AN ANALYTICAL STUDY OF PRE-TRAINED MODELS FOR SENTIMENT ANALYSIS OF SINHALA NEWS COMMENTS

M. Lishani Sadna Dissanayake

189314L

M.Sc. in Computer Science

Department of Computer Science and Engineering

University of Moratuwa.

Sri Lanka

May 2018

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

---------- ----------

M. Lishani Sadna Dissanayake Date

The above candidate has carried out research for the Masters thesis/ Dissertation under my supervision.

---------- ----------

Dr. Uthyasanker Thayasivam Date

ABSTRACT

In the area of natural language processing, due to the large-scale text data availability sentiment analysis has become a prevalence topic. Sentiment analysis is a text classification which is mainly focusing on classifying recommendations and reviews as positive or negative. Earlier for this classification task, most of methods require product reviews and label them. Using these reviews then a classifier is trained with their relevant labels. For this training procedure a huge number of labeled data is needed to train these classification models for each of the product, considering the facts that the distribution of the reviews can be different between different domains and to enhance the performance of these classification models. Nevertheless, the procedure of labeling the data is very expensive and time consuming. For low resource languages like Sinhala language, the existence of annotated Sinhala data is limited compared to the languages like English language. The need of applying classification algorithms in order to perform sentiment classification for Sinhala language is challenging. Apart from applying traditional algorithms to analyze sentiments, here using pre-trained models(PTM)s, experimenting on whether the outcome of these experiments outperform the traditional methods. In natural language processing, PTM is performing an important role, since it paves the way for applying PTMs for downstream tasks. Therefore, this research takes the step to applying PTMs such as BERT and XLnet to classify sentiments. Experiments have been done using two approaches on BERT model as fine tuning the BERT model and feature based approach. Also using the existing Roberta-based Sinhala models, named as SinBERT-small and SinBERT-large which are available in Huggingface official site which have trained using a large Sinhala language corpus.

ACKNOWLEDGEMENT

First and foremost, I would like to convey my deepest appreciation to my supervisor Dr. Uthyasanker Thayasivam for his continuous support, invaluable advice, determined efforts and assistance. I have been immmensly fortunate to have him as my supervisor, whose plentiful experience and immense knowledge have guided me to making this research a success.

Dr. Surangika Ranthunga has also guided me in some things and therefore I would like to express my gratitude towards her.

My heartfelt appreciation is given to all my friends for their assistance and encouragement given to me during this hectic and challenging endeavor.

Finally, none of this would have been achievable without the love and patience of my parents. They have been a constant source of love, strength, concern and support all these years. I am grateful to my parents for their tremendous understanding, motivating and being patient and supportive during this critical period in my academic life, which motivated me to complete the research successfully.

Table of Content

List of Figures

List of Tables

List of Equations

1 Introduction

1.1 Background

Recently, sentiment analysis is setting off as a considerable research area, because of the rapid increase of information for opinions which are available on the internet [1]. Simply, sentiment analysis can be described as the method of determining opinions, sentiments, emotions and attitudes towards a product, services, organizations, services etc [1]. Sentiment analysis also can be defined as opinion mining, evaluation extraction and emotional polarity judgement. Nevertheless, most of the machine learning techniques work properly according to the assumption where train data and test data are in same domain. If they have different domains which means not in same distribution and same feature space, then one such built statistical model may not work well for another domain. Therefore, using training data which are gathered freshly, from the scratch the statistical models need to be rebuilt. This is where the transfer learning approaches can be useful since the cost for re-collecting data and rebuild models can be reduced. Knowledge transfer or transfer learning is beneficial in the area of sentiment analysis which leverage the existing knowledge to solve different problems in different domains [2]. In relation to transfer learning later came pre training models.

1.2 Research Problem

Sentiment analysis mainly targets on predicting sentiment polarity as positive or negative automatically. Using traditional classification algorithms, the labeled text data can be trained to perform the sentiment classifiers manually. The process of labeling the text data is expensive and time-consuming to build accurate sentiment classifiers. On the other hand, in different domains most of the time users tend to use different words when they want to convey their sentiment. If we try to use a classifier directly to other

domains which was trained in another domain, then the applied classifier may not perform well because of the differences among these domains [3]. Therefore, the question which rise is that whether we can use those statistical models which have built from a certain domain to another domain [1]. When it comes to other low resource languages it is again a challenging to do sentiment classification. Since it is difficult to find large datasets with labeled texts and therefore need to infer sentiment from a small dataset. Then there is another question whether the results will be better due to the experiments are done on small datasets.

1.3 Research Objective

Huge amount of user generated sentiment data is existing on the Web with the explosion of Web 2.0 services. These data can be found as user reviews in opinion sites or in purchasing sites or in blog posts or customer feedback [4]. Sentiment analysis, is studied broadly since most of the time users don't give their sentiment directly, hence the sentiment needs to be predicted from the data indicated by users [4]. Algorithms used for supervised machine learning have used extensively and it has been proven that these algorithms give promising results in sentiment analysis [5].

Nevertheless, the performance of these machines learning methods require to provide labeled training data which are labeled manually. Moreover, these methods are domain dependent. Therefore, the objective of this research is to come up with an effective approach for transfer learning. Furthermore, after looking at related work which have done for English language and then to find out whether we can apply these approaches for Sinhala language which is a low resource language [5].

2 Literature Survey

Sentiment analysis has been performed on many languages and it has been an in demand research. Therefore it leads to finding out what are the efficient methods to carry out the experimental methods. Earlier approaches using traditional statistical models have been widely used. But after the concept of transfer learning it paves the way pre pre-trained models. Therefore, before taking into consideration the pre-trained model, it will be beneficial to look at what transfer learning is and how we can leverage transfer learning techniques. There are researches in the area of transfer learning since 1995 and it has different names such as life-long learning, learning to learn, inductive transfer, knowledge transfer, multi-task learning, context sensitive learning, knowledge consolidation, meta-learning, incremental/cumulative and learning knowledge-based inductive bias [2]. In the year of 2005, a novel operation of transfer learning has been given the Broad Agency Announcement (BAA) 05-29 of Defense Advanced Research Projects Agency (DARPA)'s Information Processing Technology Office (IPTO). Whether a given particular system can identify and put in the learned knowledge and skills to current tasks which have learned from previous tasks. They are defining that the transfer learning targets on extracting the learned knowledge from one or more source domains and then applying that learned knowledge to another target domain. When comparing with the mulit-task learning, transfer learning mainly focuses on the target task where the multi-task learning focus on learning from each and every source and target tasks at one time. In transfer learning, parts of target and source tasks are no more symmetric [2].

The comparison among the learning methods of transfer learning and traditional approaches are shown in Figure 1. According to the following figure, transfer learning approaches attempt on transmitting the learned knowledge from previous domain to another target domain where the last-mentioned part is consisting of a small amount of high-quality train data. On the other hand, traditional machine learning approaches attempt on learning each task from scratch. [2].

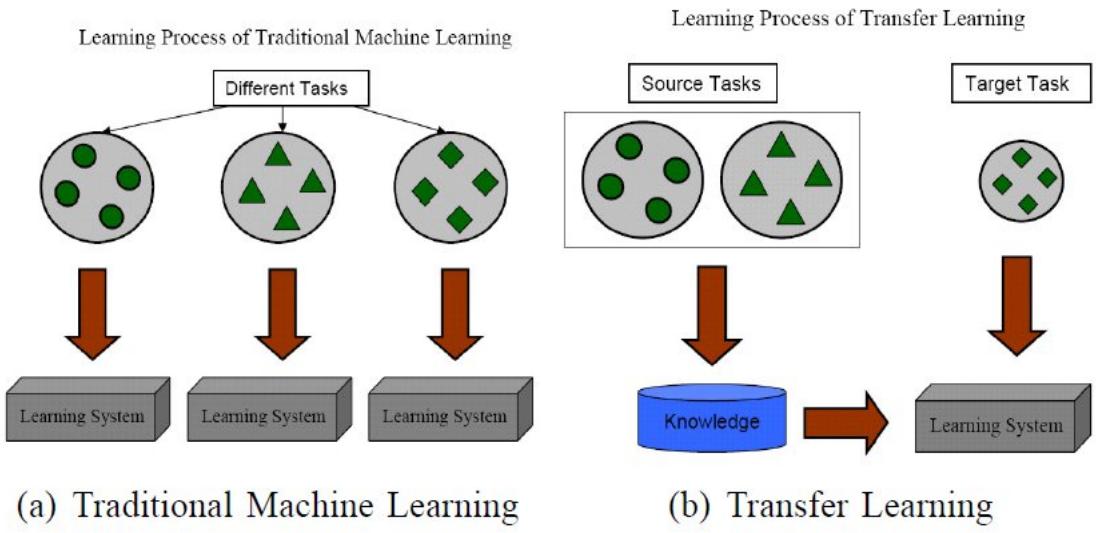

Figure 1: Different Learning Approaches, for Transfer and Learning Traditional Machine Learning. Adapted from [2]

When looking at the definition of transfer learning, in a literature survey it has summarized the association among various transfer learning and traditional machine learning settings are shown in Table 1, where they have categorized the transfer learning under three sub-settings, unsupervised transfer learning, inductive transfer learning and transductive transfer learning considering various circumstances among the source domains and target domains and also between source tasks and target tasks.

Learning Settings		Source and Target Domains	Source and Target Tasks
Traditional Machine Learning		the same	the same
Transfer Learning	*Inductive Transfer Learning* / *Unsupervised Transfer Learning*	the same	different but related
		different but related	different but related
	Transductive Transfer Learning	different but related	the same

Table 1 Relationship among Various Transfer Learning Settings and Traditional Machine Learning. Adapted from [2]

The association between the transfer learning settings and their relevant fields are showed in Table 2 and Figure 2.

Table 2 Different Settings of Transfer Learning. Adapted from [2]

Transfer Learning Settings	Related Areas	Source Domain Labels	Target Domain Labels	Tasks
Inductive Transfer Learning	Multi-task Learning	Available	Available	Regression, Classification
	Self-taught Learning	Unavailable	Available	Regression, Classification
Transductive Transfer Learning	Domain Adaptation, Sample Selection Bias, Co-variate Shift	Available	Unavailable	Regression, Classification
Unsupervised Transfer Learning		Unavailable	Unavailable	Clustering, Dimensionality Reduction

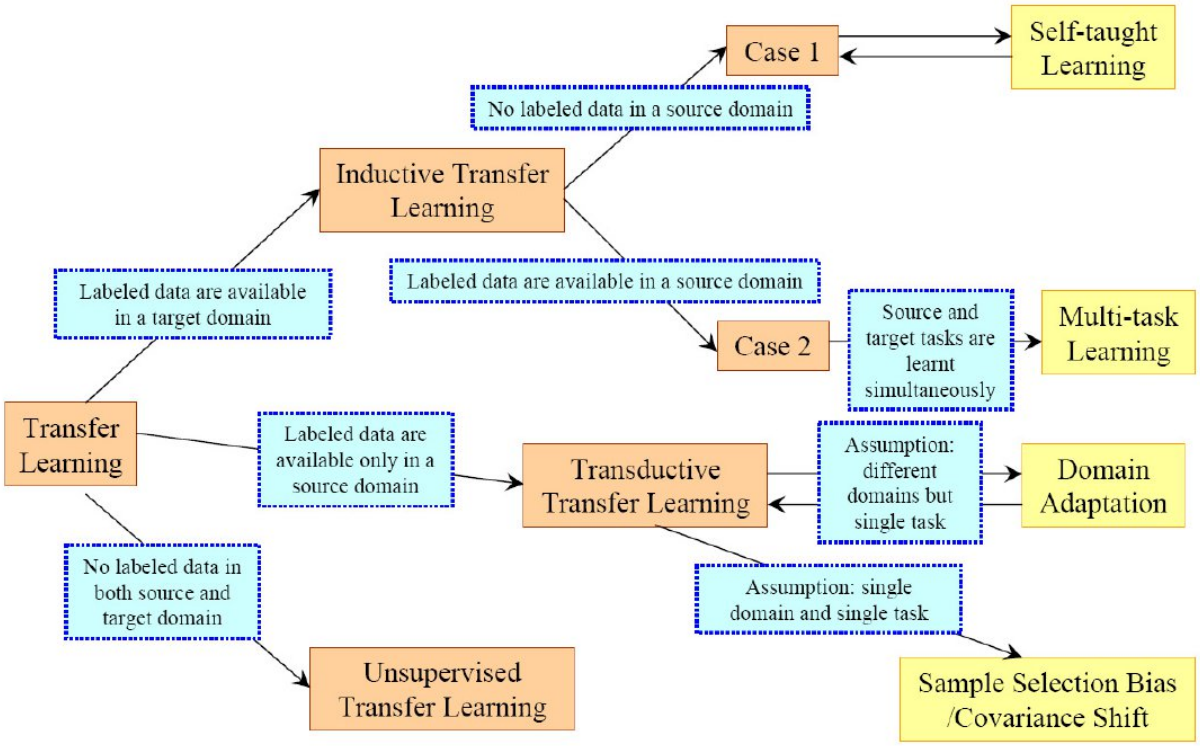

Figure 2 An Overview of Different Settings of Transfer Learning. Adapted from [2]

In the area of Natural Language Processing (NLP) there are large amount of work carried out using algorithms for domain adaptation which is a sub categorization of transfer learning. Most of the work which were conducted earlier the data of source domain was considered being "prior knowledge". Then trained a model for the target domain data applies maximum a posterior (MAP) estimation under this prior distribution [6]. Likewise, porting a parser to a new domain for which there is small or no annotated data, the enhancements that need to be applied will be enormous. Like active learning, model adaptation paves the way for reducing the number of annotations which is essential for a better performance. As a matter of fact, active learning combining with MAP may decrease the needed number of annotations [6]. Additionally, though the environment of a maximum entropy (ME) model uses source domain data when estimating prior distribution. Recently the ME model has been experimented where a mixture model is introduced for domain adaptation to learn variations among domains [7].

2.1 Sentiment Analysis

Leveraging many techniques together with semantic orientation and machine learning approaches, sentiment analysis can be carried out. When accomplishing these kinds of problems using existing models which uses machine learning techniques, the data that requires to get inserted into these models are labeled for its relevant sentiment before classifying the data. After annotating the input data, leveraging the machine learning approaches like Naïve Bayes, SVM or neural network which are supervised, then the sentiment analysis can be performed. In this approach in order to train a reasonable model it is essential for a substantial annotated corpus.

However, unsupervised classification does not require a training dataset. Turney [4] introduced an unsupervised classification technique that classifies reviews as recommended (thumbs up) or not recommended (thumbs down). Proposed solution contained three steps. First step was to extract two-word phrases from reviews which conforms to a given pattern. This was done using Part of Speech (POS) tagging. Table 2 lists the used patterns. If we take third pattern of Table 3 as an example, it indicates to extract all the phrases which the first (JJ) and second (JJ) words are adjectives and third word is not a noun (NN). Step two was to estimate Sentiment Orientation (SO) of the extracted phrases using Pointwise Mutual Information (PMI). PMI between two words is given by Equation 2. Here p (word1 & word2) refers to the probability that word1 and word2 co-occur. Therefore, PMI calculates the degree of statistical dependence among two words. Sentiment Orientation is calculated using PMI and two reference words. 'Excellent' and 'Poor' was selected as reference words as in 5-star rating scale they refer to the extreme cases of positive and negative cases.

Step 3 calculates the average Sentiment Orientation of all phrases in review. Algorithm classifies a review as positive if its average SO is positive and negative otherwise. Evaluation results shows classifier accuracy in between 65% to 85% in various application domains.

Lin et al [36] introduced a fully unsupervised method (joint sentiment/topic-JST) based on probabilistic modeling. Model employed both sentiments and topics to classify sentiment orientation of a document. JST was extended by adding a sentiment layer to the state-of-the-art topic classification model, Latent Dirichlet Allocation (LDA). Accuracy was further improved by using various sources of prior information. Preprocessed movie review dataset was used for the evaluation. Evaluation results shows accuracy values up to 85% which was very close to supervised approaches. It is was identified that use of prior information such as Mutual Information increased accuracy values significantly (up to 15%). Turney et al [37] introduced another unsupervised training method which used Pointwise Mutual Information to find out sentiment orientation. PMI was calculated using intuitively chosen seven opposing word pairs (seven positive and seven negative words). For evaluation they employed a corpus of one hundred billion words with a test word set of 3596 words (1614 positive, 1982 negative). They were able to achieve accuracy of 80%.

Even though unsupervised learning techniques are flexible [36] than their counterpart, they generate generalized models which does not fit well for a specified problem. Most of the time unsupervised techniques can be improved using prior information, making them semi-supervised or supervised. Furthermore, unsupervised techniques have poor performance compared to supervised learning techniques.

Mainly sentiment analysis is concerned towards getting insights from product reviews given by customers that can be a help for businesses to expand their customer satisfaction, to more considerate about their product quality and to reach out for many customers. Although several firms do not tend to use real-time existing tools of sentiment analysis, the majority of large tech firms are making the most from the significance of sentiment classification engines already. Researches have been done to study the many approaches to execute tasks of sentiment classification using a collection of datasets which can be accessed from public. Nowadays internet has become as something that we cannot live without and it has important part in our day to day lives, now the people can get notifications regarding medical advices through various online platforms, articles etc. Moreover, the Internet sometimes misleading information to scam people through scammers and therefore some might declare that the internet isn’t the

perfect platform. Hence there has to be a better measurement to direct the quality of the services for a better service. Using the given reviews and ratings sentiment analysis here are also focusing on implementing better sentiment analysis models.

2.2 Related Work

2.1.1 Sentiment Analysis Using Transfer Learning

There are more techniques which have been done relying on data to conquer the distribution of features differences among different domains which are not labeled in the target domain. Influencing with alternating structural optimization (ASO) algorithm which is an algorithm mostly used in multi-task learning. One of the main problems in machine learning is whether we can enhance the performance of a supervised learning algorithm making use of unlabeled data. Techniques which leverage both unlabeled and labeled data are usually indicated as semi-supervised learning. Several such methods have proposed in the recent time and in such methods trying to find a suitable way for their effectiveness. Some researches investigate a similar problem that make the way for a new procedure to semi-supervised learning. Specifically, one such research considers from multiple learning tasks, learning predictive structures on hypothesis spaces which means, finding the classifiers that give better predictive results. They propose a common framework where the problem of structural learning be able to formulate and theoretically examined and associate it to learning using unlabeled data. Using this structure, algorithms are proposed for structural learning and can be investigated for computational issues. Through their observations they are demonstrating in the semi-supervised learning setting the successfulness of the proposed algorithms [8].

Another research has been proposed using structural correspondence learning (SCL) for domain adaptation in sentiment classification focusing on online reviews for different types of products. In this research they are addressing two questions which are important in domain adaptation. First, they have showed that they can notably using the

structural correspondence learning Blitzer model for a provided source and target domains. For connecting the source domains and target domains SCL selects a set of pivot features. Source domain consists of labeled data and target domain is consisting of unlabeled data. These pivot features are chosen considering the mutual information with the source labels and the common frequency in source domains and target domains. Moreover, they revealed a method of correcting the structural correspondence misalignments by making use of only a little number of labeled target domain data. Afterwards, they were providing a procedure in order to choosing the source domains which are more suitable to adapt well for specified target domains. Due to adaptation the unsupervised A-distance measure of separation among domains with loss correlates better [9].

There has also been another work in investigating carefully domain adaptation in structuring of features. They have suggested a function that can be mapped to a high-dimensional feature space for both source and target domains data called as kernel-mapping function. Using the same domain those data points have taken as twice as closely related as the data points from various domains [11]. In another research they are scrutinizing a new machine learning approach named as translated learning. But this research has been carried out on image classification. For translated learning an instinctive idea is translating the training data in to a target feature space, where learning is experimented through a single feature space. This kind of researches have already been demonstrated in cross-lingual text classification successfully in several applications. Using this translated learning they are proposing a procedure which utilizes a language model. Through this language model class labels are linking in the source spaces to features and then again translated in the target spaces to the features. By tracing back to the instances in the target spaces this series of connection is accomplished. Making use of Markov chain and risk minimization, they are revealing that this can be modeled which is the path of linkage [12].

Further there's another research which have come up with a strategy to learn an Eigen feature representation utilizing spectral learning theory from a task graph feature representation, class labels and instances [14]. Through this research they are introducing a framework on general transfer learning which can be modeled for various

problems of transfer learning such as self-taught learning and cross domain learning. Through their structure, implemented a task graph, as the initial step to act as the transfer learning task. Afterwards, based on spectral learning theory acquire a set of Eigen vectors which is reflecting the inherent framework of the task graph. The obtained eigenvectors are using as novel features that the knowledge will be transferring from supplementary data to assist the process of classifying the target data. For the task of target transfer learning, EigenTransfer have the capability of bringing out a transfer learner appropriately, for a given a task of transfer learning and a random learner such as SVM which is non-transfer leaner. To demonstrate the unifying ability on Eigen Transfer, they are experimenting on various transfer learning tasks as self-taught learning, cross-category learning and cross-domain learning. Through their experiments they showed that the Eigen Transfer is outperforming some representative non-transfer learners [14].

In a similar way, another research has constructed a bipartite graph using the spectral feature alignment (SFA) algorithm to represent the co-occurrence association among domain-independent words and domain dependent words. SFA algorithm making use of some domain-independent words as a connection between domains. Feature clusters have been produced using co-align domain-independent words and domain-specific words. Like this method, the clusters have utilized to minimize the space among domain-specific words for a given two domains. Using these cluster train sentiment classifiers can be trained accurately in the target domain [15]. The approach of graph-based has also been investigated in which a graph is constructed with nodes indicating edges and documents. Also indicating content similarity among provided documents. Proposed algorithm allocates a score for each and every predicted document and it calculates the score recursively using the correct labels in old domain data, and also the "pseudo" labels in new domain data. Eventually, in this method the new domain data are labeled as "positive" or as "negative" considering the score. Until it converges from its nearest documents the correct labels of source domain documents and also the "pseudo" labels of target domain documents [16]. Later the mentioned procedure has been enhanced by scrutinizing associations among documents and words originating at source domains and target domains simultaneously. This technique can fully utilize the

reinforcement among words and documents by combining four kinds of associations among the documents and words. First, they have constructed three graphs considering the above associations individually. Afterwards, for every unlabeled document they have assigned a score to indicate its scope to “positive” or “negative”. Subsequently using the graphs calculate the score iteratively. Eventually, when the algorithm intersects the final score for sentiment analysis have obtained, then based on the given scores, they are labeling the target domain data [17].

At the same period in another research, they have addressed the problem where the class label of provided input data from predictive distribution of the different domain and suggested Predictive Distribution Matching SVM which learns in the target domain a robust classifier. This is achieved by using the labeled data from applicable regions of multiple sources only. To recognize the regions of relevant source labeled data iteratively, build a k-nearest neighbor graph. Through this graph the predictive distribution lines up with the target data maximally. To use these applicable sources labeled data for training the target classifier, proposed a predictive distribution equal regularization. Additionally, to deduce the label of unlabeled target data, progressive transduction is acquired for approximating the predictive distribution of the target domain [18].

A novel approach using Bayesian probabilistic model have been proposed to control different source and target domains. Using this model every word is related with three factors. These factors are domain dependence/independence, Domain label and word polarity. Furthermore, from this research they have obtained an algorithm which is efficient for deducing the parameters of the model leveraging the Gibbs sampling, from both labeled texts and unlabeled texts. They are demonstrating the effectiveness of the built model in a document polarity classification task leveraging the real data. This model is comparing with an approach not taking into account the differences among the domains. Furthermore, from their method they can show if each word’s polarity is domain-independent or domain-dependent. Because of this feature for each domain a word polarity thesaurus can be built [1].

In real-time web applications enables live discussions. Real time sentiment classification is known as the ability analyzing user sentiments and opinions automatically as discussions develop in these real time web applications. Nevertheless, this process is leading to many challenges. One such challenge is that the requirement of handling the highly dynamic textual data which is distinguished by its subjective meaning and changes in vocabulary. Another challenge is that the limited number of labeled data which is required to assist supervised classifiers. From aforementioned research, they have proposed a technique to carry out real time sentiment classification for transfer learning. They are trying to recognize a process of opinion holder bias prediction that is firmly associating with the sentiment classification. Nevertheless, when comparing with sentiment classification, it constructs models which are accurate because of the underlying relational data considers a stationary distribution. To predict content polarity rather than learning textual models for example, the procedure of the traditional sentiment analysis, as the initial step over an users connected network associated with endorsements such as Twitter retweets resolving a relational learning task, toward a topic, they are measuring the bias of users of social media. Then transferring user biases to textual features, they are analyzing sentiments. This technique performs well since while old terms may change their meaning and new terms may appear. As a fundamental thing of human behavior over time user bias tends to be more consistent. Consequently, they have adopted user bias for constructing accurate classification models as the basis [19].

Apart from the textual sentiment analysis recently, images and videos are increasingly used by users in social media to share their experiences and convey their opinions. Because of this vast-scale visual data can assist on extracting user sentiments better toward topics or events, like in image tweets. Therefore, predicting sentiment from visual data is supportive to textual sentiment classification. Using Convolutional Neural Networks (CNN) the requirement of making use of wide ranging and at same time noisy training data to resolve the very challenging issue of image sentiment classification. For image sentiment classification first, they have designed an appropriate CNN architecture. To label Flickr images by utilizing a sentiment algorithm, they have obtained half a million training samples. They perform a progressive strategy to fine-

tune the deep network to leverage these labeled noisy machine data. Moreover, by leveraging domain transfer using only a small number of manually labeled Twitter images they enhance the performance on Twitter images. They have experimented extensively on labeled Twitter images. Results of the experiments are showing that the suggested CNN architecture can perform more in image sentiment classification than other algorithms manually [20].

A sentiment classification method has been proposed which is applicable when there are some labeled data for other multiple domains and have no labeled data for a target domain, which are selected as the source domains. They have implemented a sentiment sensitive dictionary making use of both data that are labeled and unlabeled by taking different source domains to automatically identify the relationship among the words which convey similar sentiments from various domains. To train a binary classifier to expand feature vectors the built thesaurus is used [27].

Based on the subjective knowledge conveyed in the reviews, the main objective of sentiment analysis is to recognize the sentiment polarity of the reviews as positive or negative. Normally, to train a classifier, almost all learning procedures require labeled data. Nevertheless, acquiring labeled data from each domain is not practical and allocating labels for every domain is cost effective and time consuming. Furthermore, the classifier may not perform well which is trained in one domain and then applying to another domain. Therefore, to resolve this issue, utilizing dual transfer learning that learns both conditional and marginal distributions of features from source and target domain, a technique have been introduced which develops the cross-domain sentiment analysis framework. This technique is considering joint nonnegative matrix tri factorizations (NMTF). Using the decomposed latent factors which exhibit the duality property the two distributions are learned [28].

2.1.2 Sentiment Analysis Using Transfer Learning for Other Languages

Apart from the English language sentiment classification have been carried out on other languages such as Japanese, Chinese, Russian, German and Hindi. When applying classification on other languages some researches have used new approaches by utilizing the features of native language.

One such research which have been performed for Russian language and they have used Bagging, Naïve Baise classifier, Support Vector machines (SVM), POS tags and d-grams. From their experiments they are reveal that the performance of SVM is better than Naïve Bayes from a small-scale margin. They also mentioned that the lemmatization has huge impact for Russian language but not for English language [37].

Another experiment has been carried out for German language as well as for English language using an architecture based on Convolutional Neual Network (CNN) and Long Short-term Memory (LSTM). They are showing that their proposed model performs well when comparing with models that are existing recently. Moreover, they are concluding that without reducing the performance this model can be used for other languages or for mixed languages [38].

For Hindi language also there are researches that have been carried out for sentiment analysis. One of the research projects have taken three approaches and then the performances of these strategies were compared. The three strategies are, using a Hindi corpus to train a classifier to construct a classifier and then classify a given Hindi document, translating a Hindi document to English and the train a model and last approach is utilizing a majority-based classifier for Hindi SentiWordNet. Out of these approaches they are showing that the first approach outperforms other approaches [39].

Another research has performed using a Hindi lexicon by projecting SentiWordNet's synsets into Hindi language. For various features classification have been performed with stemming and using n-grams. They pointing out that there were errors when translating and because of that reason their model didn't perform well [40].

2.1.3 Sentiment Analysis Using Transfer Learning for Sinhala Language

When considering about sentiment analysis, lots of researches have been done for English language using transfer learning methods. But for Sinhala language which is a morphologically rich language with an under-resourced language, there isn't any researches which have used transfer learning methods for sentiment analysis.

There are researches which have carried out using deep learning techniques for Sinhala language. One such research performed sentiment analysis using more state-of-the-art models like capsule networks and hierarchical attention hybrid neural networks and also models like LSTM, RNN, Bi-LSTM. The text data have been collected from online websites and classified them as POSITIVE, NEGATIVE, CONFLICT and NEUTRAL classes. The dataset consists of 15059 news comments which are annotated as above-mentioned classes. For this research we are using this text data since we can then compare our implementation with their work [32].

There is another research which is much similar for the above-mentioned research but doesn't experiment using more techniques as the above research. The sentiment classification is performed on Sinhala language using deep learning techniques such as LSTM, CNN+SVM, logistic regression, Naïve Bayes, random forests, SVM and decision trees. Using Sinhala word embedding models these models were trained consequently no language-specific features were used. Comprehensive research has been done making use of models which are different regarding the dimensionality of the effect of punctuations and word embeddings [33].

As the first research that have been carried out sentiment classification on Sinhala language could be considered as the research with making use of document term frequencies which use an easy feed forward neural network [34]. The same author did another experiment for Sinhala language sentiment analysis using three different novel techniques. One such experiment is to retrieve cross linguistic features which are associated with the sentiment of Sinhala language. For this methodology used a bilingual dictionary of Sinhala and English. Linguistic features specific for Sinhala Sentiment analysis have been introduced by further analyzing this experiment. Then classification algorithms and techniques have applied for the lexicons which have generated from the

earlier steps. Support Vector Machines and Naïve Bayes machine algorithms have utilized statistical classification algorithms [35]. There exists another research for Sinhala language which uses a semi-automated approach which is based on a sentiment lexicon generation [36].

2.2 Sentiment Analysis Using Transformer Models

2.2.1 Transfer Learning and Transformer Models in NLP

2.2.1.1 Attention

Among the existing models for sequential data RNN model is most popular for analyzing text and also has been a huge success for numerous other Natural Language Processing tasks. Moreover, for autoencoders encoders and decoders of the main body, RNN models is used which is mostly used for tasks like language translation. Nevertheless, for the same task training process cannot be parallelize using these RNN based models. Even the testing environments have large memory and computing power it is difficult to making use of RNN efficiently, that can be used to decrease the training time cost. Consequently, during training RNN models can be cost inefficient specially when training tasks of language models which are large models. Furthermore, if the sequential data is too long then the RNN model will suffer. Because of the dependencies with the sequential data and from these lengthy dependencies, it will be hard to learn for RNN models. As a solution for this problem methods have been proposed such as LSTM, but still these models can be further improved. To handle this lengthy dependency issue soft memory-based attention has been introduced [41].

As shown in the figure 3, using attention technique the encoder will have the access to previous decoder output, weighted sum of all hidden states and also for the previous decoder hidden state.

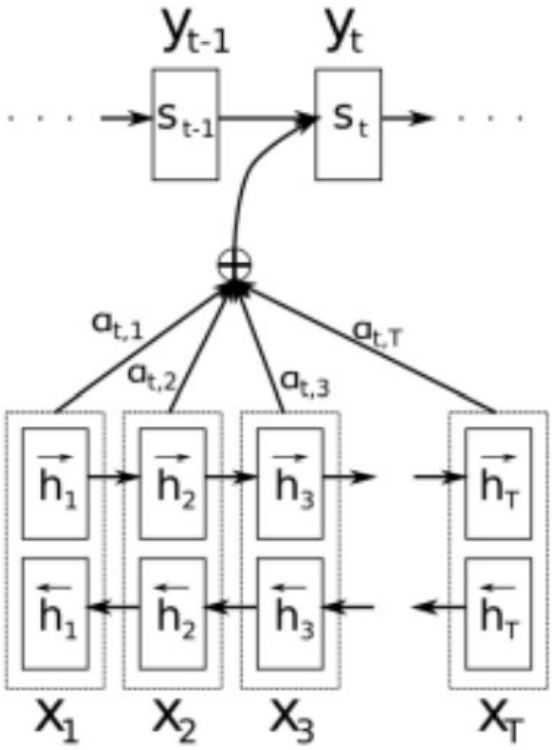

Figure 3 Graphical illustration of attention. Adapted from [41]

As shown in figure 3, in the attention technique allows the part of encoder to have the accessibility to previous decoder hidden state, previous decoder output and the weighted sum of hidden states of all encoders. It has allowed the decoder to gain more knowledge in a sequence when predicting the target word. During the prediction y_t means y at time t and it's depending on preceding outputs < y_t and also in all hidden layers, the weighted sum is taken from the layers. The layers are s_t which is the decoder hidden state and c_t is the encoder. Correspondingly, the mathematical formula for s_t, y_t and c_t can be shown as (5, 1), (5, 2) and (5, 3).

$$p(y_t | y_1, \ldots, y_{t-1}, x) = g(y_{t-1}, s_t, c_t) \quad \text{-> (5.1)}$$

Equation 1: Formula for y at time t

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \quad \text{-> (5.2)}$$

Equation 2: Formula for st

$$c_t = XNi\, \alpha_{ti} * h_i \quad \text{-> (5.3)}$$

Equation 3: Formula for ct

Leveraging the attention mechanism, the decoder can memorize or in other words for a dependency of long-term input, giving attention is enhanced successfully. In the area of NLP these discoveries have assisted on various applications, for example language translation since it has a critical problem of long dependencies.

2.2.1.2 The Transformer

There may exist many kinds of long-term dependencies for example, dependency between the sequences of input and output, dependency between the input sequences coming to itself and dependency between outputs sequences itself. These types exist when using any auto encoders when modeling a sequential data. From the attention mechanism which is explained in the earlier section attempts on resolve only the dependency between input and output sequences. Therefore, autoencoders which are based on RNN are suffering due to long-term dependency between the sequence output and sequence input which contain tokens itself also which are used by traditional attention. Additionally, from parallelizing the process of running between multiple resources, it is difficult to quicken the training. It happens due to the fact that the encoder and decoder has been blocked by used RNN units, where the each hidden unit output is depending on the previous output unit. Therefore, issue of parallelization has been there in all encoder-decoder models which leverages transitional attention techniques.

Inspired from earlier mentioned autoencoders challenges in attention mechanisms, then the transformer model has proposed to overcome these challenges. Transformer model leverages self-attention mechanisms only without RNN based encoder-decoder structure. This paves the way for the parallelization of resources in the model to training and because of that effectively decrease the training time. Moreover, all types of long sequence dependency issues are resolved as mentioned above.

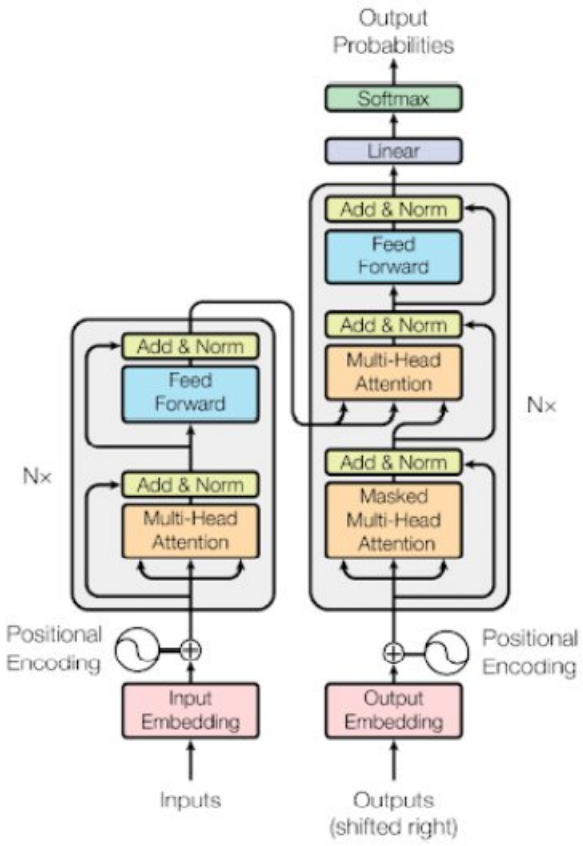

Figure 4: Architecture of transformers. Adapted from [53]

2.2.1 Sentiment Analysis Using BERT

Research has done using social media data, to detect and track the trending social media events and topics which is beneficial to found out many unanswered questions. For this topic, detection there are applying a transformer combined incremental community detection algorithm which is combining of different modules including BERT, multimodal named as entity recognizer and graph strategies. This is a graph mining technique which increases the results of the topics using a simple structural rule. There proposed solution is showing a higher precision and recall when comparing with other procedures [31].

On the other hand, there are some researches which have done using transfer learning-based approaches for other languages. One such research has investigated on Japanese language using transfer learning methods such as ELMo, ULMFiT and BERT which uses pre-training a language model through an unsupervised manner. They are concluded as the performance has increased of techniques in transfer learning than the models which trained as task specific models on three times as much as the data [42].

For the Persian language sentiment analysis have performed using BERT, version of un-normalized multilingual model which is developed for 103 languages. They have showed that their BERT model outperformed skip-gram LSTM or CNN [51].

Another research has been carried out for Italian language using BERT for sentiment analysis. They have also showed that their fine-tuned BERT model performs well than the other state of the art systems [43].

2.2.1.1 BERT

Bidirectional Encoder Representation of Transformers is shortened as BERT. It has been implemented by developers at Google AI Language and it's an unsupervised language representation model which is based on deep learning. As the firstly implemented unsupervised language model which is deeply-bidirectional is BERT. Previous language models, before BERT, knowledge is acquired by text sequences from either combined right to left or left to right contexts. Hence, these models in all layers are not bidirectional. As seen in the following diagram, when comparing with other language models, BERT consists of bidirectional architecture..

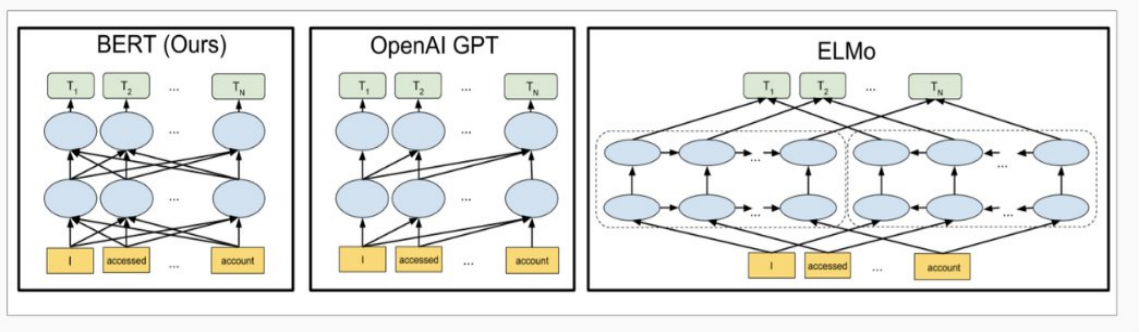

Figure 5: BERT Achitecture compared with other models: Adapted from [45]

In learning representations, BERT is associated with deep bidirectionality leveraging a new method named as Masked Language Model (MLM). This approach of deep bidirectional paves the way to take in words and also the context of words which can be taken from both sides of left words and right. When looking deeply for BERT structure, it leverages the Attention model which is a popular model for bidirectional training of transformers. By using this technique BERT declares that it has gained higher results for natural language processing tasks and understanding tasks.

It is better observe the BERT's architecture, before considering the usage of BERT for natural language processing tasks such as text classification. BERT is a Transformer encoder, which is multilayered and as well as bidirectional. The below diagrams depict a 12 layered BERT transformer model which is a BERT-Base version. Let's take a reminder for Transformer models that they are based on the Attention model.

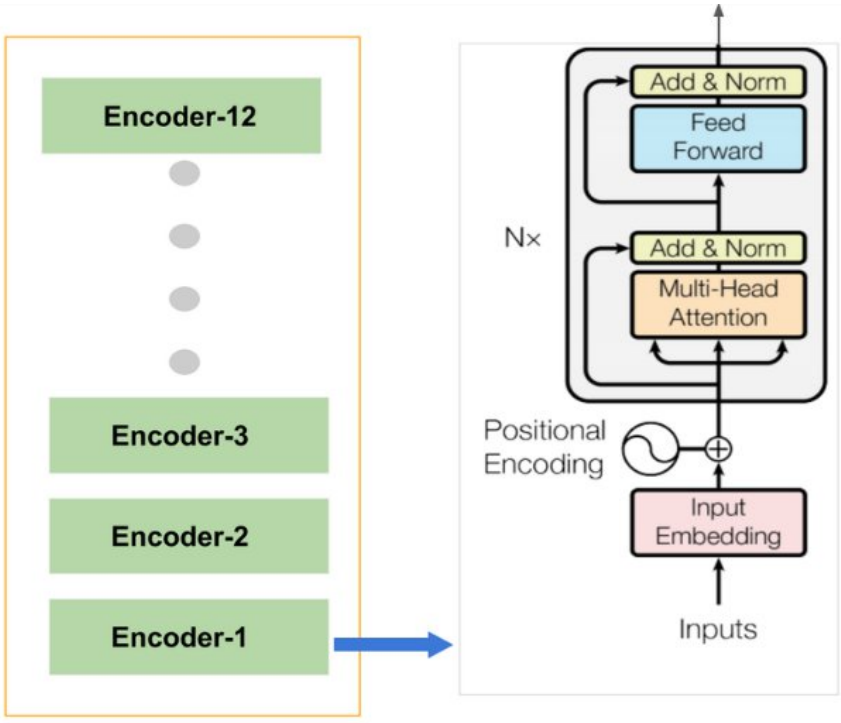

Figure 6: Overview of BERT Architecture: Adapted from [45]

These models are called pre-trained models and these are multiple pre-trained model versions which consist of changing numbers of attention heads, encoder layers and hidden size dimensions. Here follows a list that contains available variants of different models.
A = Number of self-attention heads.

H = The hidden size.

L = Number of Layers (Transformer Blocks)

2.2.1.2 Methods of Using BERT

For text classification, BERT is leveraged as a model in three different approaches.

1. **Fine Tuning Approach**: On the last layer of the pretrained BERT model, add a dense layer on top of it and after that train the entire model dataset which is a task specific.

2. **Feature Based Approach**: In feature based technique, from the pretrained model, features that are fixed have been extracted. Without fine tuning from many layers or one layer, the activations have been extracted. These embeddings which are contextual have been leveraged as input for particular tasks to downstream networks. In the BERT paper, some strategies have been mentioned for feature extraction are referenced below:

    a. Weighted Sum All 12 Layers

    b. Concat Last Four Hidden

    c. Extracting Last Hidden Layer

    d. Extracting Second-to-Last Hidden Layer

3. **As word-embedding**: In word embedding approach, using the trained model to token embedding has been generated. Token embedding is also called a vector representation of words for an end to end Natural Language Processing tasks without fine tuning. The generated token embeddings are then used for particular tasks such as text classification, summarization, topic modeling etc.

2.2.1.3 Different BERT Models

- BERT-Large

BERT-Large consists of 24 layers, 1024 dimensional output hidden vectors and 16 attention heads. There exists uncased and cased variants for each model.

- BERT-Base

This model has 12 encoder layers with 12 attention heads and has 768 hidden sized representations.

- MBERT

A multilingual BERT model using 104 highest-resource languages from Wikipedia have been pretrained. [45]

- RoBERTa

This model is an enhanced version of the BERT model. This model has been trained on a larger dataset that removes the next sentence prediction and also includes a dynamic masked language model training regimen. On multiple NLP tasks, RoBERTa model matches or exceeds the overall performance of BERT [47].

- XLM-R

This model has been trained on 100 languages and it is a transformer-based masked language model, leveraging a large data set which consists of more than two terabytes of filtered CommonCrawl data. On a various cross-lingual benchmarks, this model outperforms MBERT [48].

- LaBSE

This model combines the finest approaches for learning cross-lingual and monolingual representations comprising translation language modeling (TLM), masked language modeling (MLM), additive margin softmax and dual encoder translation ranking [49].

2.2.1.4 Fine-Tuning BERT

An experiment has been performed for reviews on Vietnamese language using mainly two fine tuning approaches on BERT. One such method is using the [CLS] token for a neural network which is an attached network as an input. The other method uses all out vectors of BERT for the input of the classification. They are experimenting on two datasets and reveal that the BERT model is moderately outperforming the other pre-trained models. For the classification these models are also using the technique GLove and FastText for word embedding. They have introduced a fine tuning BERT technique and its performance is better than the existing fine tuning method of BERT [50].

2.2.2 Sentiment Analysis Using XLNet

Using XLNet and BERT transformer models' sentiment analysis has been performed on Tigrinya language which is a low research language. First, they have created a new dataset since there was no existing dataset which can be used for analyzing sentiment for Tigrinya language. Apart from the transformer models they have introduced a new transfer learning method, where a trained model can be utilized for an unseen low resource language. They are demonstrating that the XLNet model has been outperformed the BERT model [44].

XLNet models have been used for another research which is performed for Arabic language. Arabic language is also a low resource language and due to that reason, it’s quite a challenge to perform sentiment analysis accurately. Pre-trained models can overcome this problem since it has been trained on a big dataset and then it is fine-tuned to experiment on downstream tasks. They have implemented a model named as AraXLNet which used the XLNet pre-trained and then it has been trained on a large dataset on Arabic language without annotations. Their experimental results, the newly introduced AraXLNet model, have gained higher results for the task of sentiment analysis using many datasets as benchmarks [52].

2.2.3 Sentiment Analysis Using ELMO

For many Natural Language Processing tasks, word embeddings take an important role. Word embeddings simply mean the words are represented in a form of numerical using vectors, consist of several hundred dimensions. Word2vec, Glove and FastText are the most typical word embeddings which have used in most cases. But the problem with these word embeddings are they do not identify the context of the words and therefore it is difficult to convey polysemous words. Elmo ((Embeddings from Language Models) have been introduced to overcome this problem and gives a contextual component.

Research has been carried out for seven languages Croatian, Estonian, Finnish, Latvian, Lithuanian, Slovenian, and Swedish which uses Elmo model for precomputed word embeddings. They are demonstrating that the Elmo models give better results than the non-contextual word embeddings. And also, they are showing that the volume of the datasets also an important factor when creating word embeddings [55].

2.2.3 Sentiment Analysis Using ULMFit

Since recently the people are more engaged with social media, they tend to share their information through these platforms. Sometimes, they are sharing crucial information like storms, flood, volcano eruptions, earthquakes and many more disasters. Therefore, it has shown that the social media provides great amount of data on natural disasters [29]. Because of the enormous number of unlabeled data, to filter out the data effectively is very challenging. One of the research projects has been carried out to detect localized floods utilizing Twitter, the social sensing model. With minimal labeled data this model contributes to give a reliable, accurate and efficient flood text classification. They have used inductive transfer learning method to carry out the text classification. It is a pre-trained language model named as ULMFiT and classifying the flood related data effectively they are fine tuning the model [30].

2.3 Multilingual Transformer Models

Considering the wide area of many Natural Language Processing tasks where pre-training is used, mono-lingual transformer models have accomplished better results. Nevertheless, the most regular transformer models use the languages which are highly resourced languages as English for pre-training. There can be many reasons for this such as it is easy to find large-scale datasets. On the other hand, the cost of training for transformer models which are language dependent. In contrast, from a dataset multi-lingual transformer models attempt on pre-training a transformer model that is gathered by many languages. Consequently, across many languages these models can be generalized by supplying some shared representations which are used for all languages. Most of the languages which do not have a mono-lingual transformer model on its own make most of it from this technique. As the first multi-lingual transformer models MBERT (Multi-lingual BERT) have been introduced which has been trained used by dataset that consists of104 languages. When training this mBERT 110k size of a shared vocabulary is used. The objective function of BERT and mBERT is same which is explained in the previous sections and the only difference between these models is for pre-training they use different datasets.

XLM is another multi-lingual transformer model which has performed well and it attempts on using the existing labeled dataset from introducing a novel language model named as TLM (translation language model). Like the model BERT MLM, this model leverages a language model which is auto-encoding. Nevertheless, the major objective of TLM is the need for having the novel knowledge from another translation of input sequences of language B when predicting a target word of language A.

3 Research Methodology

The main intention of this research is to perform sentiment analysis for the Sinhala language leveraging transfer learning methods. Among the existing state of the art transfer learning-based approaches like ELMo, ULMFiT and BERT, for this research the experiments done on BERT which is more sophisticated word embedding techniques and which can be utilized to grasp semantic and syntactic knowledge of the language for sentiment analysis. Then the experimentation is performed on built BERT models which is already available on HuggingFace website. These models are Roberta-based Sinhala models, named SinBERT-small and SinBERT-large which have implemented by a group students at University of Moratuwa. Apart from the BERT, using the XLNet pre-trained model it is experimented whether it provides better results for the sentiment analysis task.

In this section what kind of data have been used and how this research has been performed on the collected data will be described.

3.1 Data Collection

For this research two datasets are using to do experiments which also have been used for previous researches. The main objective of this data collection is that we can use the results of the previous experiments as a benchmark for this research. As mentioned in the section 2, most of the sentiment analysis which have been carried out for Sinhala language are based on the deep learning models, supervised learning techniques like Naïve Bayes, Support Vector Machine (SVM) maximum entropy and standard sequence models like RNN, LSTM, Bi-LSTM, and also most recently used models as capsule networks and hierarchical attention hybrid neural networks that are commonly used in sentiment analysis. Therefore, results of this research can be compared and contrast the previous researches.

The two datasets consist of news comments collected from newspaper articles. One dataset extracted comments from an online news site named as Lankadeepa.lk because this site has a lot of different categories comprising politics, economy, society, sports, culture and crime with manually moderated comments.

The other dataset acquires data from two local news websites Lankadeepa.lk as well as from GossipLanka which is a popular website among the users even though it doesn't have printed version.

For both of the datasets, the comments are manually annotated as POSITIVE, NEGATIVE, CONFLICT and NEUTRAL. Calculated Cohen's kappa value for the first dataset was 0.52 and for the dataset it was 0.65.

1st Dataset:

```
Negative comments:  1996
Positive comments:  2125
```

Figure 7: Count of Positive and negative comments

2nd Dataset:

```
Negative comments:  3868
Positive comments:  2550
```

Figure 8: Count of Positive and negative comments

3.2 Data Preprocessing

For both data sources the comments are extracted from news articles. Therefore, pre-processing and polishing data takes place as an important role before the text classification tasks. Since the texts such as comments and reviews entered by users doesn't represent the polish which can be seen in a proofread document. The following have been carried out for this research.

- Removing emoji
- Remove numbers
- Remove punctuations
- Removing stop words
- Stemming

Removing emoji

Emoji is a new way of expressing opinions on social media and there are lot of emojis can be seen in comments and reviews. Since there are different kind of emojis and it is difficult to categorize which are positive comments and negative comments. Therefore, as preprocessing technique emojis are removed.

Removing punctuations

Non-letter characters and numbers does not contribute any special meaning to comments. Including these characters in the study will make inaccurate decisions when deriving the sentiments. However, from previous research it has been found that even though the most of the punctuation marks affect the performance of the models. That research concluded that the models will perform better if we remove punctuation marks without the question mark. Since the question mark is commonly associated with negative comments. To identify negative comments from the rest it will be a much suitable feature [b]. Therefore, we remove these characters (!@#$%^&*()_+{}[]:";',./<>) from our data set except question mark (?).

Removing stop words

For Sinhala language Stop words did not have positive effect on the classification performance. Stop words are the words that are mainly used words in a document. බව /bavə/, මේ /me:/ , ඒ /e:/, නම් /nam/, හා /ha:/, හෝ /ho:/, සහ /saha/, and සමඟ /saməgə/are few examples of such words in Sinhala. Hence, we remove these types of words.

Stemming

Stemming is the process of removing affixes from a word to derive the base word or root. For an example, if we consider the word "ක්‍රීඩා /kri:da:/", it can have many variations such as ක්‍රීඩාවට /kri:da:vətə/, ක්‍රීඩාවේ /kri:da:ve:/,, ක්‍රීඩාවන් /kri:da:van/, ක්‍රීඩාව /kri:da:və/ etc. Even though all these words are related to "ක්‍රීඩා /kri:da:/", model will identify them as different words. Therefore, we need to consider only the roots of words. From a sentence the words can be tokenized using the Sinhala tokenizer from Sinling. Further we can stem the words from this library which is capable of deriving the stems of the Sinhala words. But stemming process is still in experimental stage [54].

3.3 Sentiment Analysis with BERT

For this research, sentiment analysis is performed using pre-trained BERT model. Here follows how the BERT is working and model architecture of the methodology which have used to build a sentiment classifier. BERT (Bidirectional Encoder Representations from Transformers) is an embedding Layer which has built to train deep bidirectional representations from Transformers and released in late 2018 [45]. It is a procedure for pre-training language representations which is utilized to generate models which Natural Language Processing (NLP) practitioners can download and can utilized these models freely. On the other hand, these models can be used to extract language features which are high quality from our data or else we can fine-tune these models on a given task such as entity recognition, question answering, classification etc [46]. Therefore, there are two main procedures for BERT as fine-tuning and pre-training which has visualized in the figure 5.

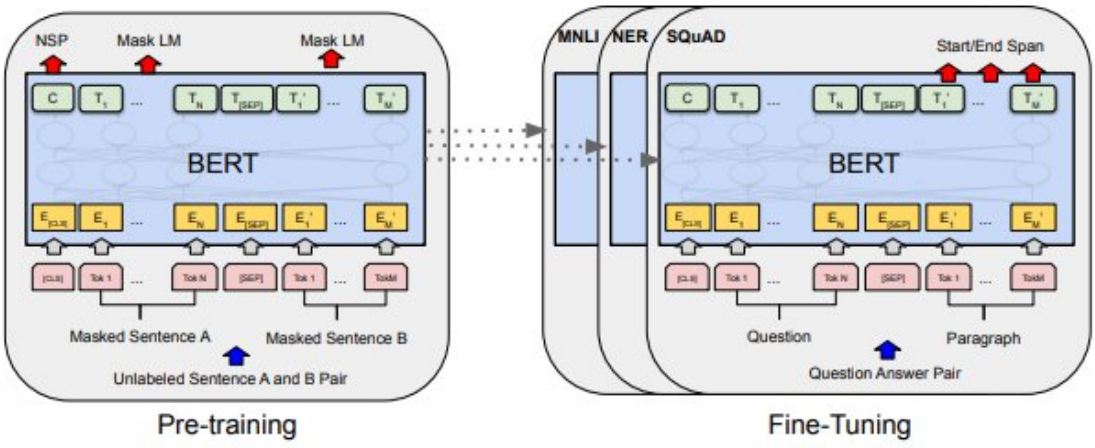

Figure 9: Pre-training and fine-tuning models in BERT: Adapted from [45]

In both procedures the same architecture has been used except for the output layers.

The attention architecture processes the all-input sequence at the same time, allowing all input tokens to be processed in parallel not like the recurrent models or traditional sequential models. Architecture of BERT with layers is depicted from Figure 6.

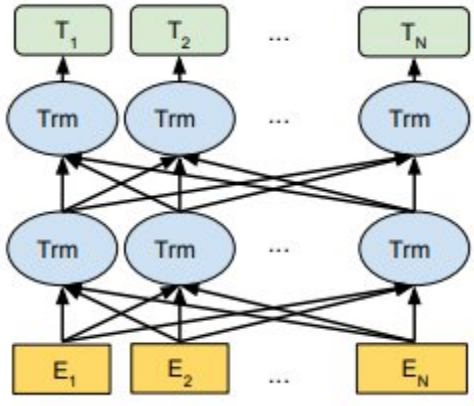

Figure 10: The layers of BERT architecture: Adapted from [45]

The main advantages of using such models as BERT can be defined as quicker development, less data and better results [43].

3.4 Fine Tuning with BERT

There are pre-trained models for English and other 103 languages are defined in BERT. Fine-tuning these mentioned pre-trained models can be done for our requirements appropriately. In this research we are slightly training a model using two datasets on top of an already trained checkpoint. This is the approach behind fine-tuning. To perform sentiment analysis, we are fine-tuning the multilingual model.

The hierarchical architecture which has been used for this research is shown in figure 11.

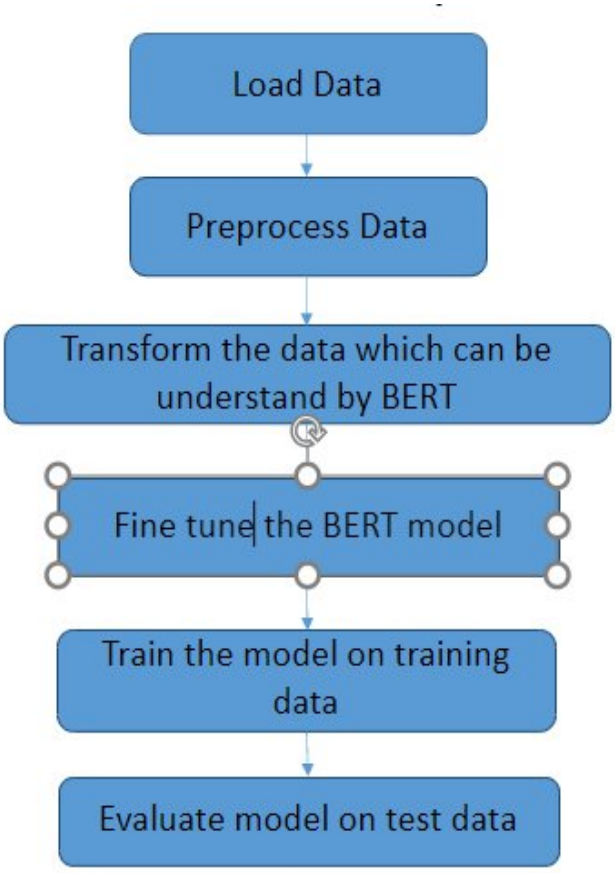

Figure 11: Architecture of the proposed model of BERT

Two data sets are using for our experiment and evaluate the model which have been used for previous researches. The first dataset consists of 2125 negative comments and 1996 positive comments. Therefore, we are creating a small dataset of to be the same size of the negative and positive comments while performing preprocessing techniques on the data.

Then we need to transform the preprocessed data to fit into BERT model which means transforming the data which can be understand by BERT. There are two steps in order

achieve this process. Transformers library provides a constructor which can be used to initiate a list of InputExample objects. Following structure needs to be there for InputExample.

- text_a: is the text data which needs to get classify
- text_b: To understand the association among the sentences when we are training a model, this parameter can be used. Can set this as blank since it is not applicable to our analysis.
- Label: is the sentiment which is given as POSITIVE or NEGATIVE

When providing inputs for deep learning models, the inputs need to be in a certain kind of format. Vectors of integers is the mostly used format each value representing a token. Therefore, each text needs to be transferred to a list of indices that can be inserted into the model. The tokenizer is using for this process and we also need to add special tokens to the list of ids.

As the next step we need execute another method named as `glue_convert_examples_to_features` to deal and manage other additional details which are required when transferring text to be eligible for BERT. Using Transformer library, we can directly transform data into features.

As the last stage we can fine-tune the BERT model after declaring some hyperparameters which can be used when performing the training such as the loss, the evaluation metric and the optimizer. For this we are leveraging BertForSequenceClassification which is a BERT model with an added single layer on top for classification which can be used as a classifier. When we insert the data into the model, the entire pre-trained model and the additional untrained classification layer is trained on our given process.

3.5 SinBERT -small and SinBERT-large

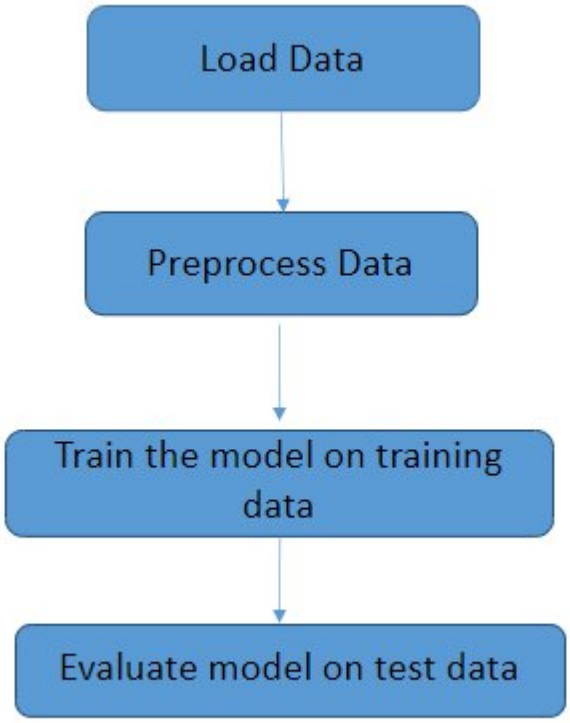

Figure 12: Architecture of Proposed Model with SinBERT

Research has been done to use already existing pre-train two monolingual Roberta-based Sinhala models, named SinBERT-small and SinBERT-large which are available in Huggingface official site. https://huggingface.co/NLPC-UOM. These models have been trained using a large corpus for Sinhala language. Since due to high usage of computational costs it is really difficult to train such models. Therefore, using these pre-trained models we can perform sentiment analysis for Sinhala and compare with other researches whether the pre-trained models are outperforming the other available models.

3.6 Architecture of Proposed Model with BERT with Feature Based Approach

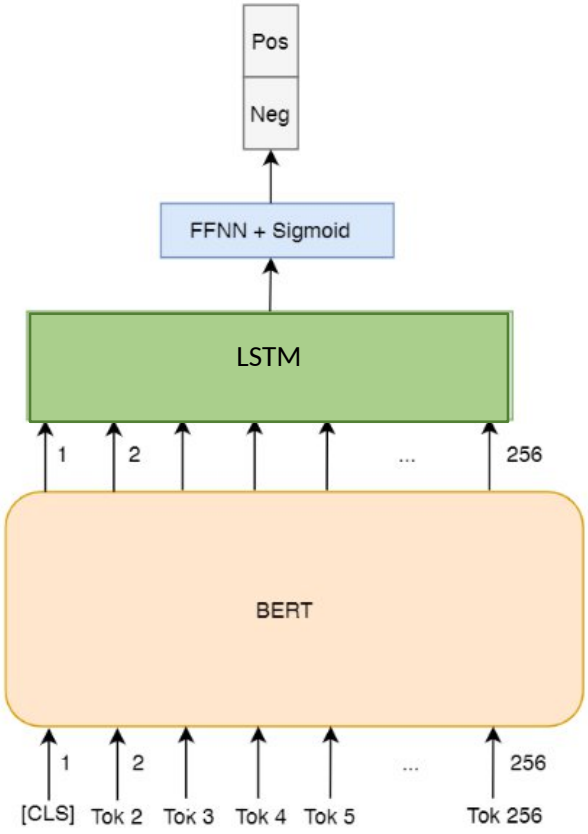

Figure 13: Achitecture of BERT Model with Feature Based

For this approach comprising the token [CLS] all of the BERT model output is used. The output consists of a format of matrix SEQ LEN* h. From the input sequence the largest length has been taken as the SEQ LEN and the length of hidden vectors is declared as h. Leveraging this output matrix, we can take this as an input for the other classification models. Here we can use the other classification models such as RCNN, TextCNN and LSTM.

For this research we are using the LSTM classification model. LSTM is shortened for Long Short Term Memory and among the existing RNN (Recurrent Neural Network) models, LSTM model is the popular method. The main advantage of this model is its focusing on the dependence distance issue of the traditional RNN model. Therefore in

this approach, the layer of LSTM is extracting the features from output acquired by the BERT model.

4 Experiments

When performing the experiments Google Colab notebook used since it would take so much time when executing the implementations on a CPU. These environments contain virtual machines consisting of 25GB high memory and high-end GPUs like P100 and T4. The implementation has been build using Python language and libraries such as Keras has used to experiment and evaluate the implementation which was defined in this paper.

Dataset 1

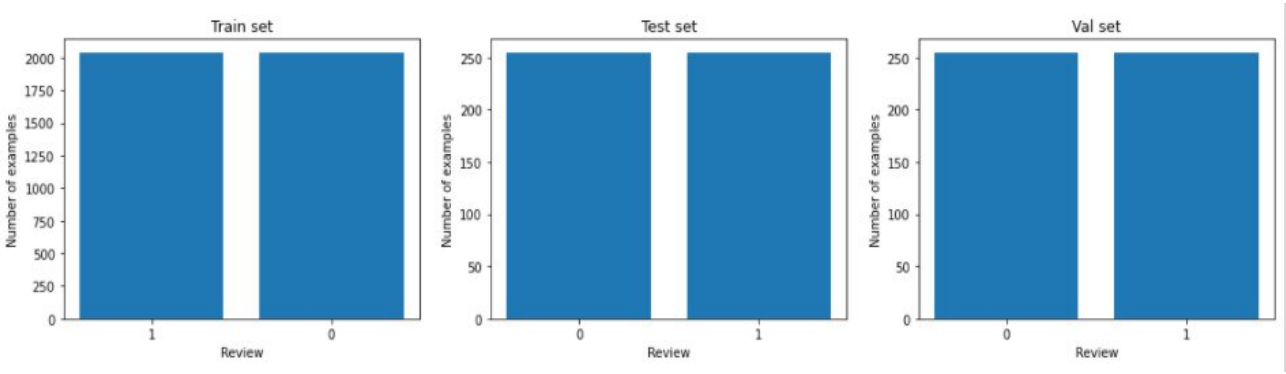

Figure 14: Analysis for Data Preparation for db1

4.1 Evaluation - BERT

The experiments performed on two data sets and the training and validation loss as well as accuracy is visualized in the figures 15 and 16.

Dataset 1:

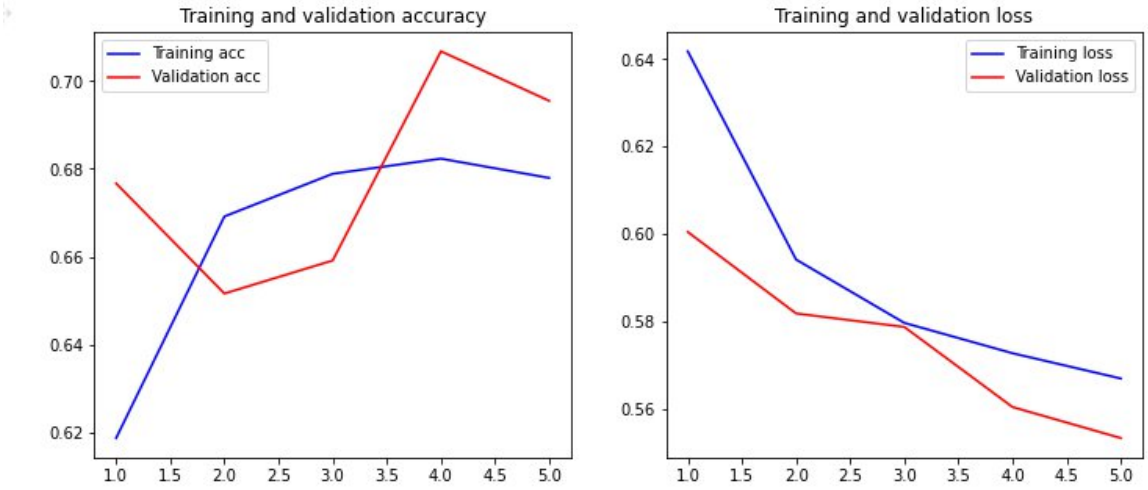

Figure 15: Training validation loss and accuracy for dataset

Dataset 2:

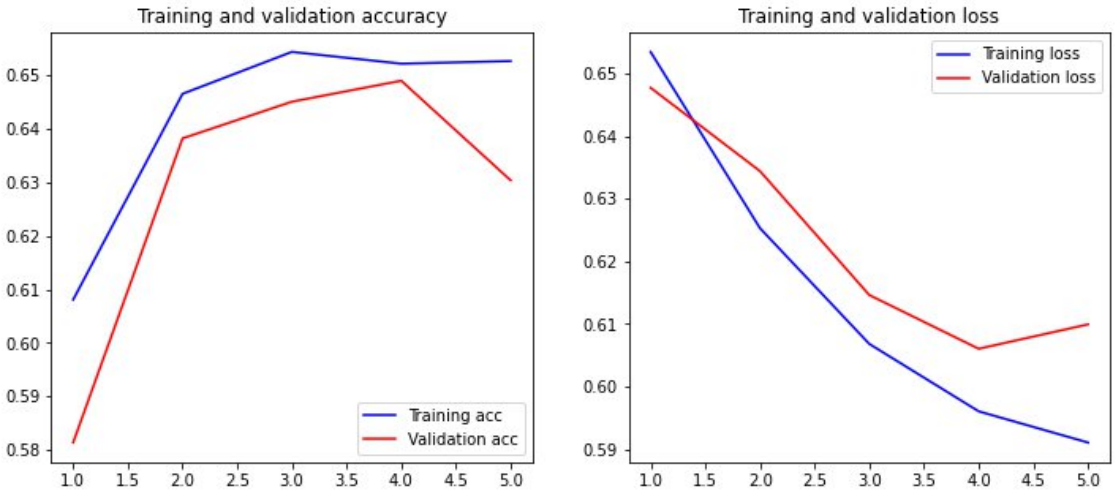

Figure 16: Training validation loss and accuracy for dataset 2

We can make predictions based on the trained model as shown in figure 17 and 18.

Dataset 1:

```
comment: සැළකියයුතු යමක් මට නම් මේ කතාවේ නොපෙනේ
, actual label: NEGATIVE, predicted label: 0
comment: කෑ ගහල ගෙනාපු ඖෂධ පනතින් කාටද සෙතක් වුණේ කියන්නේ නම් බලය ඇත අරිනවා කියල පරාද අය ගෙනත් ගෙනත්
, actual label: NEGATIVE, predicted label: 0
comment: ලක්ෂ 22 දෙකට දඩය 66 ලක්ෂයක් නම් ඉදිරියේදී මොන පක්ෂෙන් හරි දේශපාලුවෝ අහුවෙනකොට ලංකාවේ සල්ලි මදි වෙනව
, actual label: NEGATIVE, predicted label: 0
comment: නෝ බුද්ධා අරහත්ු ස්ා ස්බුදුරජාණන් හන්සේට නස්කාර ෝ
, actual label: POSITIVE, predicted label: 0
comment: ඉත්ුණ එක හොදි
, actual label: POSITIVE, predicted label: 1
comment: ේැඩේ න් ංකාට ඩ්බරක් ත් ෙද හත්තෝ පරිස්සින්
, actual label: POSITIVE, predicted label: 0
comment: තන්ගෙන් පසු දෙශපානට පරපුර නොගෙනා ක ජනනාකා ඔබ් ජේ  ර් ජර්ධන නිදුනි ඔබට අහ හ නින් සු
, actual label: POSITIVE, predicted label: 1
comment: මගේ පුතාට පාසලක් නැහැ මට ගස් නගින්න දන්නේ නැහැ මම මොකද කරන්නේ
, actual label: NEGATIVE, predicted label: 0
comment: හොඳ තරඟක් බාගන්න පුළුන්ේ
, actual label: POSITIVE, predicted label: 0
comment: මේ ගැන කතා කරලා වැඩක් නෑ
, actual label: NEGATIVE, predicted label: 0
```

Figure 17: Predictions for dataset 1

	precision	recall	f1-score	support
0	0.55	0.82	0.65	485
1	0.70	0.39	0.50	535
accuracy			0.59	1020
macro avg	0.62	0.60	0.58	1020
weighted avg	0.63	0.59	0.57	1020

Dataset 2:

```
comment: අපේ අට හොදට ගැඹුරට හාරන්න සුළුන් බ න් පේනා
, actual label: 4, predicted label: 1
comment: මම හිතන්නේ මේ කාන්තාව දුප්පත්
, actual label: 2, predicted label: 1
comment: ගුරුවරයා හිටගෙන කරන දේ ගෝලයෝ දුව දුව කරනවලු
, actual label: 2, predicted label: 1
comment: ෝකාසී සිු දෙනාට සුබ න සරක් ෝ
, actual label: 4, predicted label: 1
comment: අපේ සුභ පැතු
, actual label: 4, predicted label: 1
comment: මේ උදවියගේ ජායාරූප ප්‍රසිද්ධ කරන්නේ නැත්තේ ඇයි දැන් අර අලුත්ම මෝස්තරේට පන්සලේද දන්නේ නැහැ විවාහ වෙන්න ඉ
, actual label: 2, predicted label: 0
comment: මේවාට විරුද්ධත්වයක් නොපෙන්වන අයත් නැතුවාම නොවෙයි
, actual label: 2, predicted label: 1
comment: හරි ස්සනි පුදු හිතෙනා
, actual label: 4, predicted label: 1
comment: සිු දෙනා දක්ෂින්
, actual label: 4, predicted label: 1
comment: ත්ත ගැන ෙහෙ දැනුත් කරනට ංකාඳීපට ස්තුනි
, actual label: 4, predicted label: 1
```

Figure 18: Predictions for dataset 2

	precision	recall	f1-score	support
0	0.32	0.25	0.28	402
1	0.37	0.45	0.40	394
accuracy			0.35	796
macro avg	0.34	0.35	0.34	796
weighted avg	0.34	0.35	0.34	796

Finally, we can see that the accuracy is much higher for the second dataset which 0.63.

```
Evaluating the BERT model
13/13 [==============================] - 6s 486ms/step - loss: 0.5533 - accuracy: 0.6955
[0.5532885193824768, 0.6954887509346008]
```

Figure 19: Accuracy for dataset 1

```
Evaluating the BERT model
16/16 [==============================] - 8s 519ms/step - loss: 0.6099 - accuracy: 0.6304
[0.6099381446838379, 0.6303921341896057]
```

Figure 20: Accuracy for dataset 1

When comparing with the previous researches which have done using classification methods the BERT model outperforms most of the traditional classification techniques. The figure 19 shows the results of the earlier experiments which have been done for data set 2 for the task of sentiment analysis for the Sinhala language.

Model	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
RNN	58.98	42.93	54.98	42.30
LSTM	62.88	70.95	51.93	54.50
GRU	62.78	60.93	62.78	54.83
BiLSTM	63.81	61.17	63.81	57.71
CNN + GRU	61.59	60.41	61.59	54.19
CNN + LSTM	61.89	57.82	61.89	55.30
CNN + BiLSTM	62.72	59.54	62.72	58.53
Stacked LSTM 2	61.92	56.92	61.92	53.17
Stacked LSTM 3	62.48	54.76	62.48	53.67
Stacked BiLSTM 2	63.18	60.50	63.18	57.78
Stacked BiLSTM 3	63.13	69.71	46.63	59.42
HAHNN	61.16	71.08	48.54	59.25
Capsule-A	61.89	56.12	61.89	53.55
Capsule-B	**63.23**	**59.84**	**63.23**	**59.11**

Figure 21: Benchmark

Algorithm	Accuracy	Precision	Recall	F1_Score
Naive Bayes	0.7769461078	0.8407407407	0.6906906907	0.7529021559
Decision Tree	0.7654690619	0.7611056269	0.7597597598	0.7664015905
SVM	0.869261477	0.9230769231	0.8048048048	0.8598930481
RNN LSTM	0.8645833313	0.8917127072	0.8531468531	0.8617191671
Logistic Regression	0.8677644711	0.9160997732	0.8088088088	0.8591174907
Random Forest	0.8592814371	0.9115958668	0.7947947948	0.849197861

Figure 22: Benchmark

4.1 Experiments with SinBERT-small and SinBERT-large without Stop Words

Dataset 1

	Accuracy	F1-score	Precision	Recall
SinBERT-small	0.85	0.59	0.57	0.60
SinBERT-large	0.79	0.61	0.67	0.64

Table 3: Experiments with SinBERT without StopWords for db 1

Dataset 2

	Accuracy	F1-score	Precision	Recall
SinBERT-small	0.82	0.72	0.62	0.54
SinBERT-large	0.82	0.51	0.63	0.62

Table 4: Experiments with SinBERT without StopWords for db2

4.2 Experiments with SinBERT-small and SinBERT-large with Stop Words

Dataset 1

	Accuracy	F1-score	Precision	Recall
SinBERT-small	0.81	0.65	0.60	0.59
SinBERT-large	0.83	0.60	0.57	0.56

Table 5: Experiments with SinBERT with StopWords for db1

Dataset 2

	Accuracy	F1-score	Precision	Recall
SinBERT-small	0.82	0.69	0.63	0.66
SinBERT-large	0.88	0.61	0.54	0.54

Table 6: Experiments with SinBERT wit StopWords for db2

The above results have been obtained using both of the datasets and performing sentiment analysis using the pre-trained models named SinBERT-large and SinBERT-small. Table 3 and table 4 contain the results before analyzing the sentiment as a data preprocessing task, it was mentioned that to remove stop words. But here as the first experiment, sentiment analysis task has been carried out without stop words. As the second approach removal of stop words has been done as a data preprocessing task and analyse the sentiment. The results for that experiment included table 5 and table 6. From this it is showing that the stop words are also giving some contribution for the sentiment analysis task. Both of the models have a slightly higher value with the stop words. It might be the case where when training the models as the input from the raw corpus, models are taking the full sentences. Therefore, when analyzing sentiment stop words may also contribute when determining the sentiment.

4.3 Experiments on Feature Based Approach Using BERT

Dataset 1:

	Accuracy	Precision	Recall	F1
BERT	0.6955	0.63	0.59	0.57
BERT-LSTM	0.6820	0.73	0.75	0.74

Table 7: Experiments with Feature Based for db1

Dataset 2:

	Accuracy	Precision	Recall	F1
BERT	0.630	0.34	0.35	0.34
BERT-LSTM	0.725	0.715	0.736	0.726

Table 8: Experiments with Feature Based for db2

4.4 Evaluation – XLnet

Dataset 1:

```
Test Accuracy of the model on vla data is: 52.43055555555556 %
```

Dataset 2:

```
Test Accuracy of the model on vla data is: 63.5678391959799 %
```

When performing the experiments on XLNet models for sentiment analysis, it didn't outperform the results of BERT model. XLNet pre trained model is trained on English language and it may be a reason that it didn't perform well for Sinhala language.

5 Conclusion

This research targeted on applying transfer learning-based approaches for sentiment classification for the Sinhala language. The dataset was collected from previous researches in order to compare and contrast the most commonly used sentiment classification methods and machine learning algorithms with the transfer learning models. After going through a thorough survey on the state of the art transfer learning based methods which have used for sentiment classification, BERT takes place a huge place in sentiment classification for other languages. Using BERT some researches have been performed for other languages. Since it already contains pre-trained models for other 103 languages apart from the English language.

After performing experiments with the BERT the accuracy is high when comparing with other methods of sentiment analysis except for capsule networks. The main advantage of using these kind of pre-trained language models is, these models doesn't require resource intensive and time-consuming techniques for training the models. Even though the Sinhala language have not been experimented by fine-tuned BERT models previously, we can conclude that it can be utilized for Sentiment analysis of Sinhala language.

Apart from the existing pre-trained models, a research group has trained a model named SinBERT-small and SinBERT-large which was trained using large corpus. After using these monolingual Roberta-based Sinhala models, the evaluation results became really high. Therefore, we can conclude that these models can be used for downstream tasks as sentiment analysis.

5.1 Future Work

There are possible future enhancements and improvements can be done for this research such as,

- The dataset only consists of Sinhala news comments, this can further be improved by doing this research for data collected from social media like Twitter, Facebook etc.
- This research focused only on BERT, but this can be further extended using other models such as ELMO, ULMFit.
- For data preprocessing in this research, it has been omitted emojis. But emojis can also be useful for determining sentiment. Therefore, using emojis sentiment analysis can further improve.

References

[1] Yoshida, Y., Hirao, T., Iwata, T., Nagata, M. and Matsumoto, Y. (2011). *Transfer learning for multiple-domain sentiment analysis — identifying domain dependent/independent word polarity*. [online] Dl.acm.org. Available at: https://dl.acm.org/citation.cfm?id=2900627 [Accessed 25 Jan. 2019].

[2] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359

[3] Pan, Sinno Jialin, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. in Proceedings of International Conference on World Wide Web (WWW-2010). 2010.

[4] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135, 2008.

[5] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 79–86, 2002.

[6] Roark and M. Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In NAACL-HLT, pages 126–133.

[7] Daum´e III and D. Marcu. 2006. Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 26(1):101–126.

[8] R.K. Ando and T. Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. The Journal of Machine Learning Research, 6:1817–1853.

[9] J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In ACL, page 440–447.

[10] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In Empirical Methods in Natural Language Processing (EMNLP).

[11] H. Daum´e. 2007. Frustratingly easy domain adaptation. In ACL, pages 256–263.

[12] W. Dai, Y. Chen, G.R. Xue, Q. Yang, and Y. Yu. 2008. Translated learning: Transfer learning across different feature spaces. In NIPS, pages 353–360.

[13] N. Bel, C. Koster, and M. Villegas. Cross-lingual text categorization. In ECDL, 2003.

[14] W. Dai, O. Jin, G.R. Xue, Q. Yang, and Y. Yu. 2009. Eigentransfer: a unified framework for transfer learning. In ICML, pages 193–200.

[15] S. Thrun and L. Pratt, Eds., Learning to learn. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[16] Q. Wu, S. Tan, and X. Cheng. 2009. Graph ranking for sentiment transfer. In ACL-IJCNLP, pages 317–320.

[17] Q. Wu, S. Tan, X. Cheng, and M. Duan. 2010. MIEA: a Mutual Iterative Enhancement Approach for Cross-Domain Sentiment Classification. In COLING, page 1327-1335.

[18] C.W. Seah, I. Tsang, Y.S. Ong, and K.K. Lee. 2010. Predictive Distribution Matching SVM for Multi-domain Learning. In ECML-PKDD, pages 231–247.

[19] Guerra, P., Veloso, A., Meira Jr., W., Almeida, V.: From bias to opinion: A transfer-learning

approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD (2011).

[20] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks, arXivpreprint arXiv:1509.06041.

[21] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In ICML, 513–520, 2011.

[22] Chinchung Chang and Chinjen Lin. 2001. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[23] Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In Proceedings of ICML.

[24] Bruzzone, L., Marconcini, M.: Domain adaptation problems: A dasvm classification technique and a circular validation strategy. IEEE Trans. on PAMI 32(5), 770–787 (2010).

[25] Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. JMLR 12, 2399–2434 (2006).

[26] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41:391–407, 1990.

[27] Danushka Bollegala, David Weir, and John Carroll. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 132–141, Portland, Oregon. ACL.

[28] Rajesh, M., and J. M. Gnanasekar. "Annoyed Realm Outlook Taxonomy Using Twin Transfer Learning" International Journal of Pure and Applied Mathematics, 116.21 (2017) 547-558.

[29] Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In Twelfth International AAAI Conference on Web and Social Media.

[30] Neha Singh, Nirmalya Roy, Aryya Gangopadhyay. “Localized Flood DetectionWith Minimal Labeled Social Media Data Using Transfer Learning”.

[31] Meysam Asgari-Chenaghlu, Mohammad-Reza Feizi-Derakhshi, Leili farzinvash, Mohammad-Ali Balafar, Cina Motamed. TopicBERT: A Transformer transfer learning based memory-graph approach for multimodal streaming social media topic detection

[32] Senevirathne, L., Demotte, P., Karunanayake, B., Udyogi, M. and Ranathunga, S., 2021. *Sentiment Analysis of Sinhala News Comments using Sentence-State LSTM Networks*. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/iel7/9179991/9185188/09185327.pdf> [Accessed 6 April 2021].

[33] Isuru Udara Liyanage. 2018. Sentiment Analysis of Sinhala News Comments. (2018). Unpublished.

[34] Nishantha Medagoda. 2016. Sentiment Analysis on Morphologically Rich Languages: An Artificial Neural Network (ANN) Approach. In Artificial Neural Network Modelling. Springer, 377–393.

[35] Nishantha Medagoda. 2017. Framework for Sentiment Classification for Morphologically Rich Languages: A Case Study for Sinhala. Ph.D. Dissertation. Auckland University of Technology

[36] PDT Chathuranga, SAS Lorensuhewa, and MAL Kalyani. 2019. Sinhala Sentiment Analysis using Corpus based Sentiment Lexicon. In International Conference on Advances in ICT for Emerging Regions (ICTer), Vol. 1. 7.

[37] Nafissa Yussupova, Diana Bogdanova. 2012. Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach.

[38] Muhammad Haroon Shakeel, Safi Faizullah, Turki Alghamidi , Imdadullah Khan. 2019. Language Independent Sentiment Analysis.

[39] Joshi, Aditya, A. R. Balamurali, and Pushpak Bhattacharyya. "A fall-back strategy for sentiment analysis in hindi: a case study."Proceedings of the 8th ICON (2010).

[40] Bakliwal, Akshat, Piyush Arora, and Vasudeva Varma. "Hindi subjective lexicon: A lexical resource for hindi polarity classification." In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC), pp. 1189-1196. 2012

[41] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[42] Enkhbold Bataa , Joshua Wu . 2019. “An Investigation of Transfer Learning-Based Sentiment Analysis in Japanese

[43] Marco Pota, Mirko Ventura, Rosario Catelli, and Massimo Esposito. 2020. An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian

[44] Abrhalei Frezghi Tela, 2020. Sentiment Analysis for Low-Resource Language: The Case of Tigrinya

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[46] http://mccormickml.com/2019/07/22/BERT-fine-tuning/]

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc

[49] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. A

[50] Quoc Thai Nguyen, Thoai Linh Nguyen, Ngoc Hoang Luong, and Quoc Hung Ngo. Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews, 2020 - ieeexplore.ieee.org

[51] Soroush Karimi, Fatemeh Sadat Shahrabadi. 2019. Sentiment analysis using BERT (pre-training language representations) and Deep Learning on Persian texts.

[52] Alduailej, A., Alothaim, A. AraXLNet: pre-trained language model for sentiment analysis of Arabic. J Big Data 9, 72 (2022). https://doi.org/10.1186/s40537-022-00625-z

[53] Tsuruoka Y. Deep learning and natural language processing. Brain Nerve 2022;71:45–55

[54] https://github.com/ysenarath/sinling

[55] Matej Ulˇcar and Marko Robnik-Sikonja. High quality elmo embeddings ˇ for seven less-resourced languages. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 4731–4738, Marseille, France, May 2020. European Language Resources Association.