ANKARA UNIVERSITY

ENGINEERING FACULTY

DEPARTMENT OF COMPUTER ENGINEERING

GRADUATION THESIS

A Regression-Based Deep Learning Approach to Sentiment Analysis using BiLSTM

Berk HAYIRLI

18290099

Ayşenur CANLIDIR

18290083

Supervisor

Assistant Professor Yılmaz AR

May 2022

ABSTRACT

Sentiment analysis is an important field of study today, made possible by the exponential increase of digitalism in the last decade and vast amounts of textual data generated especially during the pandemic. Numerous studies have been conducted on varying classification models using both machine learning and deep learning approaches. However, only a few regression-based ones exist and therefore leaves much to be explored. In this research, we investigate the correlation between the Amazon customer reviews and their respective star ratings with a regression perspective using a deep recurrent neural network (RNN) with BiLSTM (Bidirectional LSTM) architecture.

INTRODUCTION

There has been an increase in efforts to study and understand sentiment of textual data in the past years because of decades-long trends of digitalisation and resulting abundance of data. This trend only escalated further with the hit of the pandemic, especially to a greater degree for social media and e-commerce (e.g., Social Networks, Online Shopping). Furthermore, in the light of recent developments of the state-of-the-art methods in the field of A.I. and Big Data, a need for customer-centric businesses to analyse and interpret consumer opinions and emotions about their products has arisen. For this reason, the study of sentiment analysis became a crucial factor for both academic researchers and digital businesses.

Moreover, the exponential increase of computational power throughout the decade made possible to apply and get better results from neural networks on NLP tasks. Traditional feed-forward networks and recurrent neural networks are still being studied in whole while more specialised networks (e.g. LSTM, GRU) are being explored. A specific implementation of LSTM architecture, namely BiLSTM, has attracted sizeable attention as it achieved state-of-the-art performances in NLP fields such as text and speech understanding [7]. This can be attributed to the fact that, unlike LSTM, BiLSTM can make use of sequential information in both forward and backward contexts.

The aim of this paper is to investigate whether a regression approach to sentiment analysis task with deep learning is feasible. The dataset we use is Amazon Customer Reviews dataset that consists of product reviews and their associated star rating labels. The multi-class labels are mapped into continuous range of values. Then we perform an empirical exploration of several BiLSTM models by training them with varying hyperparameters (e.g., model architecture, model complexity, loss function, learning rate, regularization methods, number of epochs) to tune them and analyse the results.

Related Work

Numerous studies [1][2][3][5] have been conducted in the past regarding the classification approach to sentiment analysis task. Many traditional machine learning methods (e.g. Naïve Bayes [1][2], K-nearest Neighbours [2] and Support Vector Machines [1][2]) as well as deep learning methods (e.g. CNN [3], types of RNNs [3] (LSTM [1][2], GRU)) have been experimented upon in the past. Moreover, Amazon Customer Review Dataset, which is the dataset we also opted to use [6], is pretty popular among the research [1][2][5]. With the help of pre-trained word2vec dictionaries [3], they were able to correctly classify the sentiments of reviews with accuracies as high as high as %91.

Even so, there are hardly any regression-based approaches [4]. One study [4] assesses different feature representations and hyperparameters using financial domain dataset in two state-of-the-art models - SVR and CNN using pre-defined dictionary of words and word embeddings with known sentiment knowledge such as GloVe and VADER. They, in a similar way to us, treat class labels as numeric values and map the labels into a continuous range. In a range of [-1, +1], they were able to get an MSE value of 0.09 with CNN and 0.10 with SVR.

MATERIAL AND METHODOLOGY

Dataset

The dataset we opted to use is a part of toys dataset which is a subset of Amazon product reviews dataset provided by TensorFlow. We drop the columns from the dataset that are not essential for the model implementation (e.g., customer id, product category, review id). However, due to hardware limitations and time constraints, we slice the rows to reduce the size of the dataset of about 5 million data points in a stratified manner to preserve label distributions. The final dataset consists of about half a million data points and 2 columns, being reviews and their multi-class star rating labels.

Table 1: Star Rating Label Distributions of Reviews

1 Star	2 Star	3 Star	4 Star	5 Star	Total
43904	25527	42291	84970	338375	535567

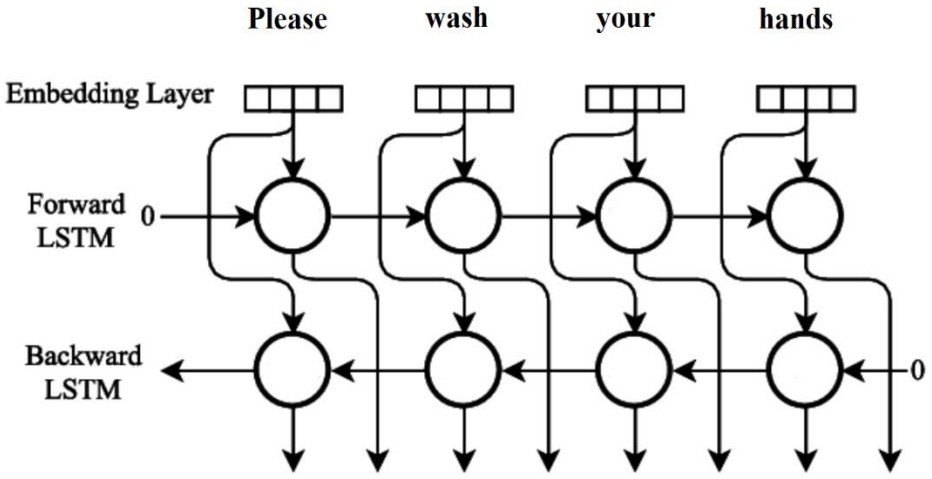
A number of conventional NLP preparation steps are applied to reviews such as stop words removal, punctuation removal, lowercasing and tokenisation with space as separator. The discrete class star labels {1,2,3,4,5} are scaled to the continuous range of [0,1]. The range [-0.5, +0.5] is also experimented upon accompanied with a change in output layer activation function and it will be discussed later thoroughly.

Review lengths were also a consideration during data preparation. Due to nature of dataset, Amazon reviews differed in length significantly – from a single word to several thousands of words. For this reason, we have empirically observed that about %98 of reviews does not indeed exceed 300-word length and have decided to slice other words that comes after it if the review exceeds this amount. For reviews shorter than 300-word, post-padding of 0s during tokenisation is applied.

Model Implementation

A BiLSTM or a Bidirectional LSTM is an RNN architecture that is composed of two LSTMs. It works in a similar fashion to LSTMs but one takes input in a forward direction and the other in a backward direction. This is convenient in text analysis because of the nature of sentences, words can have meanings depending on the future context as much as the past context. Therefore, this approach increases the amount of context available to the network.

Figure 1: Bidirectional LSTM Architecture



We have trained and tested about 70 different models, all utilising BiLSTM layers. These models are obtained by changing an hyperparameter to analyse the difference in results while holding the other hyperparameters fixed to compare models with each other. After the poor-performing model is discarded, same procedure is applied until all hyperparameter options are exhausted and the best performing model is obtained. They can be observed in Table 2.

Table 2: Explored Hyperparameters for Model Selection

(Best Performing are highlighted in green)

Embedding Layer:	input dimension = 10,000 / 20,000
	output dimension = 64 / 128 / 256
	input length = 100 / 200 / 300 / 400
BiLSTM Layer: 1st BiLSTM Layer:	Width = 32 / 64 / 128
: 2nd BiLSTM Layer:	Width = 64 / 32 / 16 / none*
Dense Layer:	Number of Neurons = 12 / 24 / 36
	Activation Function = ReLU / LeakyReLU
Dropout Layer:	Dropout Rate = 0.1 / 0.2 / 0.5
Dense Layer:	Activation Function** = sigmoid / tanh
Training:	Loss Function = MAE / MSE
	Learning Rate = 0.01 / 0.001
	Batch Size = 32 / 64 / 128

(*none refers to having no 2nd layer)
(Ranges [0,1] and [-0.5, +0.5] for **Activation Function sigmoid and tanh respectively)

Throughout the process out hyperparameter tuning, the entire dataset of about half a million data points have been split as %80 for training set, %10 for validation set and %10 for test set. After the best model in terms of hyperparameters is chosen, we applied a k-fold cross-validation split in which k is set as 5. Folds have been split in a stratified manner to preserve star rating distributions across all folds.

Results

Table 3: Cross-Validation Metrics of Each Fold and Their Average

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
MAE	0.099	0.097	0.101	0.099	0.094	0.098
MSE	0.0263	0.0269	0.0262	0.0261	0.0263	0.02636
RMSE	0.162	0.164	0.162	0.161	0.162	0.1622

Figure 2: Regression Value Distributions of Rating Labels and Their Average RMSEs

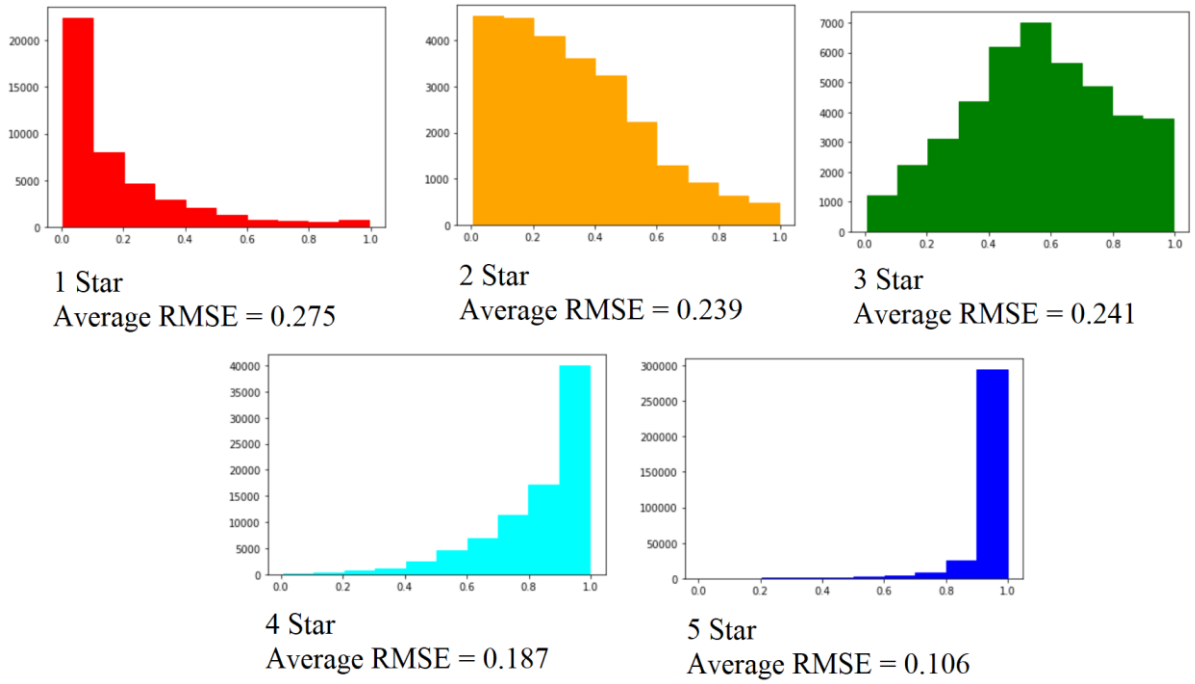
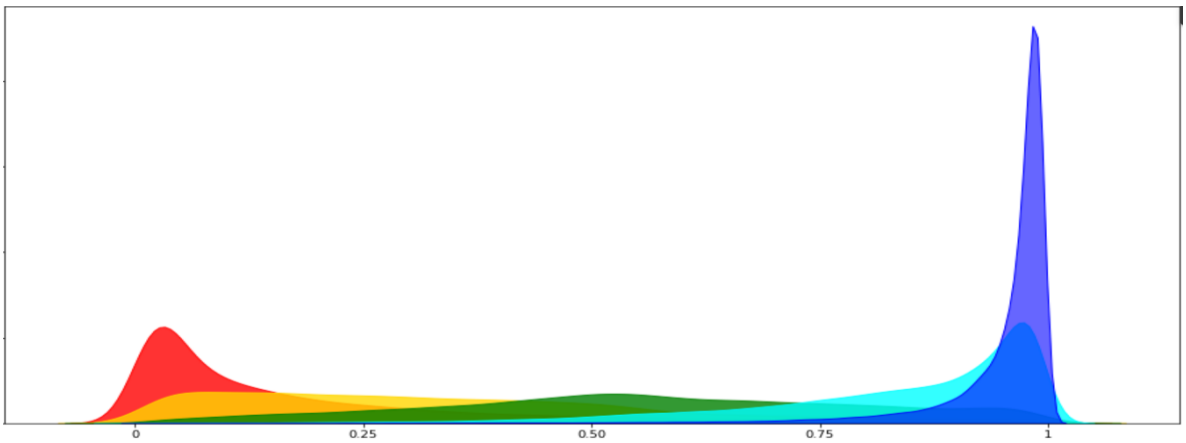


Figure 3: A Combined View of Distributions of Values in Figure 2



Conclusion and Discussion

As mentioned previously, due to hardware limitations and time constraints, we had to reduce the amount of data points used in model implementation from ~5 million to about half a million. Additionally, uneven distribution of labels in our dataset greatly affected our predictions as it can be observed in RMSE values in Figure 2. These naturally had a negative impact on our results.

It is also an option to meddle with range of continuous values depending on different input, hidden and output activation functions as data standardisation affected our results slightly whatsoever. Positive ranges of continuous values ([0, 1]) tied up with sigmoid/softmax output activation functions consistently performed better than model's tanh output function counterparts with a value range distribution with a mean of zero ([-0.5, +0.5]). Furthermore, embedding layer hyperparameters such as dictionary size (input dimension) and maximum review size (input length) had an impact to a degree in performance as our chosen values was inclusive of the input space and increasing them yielded to exponential decrease in performance improvement. Review length and distinct word distributions of the dataset can be observed in Figure 4 and 5. Next to that, change in dimension of the word vector (output dimension) had negligible effect.

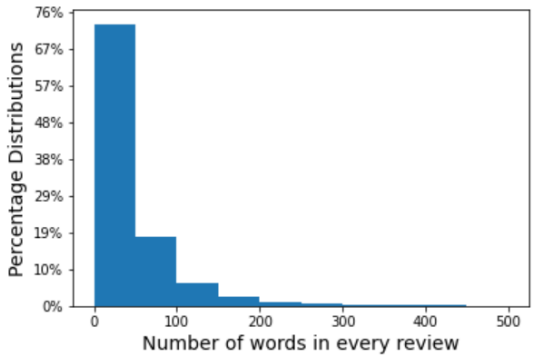


Figure 4: Input Length

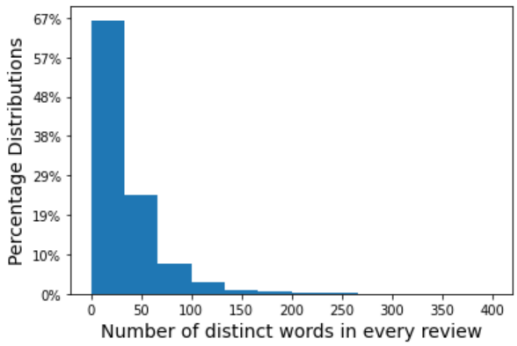


Figure 5: Input Dimension

Training is also not optimised in the best possible way as different number of epochs and batch sizes and their effect on the final results can be explored more thoroughly without the limitations of hardware and constraints on time.

Nevertheless, the results are promising. In terms of future work, the amount of data points can be increased for better results and an even distribution of star ratings class labels can help greatly to build a model that is better at generalising. Also, use of predefined lexical dictionaries and their impact on performance have been explored before [4] and can also be applied in such a context.

References

[1] Sentiment analysis for Amazon.com reviews Big Data in Media Technology (DM2583) KTH Royal Institute of Technology, Stockholm https://www.researchgate.net/profile/Levent-Guener/publication/332622380_Sentiment_analysis_for_Amazoncom_reviews/links/5cc08696a6fdcc1d49acb839/Sentiment-analysis-for-Amazoncom-reviews.pdf

[2] Sentiment Analysis for Amazon Reviews by Wanliang Tan, Xinyu Wang, Xinya Xu http://cs229.stanford.edu/proj2018/report/122.pdf

[3] Sentiment Analysis using Recurrent Neural Network by Lilis Kurniasari and Arif Setyanto 2020 J. Phys.: Conf. Ser. 1471 012018

[4] Assessing Regression-Based Sentiment Analysis Techniques in Financial Texts https://www.researchgate.net/publication/339962669_Assessing_Regression-Based_Sentiment_Analysis_Techniques_in_Financial_Texts

[5] Amazon Food Review Classification using Deep Learning and Recommender System by Zhenxiang Zhou Department of Statistics Stanford University Stanford, CA 94305 zxzhou@stanford.edu Lan Xu https://cs224d.stanford.edu/reports/ZhouXu.pdf

[6]Amazon Toys Dataset https://www.tensorflow.org/datasets/catalog/amazon_us_reviews#amazon_us_reviewstoys_v1_00

[7] Graves, A.; Jaitly, N.; Mohamed, A.R. Hybrid Speech Recognition with Deep Bidirectional LSTM. In Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU), Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.