in recurrent neural networks". In: *Advances in neural information processing systems* (2016), pp. 1019–1027.

[43] Matt Gardner et al. "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)* (2018), pp. 1–6.

[44] Maria Giatsoglou et al. "Sentiment analysis leveraging emotions and word embeddings". In: *Expert Systems with Applications* 69 (2017), pp. 214–224.

[45] John M Giorgi and Gary D Bader. "Transfer learning for biomedical named entity recognition with neural networks". In: *Bioinformatics* 34.23 (2018), pp. 4087–4094.

[46] Alec Go, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision". In: *CS224N project report, Stanford* 1.12 (2009), p. 2009.

[47] Aaron Gokaslan and Vanya Cohen. *OpenWebText Corpus*. http://Skylion007.github.io/OpenWebTextCorpus. 2019.

[48] Jose Angel Gonzalez, Lluıs-F Hurtado, and Ferran Pla. "TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter". In: *Neurocomputing* 426 (2021), pp. 58–69.

[49] Yu Gu et al. *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. 2021. arXiv: 2007.15779 [cs.CL].

[50] Maryam Habibi et al. "Deep learning with word embeddings improves biomedical named entity recognition". In: *Bioinformatics* 33.14 (2017), pp. i37–i48.

[51] Pengcheng He et al. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. 2020. arXiv: 2006.03654 [cs.CL].

[52] Ruidan He et al. "An interactive multi-task learning network for end-to-end aspect-based sentiment analysis". In: *arXiv preprint arXiv:1906.06906* (2019).

[53] Ruidan He et al. "An unsupervised neural attention model for aspect extraction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 388–397.

[54] Ruining He and Julian McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering". In: *proceedings of the 25th international conference on world wide web*. 2016, pp. 507–517.

[55] Liangjie Hong and Brian D Davison. "Empirical study of topic modeling in twitter". In: *Proceedings of the first workshop on social media analytics*. 2010, pp. 80–88.

[56] Jeremy Howard and Sebastian Ruder. "Universal language model fine-tuning for text classification". In: *arXiv preprint arXiv:1801.06146* (2018).

[57] He Hua and Jimmy Lin. "Pairwise word interaction modeling with deep neural networks for semantic similarity measurement." In: *Association for Computational Linguistics* (2016).

[58] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission". In: *arXiv preprint arXiv:1904.05342* (2019).

[59] Ming-Siang Huang et al. "Biomedical named entity recognition and linking datasets: survey and our recent development". In: *Briefings in Bioinformatics* 21.6 (June 2020), pp. 2219–2238. ISSN: 1477-4054. DOI: 10.1093/bib/bbaa054. URL: http://dx.doi.org/10.1093/bib/bbaa054.

[60] Amir Hussain and Erik Cambria. "Semi-supervised learning for big social data analysis". In: *Neurocomputing* 275 (2018), pp. 1662–1673.

[61] Forrest Iandola et al. "SqueezeBERT: What can computer vision teach NLP about efficient neural networks?" In: *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing* (Nov. 2020), pp. 124–135. DOI: 10.18653/v1/2020.sustainlp-1.17. URL: https://www.aclweb.org/anthology/2020.sustainlp-1.17.

[62] Piotr Indyk and Rajeev Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality". In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing* (1998), pp. 604–613.

[63] Niklas Jakob and Iryna Gurevych. "Extracting opinion targets in a single and cross-domain setting with conditional random fields". In: *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010, pp. 1035–1045.

[64] Deng Jia et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition* (2009), pp. 248–255.

[65] Zi-Hang Jiang et al. "ConvBERT: Improving BERT with Span-based Dynamic Convolution". In: *Advances in Neural Information Processing Systems* 33 (2020). Ed. by

H. Larochelle et al., pp. 12837–12848. URL: https://proceedings.neurips.cc/paper/2020/file/96da2f590cd7246bbde0051047b0d6f7-Paper.pdf.

[66] Kyong Hwan Jin et al. "Deep convolutional neural network for inverse problems in imaging". In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4509–4522.

[67] Wei Jin, Hung Hay Ho, and Rohini K Srihari. "A novel lexicalized HMM-based learning framework for web opinion mining". In: *Proceedings of the 26th annual international conference on machine learning*. Vol. 10. 1553374.1553435. Citeseer. 2009.

[68] Xin Jin et al. "GAD: general activity detection for fast clustering on large data". In: *Proceedings of the 2009 SIAM international conference on data mining* (2009), pp. 2–13.

[69] Yohan Jo and Alice H Oh. "Aspect and sentiment unification model for online review analysis". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, pp. 815–824.

[70] Camacho-Collados Jose and Mohammad Taher Pilehvar. "From word to sense embeddings: A survey on vector representations of meaning." In: *Journal of Artificial Intelligence Research* (2018).

[71] Mandar Joshi et al. "Spanbert: Improving pre-training by representing and predicting spans". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 64–77.

[72] Mandar Joshi et al. "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2017), pp. 1601–1611.

[73] Rafal Jozefowicz et al. "Exploring the limits of language modeling". In: *Advances in Neural Information Processing Systems* (2016).

[74] Tom Kenter and Maarten De Rijke. "Short text similarity with word embeddings". In: *Proceedings of the 24th ACM international on conference on information and knowledge management* (2015), pp. 1411–1420.

[75] J-D Kim et al. "GENIA corpus—a semantically annotated corpus for bio-textmining". In: *Bioinformatics* 19.suppl_1 (2003), pp. i180–i182.

[76] Tom Ko et al. "Audio augmentation for speech recognition". In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[77] Sosuke Kobayashi. "Contextual augmentation: Data augmentation by words with paradigmatic relations". In: *arXiv preprint arXiv:1805.06201* (2018).

[78] Flip Korn, Hosagrahar V Jagadish, and Christos Faloutsos. "Efficiently supporting ad hoc queries in large datasets of time sequences". In: *Acm Sigmod Record* 26.2 (1997), pp. 289–300.

[79] Jens Kringelum et al. "ChemProt-3.0: a global chemical biology diseases mapping". In: *Database* 2016 (2016).

[80] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[81] Matt Kusner et al. "From word embeddings to document distances". In: *International conference on machine learning* (2015), pp. 957–966.

[82] Guokun Lai et al. "RACE: Large-scale ReAding Comprehension Dataset From Examinations". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017), pp. 785–794.

[83] Guillaume Lample et al. "Unsupervised machine translation using monolingual corpora only". In: *arXiv preprint arXiv:1711.00043* (2017).

[84] Zhenzhong Lan et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations* (2020). URL: https://openreview.net/forum?id=H1eA7AEtvS.

[85] Juan J. Lastra-Díaz. "A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art." In: *Engineering Applications of Artificial Intelligence* (2019).

[86] Raymond YK Lau, Chunping Li, and Stephen SY Liao. "Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis". In: *Decision Support Systems* 65 (2014), pp. 80–94.

[87] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[88] H. H. Lek and D. C. C. Poo. "Aspect-Based Twitter Sentiment Classification". In: *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*. 2013, pp. 366–373. DOI: 10.1109/ICTAI.2013.62.

[89] Omer Levy and Yoav Goldberg. "Dependency-based word embeddings". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2014), pp. 302–308.

[90] Omer Levy, Yoav Goldberg, and Ido Dagan. "Improving distributional similarity with lessons learned from word embeddings". In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 211–225.

[91] Guoliang Li et al. "Crowdsourced data management: A survey". In: *IEEE Transactions on Knowledge and Data Engineering* 28.9 (2016), pp. 2296–2319.

[92] Jiao Li et al. "BioCreative V CDR task corpus: a resource for chemical disease relation extraction". In: *Database* 2016 (2016).

[93] Ning Li, Chi-Yin Chow, and Jia-Dong Zhang. "EMOVA: A semi-supervised end-to-end moving-window attentive framework for aspect mining". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2020, pp. 811–823.

[94] Ning Li, Chi-Yin Chow, and Jia-Dong Zhang. "Seeded-BTM: enabling biterm topic model with seeds for product aspect mining". In: *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE. 2019, pp. 2751–2758.

[95] Ning Li, Chi-Yin Chow, and Jia-Dong Zhang. "SEML: A semi-supervised multi-task learning framework for aspect-based sentiment analysis". In: *IEEE Access* 8 (2020), pp. 189287–189297.

[96] Xin Li and Wai Lam. "Deep multi-task learning for aspect term extraction with memory interaction". In: *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017, pp. 2886–2892.

[97] Xin Li et al. "Aspect term extraction with history attention and selective transformation". In: *arXiv preprint arXiv:1805.00760* (2018).

[98] Xin Li et al. "Transformation networks for target-oriented sentiment classification". In: *arXiv preprint arXiv:1805.01086* (2018).

[99] Zheng Li et al. "Exploiting coarse-to-fine task transfer for aspect-level sentiment classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 4253–4260.

[100] Bing Liu. "Sentiment analysis and opinion mining". In: *Synthesis lectures on human language technologies* 5.1 (2012), pp. 1–167.

[101] Liyuan Liu et al. "Empower sequence labeling with task-aware neural language model". In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[102] Weiyang Liu et al. "Learning towards minimum hyperspherical energy". In: *Advances in neural information processing systems* (2018), pp. 6222–6233.

[103] Yang Liu et al. "Topical word embeddings". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).

[104] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[105] Yi Luan et al. "Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3219–3232. DOI: 10.18653/v1/D18-1360. URL: https://www.aclweb.org/anthology/D18-1360.

[106] Sundermeyer M, Ney H, and Schlüter R. "From feedforward to recurrent LSTM neural networks for language modeling". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2015).

[107] Yukun Ma et al. "Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis". In: *Cognitive Computation* 10.4 (2018), pp. 639–650.

[108] Bryan McCann et al. "Learned in translation: Contextualized word vectors". In: *Advances in Neural Information Processing Systems* (2017), pp. 6294–6305.

[109] Oren Melamud, Jacob Goldberger, and Ido Dagan. "context2vec: Learning generic context embedding with bidirectional lstm". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (2016), pp. 51–61.

[110] C. H. Mendhe et al. "A Scalable Platform to Collect, Store, Visualize, and Analyze Big Data in Real Time". In: *IEEE Transactions on Computational Social Systems* 8.1 (2021), pp. 260–269. DOI: 10.1109/TCSS.2020.2995497.

[111] Zhengjie Miao et al. "Snippext: Semi-supervised opinion mining with augmented data". In: *Proceedings of The Web Conference 2020*. 2020, pp. 617–628.

[112] Risto Miikkulainen and Michael G Dyer. "Natural language processing with modular PDP networks and distributed lexicon". In: *Cognitive Science* 15.3 (1991), pp. 343–399.

[113] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* (2013), pp. 3111–3119.

[114] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[115] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations". In: *Proceedings of the 2013 conference of the north*

american chapter of the association for computational linguistics: Human language technologies. 2013, pp. 746–751.

[116] Tomáš Mikolov et al. "Recurrent neural network based language model". In: *Eleventh annual conference of the international speech communication association* (2010).

[117] Andriy Mnih and Geoffrey E Hinton. "A scalable hierarchical distributed language model". In: *Advances in neural information processing systems* (2009), pp. 1081–1088.

[118] James G Mork et al. "Extracting Rx information from clinical narrative". In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 536–539.

[119] et al. Moya Ignacio. "An agent-based model for understanding the influence of the 11-M terrorist attacks on the 2004 Spanish elections." In: *Knowledge-Based Systems* (2017).

[120] Asier Mujika, Florian Meier, and Angelika Steger. "Fast-slow recurrent neural networks". In: *Advances in Neural Information Processing Systems* (2017), pp. 5915–5924.

[121] Arjun Mukherjee and Bing Liu. "Aspect extraction through semi-supervised modeling". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2012, pp. 339–348.

[122] Martin Müller, Marcel Salathé, and Per E Kummervold. *COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter*. 2020. arXiv: 2005.07503 [cs.CL].

[123] Mark Neumann et al. "Scispacy: Fast and robust models for biomedical natural language processing". In: *arXiv preprint arXiv:1902.07669* (2019).

[124] Kamal Nigam et al. "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39.2 (2000), pp. 103–134.

[125] Benjamin Nye et al. "A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature". In: *Proceedings of the conference. Association for Computational Linguistics. Meeting* 2018 (2018), p. 197.

[126] Ankur Parikh et al. "A Decomposable Attention Model for Natural Language Inference". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016), pp. 2249–2255.

[127] Indravadan Patel et al. "Modeling information spread in polarized communities: Transitioning from legacy media to a Facebook world". In: *SoutheastCon 2017*. IEEE. 2017, pp. 1–8.

[128] Krunal Dhiraj Patel et al. "Using Twitter for diabetes community analysis". In: *Network Modeling Analysis in Health Informatics and Bioinformatics* 9 (2020), pp. 1–16.

[129] Yifan Peng, Shankai Yan, and Zhiyong Lu. *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. 2019. arXiv: 1906.05474 [cs.CL].

[130] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[131] Matthew E Peters et al. "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365* (2018).

[132] Matthew E Peters et al. "Semi-supervised sequence tagging with bidirectional language models". In: *arXiv preprint arXiv:1705.00108* (2017).

[133] Bojanowski Piotr et al. "Enriching word vectors with subword information." In: *Transactions of the Association for Computational Linguistics* (2017).

[134] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (July 2019), pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://www.aclweb.org/anthology/P19-1493.

[135] Fabian Pittke, Henrik Leopold, and Jan Mendling. "Automatic detection and resolution of lexical ambiguity in process models". In: *IEEE Transactions on Software Engineering* 41.6 (2015), pp. 526–544.

[136] Marco Polignano et al. "Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets". In: *6th Italian Conference on Computational Linguistics, CLiC-it 2019* 2481 (2019), pp. 1–6.

[137] Maria Pontiki et al. "Semeval-2016 task 5: Aspect based sentiment analysis". In: *International workshop on semantic evaluation*. 2016, pp. 19–30.

[138] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. "Aspect extraction for opinion mining with a deep convolutional neural network". In: *Knowledge-Based Systems* 108 (2016), pp. 42–49.

[139] Mohiuddin Qudar and Vijay Mago. *A Survey on Language Models*. https://www.researchgate.net/publication/344158120_A_Survey_on_Language_Models. 2020.

[140] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[141] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8 (2019), p. 9.

[142] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *arXiv preprint arXiv:1910.10683* (2019).

[143] Pranav Rajpurkar et al. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. 2016. arXiv: 1606.05250 [cs.CL].

[144] Daniel Ramage, Susan Dumais, and Dan Liebling. "Characterizing microblogs with topic models". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 4. 1. 2010.

[145] Daniel Ramage et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". In: *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009, pp. 248–256.

[146] Thrishma Reddy, Philippe J Giabbanelli, and Vijay K Mago. "The artificial facilitator: guiding participants in developing causal maps using voice-activated technologies". In: *International Conference on Human-Computer Interaction*. Springer, Cham. 2019, pp. 111–129.

[147] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. "Mctest: A challenge dataset for the open-domain machine comprehension of text". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 193–203.

[148] Alexander M Rush. "A neural attention model for sentence summarization." In: *empirical methods in natural language processing* (2015).

[149] Edunov S, Baevski A., and Auli M. "Pre-trained language model representations for language generation". In: *arXiv* (2015).

[150] Toshniwal S. et al. "A comparison of techniques for language model integration in encoder-decoder speech recognition". In: *In 2018 IEEE Spoken Language Technology Workshop* (2018).

[151] Tobias Schnabel et al. "Evaluation methods for unsupervised word embeddings". In: *Proceedings of the 2015 conference on empirical methods in natural language processing* (2015), pp. 298–307.

[152] Burr Settles. "Active learning literature survey". In: (2009).

[153] Jingbo Shang et al. "Automated phrase mining from massive text corpora". In: *IEEE Transactions on Knowledge and Data Engineering* 30.10 (2018), pp. 1825–1837.

[154] Rakshith Shetty and Jorma Laaksonen. *Video captioning with recurrent networks based on frame- and video-level features and visual content classification*. 2015. arXiv: 1512.02949 [cs.CV].

[155] Grigori Sidorov et al. "Syntactic n-grams as machine learning features for natural language processing". In: *Expert Systems with Applications* 41.3 (2014), pp. 853–860.

[156] Patrice Y Simard et al. "Transformation invariance in pattern recognition—tangent distance and tangent propagation". In: *Neural networks: tricks of the trade*. Springer, 1998, pp. 239–274.

[157] Richard Socher et al. "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing* (2013), pp. 1631–1642.

[158] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. "BIOSSES: a semantic sentence similarity estimation system for the biomedical domain". In: *Bioinformatics* 33.14 (2017), pp. i49–i58.

[159] Thomas J Steenburgh, Jill Avery, and Naseem Dahod. "Hubspot: Inbound marketing and web 2.0". In: *HBS Case* 509-049 (2009).

[160] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[161] Mikolov T et al. "Extensions of recurrent neural network language model". In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2011).

[162] Nguyen A. T. and T. N. Nguyen. "Graph-based statistical language model for code". In: *IEEE International Conference on Software Engineering* (2015).

[163] Olivier Taboureau et al. "ChemProt: a disease chemical biology database". In: *Nucleic acids research* 39.suppl_1 (2010), pp. D367–D372.

[164] Kai Sheng Tai, Richard Socher, and Christopher D Manning. "Improved semantic representations from tree-structured long short-term memory networks". In: *arXiv preprint arXiv:1503.00075* (2015).

[165] Duyu Tang, Bing Qin, and Ting Liu. "Aspect level sentiment classification with deep memory network". In: *arXiv preprint arXiv:1605.08900* (2016).

[166] Duyu Tang et al. "Effective LSTMs for target-dependent sentiment classification". In: *arXiv preprint arXiv:1512.01100* (2015).

[167] Joseph Tassone et al. "Utilizing deep learning and graph mining to identify drug use on Twitter data". In: *BMC Medical Informatics and Decision Making* 20.11 (2020), pp. 1–15.

[168] Ivan Titov and Ryan McDonald. "Modeling online reviews with multi-grain topic models". In: *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 111–120.

[169] Zhiqiang Toh and Wenting Wang. "Dlirec: Aspect term extraction and term polarity classification system". In: *Association for Computational Linguistics and Dublin City University*. Citeseer. 2014.

[170] Trieu H. Trinh and Quoc V. Le. *A Simple Method for Commonsense Reasoning*. 2019. arXiv: 1806.02847 [cs.AI].

[171] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* (2017), pp. 5998–6008.

[172] Alex Wang and Kyunghyun Cho. "BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model". In: *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation* (June 2019), pp. 30–36. DOI: 10.18653/v1/W19-2304. URL: https://www.aclweb.org/anthology/W19-2304.

[173] Alex Wang et al. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: (Nov. 2018), pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://www.aclweb.org/anthology/W18-5446.

[174] Wei Wang, Ming Yan, and Chen Wu. "Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 1705–1714.

[175] Wei Wang et al. *StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding*. 2019. arXiv: 1908.04577 [cs.CL].

[176] Wenya Wang et al. "Coupled multi-layer attentions for co-extraction of aspect and opinion terms". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[177] Wenya Wang et al. "Recursive neural conditional random fields for aspect-based sentiment analysis". In: *arXiv preprint arXiv:1603.06679* (2016).

[178] Xuerui Wang, Andrew McCallum, and Xing Wei. "Topical n-grams: Phrase and topic discovery, with an application to information retrieval". In: *Seventh IEEE international conference on data mining (ICDM 2007)* (2007), pp. 697–702.

[179] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. "Group and topic discovery from relations and text". In: *Proceedings of the 3rd international workshop on Link discovery*. 2005, pp. 28–35.

[180] Yequan Wang et al. "Attention-based LSTM for aspect-level sentiment classification". In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, pp. 606–615.

[181] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. "Neural network acceptability judgments". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 625–641.

[182] Jason Wei and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks". In: *arXiv preprint arXiv:1901.11196* (2019).

[183] Jianshu Weng et al. "Twitterrank: finding topic-sensitive influential twitterers". In: *Proceedings of the third ACM international conference on Web search and data mining*. 2010, pp. 261–270.

[184] Georg Wiese, Dirk Weissenborn, and Mariana Neves. "Neural Domain Adaptation for Biomedical Question Answering". In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (2017), pp. 281–289.

[185] Adina Williams, Nikita Nangia, and Samuel Bowman. "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (June 2018), pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: https://www.aclweb.org/anthology/N18-1101.

[186] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

[187] Ziang Xie et al. "Data noising as smoothing in neural network language models". In: *arXiv preprint arXiv:1703.02573* (2017).

[188] Hu Xu et al. "Double embeddings and cnn-based sequence labeling for aspect extraction". In: *arXiv preprint arXiv:1805.04601* (2018).

[189] Shen Y., Lin Z., and Huang C. W.and Courville A. "Neural language modeling by jointly learning syntax and lexicon". In: *arXiv* (2017).

[190] Yinfei Yang et al. "Learning Semantic Textual Similarity from Conversations". In: *Proceedings of The Third Workshop on Representation Learning for NLP* (2018), pp. 164–174.

[191] Zhilin Yang. "Breaking the softmax bottleneck: A high-rank RNN language model." In: *arXiv* (2017).

[192] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. 2020. arXiv: 1906.08237 [cs.CL].

[193] Zhilin Yang et al. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* (2019), pp. 5753–5763.

[194] Wonjin Yoon et al. "CollaboNet: collaboration of deep neural networks for biomedical named entity recognition". In: *BMC bioinformatics* 20.10 (2019), p. 249.

[195] Jason Yosinski et al. "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems* (2014), pp. 3320–3328.

[196] Adams Wei Yu et al. "Qanet: Combining local convolution with global self-attention for reading comprehension". In: *arXiv preprint arXiv:1804.09541* (2018).

[197] Li Yuhua, Zuhair A. Bandar, and David McLean. "An approach for measuring semantic similarity between words using multiple information sources." In: *IEEE Transactions on knowledge and data engineering 15.4* (2003).

[198] Kazi Zainab Khanam, Gautam Srivastava, and Vijay Mago. "The Homophily Principle in Social Network Analysis". In: *arXiv e-prints* (2020), arXiv–2008.

[199] Rowan Zellers et al. "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), pp. 93–104.

[200] Ying Zeng et al. "Scale up event extraction learning via automatic training data generation". In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[201] Haizheng Zhang et al. "Probabilistic community discovery using hierarchical latent gaussian mixture model". In: *AAAI*. Vol. 7. 2007, pp. 663–668.

[202] Zhengyan Zhang et al. "ERNIE: Enhanced Language Representation with Informative Entities". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (July 2019), pp. 1441–1451. DOI: 10.18653/v1/P19-1139. URL: https://www.aclweb.org/anthology/P19-1139.

[203] Xin Zhao et al. "Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 56–65. URL: https://www.aclweb.org/anthology/D10-1006.

[204] Yukun Zhu et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision* (2015), pp. 19–27.

Appendix A

Table of References

Table A.1: Shows the references selected for the survey, total number of citations of April'20 (TC), h-Google Index of the venue (h-i) and year of publication (Y)

Title	Venue	TC	h-i	Y
BioBERT: a pre-trained biomedical language representation model for biomedical text mining [87]	Bio Informatics	141	335	2018
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [36]	arXiv	4500		2018
A Neural Probabilistic Language Model [7]	Journal of Machine Learning Research	5862	173	2003

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Recurrent neural network based language model [116]	Eleventh annual conference of the international speech communication association	3949	65	2010
Improved semantic representations from tree-structured long short-term memory networks [164]	Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics	1604	106	2015
Deep contextualized word representations[131]	arXiv	2424		2018
A large annotated corpus for learning natural language inference [14]	Empirical Methods in Natural Language Processing	999	88	2015
Semi-supervised sequence learning[34]	Advances in neural information processing systems	542	169	2015
Universal Language Model Fine-tuning for Text Classification [56]	Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics	620	106	2015
Imagenet: A large-scale hierarchical image database [64]	2009 IEEE conference on computer vision and pattern recognition	16242	240	2009

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Empower sequence labeling with task-aware neural language model [101]	AAAI Conference on Artificial Intelligence	127	95	2018
Distributed representations of words and phrases and their compositionality [113]	Neural Information Processing Systems	18000	169	2013
GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding [173]	Empirical Methods in Natural Language Processing	312	88	2018
Exploring the Limits of Language Modeling [73]	arXiv	649		2018
SQuAD: 100,000+ Questions for Machine Comprehension of Text [143]	Empirical Methods in Natural Language Processing	1365	88	2016
Language models are unsupervised multitask learners [141]	Open AI Blog	317		2019
Semi-supervised sequence tagging with bidirectional language models [132]	Association for Computational Linguistics	239	106	2017
Deep convolutional neural network for inverse problems in imaging [66]	IEEE transactions on image processing	587	242	2017
Unsupervised machine translation using monolingual corpora only [83]	arXiv	261		2017

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Glove: Global vectors for word representation [130]	Empirical Methods in Natural Language Processing	12000	88	2014
Attention is all you need [171]	Neural Information Processing Systems	6000	169	2017
TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension [72]	Association for Computational Linguistics	262	106	2017
How transferable are features in deep neural networks? [195]	Neural Information Processing Systems	3677	169	2014
A Decomposable Attention Model for Natural Language Inference [126]	Association for Computational Linguistics	529	106	2016
A scalable hierarchical distributed language model [117]	Neural Information Processing Systems	5862	173	2003
Context2vec: Learning generic context embedding with bidirectional lstm [109]	Computational Natural Language Learning	191	39	2016
A theoretically grounded application of dropout in recurrent neural networks [42]	Neural Information Processing Systems	843	169	2016
Learning towards minimum hyperspherical energy [102]	Neural Information Processing Systems	587	169	2016

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Fast-slow recurrent neural networks [120]	Neural Information Processing Systems	39	169	2019
Transfer learning for biomedical named entity recognition with neural networks [45]	Bio Informatics	27	335	2018
Collabonet: collaboration of deep neural networks for biomedical named entity recognition [194]	BMC Bio Informatics	10	335	2019
Neural domain adaptation for biomedical question answering [184]	Association for Computational Linguistics	32	183	2017
Publicly available clinical bert embeddings [1]	Association for Computational Linguistics	45	183	2019
NCBI disease corpus: a resource for disease name recognition and concept normalization [38]	Journal of biomedical informatics	201	83	2014
Deep learning with word embeddings improves biomedical named entity recognition [50]	Bio Informatics	192	335	2017
Supervised Learning of Universal Sentence Representations from Natural Language Inference Data [32]	arXiv	680		2018

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Learned in translation: Contextualized word vector [108]	Neural Information Processing Systems	360	169	2017
State-of-the-Art Speech Recognition with Sequence-to-Sequence Models [27]	IEEE International Conference on Acoustics, Speech and Signal Processing	376	130	2018
Automatic Detection and Resolution of Lexical Ambiguity in Process Models [135]	IEEE Transactions on Software Engineering	46	151	2015
Named entity recognition with bidirectional LSTM-CNNs [28]	transactions of the Association for Computational Linguistics	743	51	2016
Automated phrase mining from massive text corpora [153]	IEEE Transactions on Knowledge and Data Engineering	86	77	2018
LSwag: A large-scale adversarial dataset for grounded commonsense inference [199]	Empirical Methods in Natural Language Processing	90	157	2018
A broad-coverage challenge corpus for sentence understanding through inference [185]	Association for Computational Linguistics	378	183	2018

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Multigranularity hierarchical attention fusion networks for reading comprehension and question answering [174]	Association for Computational Linguistics	62	183	2018
Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation [20]	Association for Computational Linguistics	210	183	2018
A context-blocks model for identifying clinical relationships in patient records [39]	BMC bioinformatics	51	183	2011
Extracting Rx information from clinical narrative [118]	JAMIA	31	132	2010
ALBERT: A Lite BERT for Self-supervised Learning of Language Representations [84]	arXiv	48		2019
RACE: Large-scale ReAding Comprehension Dataset From Examinations [82]	Empirical Methods in Natural Language Processing	169	335	2017
Improving language understanding by generative pre-training [140]	OPEN AI	680		2018

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Neural Network Acceptability Judgments [181]	Transactions of the Association for Computational Linguistics	57	80	2019
Recursive deep models for semantic compositionality over a sentiment treebank [157]	Empirical Methods in Natural Language Processing	3596	335	2013
Learning Semantic Textual Similarity from Conversations [190]	arXiv	46		2018
Scibert: Pretrained contextualized embeddings for scientific text [6]	arXiv	53		2019
Analysis methods in neural language processing: A survey [5]	Transactions of the Association for Computational Linguistics	43	168	2019
Construction of the Literature Graph in Semantic Scholar [2]	Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)	44	51	2018
Google's neural machine translation system: Bridging the gap between human and machine translation [186]	arXiv	2370		2016

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Deep biaffine attention for neural dependency parsing [40]	arXiv	243		2016
GENIA corpus—a semantically annotated corpus for bio-textmining [75]	Bioinformatics	1024	183	2003
Scispacy: Fast and robust models for biomedical natural language processing [123]	arXiv	20		2019
A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature [125]	Proceedings of the conference. Association for Computational Linguistics. Meeting	20		2018
ChemProt-3.0: a global chemical biology diseases mapping	Database: The Journal of Biological Databases and Curation [79]	40	65	2016
Allennlp: A deep semantic natural language processing platform [43]	arXiv	229		2018
Structural scaffolds for citation intent classification in scientific publications [31]	arXiv	7		2019

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Clinicalbert: Modeling clinical notes and predicting hospital readmission [58]	arXiv	17		2019
BioCreative V CDR task corpus: a resource for chemical disease relation extractions [92]	Database: The Journal of Biological Databases and Curation	88	65	2016
Efficient Estimation of Word Representations in Vector Space [114]	arXiv	14538		2013
Dependency-based word embeddings [89]	Transactions of the Association for Computational Linguistics	850	168	2014
Topical word embeddings [103]	AAAI	280	153	2015
From word embeddings to document distances [81]	International conference on machine learning	940	254	2015
Improving distributional similarity with lessons learned from word embeddings [90]	Association for Computational Linguistics	923	168	2014
Evaluation methods for unsupervised word embeddings [151]	Association for Computational Linguistics	334	335	2017

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Google's neural machine translation system: Bridging the gap between human and machine translation [186]	arXiv	2370		2016
BioCreative V CDR task corpus: a resource for chemical disease relation extractions c[92]	Database: The Journal of Biological Databases and Curation	88	65	2016
Approximate nearest neighbors: towards removing the curse of dimensionality [62]	Proceedings of the thirtieth annual ACM symposium on Theory of computing	4356	89	1998
How multilingual is Multilingual BERT? [134]	arXiv	40		2019
Scale up event extraction learning via automatic training data generation [200]	AAAI	8	95	2018
Enriching word vectors with subword information [133]	Transactions of the Association for Computational Linguistics	3000	183	2017
From word to sense embeddings: A survey on vector representations of meaning [70]	Journal of Artificial Intelligence Research	71	103	2018

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
An agent-based model for understanding the influence of the 11-M terrorist attacks on the 2004 Spanish elections [119]	Knowledge-Based Systems	7	94	2019
Pairwise word interaction modeling with deep neural networks for semantic similarity measurement [57]	Association for Computational Linguistics	140	183	2016
An approach for measuring semantic similarity between words using multiple information sources [197]	IEEE Transactions on knowledge and data engineering	1315	148	2003
Breaking the softmax bottleneck: A high-rank RNN language model [191]	arXiv	178		2017
A neural attention model for sentence summarization [148]	Empirical Methods in Natural Language Processing	1359	103	2015
Placing search in context: The concept revisited [41]	international conference on World Wide Web	75	64	2016
A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art [85]	Association for Computational Linguistics	7	183	2019

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	TC	h-i	Y
Extensions of recurrent neural network language model [161]	ACM SIGIR	78	65	2016
Graph-based statistical language model for code [162]	IEEE International Conference on Software Engineering	42	103	2016
A study of smoothing methods for language models applied to ad hoc information retrieval [16]	ACM SIGIR	53	81	2016
Unified language model pre-training for natural language understanding and generation	In Advances in Neural Information Processing Systems	42	75	2015
A comparison of techniques for language model integration in encoder-decoder speech recognition [150]	IEEE Spoken Language Technology Workshop	229		2018
From feedforward to recurrent LSTM neural networks for language modeling [106]	arXiv	7		2019
Pre-trained language model representations for language generation [149]	Association for Computational Linguistics	20	163	2019
Neural language modeling by jointly learning syntax and lexicon [189]	arXiv	38		2017