

2300244_XY_HAN_AI_7026_assi gnment_1

by Xiyu Han

Submission date: 19-Apr-2024 02:24AM (UTC+0100)

Submission ID: 229923076

File name: 50022_Xiyu_Han_2300244_XY_HAN_AI_7026_assignment_1_571450_807354664.docx (502.96K)

Word count: 3386

Character count: 16363

Sentiment Analysis for IMDB Movie Reviews

CONTENTS

I	Introduction	2
II	Problem Statement	2
III	Methodology	3
III-A	Data Collection and Division	3
III-B	Data Preprocessing and Tokenization	3
III-C	CNN	3
III-D	LSTM	4
III-E	GRU	5
III-F	BERT	5
III-G	Ensemble Learning	5
IV	Experimental Results and Discussion	6
IV-A	Datasets and Evaluation Metrics	6
IV-B	Results on CNN	7
IV-C	Results on LSTM	7
IV-D	Results on GRU	7
IV-E	Results on BERT	8
IV-F	Results on Ensemble Learning	8
IV-G	Discussion	9
V	Conclusion	9
VI	Appendix	9
	References	11

Abstract

This study delves into the different deep learning architectures that were evaluated on the IMDB benchmark dataset for movie review sentiment analysis. We employ four distinct models: Gated Recurrent Units (GRUs), Long Short-Term Memory networks (LSTMs), Convolutional Neural Networks (CNNs), and Bidirectional Encoder Representations from Transformers (BERT), each contributing unique feature extraction capabilities. The meta-features generated by these models are utilized to train a K-Nearest Neighbors Model (KNN) and Support Vector Machine (SVM) classifier. The ensemble approach aims to combine the temporal learning of GRUs and LSTMs, the spatial feature detection of CNNs, and the contextual understanding from BERT to create a robust sentiment analysis tool. Our results show that the ensemble model achieves notable accuracy with balanced precision and recall.

I. INTRODUCTION

Human creativity, social connections, and decision-making are all greatly influenced by emotion. An essential element of the human-computer interface is sentiment analysis [1]. One important technique for evaluating a film's performance is to analyze movie reviews. While star ratings or numerical rankings indicate a film's degree of success or failure, a variety of textual assessments provide more qualitative information. A closer examination of the review reveals whether the movie lives up to the critics' high standards because it highlights both the film's positives and negatives.

In this paper, we use four base models (CNN, LSTM, GRU, and BERT) to analyze the IMDB dataset and introduce an ensemble method that combines the predictive power of these diverse models to address their limitations and improve overall predictive performance. By creating meta-features from each base model's predictions and employing a KNN and SVM classifier to interpret these features, we aim to construct a model that not only achieves high performance but also maintains generalizability. We present a comprehensive analysis of the ensemble's performance and discuss its implications for future NLP applications.

II. RELATED WORK

Extensive research has been conducted on sentiment analysis of movie reviews, spanning a range of methodologies. Initially, traditional machine learning techniques such as Naive Bayes (NB), K-Nearest Neighbors (KNN), and Random Forests (RF) were explored, with Naive Bayes often emerging as the most effective classifier. This is supplemented by a lexicon-based sentiment analysis utilizing R software, as detailed in [2].

Researchers such as Baid *et al.* [3] have implemented and compared these traditional methods, finding Naive Bayes to be particularly effective. Advanced ensemble methods have also been explored for combining various machine learning approaches, as seen in [4]. Additionally, enhancements to KNN through the use of information gain were investigated in [5]. While Thongtan *et al.* [6] proposed innovative techniques involving document embeddings and feature combinations.

The domain has also embraced deep learning. Studies have analyzed the performance of RNN and CNN in semantic analysis, with findings discussed in [7]. RNNs, despite their potential, struggle with gradient issues in longer texts. To address this, new architectures like the Coupled Oscillatory RNN (CoRNN) have been proposed in [8]. Furthermore, the exploration of LSTM networks and their configurations has revealed significant insights in [9]. BiLSTM networks, combining sophisticated architectures and mixed objective functions, have shown promising results, as documented in [10].

Moreover, the application of BERT for deep sentiment analysis, especially on the SST-2 movie reviews benchmark, illustrates its effectiveness, with additional applications in areas such as stock market sentiment analysis by Sousa *et al.* [11].

Given this extensive background, our study aims to leverage four advanced deep learning models—CNN, LSTM, GRU, and BERT—to further enhance sentiment analysis of movie reviews, aiming to combine the strengths of each model for superior performance.

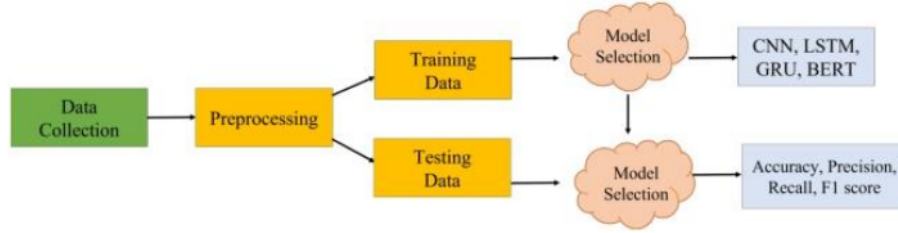


Fig. 1. Overall work procedure.

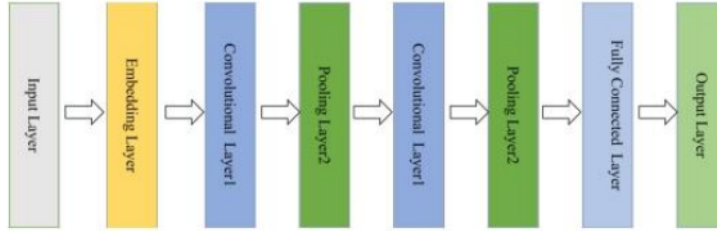


Fig. 2. CNN Network.

III. METHODOLOGY

Fig. 1 illustrates the comprehensive workflow for our project.

A. Data Collection and Division

In our research, we have allocated 60% of the dataset to the training set. The remaining 40% is split evenly, with 20% dedicated to the validation set and 20% to the test set. This allocation allows for a substantial amount of data for the model to learn from while providing equal-sized validation and test sets to thoroughly assess and validate the model's performance.

B. Data Preprocessing and Tokenization

The dataset undergoes preprocessing to transform it into an organized format, which is essential for further analysis as outlined by [12]. This study employs various methods to execute the sentiment classification task. Initially, the dataset is processed as a data frame. During tokenization, each movie review is dissected into a sequence of words. Subsequently, to standardize the text data, each word is converted to lowercase. This normalization is critical as it ensures consistency across the dataset, making it easier for the subsequent analytical models to process the text effectively.

C. CNN

As illustrated in Fig. 2, the architecture of CNN typically includes three essential layers. The initial layer is the convolution layer, where input data is processed using multiple filters, also referred to as kernels. These filters are crucial for extracting specific features from the input matrix. Following the convolution layer is the pooling layer. This layer simplifies the output from the convolution layer by merging the

outputs of clusters of neurons into a single neuron in the subsequent layer. The sequence culminates with a fully connected layer that follows the convolution and pooling layers. This layer performs the critical function of integrating the low-level features detected previously into high-level reasoning, which is necessary for effective classification. The network's output layer, which is used for binary classification tasks, consists of a single neuron equipped with a sigmoid activation function to output the probability of the input belonging to one of two classes.

D. LSTM

As demonstrated in Fig. 3, LSTM is an enhanced version of the RNN designed to handle long-term dependencies. An LSTM cell includes three primary gates that regulate its operations. featuring a sigmoid layer, selects the information to be discarded from the cell. Following this, the Input Gate determines the new information to be added to the memory cell. This gate consists of two layers: a sigmoid layer that selects the values to update, and a tanh layer that generates new candidate values. The third gate, the Output Gate, also uses a sigmoid layer to decide which components of the cell state are passed on to the output. Together, these gates function collaboratively, enabling the LSTM to efficiently manage long-term data dependencies. The formulate are as follows:

$$F_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$I_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3)$$

$$h_t = O_t * \tanh(C_t) \quad (4)$$

where, F_t , I_t , and O_t denote the layers corresponding to the forget, input, and output gates, respectively. The biases for these gates are symbolized as b_f , b_i , b_o . The variables x_t and h_t represent the input to the current cell and its output, while h_{t-1} corresponds to the output from the preceding cell associated with input x_{t-1} . The sigmoid function is signified by σ , dictating the gate activations, and $\tanh$ indicates the hyperbolic tangent function, applied within the cell's activation.

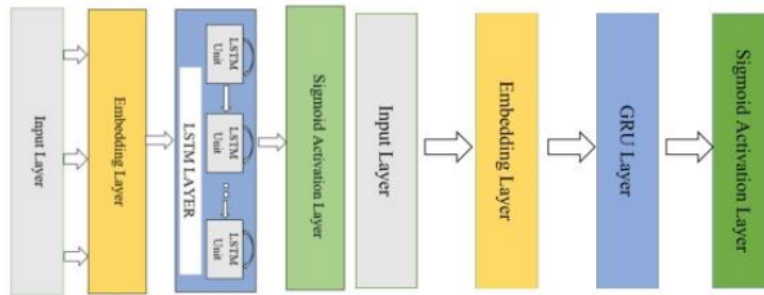




Fig. 5. Simplified diagram of BERT.

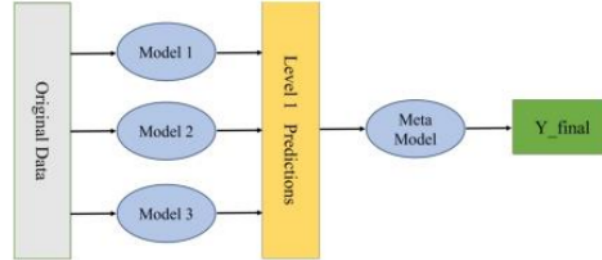


Fig. 6. Ensemble learning technique of stacking.

E. GRU

To address the limitations of RNNs, LSTM and GRU present improved alternatives. GRU simplifies the structure by incorporating two gates: a reset gate and an update gate. The update gate merges the functionalities of the forget and input gates found in LSTM, and it operates without a separate memory cell. Although the GRU structure is less complex than that of LSTM, both architectures demonstrate similar efficacy in training, as depicted in Fig. 4.

F. BERT

As depicted in Fig. 5, BERT employs WordPiece embeddings and requires its specific tokenizer to preprocess text appropriately. During data preprocessing, text is tokenized and converted into IDs from BERT's vocabulary. This tokenized text is then fed directly into BERT, which generates sentence vector representations within its embedding space. These representations are subsequently used for classification tasks.

G. Ensemble Learning

Stacking is an ensemble learning technique that combines the predictions of multiple different base models to improve the overall predictive accuracy of the model, as shown in Fig. 6. This method involves two main steps:

(1) Training Base Learners: Initially, several distinct learning models are trained on the same dataset. The key is that they independently learn from the data.

(2) Training a Meta-Learner: Subsequently, a new model (known as the meta-learner or secondary learner) is trained using the predictions of the base learners as input. The purpose of the meta-learner is to learn how to best combine the predictions of the base models to achieve better final predictions than any individual base model.

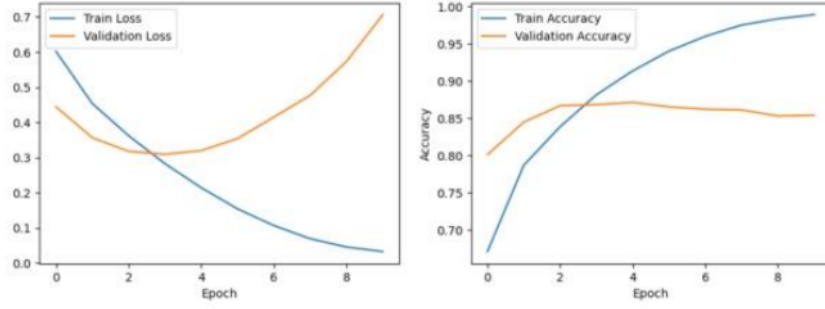


Fig. 7. Results on CNN.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Datasets and Evaluation Metrics

Datasets: This dataset, sourced from the IMDB website—a repository of professional movie reviews—aims to classify texts into positive or negative sentiments based on features extracted from the reviews. This balanced dataset consists of positive and negative sentiment reviews, all in English. It includes various symbols, such as hashtags and exclamation marks, which are to be removed during preprocessing to ensure cleaner data for analysis.

Evaluation Metrics: We use metrics like F1 score, recall, precision, and classification accuracy to evaluate the effectiveness of the suggested model. The mathematical formulations are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$F_1 = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

where TP and TN represent the correct predictions of positive and negative classes, respectively, while FP and FN denote the incorrectly predicted positive and negative classes.

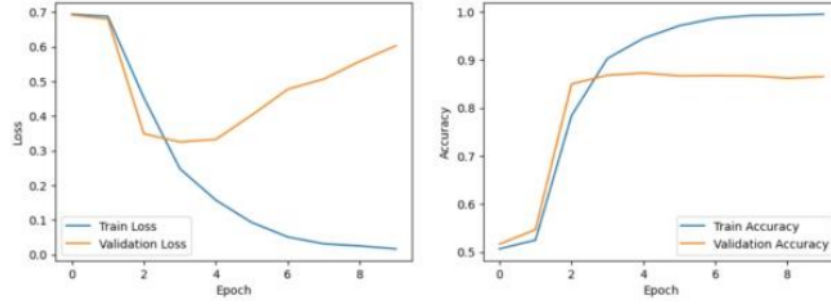


Fig. 8. Results on LSTM.

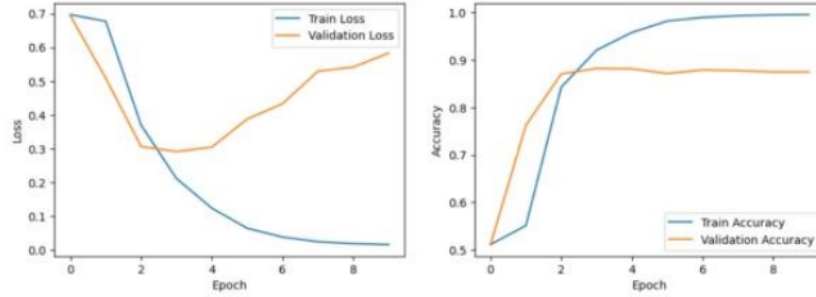


Fig. 9. Results on GRU.

B. Results on CNN

As shown in Fig. 7, The training output shows a consistent decrease in training loss and an increase in training accuracy over 10 epochs, suggesting good learning progress. However, the validation accuracy plateaus around 85%, indicating the model may be overfitting. The accompanying graphs support this, with training performance outpacing validation, especially after the initial epochs. Measures to combat overfitting might be needed to improve model generalization.

C. Results on LSTM

As shown in Fig. 8, the training loss decreased from about 0.694 to 0.017, and the training accuracy increased from 50.69% to 99.53%. Validation accuracy started at 51.67% and increased to 86.51%, showing good progress but a clear gap with the training accuracy, hinting at some overfitting. Overall, the model demonstrates strong learning capability, but the discrepancy between training and validation metrics indicates overfitting. Regularization or more data might be needed to improve generalization.

D. Results on GRU

As shown in Fig. 9, the training process has gone through 10 epochs, with both training loss and validation loss decreasing over time, which is indicative of learning. The training loss started at approximately 0.697 in epoch 1 and decreased to about 0.015 by epoch 10. Correspondingly, the training accuracy improved from around 51% to nearly 99.6%. This pattern of decreasing loss and increasing accuracy is typical for a converging training process.

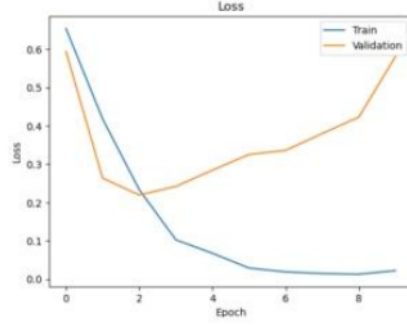


Fig. 10. Results of BERT.

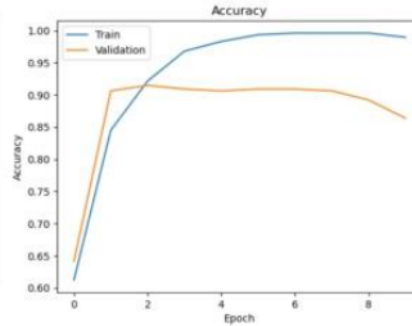
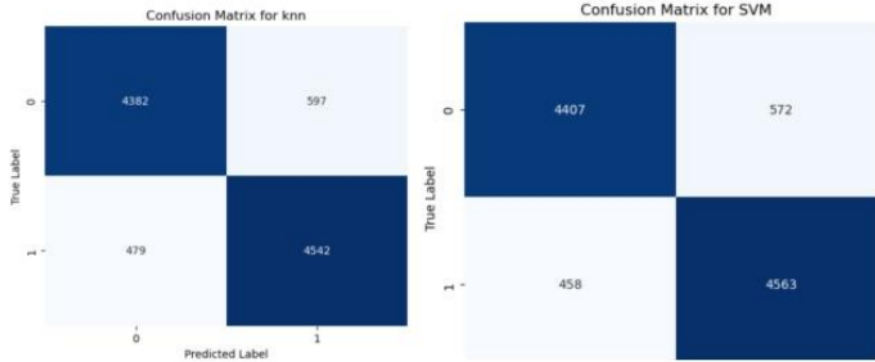


Fig. 11. Results of BERT.

E. Results on BERT

As shown in Fig. 11, the training performance improves substantially across 10 epochs, with training accuracy reaching nearly 99%. However, validation accuracy peaks at around 91% in the third epoch and then declines to approximately 86% by the tenth epoch, while validation loss increases, signaling overfitting. This suggests the need for early stopping or regularization to improve the generalization to unseen data.



F. Results on Ensemble Learning

(1) KNN Performance on Integrated Meta-Features: The dataset used for training in this work was likely made up of meta-features from several different models. When the number of neighbors is set to 5, the KNN classifier obtains an accuracy of 88.51% for validation and around 99.9% for training. The model produces an accuracy of 89.24%, a recall of 90.46%, and an F1 score of 89.41% when tested on a test set. And a confusion matrix in Fig. 12.

(2) SVM Performance on Integrated Meta-Features: The SVM model achieves an accuracy of 89.70%, a recall of 90.88%, an F1 score of 89.86%, and a precision of 88.86%. The confusion matrix in Fig. 13 shows the high accuracy and F1 score imply that the model is robust for this particular sentiment analysis task, especially considering it uses an ensemble of meta-features from various models.

G. Discussion

(1) The individual models showed varying degrees of effectiveness, with BERT presenting a strong performance, though it displayed signs of overfitting. However, when meta-features were combined and fed into an SVM classifier, the ensemble model achieved high accuracy and balanced precision and recall.

This suggests that the ensemble model effectively leverages the individual strengths of each base model while mitigating their weaknesses.

(2) Nevertheless, the presence of false positives and negatives indicates that there is a potential mismatch between the features learned by the model. This might be addressed by exploring more complex ensemble methods, experimenting with different sets of meta-features, or tuning the hyperparameters of the base models and the meta-classifier. Additionally, since ensemble models can become complex and less interpretable, there is a trade-off between performance and explainability that needs to be considered in practical applications.

(3) Overall, the ensemble method appears to be a promising strategy for sentiment analysis, capitalizing on the complementary predictive capabilities of diverse models. Future work could focus on refining the ensemble strategy and investigating how it performs across different domains or with datasets that present a wider range of sentiments and linguistic expressions.

V. CONCLUSION

This research investigated various deep learning and ensemble learning models utilizing the IMDB benchmark dataset. Future studies aim to employ a wider array of algorithms and word-embedding techniques to take advantage of continual improvements in deep learning models and embedding methods, especially in multilingual contexts.

VI. APPENDIX

The other results of CNN, LSTM, GRU, and BERT models are as follows:

TABLE I
PERFORMANCE METRICS FOR VARIOUS MODELS

Model	Accuracy	Recall	Precision	F1 Score
CNN	86.59	86.58	86.70	86.64
LSTM	86.97	88.21	86.17	87.18
GRU	87.79	87.39	88.18	87.79
BERT	85.20	94.88	-	86.56

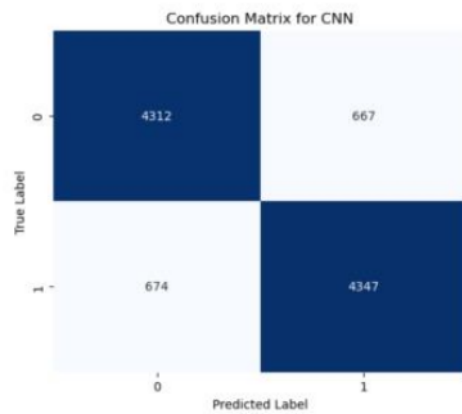


Fig. 14. Confusion matrix on CNN.

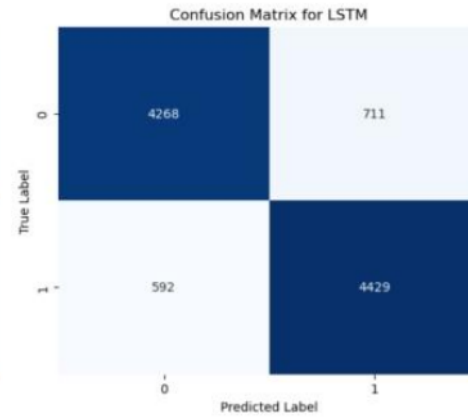


Fig. 15. Confusion matrix on LSTM.

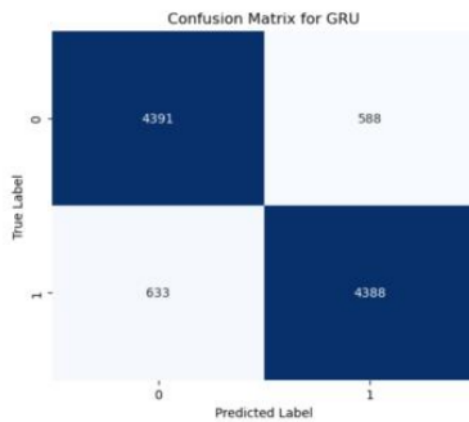


Fig. 16. Confusion matrix on GRU.

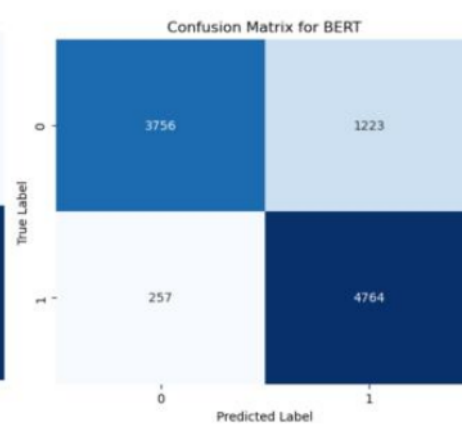


Fig. 17. Confusion matrix on BERT.

Video of code will be available at <https://youtu.be/msUzGFhjLVc>.

REFERENCES

- [1] C. Lin and M. S. Obaidat, "Behavioral biometrics based on human-computer interaction devices," *Biometric-based physical and cybersecurity systems*, pp. 189–209, 2019.
- [2] R. Jaichandran, C. Bagath Basha, K. Shunmuganathan, S. Rajaprakash, and S. Kanagasuba Raja, "Sentiment analysis of movies on social media using r studio," *Int. J. Eng. Adv. Technol.*, vol. 8, pp. 2171–2175, 2019.
- [3] P. Baid, A. Gupta, and N. Chaplot, "Sentiment analysis of movie reviews using machine learning techniques," *International Journal of Computer Applications*, vol. 179, no. 7, pp. 45–49, 2017.
- [4] G. Mesnil, T. Mikolov, M. Ranzato, and Y. Bengio, "Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews," *arXiv preprint arXiv:1412.5335*, 2014.
- [5] N. O. F. Daeli and A. Adiwijaya, "Sentiment analysis on movie reviews using information gain and k-nearest neighbor," *Journal of Data Science and Its Applications*, vol. 3, no. 1, pp. 1–7, 2020.
- [6] T. Thongtan and T. Phientrakul, "Sentiment classification using document embeddings trained with cosine similarity," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 407–414.
- [7] H. Shirani-Mehr, "Applications of deep learning to sentiment analysis of movie reviews," in *Technical report*. Stanford University, 2014.
- [8] T. K. Rusch and S. Mishra, "Coupled oscillatory recurrent neural network (cornn): An accurate and (gradient) stable architecture for learning long time dependencies," *arXiv preprint arXiv:2010.00951*, 2020.
- [9] J. D. Bodapati, N. Veeranjanyulu, and S. N. Shareef, "Sentiment analysis from movie reviews using lstms," *Ingenierie des Systems d'Inf.*, vol. 24, no. 1, pp. 125–129, 2019.
- [10] D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Revisiting lstm networks for semi-supervised text classification via mixed objective function," in *Proceedings of the aaai conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6940–6948.
- [11] M. G. Sousa, K. Sakiyama, L. de Souza Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, "Bert for stock market sentiment analysis," in *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)*. IEEE, 2019, pp. 1597–1601.
- [12] S. K. Dwivedi and B. Rawat, "A review paper on data preprocessing: A critical phase in web usage mining process," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*. IEEE, 2015, pp. 506–510.

FINAL GRADE

GENERAL COMMENTS

69/100

Demonstrated the ability to build and evaluate AI models. However writing skill and the presentation of the work need to be improved.

PAGE 1

PAGE 2



Comment 1

it would be better to state that why these models ?



Comment 2

you do not need a separate title like this for a paper/report. This should be part of the introduction

PAGE 3

PAGE 4



Comment 3

?

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

GRADING FORM: ACCA 7026 GRADING

XIYU HAN

69

ENGAGEMENT WITH LIT. SKILLS

(including current research, referencing, academic conventions and academic honesty)



Introduction is not well written, lack of literature support. Also it lacks recent references.

9

KNOWLEDGE&UNDERSTANDING SKILLS

(A systematic, conceptual understanding of knowledge, and a critical awareness of current problems and/or new insights, much of which is at, or informed by, the forefront of the discipline; a comprehensive understanding of techniques applicable to their own research)



Knowledge has a well defined focus and concepts are well explained in Methodology section.

7

COGNITIVE&INTELLECTUAL SKILLS

(Critical evaluation of current research and methodologies and develop critiques of them and, where appropriate, to propose new hypotheses; make sound judgements in the absence of complete data.)



Methodology is well explained. Provided a detailed explanation of the evaluation metrics. However there are some typo errors in the equations provided. Evaluation of the results is not presented well.

16

PRACTICAL APPLICATION SKILLS

(Originality / creativity in the application of knowledge, tools and techniques and in tackling and solving problems in complex and unpredictable professional situations; practical use of established techniques of research and enquiry to create and interpret knowledge in the discipline.)



An appropriate application of methods such as LSTM, GRU , CNN , BERT and stacking provided. However the results are not well presented/discussed

30

TRANSF. SKILL & PROF.PRACTICE

(Exercise of self-direction, autonomy and personal responsibility; plan and implement tasks at a professional level; independent learning; use appropriate media to communicate effectively and professionally to a variety of audiences; fluency of expression; systematic approach; clarity and effectiveness in presentation and organisation.)



Need improvement in presentation of the abstract, methodology , results and discussion. Headings provided for some sections are not relevant.

7