Some writing issues:

Correction required: The number of unique argument types per event type ranges from **four to 21**, with an average of nine argument types.

Corrected: The number of unique argument types per event type ranges from **4 to 21**, with an average of nine argument types.

Visual representation of whole work is necessary to give the overview of research task instead of lot of paragraphs.

Table 1. Event detection results for hard news events from the DocEE dataset. The best results are underlined.

Model	Precision	Recall	F1
SVM (Linear)	0.929	0.908	0.915
RoBERTa (Base)	0.955	0.938	0.945
RoBERTa (Large)	0.954	0.941	0.947
DeBERTa-v3 (Base)	0.932	0.910	0.917
DeBERTa-v3 (Large)	0.944	0.935	0.938
ALBERT-v2 (Base)	0.949	0.933	0.939
DistilRoBERTa (Base)	0.948	0.938	0.942

The content should be presented like:

Model	Precision (%)	Recall (%)	F1
SVM (Linear)	93	90.80	91.50%

Technical Questions

Question: Accuracy and inference speed. How its trade of measured?

What are the practical issues to implement existing models?

"To this end, we first evaluated and compared the performance of one shallow and a range of deep CDMEE models trained for ED and AE subtasks on hard news articles from DocEE Tong et al. (2022),a 4 **standard CDMEE dataset for the English language**."

Using **Dataset ASHNEE**, to evaluate the model trained on DocEE dataset is hard to justify because of the structure and nature of the dataset. It may be affected the performance of develop model. Although

approach is appreciable to train and test the model on different datasets but the performance of the model remains compromised.

Question: **How to improve the performance of model as well as to handle the structural difference of the dataset?**

Overall Summary

The research questions are interesting to tackle as covering three major aspects i.e, selection, robustness and scalability.

A standard data for English language is used to measure and analyze the three-above mention aspects.

- Dataset preparation is one of the major contributions of this work as reported by the author. But it is modification of existing ones.

The use of term "Shallow" is misleading to think about neural networks. While using linear Support Vector Machine is traditional statistical model.
Recommendations: Please use the relevant terms to avoid the confusion i.e. Machine Learning classifier, traditional classifier, statistical model etc.
Need to explore the difference between shallow and deep neural network.