Applied Artificial Intelligence:

Sentiment Analysis of Titled Hotel Reviews with Shallow Gated Recurrent Units

MSc in Artificial Intelligence and Data Science

Osagie Elliot Aibangbee

202115576

May 2022

ABSTRACT

The growing volume of existing online reviews today far exceeds the reading capacity of humans. Therefore, there is an urgent need for the introduction of more innovative methods to automate the task of understanding customers' mindsets through the sentiment analysis of customer reviews (Shi et al 2011). A big part of business success in the hospitality industry (like most industries) is derived from having a good market reputation. Sentiment analysis now provides a way for HSPs to gain a better understanding of how they are perceived in the marketplace by existing and potential customers. In this paper, the focus was on optimizing the hyperparameters of a shallow GRU that can perform sentiment analysis of titled hotel reviews with comparable results to other state-of-the-art algorithms and architectures. As presented in the experimental results, with optimized hyperparameters, a GRU model with one hidden layer improved tremendously in performance and outperformed an SVM classifier. In addition, when tuning the GRU model more attention should be given to the number of units, batch size, and learning rate than the number of layers.

I. INTRODUCTION

Since the advent of the internet, people have begun to communicate at an increasingly faster rate, and in the last decade, online social interactive platforms have become one of the most ubiquitous methods of communication worldwide. This has resulted in an explosion in the volume of textual data from platforms such as discussion and product-review forums, blogs, social media, etc. Human interaction essentially involves the expression of diverse individual opinions, which often convey emotions in relation to the topic being addressed. The growing volume of existing online reviews today far exceeds the reading capacity of humans. Therefore, there is an urgent need for the introduction of more innovative methods to automate the task of understanding the customers' mindset through sentiment analysis of customer reviews (Shi et al 2011).

A big part of business success in the hospitality industry (like most industries) requires a good market reputation. Potential customers, hoteliers, and other stakeholders in the hospitality industry now directly access the market reputation of hospitality service providers (HSP) in ways that used to be considered impractical but have become virtually a routine due to advancements in tools and computing resources. As more customers share their experiences about HSP, there's a growing need for the interpretation of these diverse opinions into an overall assessment. HSP can take advantage of this trend by analyzing online reviews using machine learning algorithms.

Sentiment analysis involves using a machine-learning algorithm to identify the emotional value of textual opinionated data. Generally, analyzing reviews means classifying documents into a positive, neutral, or negative sentiment class. The task of categorizing reviews based on sentiment polarity can be complicated due to linguistic nuances, hence the necessity for improvement in sentiment analysis methods. According to Zhang et al (2018), sentiment analysis is one of the active research areas of natural language processing (NLP) today.

Researchers have proposed novel ideas for the efficient implementation of algorithms for sentiment analysis using mainly two different techniques namely the lexicon-based (Piryani et al, 2017; Maity et al, 2020) and machine learning approaches (Shi et al, 2011; Hsieh et al, 2014; Zvarevashe and Olugbara, 2018; Ray et al, 2021, Wu et al, 2021). Earlier, researchers used various supervised machine learning techniques such as support vector machine (SVM), naïve Bayes, etc., and feature combinations (Zhang et al, 2018). However, in recent years, deep learning methods have received a lot of attention, which has led to the development of several advanced algorithms for sentiment analysis featuring long short-term memory units (LSTM), gated recurrent units (GRU), Bi-directional LSTM (Bi-LSTM), etc.

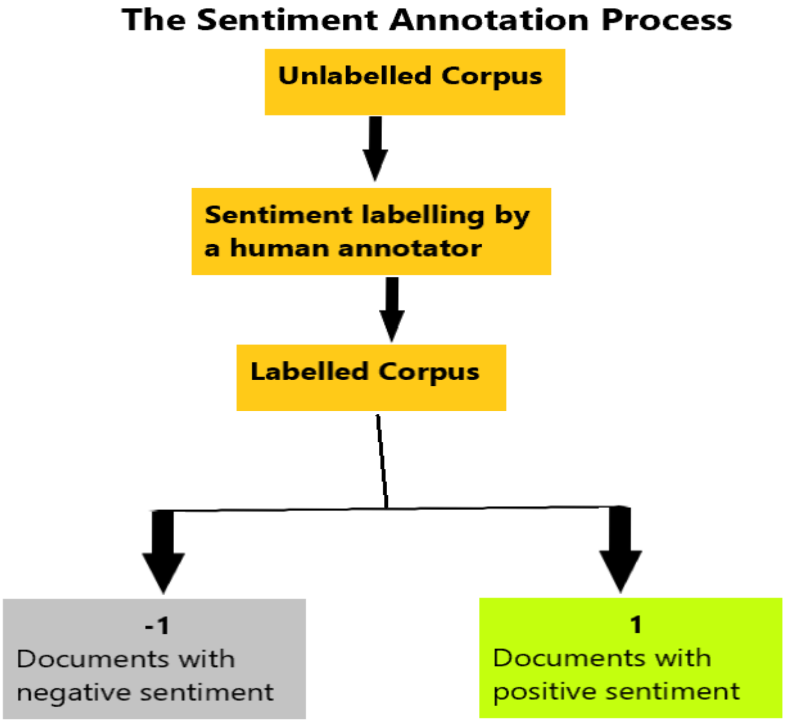
This study will focus on providing a relatively simple, computationally inexpensive approach to sentiment analysis of hotel reviews. Most of the architectures that have been proposed are resource-intensive and computationally expensive to implement. Hence, a shallow GRU with competitive performance would be a novel approach that could be useful in the domain of opinion mining. There will be manual hyperparameter optimization of the shallow GRU, as we look to develop an efficient GRU architecture for document-level sentiment classification of titled hotel reviews. Also, there will be a comparison of two supervised learning frameworks for sentiment analysis, SVM and GRU. Both algorithms will be evaluated on the same data. It is expected that the deep learning model with effectively optimized hyperparameters would outperform the SVM model.

II. RELATED LITERATURE

Sentiment Annotation

When customers submit reviews, they generally do not follow a defined pattern or writing style which makes it difficult to automatically detect the emotional dimension of the comment (document) without human intervention. Thus, human annotators play a key role in the entire process of opinion mining, as they are responsible for interpretating and assigning an overall sentiment to each document. However, manual sentiment annotation can be very inconsistent due to domain expertise, mood, personal assumptions, etc. (O'Hare et al, 2009).

Zvarevashe and Olugbara (2018) proposed a technique for labeling reviews without human involvement. In this research, however, the labels were manually assigned by a human agent due to linguistic nuances. Keywords were used as a guide during document annotation and special attention was given to occurrences of negation (of positive and negative expressions) like "not bad", "not okay", "not good", etc. An overview of the entire process of labeling the corpus is illustrated in the diagram below.


The Sentiment Annotation Process
Unlabelled Corpus
Sentiment labelling by a human annotator
Labelled Corpus
-1
Documents with negative sentiment
1
Documents with positive sentiment

Sentiment Classification

Sentiment analysis can be executed at three levels of granularity: document, sentence, and aspect level (Zhang et al 2018). This study will be performed under the assumption of topic relevance because each author's opinion was completely dedicated to the hotel under review and there was no room for a topic shift.

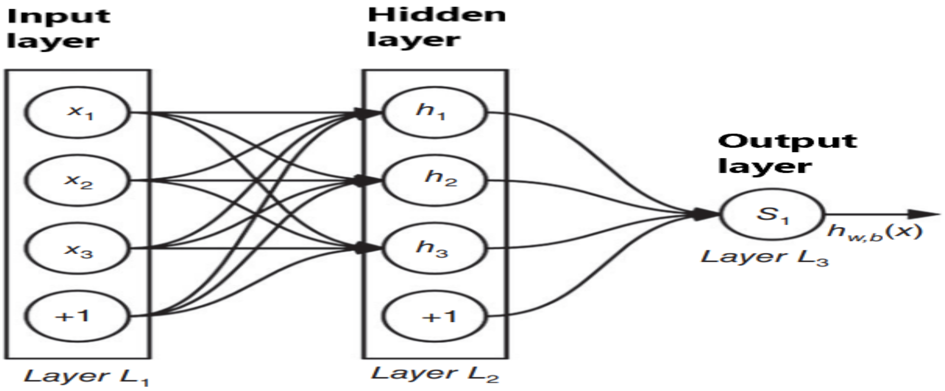
SVM

SVM is a supervised learning algorithm that learns linear and nonlinear relationship(s) between the input variables and outcome using a decision boundary. At the center of the decision boundary lies the separating hyperplane which can be multi-dimensional (depending on the existing relationship between input and target) and separates the data points of both classes. In this research, an SVM classifier with a radial basis function (RBF) kernel will serve as the baseline model.

Shi et al (2011) used SVM to classify documents into polar sentimental categories and found that a TF-IDF is more effective than a frequency word vector. Ahmad et al (2018) recommended using a grid search along with the SVM algorithm for better performance. In these studies, the strength of the SVM algorithm in both classification and regression tasks was showcased. However, in this research, there will be no hyperparameter optimization for the SVM model.

Deep Learning-based Sentiment Analysis

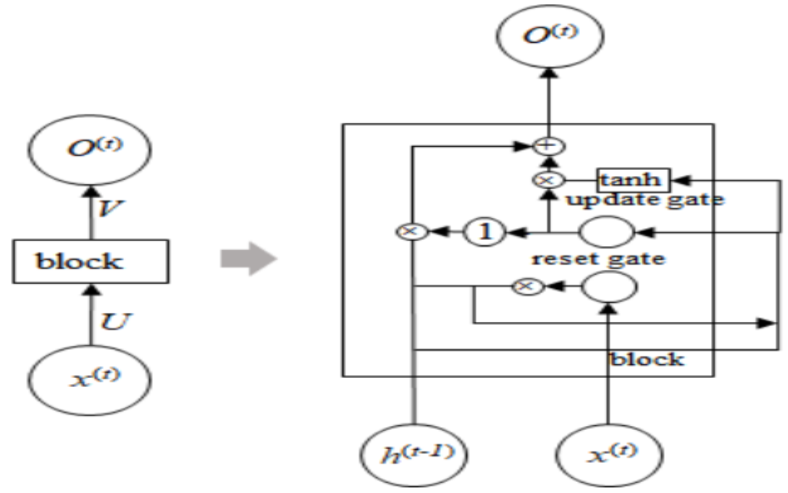
Deep learning is the subset of machine learning that concerns the application of artificial neural networks (ANN) in learning relationships between the input and outcome using neurons (data processing units) organized in layers. A layer takes one or more tensors as input and returns one or more tensors as output. ANNs can learn to perform complex tasks such as computer vision, speech recognition, and time-series data analysis, by adjusting each weight associated with the connections between neurons. Types of ANN include the feedforward neural networks, convolutional neural networks (CNN), and recurrent neural networks (RNN) which can be combined or used separately depending on the task at hand (Chollet, 2021; Zhang et al, 2018).



A feedforward neural network

A neural network with two-dimensional convolutional layers (CNN) is most suitable for processing image data, while an RNN is best suited for processing three-dimensional sequence data such as time-series, text, and audio data (Chollet, 2021). The RNN contains a memory component that enables it to process data with elements that occur in sequence. This mechanism allows the RNN to repeat an operation for each element of an input sequence accounting for previous outputs. However, the RNN has some limitations such as retaining information from only a few timesteps back caused by the diminishing gradient or exploding gradient problem (Bengio et al, 1994). Addressing this challenge, several variants of the RNN have been developed namely, Bidirectional RNN (two-way or Bi-RNN), long short-term memory units (LSTM), and gated recurrent unit (GRU).

The LSTM and GRU are somewhat similar in function, as they both retain information over longer timesteps through a gated mechanism. These gates are used to regulate information flow within each unit. LSTM has four gates and two states (hidden state and cell state), while GRU has only two gates (reset gate and update gate) and a hidden state. This makes the GRU a much simpler in computational requirements than the LSTM, and this is responsible for its growing popularity since it was introduced in 2014 (Shewalkar, 2019; Cho et al, 2014). The GRU is shown below (researchgate.com).



A gated recurrent unit (GRU)

Wu et al (2021) used a combination of Bi-LSTM and GRU to tackle semantic dependence and vanishing gradient in sentiment analysis of hotel reviews and found an improvement in model accuracy. Agarap (2018) performed an unusual modification to the normal GRU architecture by replacing the output layer with a linear SVM. Shewalkar (2019) after evaluating the performance of RNN, LSTM, and GRU on speech recognition tasks concluded that tuning the GRU hyperparameters was faster than LSTM and RNN. This finding supports our decision to opt for a GRU over an LSTM due to its better hyperparameter optimization capacity. Compagnon et al (2019) used a shallow GRU model (two hidden layers) for fall detection binary classification on two sequential datasets. Although the result was inconclusive due to insufficient data for training, this is the only research that was found to involve a shallow GRU model. This research was inspired by the previously mentioned works in terms of choice of the algorithm used, model architecture, hyperparameter optimization, and model comparison.

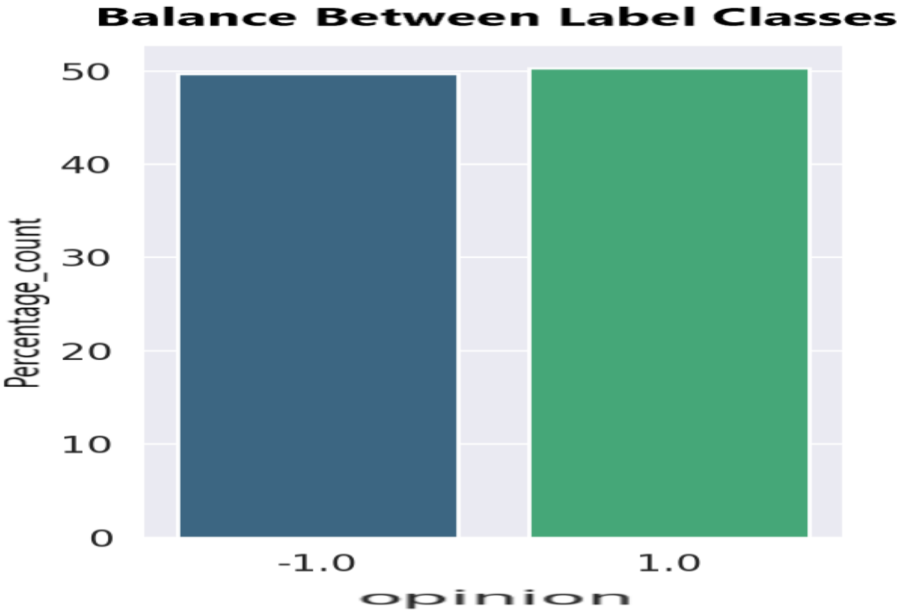
Therefore, the aim of this research is to propose a relatively simple computationally cheap approach to sentiment analysis as an alternative to the more resource-intensive, computationally expensive models and architectures out there. It would be relevant to also compare the performance of an SVM model to that of a shallow GRU model on the sentiment analysis of titled hotel reviews. To achieve the aim of proposing a shallow GRU architecture, the initial state of hyperparameters will be adopted from the work of Agarap (2018), then there will be the manual tuning of hyperparameters based on research recommendations. To achieve the aim of making a direct comparison between the SVM and GRU algorithms in due time, the hyperparameters of the SVM algorithm will not be tuned.

III. METHODOLOGY

Data Set

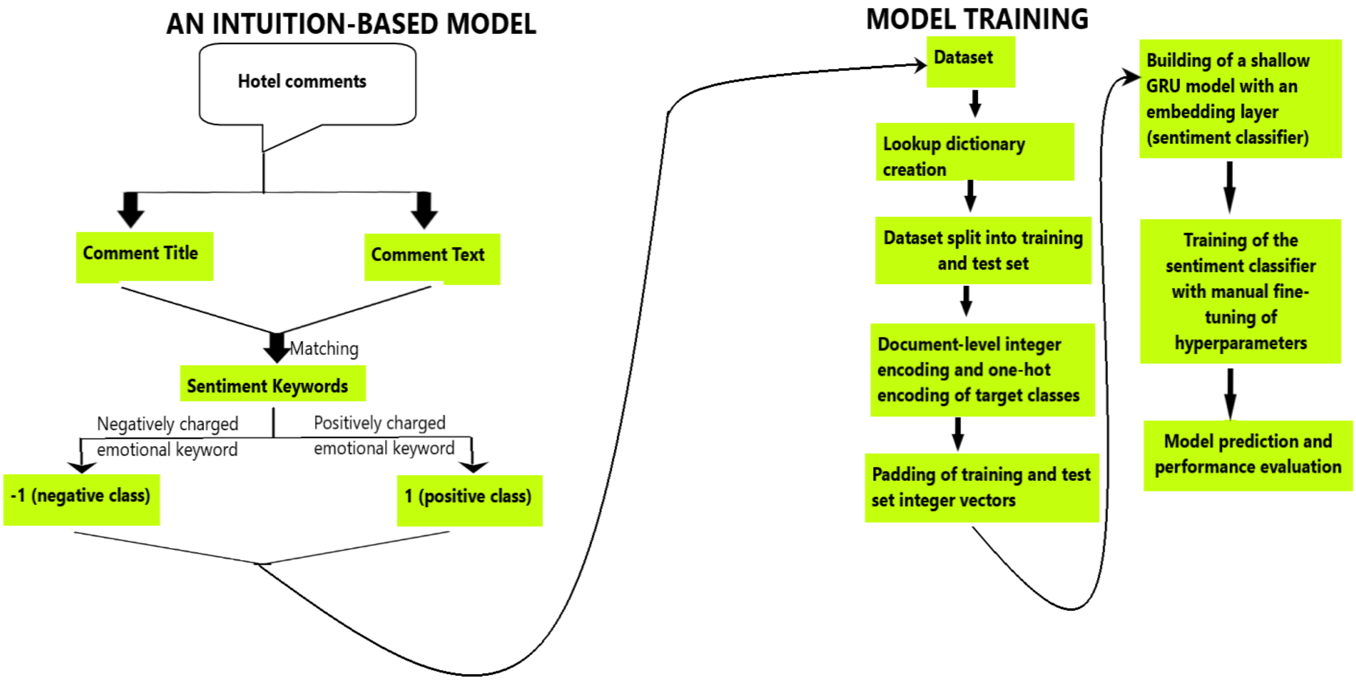
Hotel review data used in this research was collected from Opin Rank datasets (Ganesan and Zhai, 2012) comprising titled comments from guests about hotels in many cities worldwide. Only customer reviews of London hotels were used in this study. Due to data volume, it was impractical for the human annotator to carefully read through each document and correctly assign a sentiment. Therefore, the annotator assigned sentiment labels to comments by the emotional meaning of review keywords. -1 was assigned to the negative reviews and 1 to the positive reviews. Keywords were determined by selecting only matching words between each document's topic and its body of text. This method helped to greatly narrow down the search space for the annotator. Then careful consideration was given to the occurrence of word negation and sarcasm. 8,532 instances were carefully selected from the labeled data with balanced representation from both sentiment classes as shown in the following table and chart:

Sentiment Class	Count	Percentage Count
Positive (1)	4,293	50.32%
Negative (-1)	4,239	49.68%



Intuition-Based Model

The diagram below illustrates the flowchart of this research.


AN INTUITION-BASED MODEL
Hotel comments
Comment Title
Comment Text
Matching
Sentiment Keywords
Negatively charged emotional keyword
Positively charged emotional keyword
-1 (negative class)
1 (positive class)
MODEL TRAINING
Dataset
Lookup dictionary creation
Dataset split into training and test set
Document-level integer encoding and one-hot encoding of target classes
Padding of training and test set integer vectors
Building of a shallow GRU model with an embedding layer (sentiment classifier)
Training of the sentiment classifier with manual fine-tuning of hyperparameters
Model prediction and performance evaluation

Since sentiment labels were assigned to the unlabeled data at the discretion of a human agent, this research method can be categorized under the intuition-based approach. After the labeling process, the dataset was used to create the word-to-index and index-to-word lookup dictionaries, both containing (in reverse order) the unique words in the corpus paired with their index position in the dictionary. Then the data was split into training, validation, and test samples at a ratio of 70: 18: 12 respectively. Hyperparameters were fine-tuned using performance on the validation set, while the test set was used to compute performance metrics for the final report. Integer vectors were then created from input and target variables using the word-to-index dictionary, and one-hot encoder respectively. Then each integer vector was padded to ensure consistency in the input size being fed to the predictive model. After building and training the classification algorithms, only the hyperparameters of the GRU classifier were fine-tuned.

Most of the preprocessing was performed using the Scikit-Learn library (Pedregosa et al, 2011), while the GRU classifier was implemented through the deep learning library called Keras (Chollet, 2021). The Adaptive moment (ADAM) optimizer runs on a stochastic gradient descent method and is quite resilient to learning rate and other training parameters which makes it less reliant on hyperparameter optimization (Shewalkar, 2019). Hence, this was chosen as the optimizer for the shallow GRU model. Only the learning rate of Adam was tuned in this study. The binary cross-entropy loss function computes the distance between predicted and actual values during training.

For the SVM classifier (RBF kernel), the input variable consisted of uni-gram, 2-gram, and 3-gram vectors derived from the operation of a count vectorizer. Both the SVM model and the count vectorizer algorithm were implemented also through the Scikit-Learn machine learning library.

IV. EXPERIMENT (SET-UP)

In this experiment, SVM (RBF kernel) with default hyperparameters was used as the baseline, and a shallow GRU model with a single hidden layer was manually tuned based on performance on the validation set.

The architecture of the baseline classifier:

Hyperparameters	**SVM**
Kernel	RBF
C	1 (default)

Hyperparameters used in the Keras GRU classifier:

Hyperparameters	**GRU-unoptimized**	**GRU-optimized**
Embedding layer dimension	300	300
Training batch size	128	64
Test batch size	32 (default)	32 (default)
Output activation	Softmax	Softmax
Units	128	128
Dropout Rate	0.5	0.85
Epochs	10	19
Learning Rate	0.001 (default)	2.8e-4

Optimizer	Adam
Training Loss	Binary cross-entropy
Training Metrics	Accuracy
Test Metrics	Accuracy, Precision, Recall, F1-score

Performance evaluation

Each model's performance was evaluated based on accuracy, precision, recall, and f1-score. Accuracy is a valid metric in this study because of a balance between corpus labels, while the precision, recall, and f1-score were used to assess more specific aspects of the model's performance.

They are calculated as:

		ACTUAL	
		Positive review	Negative review
PREDICTED	Positive review	True Positive (TP)	False Positive (FP)
	Negative review	False Negative (FN)	True Negative (TN)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision (positive review) = TP / (TP + FP); ***Precision (negative review)*** = TN / (TN + FN)

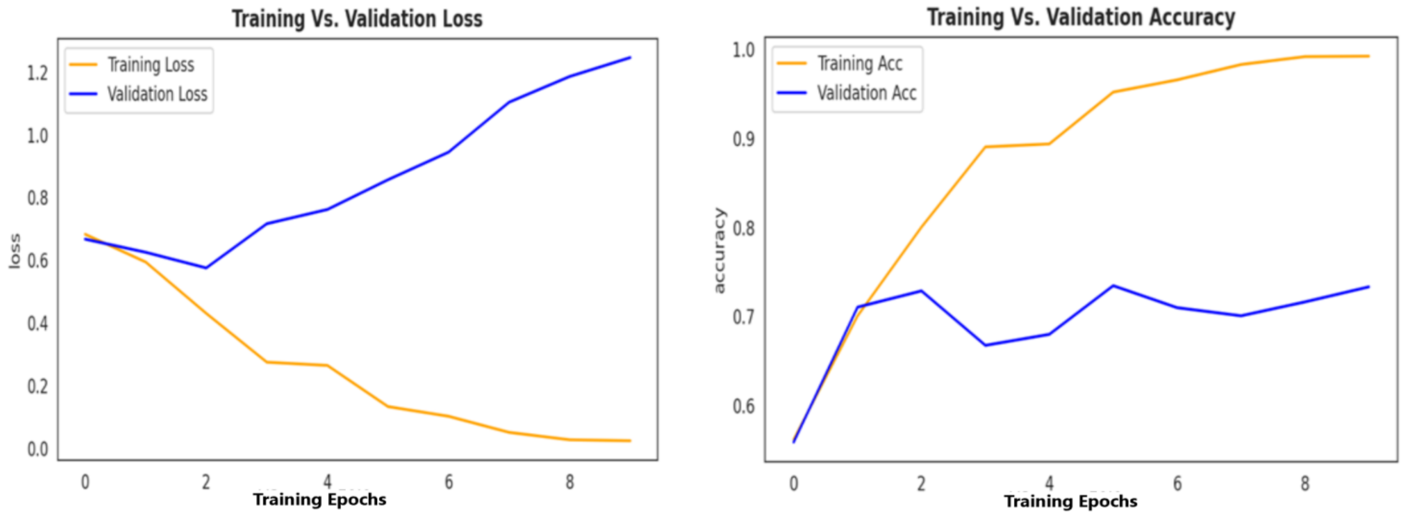
Recall (positive review) = TP / (TP + FN); ***Recall (negative review)*** = TN / (TN + FP)

F1-score = (2 * Precision * Recall) / (Precision + Recall)

V. RESULTS AND INTERPRETATION

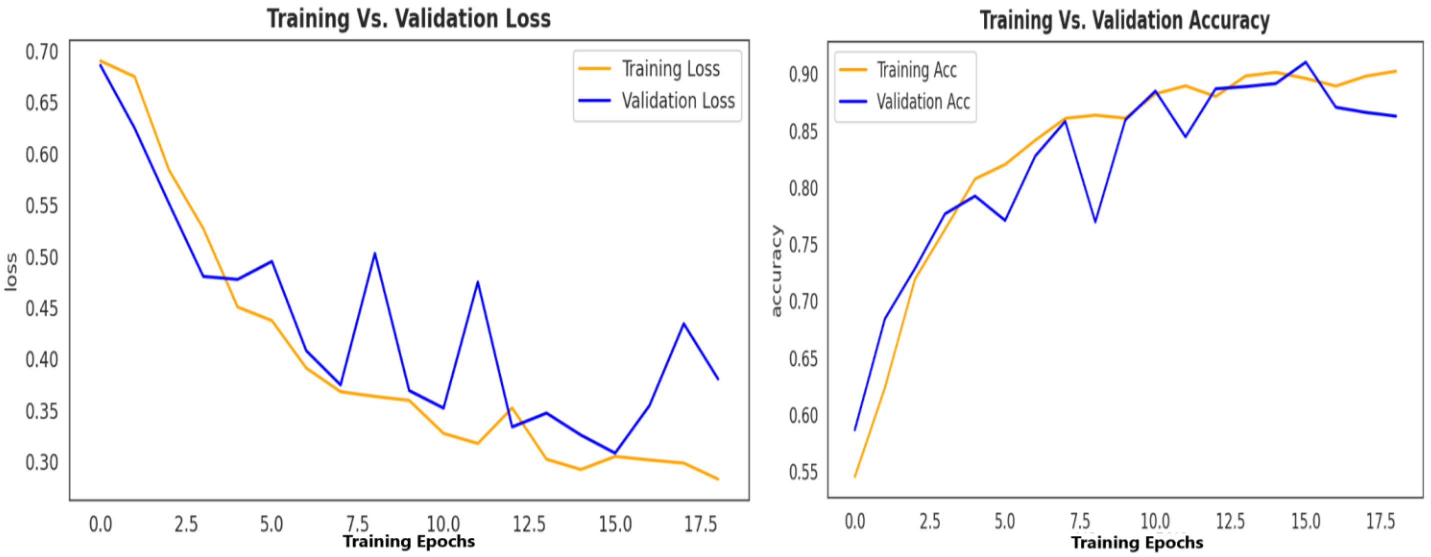
This experiment was conducted on the GPU infrastructure of Google Colab virtual platform (https://colab.research.google.com/).

Training and Validation of Unoptimized GRU



The above plot shows a steady rise in validation loss while the reverse was the case for the training loss clearly indicating overfitting began after epoch 3.

Training and Validation of Optimized GRU

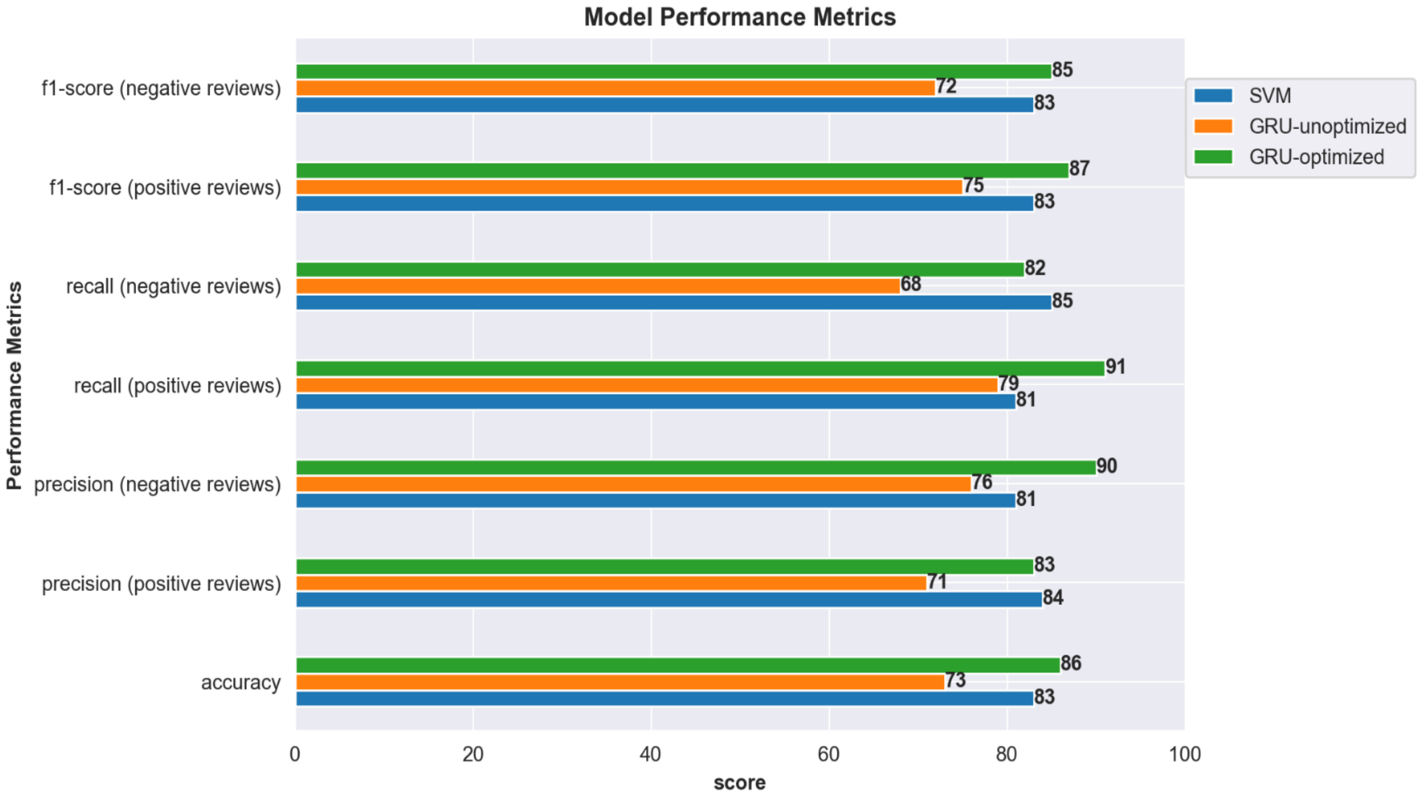


The above plot shows a steady drop in training loss while validation loss was minimum at epoch 16. The validation accuracy being slightly higher than the training accuracy (epoch 16) indicates no overfitting.

Summary of experiment results on SVM and GRU classifiers are given in the following table

Parameter	**SVM**	**GRU-unoptimized**	**GRU-Optimized**
No of training samples	6,399	5,972	5,972
No of validation samples	N/A	1,536	1,536
No of test samples	2,133	1,024	1,024
Epochs	N/A	10	19
Testing Accuracy	83%	73%	86%
Training Duration (secs)	N/A	463	751
Precision (positive reviews)	84%	71%	83%
Precision (negative reviews)	81%	76%	90%
Recall (positive reviews)	81%	79%	91%
Recall (negative reviews)	85%	68%	82%
F1-score (positive reviews)	83%	75%	87%
F1-score (negative reviews)	83%	72%	85%

The result presented in this paper proves that a well-tuned shallow GRU has the capacity to compete with or outperform some state-of-the-art algorithms like the SVM. However, as shown by the relatively weaker accuracy of the unoptimized GRU, getting the best performance from the GRU model is heavily reliant on selecting the right values for the hyperparameters which often requires research and expertise. The number of units, learning rate, and batch size were highly influential in model performance.



For each metric, the optimized GRU outperformed the SVM, except for recall (negative reviews) and precision (positive reviews) where SVM performed slightly better. It is also important to note the difference between the performance of the GRU model with and without optimization. Therefore, this underlines the importance of fine-tuning hyperparameters in deep learning-based sentiment analysis.

VI. CONCLUSION

This paper proposes a relatively simple, computationally inexpensive but competitive approach to sentiment analysis of hotel reviews. A lot of sophisticated state-of-the-art approaches have been proposed which are resource-intensive and computationally expensive to implement at a broader scale. In this paper, a shallow GRU with one hidden layer after manually fine-tuning its hyperparameters improved tremendously and outperformed the SVM classifier on five out of 7 performance metrics. This also highlights that fine-tuning hyperparameters is greatly beneficial to the GRU model in the sentiment analysis domain.

In the future, we would like to further improve the performance of the shallow GRU by using a grid search algorithm to automate the optimization process or adding a second hidden layer for more learning capacity. It would also be interesting to see how the combination of the GRU-SVM (RBF kernel) algorithm would perform in sentiment analysis of unlabeled titled hotel reviews due to the promising potential each possesses separately.

REFERENCES

Agarap, A.F.M. (2018) A neural network architecture combining gated recurrent unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data. *Proceedings of the 2018 10th international conference on machine learning and computing*, 26-30

Ahmad, M., Aftab, S., Bashir, M.S., Hameed, N., Ali, I. and Nawaz, Z. (2018) SVM optimization for sentiment analysis. *Int. J. Adv. Comput. Sci. Appl*, 9(4), 393-398.

Bengio, Y., Simard, P. and Frasconi, P. (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2),157-166.

Cho, K., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine transla tion. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Chollet, F. (2021) *Deep learning with Python*. Simon and Schuster.

Compagnon, P., Lefebvre, G., Duffner, S. and Garcia, C. (2019) Personalized posture and fall classification with shallow gated recurrent units. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*114-119

Ganesan, K. and Zhai, C., 2012. Opinion-based entity ranking. *Information retrieval*, 15(2), 116-150.

Hsieh, H.Y., Klyuev, V., Zhao, Q. and Wu, S.H. (2014). SVR-based outlier detection and its application to hotel ranking. *IEEE 6th International Conference on Awareness Science and Technology* 1-6.

Lung Sound Recognition Algorithm Based on VGGish-BiGRU - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-structure-of-GRU_fig3_336093319 [accessed 9-May-2022]

Maity, A., Ghosh, S., Karfa, S., Mukhopadhyay, M., Pal, S. and Pramanik, P.K.D. (2020) Sentiment analysis from travellers' reviews using enhanced conjunction rule based approach for feature-specific evaluation of hotels. *Journal of Statistics and Management Systems*, 23(6) 983-997.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C. and Smeaton, A.F. (2009) Topic-dependent sentiment analysis of financial blogs. *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* 9-16.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825-2830.

Piryani, R., Gupta, V., Singh, V.K. and Ghose, U. (2017) A linguistic rule-based approach for aspect-level sentiment analysis of movie reviews. *Advances in computer and computational sciences* 201-209.

Ray, B., Garain, A. and Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing* 98, 106935.

Shewalkar, A. (2019) Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, 9(4), 235-245.

Shi, H.X. and Li, X.J., (2011). A sentiment analysis model for hotel reviews based on supervised learning. *International Conference on Machine Learning and Cybernetics*. 3, 950-954.

Wu, T. and Zheng, G. (2021) Research on Hotel Comment Emotion Analysis Based on BiLSTM and GRU. *IEEE 4th International Conference on Robotics, Control and Automation Engineering (RCAE)* 143-146.

Zhang, L., Wang, S. and Liu, B. (2018) Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), p.e1253.

Zvarevashe, K. and Olugbara, O.O. (2018). A framework for sentiment analysis with opinion mining of hotel reviews. *IEEE 2018 Conference on information communications technology and society,* 1-4.