TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT MODEL ARCHITECTURES

Model	Precision	Recall	F1-Score
BiLSTM-CRF	0.8693	0.7705	0.8060
BiLSTM	0.8347	0.7301	0.7712
ForwardLSTM	0.7643	0.7373	0.7473
BackwardLSTM	0.8076	0.7227	0.7540

recall, and 80% for F1-Score. These findings suggest that the BiLSTM-CRF model outperforms other LSTM models.

B. Analysis Test Result

The results demonstrate that the BiLSTM-CRF model exhibits superior performance in terms of precision, recall, and F1-score compared to other models. The results demonstrate that the BiLSTM CRF model outperforms other models in terms of precision, recall, and F1-score.

However, the model exhibits relatively high precision, with a notable discrepancy in recall values. This is evident in Table 3, where labels such as I-Brand, I-Specifications, and I-Product demonstrate low recall due to the uneven distribution of labels across the utilized dataset. This results in the model rarely, if ever, encountering the data with the aforementioned labels during the training process, thereby preventing it from providing accurate predictions.

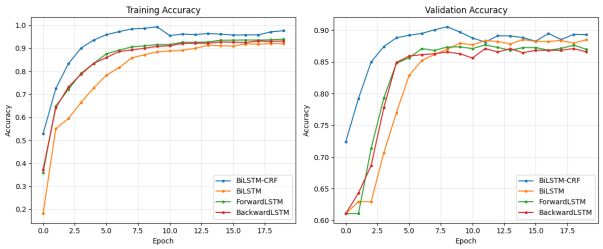

Fig. 5. Training Accuracy Validation Accuracy

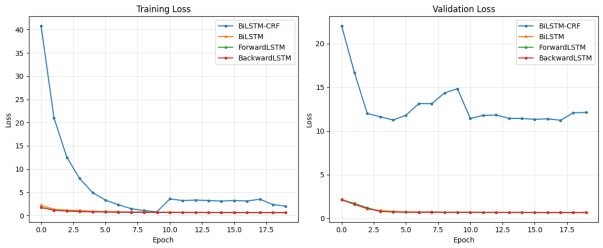

Fig. 6. Training Loss Validation Loss

Based on a comparison of the accuracy and loss values obtained during the training process, as illustrated in Figures 5 and 6, all models demonstrated convergence after epoch 10. The loss validation for BiLSTM-CRF exhibited greater fluctuations in comparison to the other models. This indicates that the model is more sensitive. Conversely, the models without CRF demonstrated more stable losses but lower performance, as illustrated in Figure 6. The BiLSTM-CRF model exhibited signs of overfitting on the data set, as evidenced by the discrepancy between the training accuracy and the validation accuracy, as illustrated in Figure 5.

V. CONCLUSION

In this study, implement information extraction using BiLSTM-CRF. The model demonstrated an average performance of 80% for F1-Score. The implementation of the model utilized hyperparameters with a dimension of 200 in the embedding layer, 16 for the batch size, 20 for the epoch size, 5 for the patience size, and 20 epochs for the training. The resulting performance exceeded expectations. The remaining LSTM models exhibited average F1-scores of 77% for BiLSTM without CRF, 74% for Forward LSTM, and 75% for Backward LSTM. The incorporation of the CRF layer has resulted in a notable enhancement in performance. The precision value (87%) and lower recall value (77%) observed for the BiLSTM-CRF model were attributed to the unequal distribution of labels across the dataset. This resulted in instances where models failed to identify labels with limited data during the training phase, leading to the generation of expected predictions. During the training phase, the model exhibited overfitting tendencies when comparing accuracy and loss values between validation and training data, which subsequently impacted the prediction outcomes.

It is recommended that future research employ a greater number of datasets, with more evenly distributed labels, in order to obtain a higher level of accuracy in the classification process. Additionally, further adjustments can be made to the parameters to improve the model's performance. Modifications can also be made to the layers of the utilized architecture to enhance its operational efficiency. Furthermore, additional observations and refinements during the construction of the model are necessary to prevent the occurrence of overfitting.

REFERENCES

[1] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani, "TwitIE : An Open-Source Information Extraction Pipeline for Microblog Text," no. September, pp. 83–90, 2013.

[2] W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri, and C. Haruechaiyasak, "Discovering consumer insight from twitter via sentiment analysis," J. Univers. Comput. Sci., vol. 18, no. 8, pp. 973–992, 2012.

[3]] M. B. Habib and M. van Keulen, "Information Extraction for Social Media," SWAIE 2014 - 3rd Work. Semant. Inf. Extr. Proc. Work., pp. 9–16, 2014.

[4] S. Jain, M. van Zuylen, H. Hajishirzi, and I. Beltagy, "SCIREX: A challenge dataset for document-level information extraction," Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 7506–7516, 2020.

[5] N. Limsopatham and N. Collier, "Bidirectional LSTM for named entity recognition in ttwitter messages," Proc. 2nd Work. Noisy User-generated Text, pp. 145–152, 2016.

[6] A. Milani, N. Rajdeep, N. Mangal, R. K. Mudgal, and V. Franzoni, "Sentiment extraction and classification for the analysis of users' interest in tweets," Int. J. Web Inf. Syst., vol. 14, no. 1, pp. 29–40, 2018.

[7] E. Oliveira, G. Dias, J. Lima, and J. Pirovani, "Extracting Named-Entities and their Relationships," J. Inf. Data Manag., vol. 13, no. 6, pp. 555–565, 2022.

[8] Sugiarto, "Named Entity Recognition in Social Media Data," vol. 4, no. 1, pp. 1–23, 2016.

[9] M. Xu, X. Zhang, and L. Guo, "Jointly detecting and extracting social events from twitter using gated BiLSTM-CRF," IEEE Access, vol. 7, pp. 148462–148471, 2019.

[10] L. Derczynski, “Complementarity, F-score, and NLP evaluation,” Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016, pp. 261–266, 2016.
[11] Y. Li, C. Lockard, P. Shiralkar, and C. Zhang, “Extracting Shopping Interest-Related Product Types from the Web,” Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 7509–7525, 2023.
[12] A. Brinkmann, R. Shraga, R. C. Der, and C. Bizer, “Product Information Extraction using ChatGPT,” 2023.