A study and comparative analysis of various use cases of NLP using Sequential Transfer learning techniques

Natural language is used by humans for every day communication. Every human language is evolving by addressing its own inherit challenges and ambiguities. Natural language processing is a domain presents various text manipulation techniques for human languages. Since the development of human language is a continuous process, it is inevitable for the linguistics and techies to present new kind of techniques for complement the process. This study analysis the impact of transfer learning techniques for solving the two bench marking problems. Sentiment analysis (SA) and Named Entity Recognition (NER) have grown into a popular applications of natural language processing (NLP). The objective of sentiment analysis is to get to know a user or audience opinion on a target object by analysing a huge amount of textual information from various sources. Segmenting the various entities and establishing relationship among them is tedious task. Though reasonable amount of effort put up in identifying the various entities, still there is requirement in the picture for domain based entity segmentation. One of the major problems of Machine Learning is that the models are highly dependent on large amounts of high-quality data. Unfortunately, these data are rarely available, and if it does exist, they are very highly expensive to access. With the help of Transfer learning, there is less need of high-quality data. This chapter address the trends of SA and NER problem, and demonstrates the various solution derived through Machine learning and transfer learning model along with an insight how sequential transfer learning influences the tasks.

Keywords: nlp, sentimental analysis,NER, deep learning, sequential transfer learning

1.Introduction

Natural language is used by humans for every day communication. Every human language is evolving by addressing its own inherit challenges and ambiguities. Natural language processing is a domain presents various text manipulation techniques for human languages. Since the development of human language is a continuous process, it is inevitable for the linguistics and techies to present new kind of techniques for complement the process. The increasing use of Internet, produces huge amount data. Analysing those data and inferring insight become a challenging tasks for data analysts. In the literature [1-10] many researchers contribute various Machine learning and deep learning, techniques to solve the various use cases of natural language problem. Among the various use cases of NLP the Sentiment analysis and Named entity Recognition has become the most crucial application in the trend. The objective of sentiment analysis is to get to know a user or audience opinion on a target object by analysing a huge amount of textual information from various sources. The insight from the huge amount of data in terms of sentiment used in many applications [11-13]. The objective of NER system is to extract the entities from raw data and determines their corresponding category. This information is useful in a variety of NLP tasks such as Information Extraction Systems, Question-Answer Systems, Machine Translation Systems, Automatic Summarizing Systems, and Semantic Annotation. Segmenting the various entities and establishing relationship among

them is tedious task. Though reasonable amount of effort put up in identifying the various entities, still there is requirement in the picture for domain based entity segmentation [14-16].

The researcher analysed the sentiments from the raw text in three different level in detail. The document level[17],sentence [18],aspect level[19].Sentiment analysis or classification at the degree of document is to predict an overall sentiment to an opinion document. The input to the system is textual information conveyed in the document and expected output is the measure of the insight in terms of positive or negative. In the literature many researchers address the sentiment analysis at the level of document. Here document representation plays an important role. Most of the authors use BoW model [20,21] to represent the textual data. Whereas the technique [22] learns the sentence representation using the neural network architectures such as CNN or LSTM from word embeddings. Then it use GRU to adaptively encode semantics of sentences and their inherent relations. [23] use word embeddings to represent text and designed LSTM architecture for cross-lingual sentiment classification. Some of the authors [24,25] address the sentiment analysis at the degree of sentence level that is finding the sentiment expressed in the given sentence. In [26], the authors presented the recursive neural tensor network (RNTN) to understand the relations between elements. And in [27] authors applied LSTM for Twitter sentiment classification. The third category of technique [28, 29,30] which address the sentiment at the degree of target aspect. Such category technique consider the sentiment and target aspect from the text. Among the three different level in handling the SA, the proposed SA technique determines the sentiment at the sentence level. At the outset the techniques addressed in this section present various approaches to solve SA and NER problem using Machine learning, Deep learning models.

One of the major problems of Machine Learning is that the models are highly dependent on large amounts of high-quality data. Unfortunately, these data are rarely available, and if it does exist, they are very highly expensive to access. With the help of Transfer learning, there is less need of high-quality data. Transfer learning is a pre-trained model which has already been trained on a task for which labelled training data is enormous, which can handle similar tasks with less data. These pre-trained models are often faster than our traditional models. These pre-trained models are contributing to the SOTA on variety of the task. This chapter address the trends of SA and NER problem, and demonstrates the various solution derived through Machine learning and transfer learning model along with an insight how transfer learning influences the tasks. Section 2 address the existing works section 3 discuss the Background of the research work section 4 discuss the impact of transfer learning for the NLP problem such as sentimental analysis and NER. Section 5 drawn the conclusion.

2. Literature review

The authors in [31] have classified the Recent trends of Transfer learning into three broad settings which is Inductive transfer, Transductive transfer and Unsupervised transfer. They have inadequately elaborated about unsupervised learning. If some prior assumptions like the source and target domains are not related to each other doesn't hold good, then negative transfer can happen. The techniques mentioned in the research paper have only been applied to small scale applications but they can be applied to large applications such as social networking analysis, video classification.

An integrated lexicon and rule-based approach [32] is employed to extract explicit and implicit aspect as well as sentiment classification for these aspects. This approach confirmed that integrating sentiment and aspects lexicons with various rules settings that handle various challenges in sentiment analysis. For improving future work and enhancements, proper classification and categorization should be done so that it is easily understandable to the governments.

The technique in [33] presents a Bidirectional Long Short-Term Memory networks (Att-BiLSTM) networks and conducted cross datasets training/evaluation, in order to see the generalizability of this approach. In clinical NLP one challenge is there that incorrect assertion can cause incorrect diagnosis of patients and can further lead to major issue. So to avoid it, this model is proposed for assertion detection.

The authors is [34] proposed a domain-adaptive transfer learning method. They have also implemented transfer tests among different corpuses so that to testify the effectiveness of the proposed methodology. They have proposed only a few strategies to acquire transfer knowledge and has less linguistic information for transfer learning methods.

[35] implemented Transfer learning by pretraining a model for a certain NER task and than fine-tuned the learned model for another NER task for which there are few labelled training data. They have also built several variant models that integrate word embeddings, attention mechanisms and the BERT language model[5].

In [36] a system using transfer learning with ANNs for NER is proposed, specifically patient note de-identification, by transferring ANN parameters trained on a large labelled dataset to another dataset with limited human annotations.

[37] Presents experiments on seven new source/target corpus pairs, nearly doubling the total number of corpus pairs that have been studied in all past work combined. They have compared three existing methods that can be applied to the setting of transfer learning with novel entities in the target domain. These methods have not been compared against each other before in the literature.

Authors in [38] proposed two main aspects of NER research - target languages and technical approaches with statistical analysis. The only problem with this is the models based on deep learning need large amount of corpus so that low resource languages and some domains that lack large amounts corpora could not get high performance. Knowledge Bases enhances current NER system performance, Transfer Learning can partly overcome the scarcity of labelled training data.

[39]Proposed a system for detecting valence task using transfer learning BLSTM technology. To avoid the overfitting, layers of the pre-trained model were frozen. It primarily focus on single transfer but can focus on multiple transfers, to increase the amount of data used in the process. We will perform transfers from two classes (positive and negative) to three classes (adding neutral), then five classes and finally seven classes.

[40]Proposed unlabelled bilingual corpora to extract useful features from transferring information from resource-rich language toward resource-poor language and by using these features and a small training data, make a NER supervised model. The problem with this technique is on one hand the efficiency of supervised method will decrease massively along

with changing the domain of texts and so because of that we get a poor result when, we train the Stanford NER tagger on the IUST training set and test it on bilingual test set.

[41] implemented clickbait detection is taken as an example to study the sentence classification with a transferring network. The paper trains the source model on English corpus and transfers it to corpus in Chinese. Experimental results show that transfer learning model in this paper can achieve similar performance on the target language using less annotation, showing the effectiveness and robustness of this model. The problem with this model is it has very few annotations.

[42] presents the complete picture of SA techniques and the related fields with brief details. The main contributions of this paper include the sophisticated categorizations of a large number of recent articles and the illustration of the recent trend of research in the sentiment analysis. The problem is other than English language there is still a lack of resources and researches concerning all those other languages. Also, it has been noticed that there is lack of benchmark data sets in this field[12]

[43] Facilitates the researcher by summarizing the relevant research results of the sentiment analysis in recent years and focuses on the algorithms and applications of transfer learning in the sentiment analysis. And also the authors discussed the sentiment analysis and put forward the prospect of it, like The application of cross-domain transfer learning in aspect extraction has not been fully explored and how to solve the negative transfer problem in transfer learning becomes the difficulty of using transfer learning for text analysis.

The authors [44] presented two approaches that use unlabelled data to improve sequence learning with recurrent networks. The first approach is to predict what comes next in a sequence, which is a conventional language model in natural language processing. The second approach is to use a sequence autoencoder, which reads the input sequence into a vector and predicts the input sequence again. The main problem with this approach is that it is unstable: if we were to increase the number of hidden units or to increase the number of backprop steps, the training breaks down very quickly, the objective function explodes even with careful tuning of the gradient clipping. This is because LSTMs are sensitive to the hyperparameters for long documents.

[45] demonstrated the contextualized word representations can be easily added to existing models and significantly improve the state of the art across six challenging NLP problems, including question answering, textual entailment and sentiment analysis. They have also showed that exposing the deep internals of the pre-trained network is crucial, allowing downstream models to mix different types of semi-supervision signal. When applying ELMo to a broad range of NLP tasks, the model shows large improvement.

[46] implemented an effective transfer learning method that can be applied to any task in NLP, and introduced techniques that are key for fine-tuning a language model. It significantly outperforms the state-of-the-art on six text classification tasks, reducing the error by 18- 24% on the majority of datasets. They have also proposed several novel fine-tuning techniques that in conjunction prevent catastrophic forgetting and enable robust learning across a diverse range of tasks.

[47] They have proposed that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing

a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. It is a new approach and they have aimed to achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. The one challenge is to better handle unknown, or rare words. This will be required for the model to be more widely used and to match the performance of current state-of-the-art machine translation systems in all contexts.

The authors in [48] combined the benefits of both approaches by integrating data mining and information extraction methods. Their aim is to provide a new high-quality information extraction methodology and at the same time to improve the performance of the underlying extraction system and hence shorten the life cycle of information extraction engineering. If we integrate data mining methods it will enable more precise feature selection for IE, which in turn reduces the feature space to the most significant information for mining new knowledge.

[49] They have made a survey of event extraction technology, describing the tasks and related concepts of event extraction, analyzing, comparing and generalizing the relevant descriptions in different fields. Then analyzed, compared and summarized the three main methods of event extraction. The problem with this methodology is that research on related technologies such as entity, relationship identification, and syntax analysis is not mature enough, leading to cascading errors. The field scalability and portability of the event extraction system are not ideal. For example, the relevant research on Chinese event extraction mainly focuses on biomedicine, microblog, news, emergencies, etc. Also, lack of large-scale mature corpus and labeling corpus, the corpus needs further improvement.

[50] explores the topic modelling from different tools and techniques, such as the Python libraries Gensim and Mallet in order to compare and contrast the relevance of those models to our dataset. The impact that these techniques have on the humanities fields can be astoundingly influential, but severely limited by the availability, size, and domain of the training dataset. The only problem noticed is a huge disparity for Dates as most of the years in the text were unrecognized. There is a huge opportunity for NER algorithms to begin considering more cultural contexts during the training phase of creation. One of the major problems of Machine Learning is that the models are highly dependent on large amounts of high-quality data. Unfortunately, these data are rarely available, and if it does exist, they are very highly expensive to access. With the help of Transfer learning, there is less need of high-quality data. Transfer learning is a pre-trained model which has already been trained on a task for which labelled training data is enormous, which can handle similar tasks with less data. These pre-trained models are often faster than our traditional Machine learning models. These pre-trained models are contributing to the SOTA on variety of the task. This chapter demonstrates the transfer learning models in designing the predictive modelling tasks such as SA and NER.

3. Empirical study

In the past few years transfer learning gain its application in object detection and some of tasks of NLP. The main highlight of transfer learning is reducing the burden in designing the classifier from the scratch. One more highlight is transfer learning facilitates the researchers reuse or customize the model which has been designed for some other task to their required

task. By doing so it reduces lot of time need to be spend for development and training. In contrast in tradition learning a new model need to be generated for every new task as shown in figure1. The transfer learning mainly divided into transductive and inductive learning. Then it is further narrowed into various sub categories to tackle the some variants such as domain adaption, cross-lingual learning, multi-task learning(MTL) and sequential transfer learning(STL). This section aims to analyse the impact of sequential transfer learning in solving the bench marks sentiment analysis and NER. As the name implies STL transfers the insight sequentially in order to accomplish the tasks. Sequential transfer learning is the form that has led to the biggest improvements so far. The general practice is to pretrain representations on a large unlabelled text corpus using your method of choice and then to adapt these representations to a supervised target task using labelled data as can be seen below in figure 2.

The target task is accomplished per sequential transfer learning (STL) in two phases such as pretraining and adaptation. The Pretraining phase of STL is costlier than the adaption phase since more training is required because if it's one time execution on the source model. The pretraining is accomplished in three approaches distant supervision, traditional supervision and no supervision. Where in distant supervision data obtained from heuristics and domain expertise. The traditional supervision required manually annotated training samples and no supervision required large unannotated samples. In adaption phase is accomplished through two methods feature extraction and fine tuning. Feature extraction uses the representations of a pre-trained model and feeds it to another model while fine-tuning involves training of the pre-trained model on target task. The feature extraction and fine tuning phase can be represented as mentioned in eq.1 and eq 2.

$\eta_t^{(l)} = 0, \forall l \in [1, L_s] \forall t$-----eq.1

Fine-tuning on the other hand, requires updating at least one of the source layers during adaptation:

$\eta_t^{(l)} > 0, \exists l \in [1, L_s] \exists t$----eq.2
where $\exists l$ means there exists an l

Especially, Sequential transfer learning is useful in the following cases, where the Source and the target task data is not available at the same time, and where the source task has more data than the target task and also Adaptation to many target tasks is required.

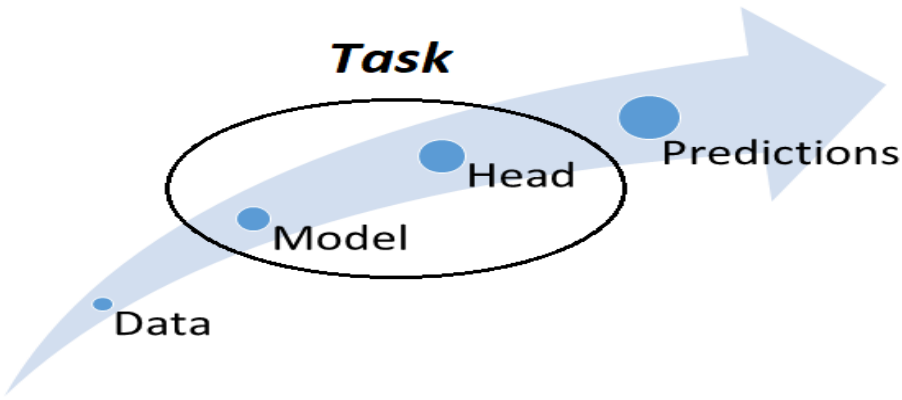


Figure 1-Task learning using traditional model.

3.1 Sequential transfer learning model for sentiment analysis

This section discusses the various STL models in the process of designing SA application. The bench marking models ULMFIT, RoBETa, XLNET and DistilBERT have exploited well to influence the task performance. Highlights of STL models are as follows.

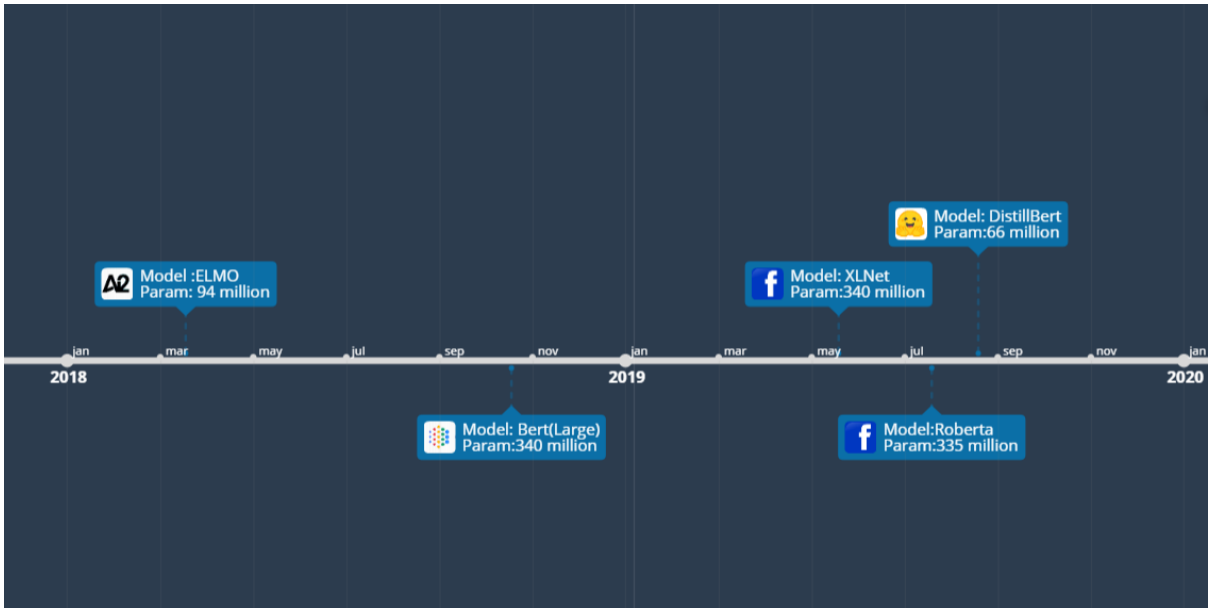


Figure 2: Timeline of Various Models

3.1.1 ULMFIT

Jeremy Howard and Sebastian Ruder [41] designs a language model called ULMFIT. It is a method more than just embeddings and contextualized embeddings. ULMFIT language model can be fine-tuned for various NLP tasks. [42] introduce a LM which require huge amount of data to attain good performance. In contrast ULMFIT able to complete the task with small corpus. ULMFIT is more effective than the another TL model called Elmo. Which makes the use the Language Model in the Fine-Tuning Process. The readers interested in ULMFIT model are suggested to refer [41] to get more information.

3.1.2 RoBERTa

RoBERTa was established at Facebook. RoBERTa uses 1000% of more textual data than BERT model as well as compute power. In order to improve the training, RoBERTa eliminate the Next Sentence Prediction (known as NSP) task from BERT pre-training and introduces another masking, i.e dynamic masking so that masked token during each epochs keep on changes.RoBERTa uses 160GB of text for pre-training which includes 16Gb of Books Corpus English Wikipedia used in BERT.

3.1.3 XLNet

XLNet is an improvised version of BERT model which has more computation power than Bert and increased efficiency in terms of accuracy. XLNet works on a different architecture

than Bert model, it uses permutation language modeling. In this model tokens are taken and predicted in a random fashion, because of this variability in the architecture the throughput of learning bidirectional relationships increases thereby enhancing the dependencies to deal with and relations between the words. We can see XLNet as an enhanced version of Bert Model, as Bert is an autoencoder(AE) language, model. Bert model aims to reconstruct the sentence using Mask. This model can see the context in both forward and backward direction . But this has several disadvantages first it will cause fine-tune disparency as the mask variable is not present during the fine-tuning process in our dataset. Second is that masked tokens assume that they are independent of other tokens in the sentences. To overcome those things XLnet uses permutation Language Modeling which provides better results than that. For training, this whole model about 130GB of data is used.

3.1.4 DistilBERT

DistilBERT is a small, fast in execution, economically cheap and light Transformer model trained by distilling Bert base. This modelhas 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of Bert's performances. DistilBERT - has the same general architecture as BERT. The token-type embeddings and the pooler are removed while the number of layers is reduced by a factor of 2. Most of the operations used in the Transformer architecture are highly optimized in modern linear algebra frameworks and investigations showed that variations on the last dimension of the tensor (hidden size dimension) have a smaller impact on computation efficiency (for a fixed parameters budget) than variations on other factors like the number of layers. Thus we focus on reducing the number of layers. DistilBERT is distilled on very large batches leveraging gradient accumulation (up to 4K examples per batch) using dynamic masking and without the next sentence prediction objective.

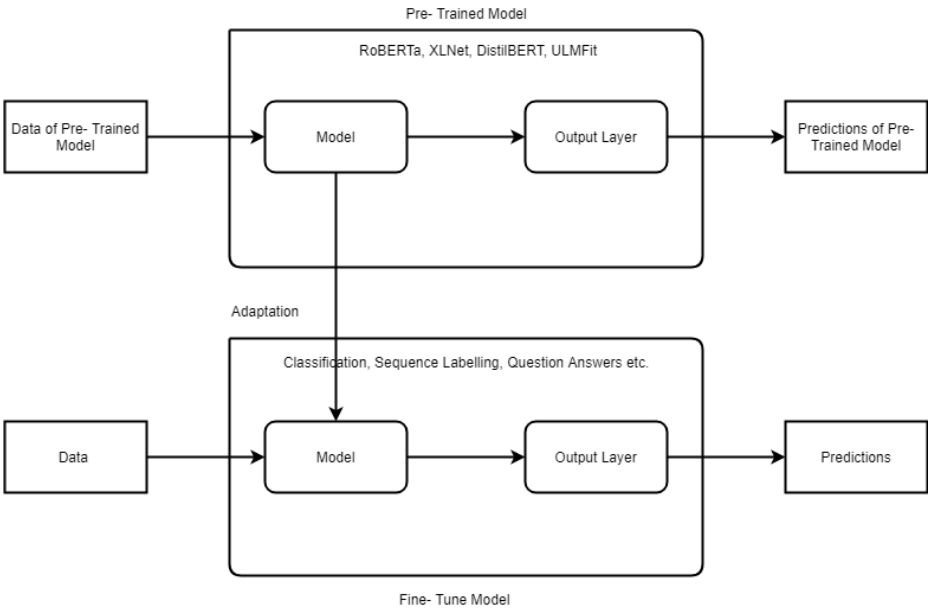


Figure 3 Sequential Transfer Learning Process

3.1.5 Methodology used in SA:

The steps used to implement have been given as pseudo code and the same is depicted a architecture diagram.

Pseudo code
Step1: Accept each sample from the corpus
Step2: Tokenize the sample
Step 3: Sentence splitting
Step 4: Pre-train transfer learning model
Step 5: Adopt the pre-trained model on the target task
Step 6: Evaluate the model
Step 7: summarize the performance using quantitative metrics

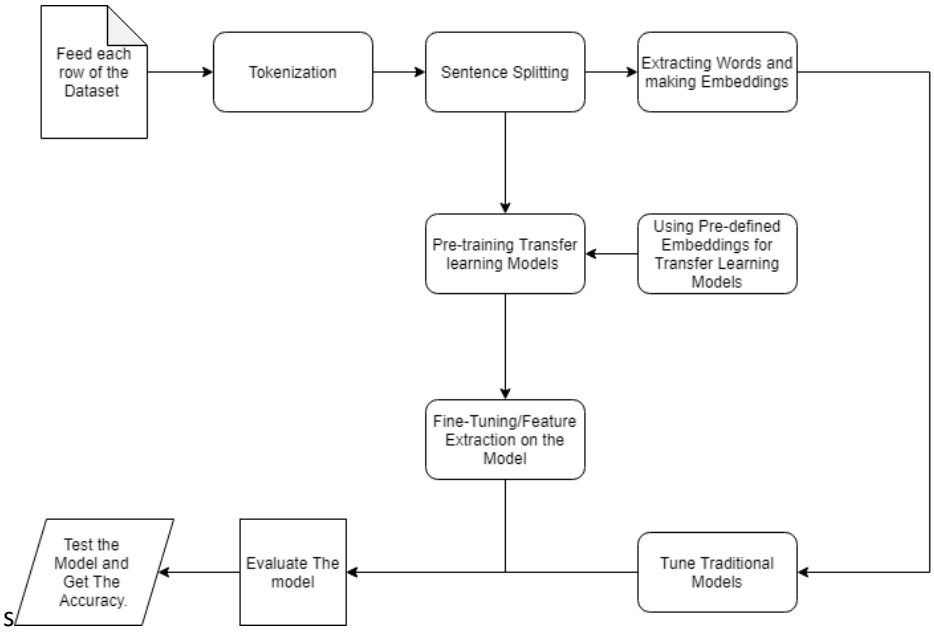


Figure 4. Sentiment analysis using STL

3.1.6 Results and discussion

All the notebooks are run in Kaggle kernel and have the default aspects of the GPU as of 2020. The python library version are also default version of Kaggle kernel library. No changes of version was made. For Neural Network models we have used learning rate of 2*10^-3 with training upto 3 epochs. No additional constraints were used in NN and other ML approaches.

1) **IMDB Dataset**

The below are the results of IMDB reviews. This dataset contains 50K movie review of which 25000 are positive and 25000 are negative.

Table 1 : Comparative results from various STL model Vs Traditional machine/deep learning Model in terms of amount of time(Seconds)

(in minutes)	RoBERTa	XLNet	DistilBERT	ULMFit	LSTM	Naïve Bayes	SVM	Logistic Regression

100% of Dataset Time	30:40	27:41	15:13	93:04	274:32	03:12	30:13	13:41
50% of Dataset Time	13:41	14:37	07:33	49:31	111:23	01:46	09:14	06:48

Table 2:Comparative results from various STL model Vs Traditional machine/deep learning Model in terms of a metric

	RoBERTa	**XLNet**	**DistilBERT**	**ULMFit**	**LSTM**	**Naïve Bayes**	**SVM**	**Logistic Regression**
100% of Dataset Accuracy	0.904	0.870	0.871	0.922	0.945	0.866	0.903	0.842
50% of Dataset Accuracy	0.892	0.863	0.871	0.914	0.949	0.856	0.898	0.834

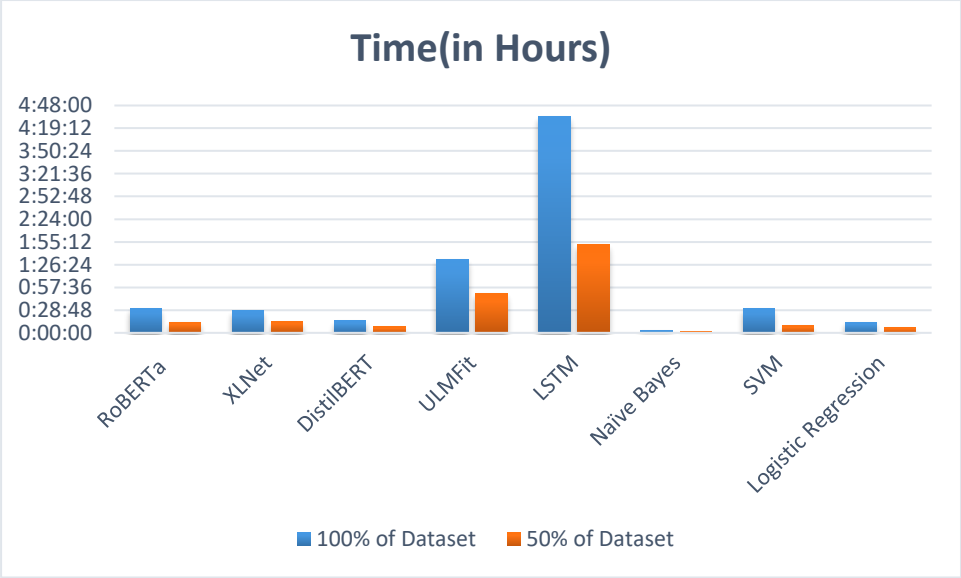


Figure 1:Time Comparison Between Transfer Learning Model and Traditional Model

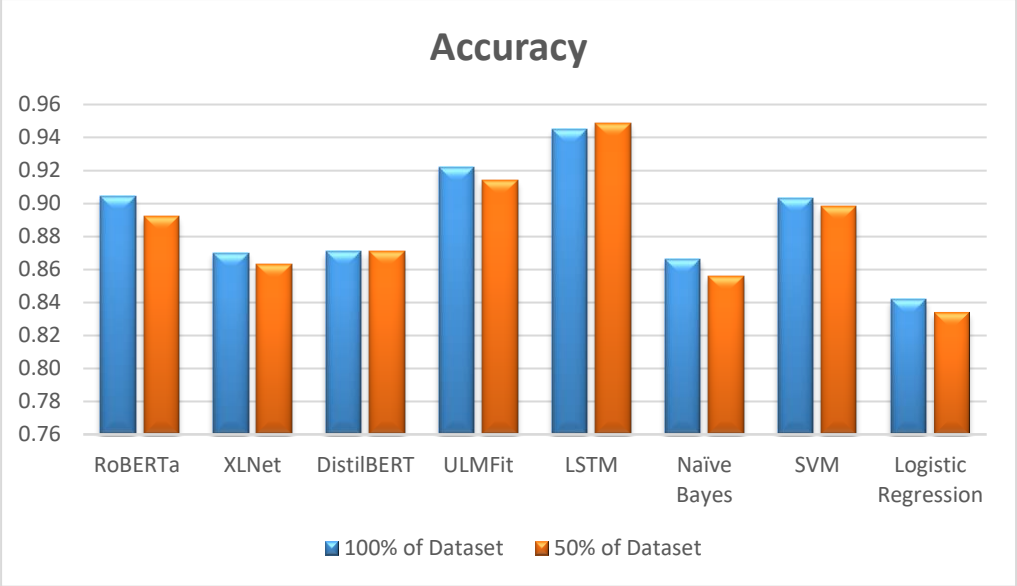


Figure 2 :Accuracy Comparison Between Transfer Learning Model and Traditional Model

2) YELP Review dataset:

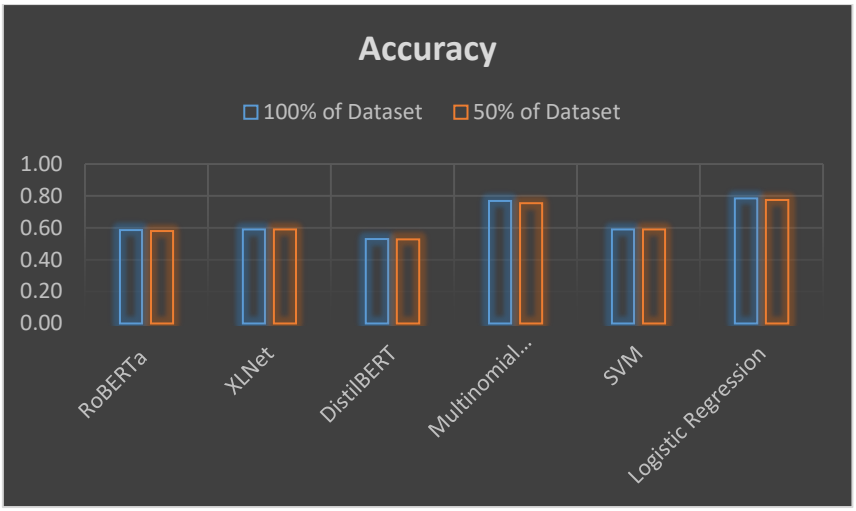
This dataset contains 10000 values which are labelled from 1 to 5 according to user review.

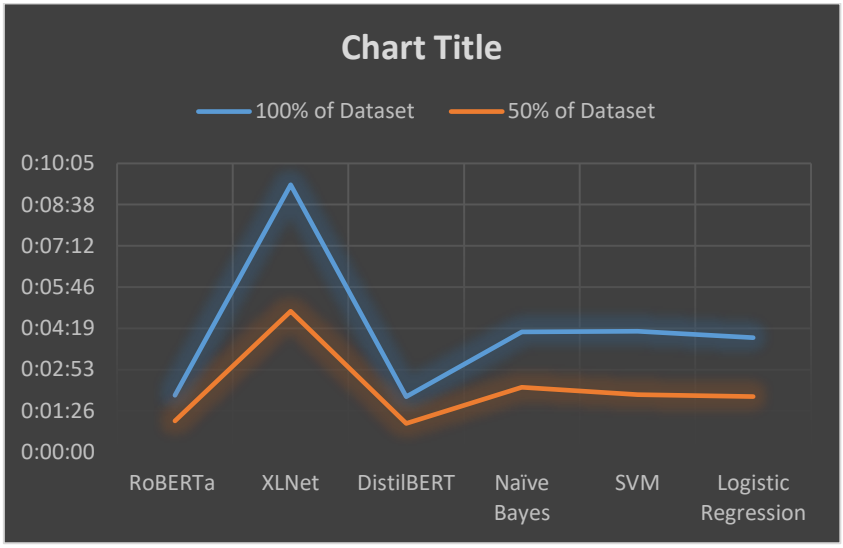
Table 1 : Comparative results from various STL model Vs Traditional machine/deep learning Model in terms of amount of time(Seconds)

(in minutes)	RoBERTa	XLNet	DistilBERT	Multinomial Naïve Bayes	SVM	Logistic Regression
100% of Dataset Time	01:59	09:20	01:56	04:12	04:13	4:00
50% of Dataset Time	01:05	04:55	01:00	02:16	02:16	01:56

Table 2:Comparative results from various STL model Vs Traditional machine/deep learning Model in terms of a metric

	RoBERTa	XLNet	DistilBERT	Multinomial Naïve Bayes	SVM	Logistic Regression
100% of Dataset Accuracy	0.59	0.59	0.53	0.7694	0.59	0.785
50% of Dataset Accuracy	0.58	0.59	0.528	0.754	0.59	0.774





3.2 Sequential transfer learning Model for NER

Named Entity Recognition – Named Entity recognition is a necessary part of NLP tasks such as IR and IE. It is used to find the entity-type of words in a given dataset. This section demonstrates the NER detection using STL's two bench marking models ELMO as well as BERT. In NER, we have used sequential transfer learning models like BERT and Elmo along with the basic models which uses feature vector as well as embeddings (e.g. LSTM, LSTM-CRF, Random Forest Classifier etc.). To facilitate the demonstration the authors used NER dataset which is taken from GMB corpus[44]. This dataset contains four variables called sentence number, words in a sentence (distributed in a row for specific sentence), parts of speech tagging and the Tags (target attribute). Target attribute uses BIO notation ('O' is used for non-entity tokens). ELMO is an approch in transfer learning that is learns from both from front and back using LSTM architectures. To identify the relationships between entities ELMO uses Contextual learning which is better than word Embeddings). Due to above the ELMO learn the word meaning and the context in which word is used that is instead of assigning fixed

embedding the ELMO first looks in the sentence than assigns the word embedding to the word. BERT Model: BERT model was first the unsupervised, pure Bidirectional system used for pre-training model for NLP tasks. Bert uses three types of embeddings for computing it's input representations such as token embeddings, segment embeddings and position embeddings. Since Bert is pre-trained on unlabelled data on different tasks. It can be fine-tuned on labelled data to get desired results. The advantage of BERT is that it was built upon training in contextual representations and was purely bidirectional unlike ELMO and ULMFit which are unidirectional or not completely bidirectional. The steps for detecting the NE using ELMO and BERT is given as a pseudo code in section 3.2.1 and 3.2.2

3.2.1 NER detection using ELMO model

Read the sample from the GNB corpus and pre-process it as required for ELMO,then include the residual LSTM network with ELMO embedding layer,then fine tune and fit the model.

Step 1: Read the data from GNB corpus

Step 2: Pre-process the data as required for ELMO model

Step 3: include residual LSTM network with an ELMo embedding layer.

Step 4.Fit the model

Step 5: Summarize the Model performance.

3.2.2 NER detection using BERT model

Step1: Read the data

Step2 : Pre-process the data as required by the BERT model

Step3: Define data loaders

Step 4: use Bert for token class for tokenization

Step 5 :Fine tune the BERT model by adjusting the parameters

Step 6: Fit the model

Step 7. Evaluate the model

We have used Sequential transfer learning models like RoBERTa, DistilBERT, XLNet and ULMFit as well as Bi-LSTM, Naïve Bayes, SVM, Logistic Regression on the Dataset provided by Stanford i.e. IMDB dataset. We have implemented all these approaches in 100% of the dataset as well as 50% of the dataset to get the results. We have found that transfer learning models give decent accuracy with

reasonable period of training (Time Taken). We can infer the results from the graph which is shown below:

The dataset is Annotated Dataset from Kaggle. To differentiate between transfer learning models with traditional models we have used 100% of dataset as well as 50% of dataset, and came to a conclusion that these transfer learning models gives fair results with adequate amount of time taken.

3.2.2 Results And Discussion:

We have used NER dataset which is taken from GMB corpus. The environment which we have used here is same as we have discussed above for Sentimental analysis. This dataset is annotated and tagged. It is used to train the models to predict Named entities. These entities are:

geo = Geographical Entity

org = Organization

per = Person

gpe = Geopolitical Entity

tim = Time indicator

art = Artifact

eve = Event

nat = Natural Phenomenon

It contains Total word count of 1354149

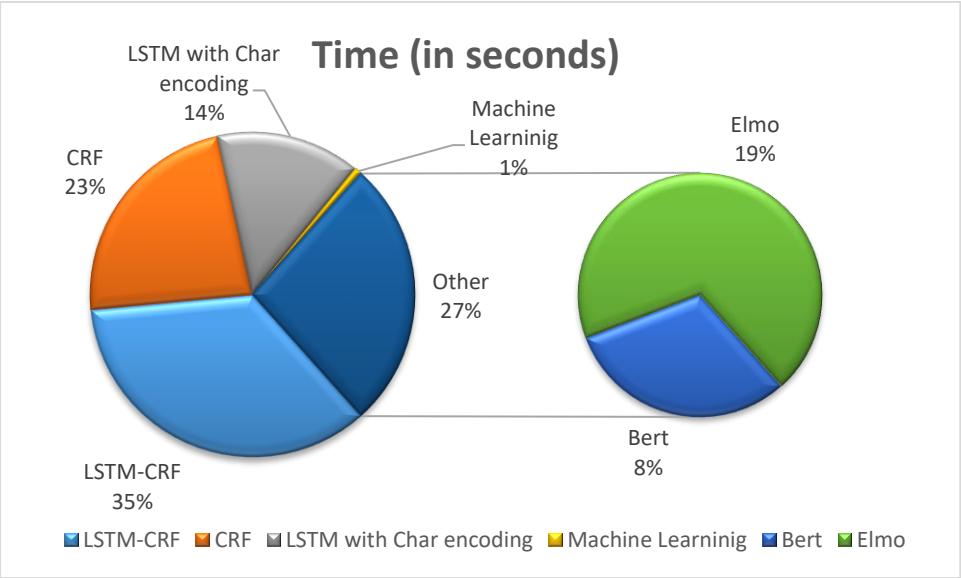


Figure 3:Time Comparison Between Transfer Learning Model and Traditional Model in 100% datasets

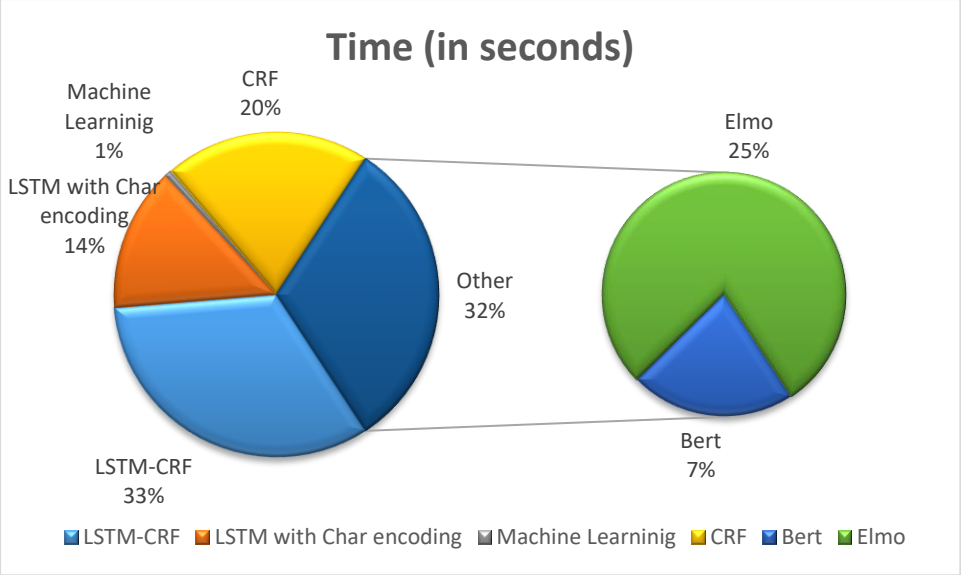


Figure 4:Time Comparison Between Transfer Learning Model and Traditional Model in 50% dataset

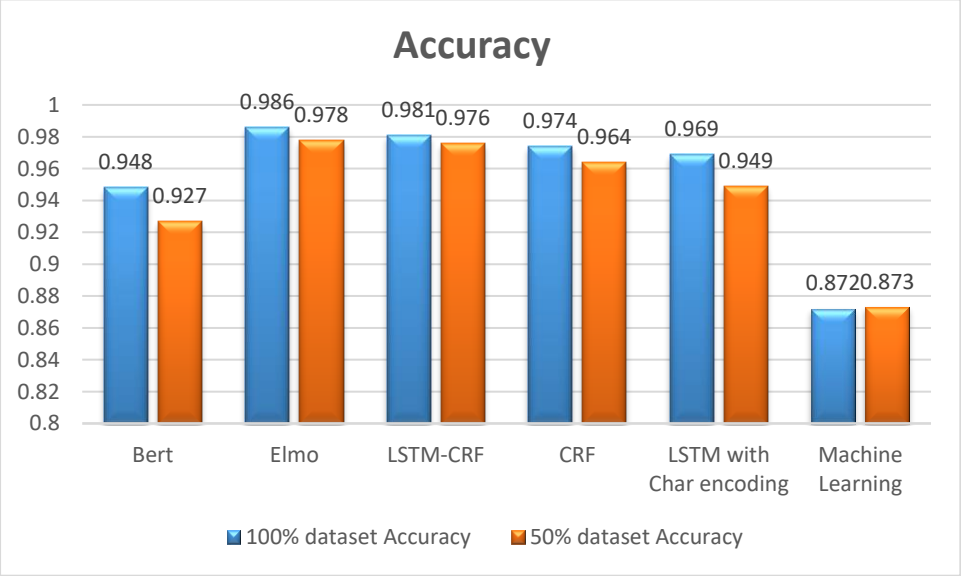


Figure 5:Accuracy Comparison Between Transfer Learning Model and Traditional Model

From above results we can clearly see that transfer learning outperforms the traditional models and it is surely going to be driving industry in future.

Conclusion:

In this chapter, we have evaluated certain Sequential transfer learning models as well as traditional models on basis of two datasets, namely Stanford IMDB dataset and Named Entity Recognition Dataset as stated above. As clearly shown in results and discussion section, these pre-trained models can give high accuracy with less data as well, however traditional models fails to do so. Thus the above results clearly show us that transfer learning is achieving SOTA in NLP. Also we have to tackle the problem of negative transfer learning i.e we have to

carefully see transferability between source and domain. We can one issue that when entire domain cannot be used for transfer learning .Also we can see that see each model is pre-trained to perform some tasks and not all the task in NLP.