National College of Ireland

Multilingual Amazon Review Rating Classification using Bi-directional LSTM

MSc Research Project
Data Analytics

Priyanka–
Student ID: x20192037

School of Computing
National College of Ireland

Supervisor: Aalok Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Priyanka–
Student ID:	x20192037
Programme:	MSc. Data Analytics
Year:	2021-22
Module:	MSc. Research project
Supervisor:	Aalok Anant
Submission Due Date:	31/01/2022
Project Title:	Multilingual Amazon Review Rating Classification using Bi-directional LSTM
Word Count:	6200
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Priyanka–
Date:	29th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multilingual Amazon Review rating Classification using Bi-directional LSTM

Priyanka–
x20192037

Abstract

Today Online shopping has become an essential part of the everyone's life. The rating of the product influences the purchasing decision of the customer. Every day, millions of the customers are posting their reviews online about the products. The posted reviews can be written in the multiple languages. Therefore, identifying the rating from the reviews becomes a tedious task. To solve such issues, we have proposed a word embedding based framework for identifying the rating from multilingual amazon review dataset. The dataset consists of the reviews in English and French language. After certain steps of data pre-processing and feature engineering, all the four models have been trained over the large subset of review data and the performance of each model has been evaluated in terms of Accuracy, PRF Score and Validation loss. After the evaluation, we have identified the better outcomes using Bi-LSTM and BERT Model.

1 Introduction

With the advancement of technology over the decades and rapid digitalization boom led to the initiation of multiple e-commerce stores and online shopping websites. Creation of mega e-commerce sites like Amazon has proven to be very useful, cost-effective, and beneficial for people to make their purchase in the comfort of their homes. As online commercial stores and product selling websites have been well working during the previous many years, the internet-based venders and e-commerce shop owners request that their buyers share their viewpoints and authentic reviews about the items they have purchased. By doing so, it will increase the trust worthiness of the seller and their products. Regularly there are a huge number of virtual accounts created all around the Internet by the potential customers who are planning to buy from the e-commerce store. After purchase, there will be millions of reviews created everyday by such customers. It gives an in-depth clarification and detailed evaluation of the purchased various items, their quality, performance, and longevity of the product. The trend of review writing has made the Internet the main and trustable source of opinion about an item when it comes to shopping online. Although there are more benefits to it, the rate of accumulation of numerous reviews makes it very hard for the potential customer to check the authenticity of the product and purchase it. Getting to read different opinions from the people who have purchased the products will leave the customer more edgy and confused. This will make refrain them from buying the product. To help the customers make the right decisions, there is a need for analysis and classification of these reviews in a categorical manner.

This will in-turn help the e-commerce stores and business owners to actually grow their business. It is very essential in a wide space such an Amazon and other international ecommerce platforms. For this, the method of Review Classification can be used. It is an all-inclusive analysis that helps to classify and curate the listed reviews by considering some of the subjective data that is mentioned there. It considers the language of the customer, certain keywords, characteristics, and opinions to classify the reviews. Apart from this technique there are many other distinctive methods that can be employed to classify the customer reviews in order, which are biometrics, computational linguistics, natural language processing and text analysis. However, these methods are so intricate, requires more additional input and are not as effective as compared to machine learning algorithms. In the past years, much research have been done in the domain of review analysis employing machine learning and deep learning algorithms. They have proven to produce accurate results and moreover the process of application is very simple compared to other methods.

In this research work, we consider the e-commerce platform of Amazon where millions of reviews are written and published in a day. Classifying and sorting these reviews based on the customer input will be very useful for the potential customer to get to know about the detailed features, application and working technique of the stipulated product. Here, the advanced deep learning algorithms like Bidirectional LSTM, Convolutional Neural Network, Conv-LSTM and Transformer Model BERT are employed to perform the Multilingual Classification of the amazon reviews into positive and negative criteria using supervised approaches. The input data used for this analysis is sourced from the dataset of AWS S3 which contains multilingual amazon reviews. The complete intention of this research work is to analyze and classify the given review, irrespective of the language in stipulated order for the hassle-free shopping experience of the potential customers. The outcomes of the research work showcase a beneficial result.

1.1 Research Question

Which algorithm efficiently classifies the rating on Multilingual amazon review dataset ?

2 Literature Review

There are a lot of existing application and current exploring trends regarding the Multilingual Classification of Amazon Reviews using Advanced deep learning algorithms. This section of the research paper contains the various existing research papers that has shed light on the different methodologies which have been used for a long time.

2.1 Using the Machine Learning Technique

In a research work done by Paknejad (2018) the author explains all the underlying terms in the analysis and classification of amazon reviews that are listed on the website. They carry out a classification method using machine learning algorithms like Support Vector Machine and Naïve Bayes. The grouping is done into positive and negative reviews to help the user to gain more in-depth understanding of the product. The supervised learning methods is also used here. A comparative study is carried out in this research work and the best algorithms that works good in the classification of reviews in listing

out progressively. An exclusive sentiment analysis of the review classification is carried out in the research study done by Fang and Zhan (2015). In their paper, the author uses the natural language processing technique to examine the reviews and curate them in the specified order. The problem of categorising the polarities of the reviews is also faced and effectively tackled in this work. Information utilized in this review are online item audits gathered from Amazon.com. Tests for both sentence-level classification and survey level arrangement are performed with promising results. Finally, an additionally understanding of the process is also obtained. This paper also sheds light on the other obstacles and challenges involved in this research work. The brief research done by Stanford researchers on twitter classification can be observed from the works of Go et al. (2009). In this research, the authors use the distant supervision approach. They also use machine learning algorithms like Naïve Bayes, Support Vector Machine and Maximum Entropy to evaluate and make the classification. These messages are named either certain or negative concerning an inquiry term. This is helpful for buyers who need to explore the feeling of items before buy, or organizations that need to screen the public opinion of their brands. The accuracy level of this examination is also high and seen to be around 80 percent. A very important step of review classification is the pre-processing. It's significance and emphasis is given out in the research work done by Haddi et al. (2013), where the authors give out the entire role of the pre-processing step in the review classification. They investigate the task of review pre-handling in examining the reviews, and report on test results that show that with suitable element choice and portrayal, opinions, and exactness. For this analysis work the authors utilize support vector machines in this space to improve the devised model. The degree of precision accomplished is demonstrated to be tantamount to the ones accomplished in theme categorisation despite the fact that feeling examination is viewed as a lot more difficult issue in the analysis. Obtaining or outsourcing the customer reviews are not an easy task. In this analysis done by Hu and Liu (n.d.), the authors mean to extract and to sum up all the review listings of an item. This synopsis task is not the same as customary text outline since it is just mining the highlights of the item on which the clients have communicated their suppositions and regardless of whether the assessments are positive or negative. They authors do not gather the reviews by choosing a subset or modify a few of the first sentences from the audits to catch the principle focuses as in the exemplary text rundown. The undertaking is performed in three stages. It proposes a few novel procedures to play out these undertakings. The test results utilizing the reviews of various items sold online exhibit the adequacy of the strategies. Because of overpowering measure of client's viewpoints, perspectives, input and ideas accessible through the web assets Khairnar and Kinikar (2013), it's especially fundamental to investigate, break down and arrange their perspectives and reviews for better direction. Review classification or on the other hand Sentiment Analysis is a Natural Language Processing and Data Extraction task that recognizes the client's perspectives or suppositions clarified as sure, negative or unbiased remarks and statements basic the text. Different regulated or information driven procedures to Sentiment examination like Support Vector Machine and Naïve Bayes through which the task additionally consider opinion grouping exactness and performance of the model is also estimated. The process of review analysis is very important to make the potential customer go for the product they are intending to buy. For that the classification must be done in a very prudent manner which is clearly mentioned in detail in the work done by Liu (2012). The objective of this research work is to give an inside and out prologue to this captivating issue of the review classification and to introduce a far-reaching study of immeasurably

significant examination themes and the most recent improvements in the field. As proof of that, this research covers in excess references from every single significant gathering. Even though the field manages the normal language text, which is regularly thought about the unstructured information, it adopts an organized strategy in presenting the issue determined to connect the unstructured and organized universes and working with subjective and quantitative investigation of the reviews collected. This is essential for pragmatic applications. From the reflection, it is essential to see its key sub-issues. The resulting parts talk about the current methods for tackling these sub-problems.

2.2 Using Deep Learning Technique

Another similar research work that deals with the classification and listing of food review of amazon done using deep learning algorithms is presented by Zhou and Xu (2016). Here, the author employs Long short-term memory and feed forward neural network to carry out this research. It was used to create the baseline of the model presented in the work. The training of dataset is carried out in a proper manner. The performance and accuracy of the model is estimated by taking into account the false errors, and F1-score. The model is also compared and examined with techniques such as matrix factoring and collaborative filtering on the basis of RMSE rating model. The use of Convolutional neural organizations is well known for creating best in class recognizers for archive handling and classification of reviews as said by Chellapilla et al. (2006). In any case, they can be troubles to carry out and are normally more slow than conventional multi-facet perceptrons. We present three module ways to deal with accelerating CNNs. By using the convolution mechanisms and utilizing essential direct polynomial math subroutines along with utilizing other graphic processing units. The Unrolled convolution changes over handling each convolutional layers in the both forward-spread and back-spread into a network grid item.CNNs makes their execution as simple as MLPs. It is utilized to effectively register network items on the CPU. The comparative results demonstrate that unrolled convolution produces a better outcome compared to machine learning models. In another research done by Agüero-Torales et al. (2021), the author portrays the effective method of performing review classification using the deep learning algorithms. The multilingual sentiment analysis is also done in this paper. Numerous spoken languages and social media adjustives are also taken into account for this research work. For this analysis, the Convolutional neural network and long short-term memory is also employed in this paper. Three years of reviews are data are collaboratively used to perform this analysis. Intriguing discoveries of the exploration are the shift of examination interest to cross-lingual and code-exchanging approaches, and the clear stagnation of the less intricate models got from a spine highlighting an inserting layer, a component extractor. From this work, it is evident that the deep learning algorithms are comparatively better than the machine learning methods. Another research by Minaee et al. (2021) gives a forward understanding of the review and text classification that is done using the deep learning techniques. These models have outperformed traditional machine learning methodologies in different message grouping assignments, including analysis, news arrangement, question responding to, and regular language derivation. In here, the authors give an exhaustive survey of in excess of deep learning-based models for text order created as of late, and we talk about their specialized commitments, likenesses, and qualities. At last, they give a quantitative investigation of the exhibition of various deep learning models on well-known benchmarks, and examine future examination bearings. The aspect-based

style of review classification is also done . It is briefly given out in the research work carried out by Do et al. (2019). Existing research for the investigation and classification of reviews is the improvement of granularity at perspective level, addressing two particular points. It is nothing but the extraction of data and classification of item reviews using feeling by characterization of target-subordinate reviews. The use of algorithms of deep learning have arisen as a possibility for accomplishing these points with their capacity to catch both syntactic and semantic elements of text without necessities for undeniable level component designing, just like the case in prior strategies. In this work, we intend to give a relative survey of deep learning for angle-based feeling investigation to put various methodologies in setting. From this research it is known that the deep learning algorithm tend to produce effective outcome in the review classification technique. The product genre specific review is also mentioned in the recent research carried out. In the study presented by Abah (2021) the research work tries to understand the perspective of the customers in dealing with the reviews and understand their mindset. To successfully carry out the analysis, the author makes use of deep learning algorithms like long- and short-term memory and convolutional neural networks. This is done to consequently distinguish and order opinion extremity in Amazon Electronic audit dataset. The crude message is handled into their particular word vector portrayal utilizing GloVe Embeddings. Exactness, Precision, Recall, and F1 Score are utilized to evaluate the chose models. Both models get the Accuracy range of more than 90 percent. The outcomes exhibit that the models can precisely characterize the customer reviews. The various applications and multi-benefits of deep learning on the technique of review classification is explained in-depth in the research work done by Pathak et al. (2020). The classification of reviews is a robotized cycle of extracting information and arranging the opinions as good, negative, and impartial. Absence of enough named information for opinion examination is one of the vital difficulties in the method of Natural Language Processing. The method of deep learning has exuded as one of the exceptionally pursued answers for address this test because of computerized and progressive learning ability innately upheld by the algorithms. Thinking about the use of deep learning approaches for product review classification, intends to advance scientific classification of attributes to be considered for the deep learning-based opinion investigation and complete the job of review sorting.

2.3 Using other Techniques

Apart from the machine learning and deep learning methods, there are some of the exclusive techniques carried out to perform the multilingual review classification. One such research work is given out by Gindl et al. (2010). In this paper, the author uses a very different approach of using a generic high throughout approach for the sentiment analysis of the product review mentioned in the listing. Since it is language specific, the paper uses the method of language analysis. Here the author considers only two languages which are French and English. The best mix of high-throughput strategies and more precise approaches relies upon the particular prerequisites of an application. To oblige a wide scope of potential applications, this paper presents a versatile technique, adjusting exactness and adaptability of multilingual printed sources, a nonexclusive methodology for creating language-explicit punctuation designs and multilingual labelled word references, and a broad assessment confirming the strategy's presentation dependent on Amazon item audits and client assessments from Sentiment Quiz, a "game with a reason" that welcomes

clients of the Facebook interpersonal interaction stage to evaluate the obtained outcomes, which are better. In another research work, the deep learning algorithm is also used with a very unique model which is presented by Shehu et al. (2021). The most commonly used Long and short-term memory is also employed here as a part of the deep learning algorithms. As a unique addition to this research, a Optimization Algorithm is also used for the review classification and the sentiment analysis. The model was utilized to fetch opinions of clients recovered from the reviews of Amazon item. The presentation of the created collaborative model shows an ideal exactness, accuracy, review and F1 proportion of very high amount but separately, when contrasted and LSTM model with a very good rate individually.In another such research, the reviews and comments of social media are taken for analysis. This works is done by Zola et al. (2019), who presents the cross domain and cross source type of review classification method. The author proposes a clever cross-source cross-area review characterization, in which cross-space marked Web sources like Amazon is utilized to prepare managed models using deep learning algorithms, that are tried on regularly non labelled online media audits which are Facebook and Twitter. They also investigated a three-venture strategy, where particular adjusted preparing, text pre-processing and machine learning strategies were tried, utilizing two dialects. They are English and Italian. The best outcomes were accomplished utilizing under sampling preparing and a Convolutional Neural Network. Instead of combining, the deep learning algorithms are normally tends to give a better performing outcome other techniques.

3 Methodology

After the Pandemic, the trend of online shopping has increased tremendously. In the online shopping, the reviews of the product influence the purchasing decision made by the customer. Text classification is one of the most exciting area of Natural language processing, where the sentiments of review can be classified and prediction for the rating can be made.In this research, our main objective is to identify the best deep learning-based model capable of performing multilingual classification on amazon site reviews based on ratings (1-5) from collected data. To achieve this task a consistent flowchart/ Framework is presented which consists of some procedures that include data collection, data cleaning, data preprocessing, exploratory data analysis, visualization, feature extraction, model initialization, model training, and testing. In this section, a detailed explanation of every step is covered. The proposed framework for amazon review classification is shown in Figure 1

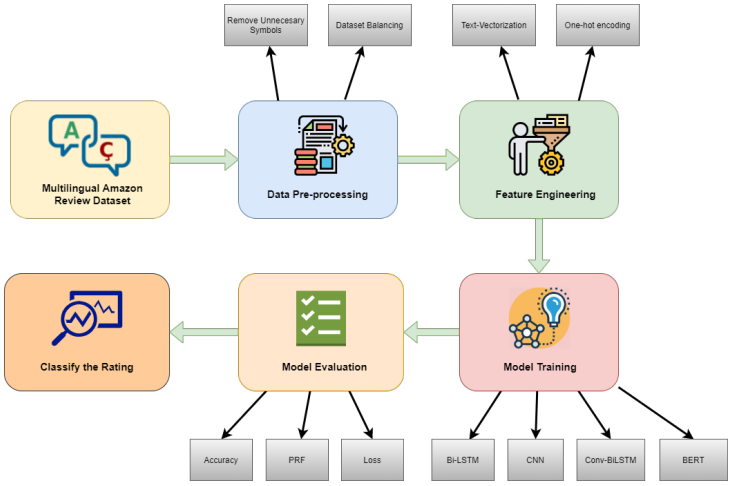

Figure 1: Proposed Framework for Amazon Review Rating Classification

3.1 Data Set Description

To multi-lingual classification, we have collected the data in two different languages that are English and French. The English language dataset consists of 1.7 million reviews. Whereas the number of reviews in French language are 2,50,000. Both datasets contain the same number of columns, and the star rating column is designated as the target column which contains ratings from 1 to 5.

3.2 Data Pre-Processing

Since obtained data has many columns, so unnecessary columns have been dropped from both datasets. Only Review-headline, review-body, and star-rating have remained columns in the dataset. The size of both datasets is large and imbalanced; therefore, 250k shuffled data points are taken to balance both datasets. To make predictions without biasing, each dataset is shuffled first and then concatenated in a single data frame. Columns 1 (review-headline) and column 2 (review-body) are concatenated together to make a single 'review' column and Regex is applied to eliminate unnecessary symbols from review column. Since the target variable contains five classes (1-5) that are unbalanced, therefore a balanced dataset is created by considering the 20,000 reviews from each class. At the final step, the review column value is scaled by performing division by the max value of the review column as shown in Figure 2.

	review	star_rating
118108	du copyright du pur copyright c est tres bi...	1.0
60559	trop bon tout comme shakatak swing out sist...	1.0
1591000	great debut cd if you re expecting the usual p...	1.0
250010	eggleston s best probably the greatest book bi...	1.0
349129	camelot very satisfied with this buy very goo...	1.0

Figure 2: Pre-Processed dataset

3.3 Exploratory Data Analysis

After Pre-Processing of the dataset an analysis and visualization is performed on the processed dataset. Initially, dataset was imbalanced as the label feature i.e. rating is given in between 1 to 5 and data rows associated with each label are different as shown in figure 3.

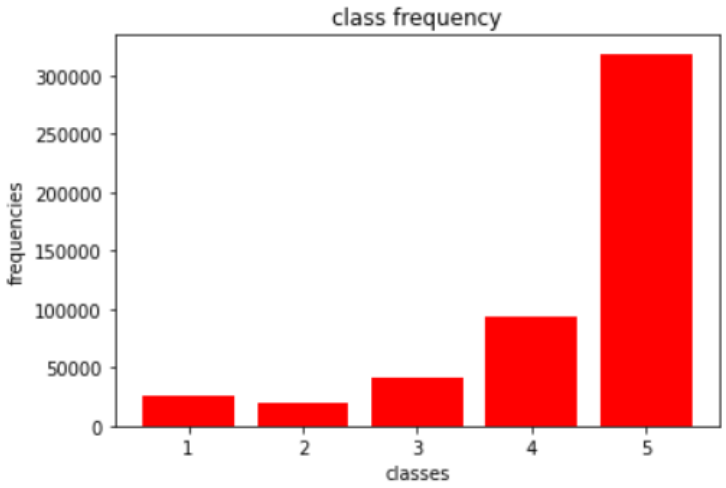

Figure 3: Bar plot showing counts of imbalanced label

From the plot it is observed that rating 5 has the highest number of data rows while rating 2 has the lowest number of data rows. So, this needs to be balanced. After balancing bar plot is plotted that is shown below in figure 4.

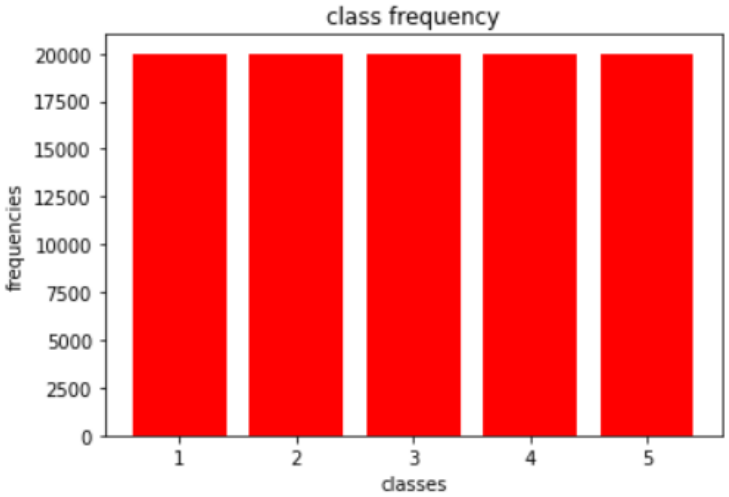

Figure 4: Bar plot showing counts of balanced label

3.4 Feature Engineering

Our task is to classify amazon reviews in two different languages i.e., English and French hence feature extraction and selection is an important step for further tasks. After preprocessing, the whole dataset is dumped into a .pkl file. Text Vectorization is executed

to separate the reviews and ratings which will make features and labels for model building followed by encoder adaptation to encode reviews. Word list and word length is calculated to get a ratio of reviews having ¡ 512 words. One- Hot encoder is applied on labels to encode them. After performing one-hot encoding technique, labels and features are dumped into .pkl files for further use. Final Processed data is shown in figure 5.

	review	star_rating	word_list	word_len
151454	harry potter tome 1 le livre est assez bon dan...	0.8	[harry, potter, tome, 1, le, livre, est, assez...	232
173190	spirituel oui hé bien oui on retrouve le mon...	0.8	[spirituel, oui, hé, bien, oui, on, retrouve, ...	65
225639	joli original bravo à cet éditeur qui propos...	0.8	[joli, original, bravo, à, cet, éditeur, qui, ...	57
70905	trop hétéroclique au niveau de la qualité 3 ...	0.6	[trop, hétéroclique, au, niveau, de, la, quali...	169
549395	i wanna to be better do i wanna know album op...	0.4	[i, wanna, to, be, better, do, i, wanna, know,...	91

Figure 5: Processed dataset with new features

3.5 Model Training

In this work, we have used 4 different algorithms, that are based on Transformer and non-transformer. In this research the non-transformer model are CNN, Bi-LSTM and convolutional Bi-LSTM. On the other hand, for transformed models BERT is considered. Where, for each algorithm, the 80% of the data samples were used for training purposes and the remaining 20% of the data is used as a testing set. All the above models are trained over the cloud instance (Google Colab) with the epoch value as 3.

3.6 Model Evaluation

In this work, our task is to find the best model for analysis and classification of amazon consumer reviews based on historical data. Since this is a classification task, more accurately a multi-lingual classification task, we have used metrics such as accuracy, precision, recall, and F1 score of test data for evaluate different deep learning architectures. After training the model, each model is evaluated on the test dataset and for each model, all metrics have been calculated. The model which has a high value of accuracy, precision, recall, and f1 score can be selected as the best model for further prediction purposes. Loss is another metric, which also has been considered for evaluation.

4 Design Specification

Since 4 different models are deployed for analysis hence the architecture of all models are different. These models are BILSTM, CNN, Conv-LSTM and a BERT Model. In further subsections brief description of all models is provided.

4.1 Bidirectional LSTM Model

BILSTMs stands for bi-directional Long short-term model. It is comprised of two LSTM models and works in a similar manner as LSTM. It stores information that is important for predicting the end outcome. The only difference is that it is bi-directional in nature, so it runs an LSTM algorithm from the beginning and from the end as well. BILSTMs

are widely used in sequential data prediction as it is more efficient than LSTM models. BILSTM takes input, checks whether the current word is contributing something to the outcome, if it is contributing then that word is saved into the storage cell and if not, it is neglected by the gates. It does that one from start to end as well as from end to the beginning. As the model is running bidirectionally it will have more context about the input and will be able to predict more accurately as compared to other sequential models. Now, another task is to merge the outcomes of the two LSTM models, to do that BILSTM uses four techniques that are averaging, multiplication, addition, and concatenation. Based on the usage Bi-LSTM algorithm uses these techniques to merge the two streams and predict the result.

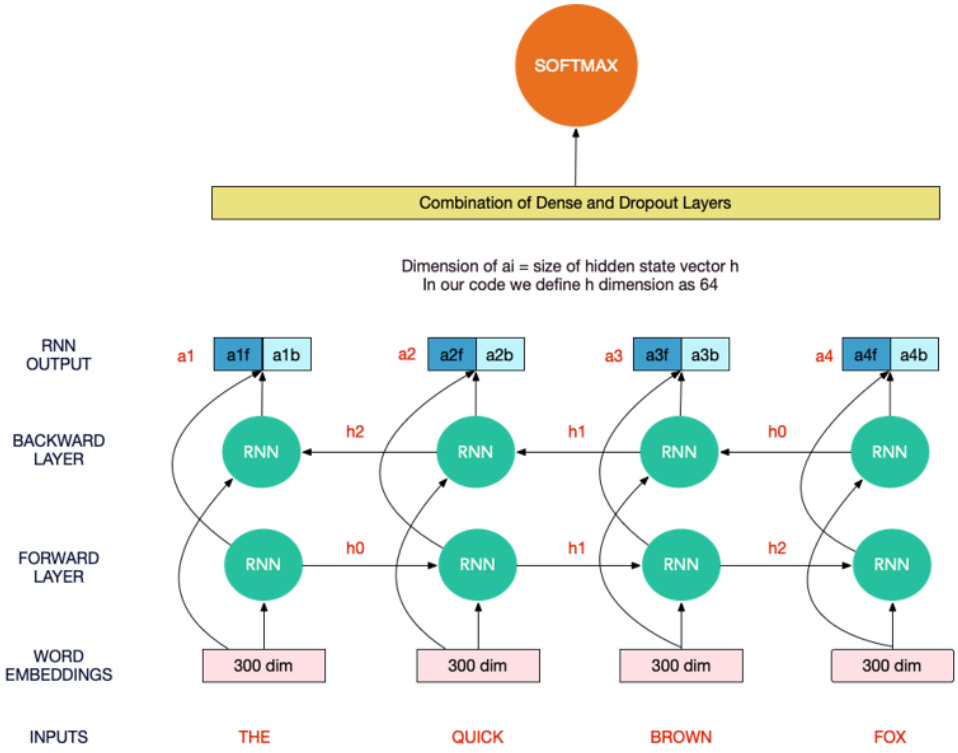

Figure 6: Bi-LSTM Model Architecture

4.1.1 Implemented BILSTM Architecture

The first model being used is a Bi-directional model. In this model we have trained a neural network that is a sequential network. First the embedding layer with encoder where output dimension is 128 is used. Next, a bidirectional layer is used, after these 3 dense layers are used. The first 2 use ReLU activation function functions with 128 and 64 neurons respectively. The last dense layer uses Softmax activation function with 5 neurons. Here the optimizer used is Adam optimizer and Loss function used is the Categorical Cross entropy function.

```
Layer (type)                 Output Shape              Param #
=================================================================
text_vectorization (TextVec  (None, None)              0
torization)

embedding (Embedding)        (None, None, 128)         6400000

bidirectional (Bidirectiona  (None, 256)               263168
l)

dense (Dense)                (None, 128)               32896

dense_1 (Dense)              (None, 64)                8256

dense_2 (Dense)              (None, 5)                 325

=================================================================
Total params: 6,704,645
Trainable params: 6,704,645
Non-trainable params: 0
```

Figure 7: Implemented Bi-LSTM Architecture

4.2 Convolutional Neural Network (CNN 1-D)

A convolutional neural network does the work of the visual cortex for computers. Just like a human brain takes input from whatever the eyes see and processes it to classify and differentiate from other images CNN takes input in the form of images breaks it in the form of a matrix, add biases to each section of the matrix to classify it from different images and then gives output as the result. CNN is used for image recognition and classification. CNN tries to mark and remember features of an image and later compare those features with the features of another image to differentiate between the two. The part with more prominent features will have more bias added to it and the sections that are empty or have fewer features will have reduced bias as it is not accounting for the main image as much as the sections having more features. This helps the CNN model in classifying images.

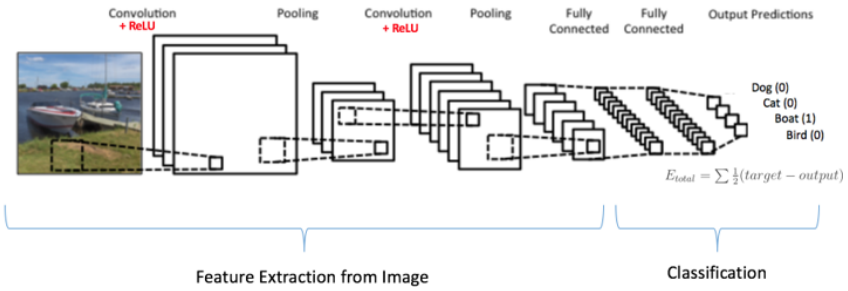

Figure 8: Convolutional Neural Network Model Architecture

4.2.1 Implemented CNN Architecture

The second model used is CNN(1D) model. In this model we have trained a sequential neural network. The first layer is an embedding layer where the output dimension is 128. Next 3 Convolutional (1D) layers are used with 32,64,64 filters and each layer use ReLU activation function. After this flattening layer is used. After flattening, three dense

layers are used. The first two use ReLU activation function with 128 and 64 neurons respectively. The last dense layer uses Softmax activation function with 5 neurons. Here, Also the optimizer used is Adam optimizer and Loss function used is the Categorical Cross entropy function.

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 3725, 128)	6400000
conv1d (Conv1D)	(None, 3723, 32)	12320
conv1d_1 (Conv1D)	(None, 3721, 64)	6208
conv1d_2 (Conv1D)	(None, 3719, 64)	12352
flatten (Flatten)	(None, 238016)	0
dense (Dense)	(None, 64)	15233088
dense_1 (Dense)	(None, 32)	2080
dense_2 (Dense)	(None, 5)	165

Total params: 21,666,213
Trainable params: 21,666,213
Non-trainable params: 0

Figure 9: Implemented CNN Architecture

4.3 Conv-LSTM Model

Convolutional LSTM models are used for the sequential image recognition process. It is used when there is a need to extract information from a sequence of images like a video or images of a report. Images are passed as input in a sequence, now these images when going through the convolutional layers gets filtered and important information is stored in the cells of the LSTM layer, this information is used for predictions and result. Images are processed in the convolutional layers and the work of information extraction and storage is done in LSTM layers. As the ConvLSTM model uses LSTM for data extraction and storage it consumes less computational power and is fast in operation.

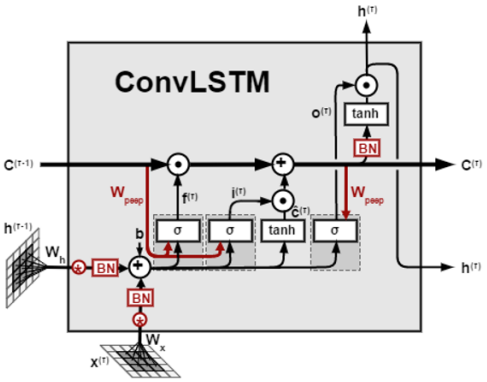

Figure 10: Conv-LSTM Model Architecture

4.3.1 Implemented Conv-LSTM Architecture

The third model being used is the Convolutional LSTM. In this model we have trained a sequential neural network. The first layer is an embedding layer where the output

dimension is 128. The next two layers are the Convolutional (1D) layers with 32,64 filters and each layer uses ReLU activation function. Next, a bidirectional layer is used. After this flattening layer is used. After flattening three dense layers are used.The first two use ReLU activation function with 64 and 32 neurons respectively. The last dense layer uses Softmax activation function with 5 neurons. Here the optimizer used is Adam optimizer and Loss function used is the Categorical Cross entropy function.

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 3736, 128)	6400000
conv1d (Conv1D)	(None, 3734, 32)	12320
conv1d_1 (Conv1D)	(None, 3732, 64)	6208
bidirectional (Bidirectional)	(None, 256)	197632
flatten (Flatten)	(None, 256)	0
dense (Dense)	(None, 64)	16448
dense_1 (Dense)	(None, 32)	2080
dense_2 (Dense)	(None, 5)	165

Total params: 6,634,853
Trainable params: 6,634,853
Non-trainable params: 0

Figure 11: Implemented Conv-LSTM Architecture

4.4 Transformer BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning algorithm which was created and published by Jacob Devlin and his colleagues in 2018. It is a transformer based deep learning for natural language processing (NLP). The original BERT which was used for English language has 2 models :- the BERT BASE: 12 Encoders with 12 bidirectional self-attention heads, and the BERT LARGE: 24 Encoders with 16 bidirectional self-attention heads. With the help of BooksCorpus with 800M words 2,500M words from English Wikipedia these models are pre-trained using unlabeled data. It has an inconsistent number of self-attention heads and varying number of encoders.At its core it is a Transformer language model and the architecture is comparable to the original implementation of the transformer in Vaswani et al. The two tasks on which BERT was trained on were modeling of language where it was trained to predict the words from context the masked tokens which were 15 percent and prediction of next sentence, where BERT was trained based on the first sentence whether the next sentence is admissible or not. Due to this training contextual word embedding is learnt by BERT. BERT can be fine tuned with less resources using smaller database after this pre-training which is usually expensive computationally. It is a state of the art model which gives great performance on many natural language understanding tasks like SWAG(Situation with Adversarial Generation) GLUE (General Language Understanding Evaluation) task set which consists of 9 tasks and SQUAD (Stanford Question Answering Dataset) v1.1 and v2.0.

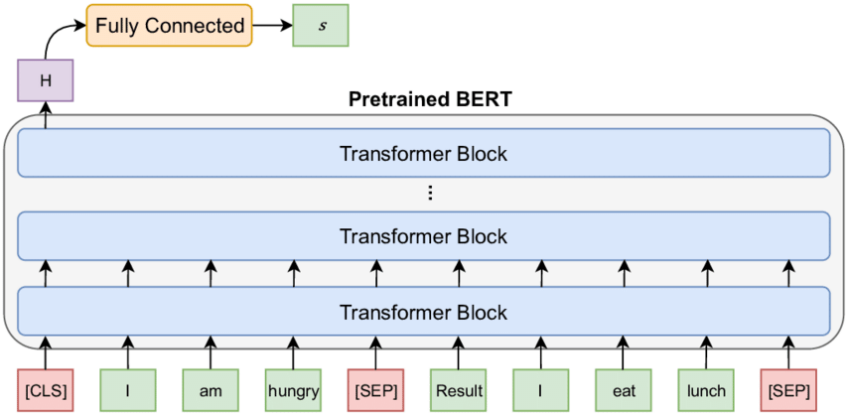

Figure 12: BERT Model Architecture

4.4.1 Implemented BERT Architecture

The fourth model being used is the Transformer BERT model. In this model we have trained a sequential neural network. The first layer is the input layer which takes text as the input. The next two layers are Keras Layer, first one is used for preprocessing and the next one is the BERT-encoder. The next one is the dropout layer and the last one is the dense layer (classifier).

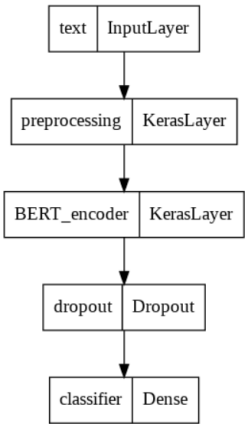

Figure 13: Implemented Bert Architecture

5 Implementations

We have implemented four deep learning models (BILSTM, CNN, Conv-LSTM and BERT model)We have implemented four deep learning models (BILSTM, CNN, Conv-LSTM and BERT model) and best model is selected whose value of accuracy, Precision, Recall and f1score on validation data is highest for same data. All Models with ADAM optimization function and categorical cross entropy as loss function is used and trained on the training dataset. All models are fitted with training data and trained for 3 epochs and tested on test dataset. There are numerous numbers of libraries utilized to implement the proposed work which includes pandas, NumPy, matplotlib, seaborn, TensorFlow,

keras, sklearn etc. The whole experiment is performed into google colab for training purpose with python as programming language. The following specification were required to implement the model.

- Operating System : windows 10
- Random Access Memory (RAM) : 12GB (Provided by Colab)
- Hard disk : 15GB (Provided by Colab)
- Languages : Python
- Cloud Platform : Google Colab
- Python libraries : numpy, Pandas, matplotlib, tensorflow, numpy, seaborn and keras.

6 Evaluation

In this work, we aimed to select the best optimal model for classifying the review ratings from amazon reviews, so it is necessary to evaluate each model based on some metrics. Since this is a classification task, more accurately a multilingual classification task, therefore test accuracy, test precision, test recall, and test F1 score are some different metrics to evaluate different deep learning algorithms. After training, each model is evaluated on the test dataset and for each model, all metrics have been calculated. The model which has a high value of accuracy, precision, recall, and f1 score can be selected as the best model for further prediction purposes. A comparison is made with bar plots to visualize the same.

6.1 Experiment 1 / Evaluation Based on Accuracy

The BILSTM model is trained on 3 epochs. During training, it has been observed that at the end of three epoch the model's validation loss starts to increase which indicates the overfitting of the model therefore early stopping of the model is implemented to prevent the model from overfitting. During training, the highest accuracy score using Bi-LSTM is 57.02%. The accuracy of Bi-LSTM model over every epoch is represented with the help of line graph in Figure 14. On observing the graph, training accuracy was found to be increasing on every epoch but after second epoch the validation accuracy was found to be decreasing this represents the overfitting behaviour of Bi-LSTM model after 2 epochs.

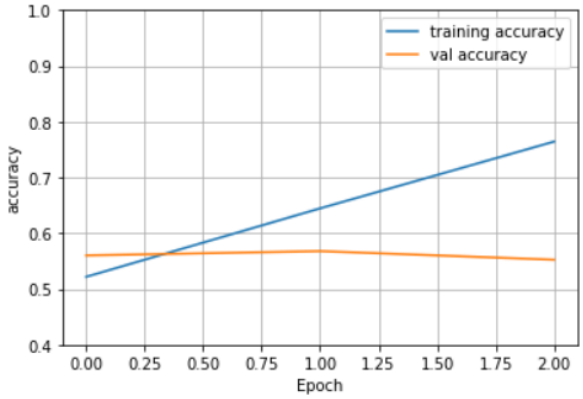

Figure 14: Bi-LSTM Training and Validation Accuracy

The Convolutional neural network (1-d) Model is trained on three epochs and within three epoch the model starts to overfit therefore, to prevent it from over-fitting, model is trained only for three epochs. Highest validation accuracy obtained using convolutional neural network is is 49.18%. The accuracy score of CNN model over every epoch is shown in Figure 15.

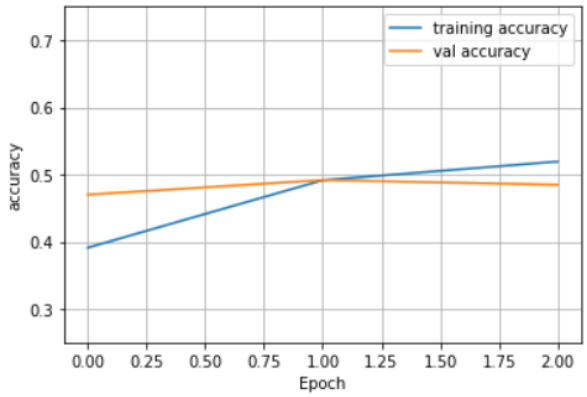

Figure 15: CNN Training and Validation Accuracy

The Conv-BiLSTM model is trained over 3 epochs and it starts overfitting for even 3 epochs, therefore the model is stopped from training further. Validation accuracy of this model is 49.76%. The accuracy score of Conv-BiLSTM model over every epoch is shown in Figure 16.

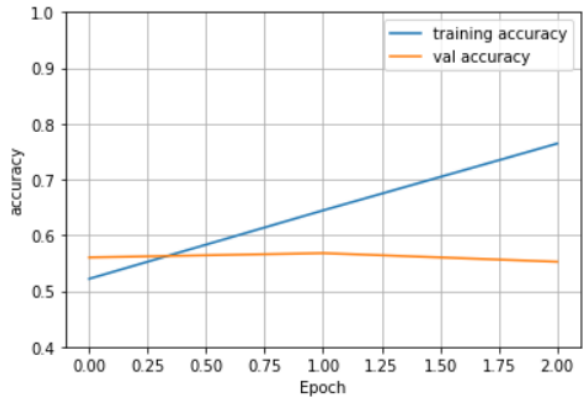

Figure 16: Convolutional Bi-LSTM Training and Validation Accuracy

The Transformer BERT model is utilized for multilingual text classification, as this is a pretrained model and after training of this model for 10 epoch, it starts overfitting. Therefore, model has been trained over 10 epochs. Highest accuracy obtained using BERT Model is 53.51%.

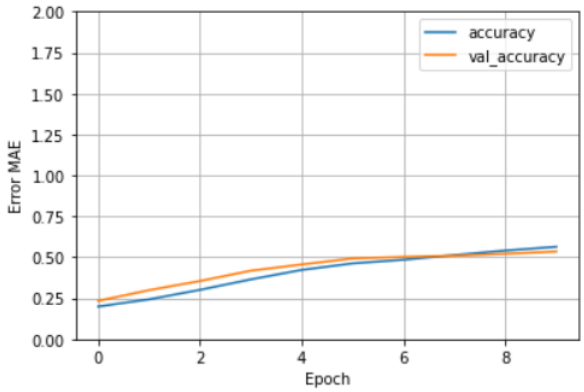

Figure 17: BERT Training and Validation Accuracy

Comparison graph of all implemented models based on accuracy is shown in figure 18. After analyzing the graph it has been observed that using Bi-LSTM architecture, a better accuracy score of 57.02% is achieved, followed by the BERT model with the accuracy score of 53.51%. The other models CNN and Convolutional-BiLSTM generates the poor results with accuracy score of 49.87% and 49.76%.

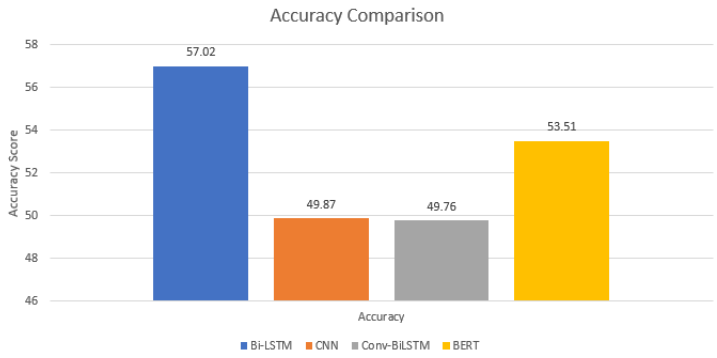

Figure 18: Comparison of models based on Accuracy

6.2 Experiment 2 / Evaluation Based on PRF Score

Precision, recall and F1-score informs about the Correctly and incorrectly identified classes between the actual and predicted results. The PRF Score (precision, recall and F1-score) has been calculated for all the algorithms For the BILSTM,the values of Precision, recall and f1 scores on validation data are 0.549, 0.545 and 0.546 respectively. For the CNN 1-d model, the values of Precision, Recall and f1 score are 0.487, 0.491 and 0.488 respectively. For the Conv-LSTM the values of precision, recall and f1 score are 0.482, 0.492, and 0.482 respectively for each metric. For BERT model the values of precision, recall and f1 score are 0.537, 0.535 and 0.533 respectively. Comparison of all implemented models based on PRF scores is shown in figure 19. The highest PRF score has been obtained using Bi-LSTM algorithm, followed by transformed architecture (BERT Model).

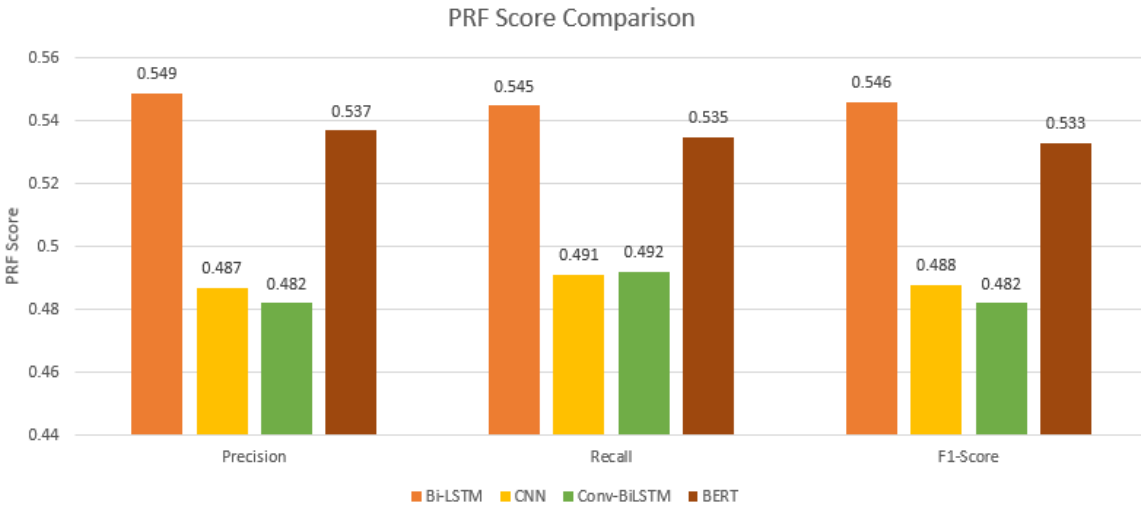

Figure 19: Comparison of models based on PRF Score

6.3 Experiment 3 / Evaluation Based on the Validation Loss

In this section validation loss is used to evaluate the model. For better prediction, the values of validation loss should be minimum. For the BILSTM model, the value of val-

idation loss is 1.08. For Convolution Neural Network (CNN 1-d) the value of validation loss is 1.16. For Conv-Long short term memory (Conv-LSTM) model, the value of validation loss is 1.17. For the Transformer (BERT) model, the value of validation loss is 1.08. Comparison of all implemented models based on validation loss is shown in figure 20. We have obtained the minimum loss using the Bi-LSTM model, followed by BERT, Conv-BiLSTM and CNN.

Figure 20: Comparison of models based on Validation Loss

6.4 Discussion

After certain set of experiment over the multilingual amazon review dataset, the best model is identified as Bi-LSTM model, which is non-transformed type of model. The model outperforms in all the aspects which includes the precision, recall, F1-score, accuracy and validation loss. However, the Bidirectional Encoder Representations from Transformers (BERT) also produced the very close results to the Bi-LSTM model. As the size of dataset is large, the both the consumed more than 4 hours of time for training. In order to analysis the miss-classified classes using Bi-LSTM model, we have calculated the confusion matrix for every class over the test data, the image for which is shown in Figure 21. The diagonals in the confusion matrix represents the correct classification. Where non-diagonal elements are not classified correctly. For example: there are 79 such reviews, whose actual rating was 4 star but our Bi-LSTM model predicted the rating as 0. Similarly, there are 973 such reviews, which actually belong to class 3. but predicted by Bi-LSTM model as class 2.

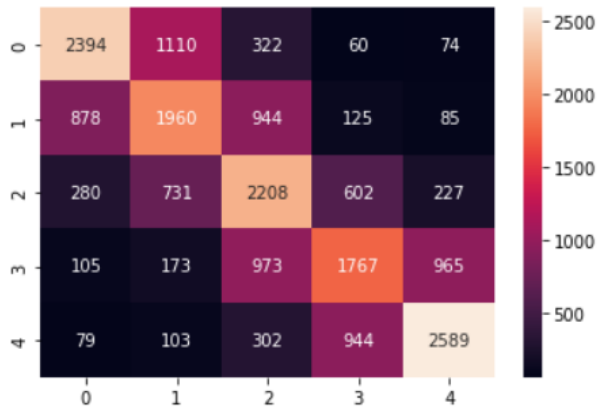

Figure 21: Confusion Matrix of Bi-LSTM over Test data

7 Conclusion

In today's era, Natural language processing (NLP)is building a strong capability to deal with the multi-lingual dataset. However, analyzing and identifying the rating from Multilingual amazon review dataset is a challenging task. Therefore, in this research we have utilized the capabilities of word-embedding technique for classifying the review-based amazon multi-lingual dataset. We have mainly collected the amazon reviews dataset for two languages that are English and French. Our proposed framework using Bi-LSTM algorithm can correctly identify the ratings with accuracy score of 57.02%. The main advantage of using the word-embedding approach is that we don't need to train the data for every language separately. Whereas the combined data of any language can be trained, and rating can be predicted. Our current analysis has been performed over 100k reviews that includes both English and French language. The current research fulfills its goal to correctly identify the rating of amazon review for multiple languages. However, in the future work the review data for more languages as such as Chinese, Japanese, Russian can be Incorporated. There are some large complex models which contains the millions of features vector parameter can be utilized for better results. However, for the complex model the training time will be very high and large amount of computing resources will be required to train the model.

References

Abah, J. O. (2021). *Sentiment Analysis of Amazon Electronic Product Reviews using Deep Learning*, PhD thesis, Dublin Business School.
URL: *https://esource.dbs.ie/handle/10788/4291*

Agüero-Torales, M. M., Abreu Salas, J. I. and López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview, *Applied Soft Computing* **107**: 107373.
URL: *https://www.sciencedirect.com/science/article/pii/S1568494621002969*

Chellapilla, K., Puri, S. and Simard, P. (2006). High Performance Convolutional Neural Networks for Document Processing, p. 7.

Do, H. H., Prasad, P., Maag, A. and Alsadoon, A. (2019). Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review, *Expert Systems with Applications* **118**: 272–299.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417418306456*

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data, *Journal of Big Data* **2**(1): 5.
URL: *https://doi.org/10.1186/s40537-015-0015-2*

Gindl, S., Scharl, A. and Weichselbraun, A. (2010). Generic high-throughput methods for multilingual sentiment detection, *4th IEEE International Conference on Digital Ecosystems and Technologies*, pp. 239–244. ISSN: 2150-4946.

Go, A., Bhayani, R. and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision, p. 6.

Haddi, E., Liu, X. and Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis, *Procedia Computer Science* **17**: 26–32.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050913001385*

Hu, M. and Liu, B. (n.d.). Mining and Summarizing Customer Reviews, p. 10.

Khairnar, J. and Kinikar, M. (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classification, **3**(6): 6.

Liu, B. (2012). Sentiment Analysis and Opinion Mining, p. 168.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. (2021). Deep Learning–based Text Classification: A Comprehensive Review, *ACM Computing Surveys* **54**(3): 62:1–62:40.
URL: *https://doi.org/10.1145/3439726*

Paknejad, S. (2018). Sentiment klassificering pa Amazon recensioner med hj¨alp av maskininl¨arningstekniker, p. 25.

Pathak, A. R., Agarwal, B., Pandey, M. and Rautaray, S. (2020). Application of Deep Learning Approaches for Sentiment Analysis, *in* B. Agarwal, R. Nayak, N. Mittal and S. Patnaik (eds), *Deep Learning-Based Approaches for Sentiment Analysis*, Algorithms for Intelligent Systems, Springer, Singapore, pp. 1–31.
URL: *https://doi.org/10.1007/978-981-15-1216-2*$_1$

Shehu, S. A., Mohammed, A. D. and Abdullahi, I. M. (2021). An Optimized Customers Sentiment Analysis Model Using Pastoralist Optimization Algorithm (POA) and Deep Learning, *2020 IEEE 2nd International Conference on Cyberspac (CYBER NIGERIA)*, pp. 132–139.

Zhou, Z. and Xu, L. (2016). Amazon Food Review Classification using Deep Learning and Recommender System, p. 7.

Zola, P., Cortez, P., Ragno, C. and Brentari, E. (2019). Social Media Cross-Source and Cross-Domain Sentiment Classification, *International Journal of Information Technology & Decision Making* **18**(05): 1469–1499. Publisher: World Scientific Publishing Co.
URL: *https://www.worldscientific.com/doi/abs/10.1142/S0219622019500305*