Improving Interpretability in Document-Level Polarity Classification by Applying Attention

Kato, Shingo
Kyushu University

Ikeda, Daisuke
Kyushu University

https://hdl.handle.net/2324/7378126

出版情報：IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp.21-26, 2024. Institute of Electrical and Electronics Engineers (IEEE)
バージョン：

Improving Interpretability in Document-Level Polarity Classification by Applying Attention

Shingo Kato
Kyushu University, Fukuoka, Japan

Daisuke Ikeda
Kyushu University, Fukuoka, Japan

***Abstract*—Document-level polarity classification has attracted interest in the real world. While LLMs have made it possible for accurate classification, these complex models have the problem of *interpretability*. Our contribution is to apply inter-sentence attention, which captures the relationship between sentences, to a more practical interpretable model. By utilizing high inter-sentence attention scores, meaning corresponding sentences are related to each other, we attempt to capture the context of sentences and make them more similar to the human judgment process. With two real datasets, we compared our model with prior models in terms of classification performance and interpretability and found that our model is more accurate on both datasets. In addition, to assess interpretability, we examined the overlap between sentences that contribute to the model's predictions and those annotated by humans for the same document. The results show that our model has a larger overlap and is more likely to extract interpretive sentences that humans intuitively consider important. In addition, our result partially captures the polarity of "implicit" sentences that do not contain direct expressions, which could not be captured by prior models, suggesting that our model may lead to a more natural interpretation.**

***Index Terms*—Interpretability, Inter-Sentence Attention**

I. Introduction

Document-level polarity classification, which classifies whether a document is positive or negative, plays an important role in a wide range of areas, including the medical and financial sectors. For example, it has been shown that there is a relationship between the polarity of the text of Management Discussion and Analysis (MD&A) and stock prices [1].

Polarity classification has been widely applied, from simple methods using polar word dictionaries to new methods such as machine learning and deep learning, and the accuracy of classification has been improving year by year. In particular, in recent years, large language models, LLMs for short, such as BERT have dramatically improved accuracy. LLMs' ability to capture context has enabled them to correctly predict the polarity of complex documents, even when the document contains a mixture of negative and positive sentences.

However, these complex models have "interpretability" issues, and the models may not explain the basis for the forecast results. In general, there is a trade-off between the performance of a model and its interpretability [2]. Interpretability is as important as model performance, especially in fields such as medicine and finance, where there is a great deal of responsibility for the predictive results by models.

In text classification, such as polar classification, many interpretation methods extract and visualize the features that contributed to the model prediction. Visualizing the important parts of the document allows the reader to know the main points and improves the convincing of the model's predictions. Features are extracted at various granularities, including words, phrases, and sentences. However, Mosca et al. pointed out that interpretation at the word or phrase level is not intuitive because the meaning can change dramatically with the surrounding context [3]. Yan et al. also pointed out that because LLM analyzes the meaning of features hierarchically, interpretation by a single feature may be insufficient to explain the model's predictions [4]. In order to correctly interpret the model's predictions, it is necessary to hierarchically capture the shifting meanings of the text depending on the context.

In addition, document-level polarity classification tasks require more complex interpretation methods. Luo et al. noted that real-world documents often have mixed polarity, and classifying overall polarity at the document level is not suitable for real-world applications [5]. An interpretation limited to document-level polarity may not be practical. Even when polarity is mixed in a document, visualization of the important parts in both polarities is expected to improve understanding of the overall document content and the model's final predictions.

Based on these points, the goal of this study is to realize a polarity classification model that (1) can visualize mixed polarity and (2) can hierarchically capture shifts in textual meaning depending on context. To achieve this goal, we have applied inter-sentence attention, which captures the relationship between sentences. By incorporating inter-sentence attention into the previous model, we have attempted to propose a more practical and interpretable classification model.

II. Related Work

A. The Context-Aware Polarity Shift of Texts

Ito et al. proposed a CSNN (Contextual Sentiment Neural Network) that captures word polarity shifts in polarity classification and can naturally explain the process of prediction [6]. Polarity shift is a phenomenon in which the polarity changes depending on the surrounding context, such as "good" and "not good". Using the output results of each layer of the CSNN, which consists of a layer that calculates the original polarity of a word, the presence or absence of polarity shift, and the polarity of the word with context, they attempted to provide a more natural description of the polarity classification process.

The current affiliation of the first author is NTT Docomo, Inc.
mailto: ikeda.daisuke.899@m.kyushu-u.ac.jp

They succeeded in capturing contextual word polarity shifts and having output them as an interpretation of the prediction process, but their model can not capture polarity shifts at the sentence-level, which are dependent on document structure. Capturing sentence-level polarity shifts due to relationships with surrounding sentences may provide a more natural understanding for the prediction process.

B. The Interpretation Using Inter-Sentence Attention

Lu et al. proposed a model that uses a hierarchical Transformer to capture the relationship between each sentence and extracts important sentences as interpretations based on inter-sentence attention [7]. The Transformer is a model that captures the relationship between tokens using a mechanism called attention [8]. Similarly, a hierarchical Transformer consists of a token-level Transformer and a sentence-level Transformer that inputs the representation of each sentence obtained from the token-level transformer, and captures the relationship between sentences. They proposed a document classification model HBM (Hierarchical BERT Model) using this hierarchical Transformer, and it showed higher performance than conventional models. In addition, they analyzed the inter-sentence attention of the sentence-level Transformer and extracted the sentence that gathered the most attention from the other sentences in the document as the interpretation.

However, the interpretation of HBM is limited to visualizing sentences that received a lot of attention and does not visualize the mixed polarity in the document.

C. The Interpretation by Ranking Each Sentence

Bacco et al. introduced SCC (Sentence Classification Combiner model), which determines the polarity of an overall document by calculating the polarity of each sentence independently and then averaging them together [9]. It is interpretable by extracting the sentences with the highest polarity score corresponding to the polarity of the entire document as the interpretation. Furthermore, since the polarity of all sentences is calculated, the sentences with the highest score for each polarity can be identified even if the document contains mixed polarity. Although SCC is a very simple model, it showed classification performance near that of state-of-the-art models. They also conducted quantitative evaluations of interpretability, comparing interpretations extracted from it with sentences that humans consider important on the same document.

However, SCC's interpretability has significant challenges in that it evaluates polarity independently for each sentence. As mentioned, the meaning of a text is highly dependent on its context, and this is also true with regard to sentences. As long as the polarity ignores the interaction of each sentence, SCC cannot properly evaluate the polarity of these sentences and may not be able to extract them as interpretations.

Therefore, this study attempts to improve SCC by changing from separate polarity to context-sensitive polarity. We propose a method to calculate the polarity of each sentence that reflects the interaction between sentences using the inter-sentence attention introduced in Section II-B.

III. Model

In this section, we describe a model that can output inter-sentence attention, which captures the relationship between sentences, in addition to SCC. Since HBM is trained on a classification task that is easy to learn, we will attempt to obtain an inter-sentence attention that captures more complex relationships between sentences by using a model trained on difficult tasks. Then, we describe a proposed model that combines these methods.

A. SCC

SCC averages the polarity probabilities of each sentence to determine the polarity of the entire document.

1) Model Description: SCC calculates the polarity probability of each sentence by using RoBERTa [10]. Let document D be denotated by $(S_1, \ldots, S_{|D|})$, where each sentence S_i is represented by $(w_0^i, \ldots, w_{|S_i|+1}^i)$ of tokens w_j^i, where w_0^i and $w_{|S_i|+1}^i$ represent special tokens for the beginning and end of each text, respectively. Each tokenized sentence is independently input to RoBERTa to obtain a contextual embedding $(\mathbf{h}_0^i, \ldots, \mathbf{h}_{|S_i|+1}^i)$. By inputting $\mathbf{h}_0^i$ into the classification layer, the polarity probability of S_i, $\mathbf{P}_i \in \mathbb{R}^3$ is output, where $\mathbf{P}_i$ is a vector representing the probability that the sentence is negative, neutral, or positive, and the sum is always 1. Finally, the polarity probabilities of all sentences are averaged, and the polarity that shows the largest probability is determined as the polarity of the entire document.

2) Interpretation: Since SCC determines the polarity of the entire document by averaging the polarity probabilities of each sentence, it is clear that sentences with high probabilities contribute significantly to the model determination. Therefore, after determining the polarity of the entire document, we can interpret SCC by ranking all sentences in order of their polarity probabilities and extracting the top sentences.

B. STAS

Now we introduce STAS [11] and explain inter-sentence attention as the key feature of STAS, and its training method.

1) Model Description: STAS is a model for unsupervised extractive summarization tasks that aims to capture the hierarchical structure of documents. To capture the relationship between both tokens and sentences, STAS uses a hierarchical Transformer that concatenates a token-level Transformer (Trans_T) and a sentence-level Transformer (Trans_S).

As well as SCC, a document D is devided into tokens $(S_1, \ldots, S_{|D|})$ and they are input into STAS. STAS differs from SCC in that tokenized documents are entered together. Tokenized documents are input to Trans_T to obtain a contextualized embedding for each token $\mathbf{h}_0^1, \ldots, \mathbf{h}_{|S_{|D|}|}^{|D|} \in \mathbb{R}^{d_e}$. The embedding of the special token $<s>$ is used for the embedding of each sentence, and the corresponding $\mathbf{H} = (\mathbf{h}_0^0, \mathbf{h}_0^1, \ldots, \mathbf{h}_0^{|D|}) \in \mathbb{R}^{|D| \times d_e}$ is input into Trans_S. The architecture of Trans_S is the same as BERT, it stacks multiple encoder layers. Each encoder layer consists of *Multi-Head Attention*, *Add&Norm*, *Feed Forward* layers. When the embedding of each sentence $\mathbf{H}$ is input, it is calculated in the *Multi-Head Attention* layer as $Concat(head_1, head_2, \ldots, head_h)\mathbf{W}^O$, where

$$head_i = Softmax(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_k}})\mathbf{V}_i, \text{and}$$

$$\mathbf{Q}_i = \mathbf{H}\mathbf{W}_i^Q, \mathbf{K}_i = \mathbf{H}\mathbf{W}_i^K, \mathbf{V}_i = \mathbf{H}\mathbf{W}_i^V.$$

$Trans_S$ uses h heads similar to BERT, and when $d_k = d_e/h$, $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d_e \times d_k}$ are weight matrices at the ith head, and $\mathbf{W}^O \in \mathbb{R}^{hd_k \times d_e}$ is also a weight matrix. Here, the (i, j) component of $Softmax(\mathbf{Q}_i\mathbf{K}_i^\top/\sqrt{d_k}) \in \mathbb{R}^{|D|\times|D|}$, called the self-attention matrix, is the attention that S_i pays to S_j, and its value varies according to various relationships between sentences. *Add&Norm* layer performs residual connection and layer normalization, and *Feed Forward* layer consists of fully-connected layers and activation functions. Finally, $Trans_S$ outputs contextualized embeddings $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_{|D|})$ of sentences and inter-sentence attention $\mathbf{A} \in \mathbb{R}^{|D|\times|D|}$ that captures the relationship between sentences from each head in each encoder layer.

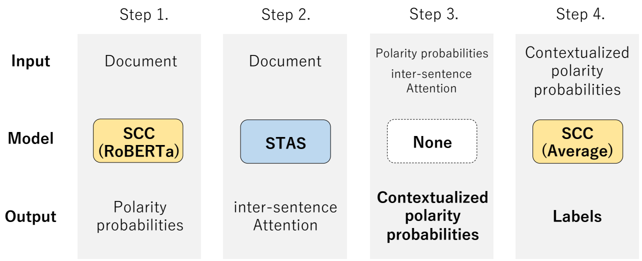
2) Pre-training: STAS is pre-trained STAS with two tasks to capture relationships between sentences. The first task, called Masked Sentence Prediction (MSP), estimates the original masked sentence. This is similar to the task used in BERT pre-training, but STAS differs in that all tokens in the masked sentences are replaced by [MASK] tokens. When estimating the masked S_m, $\mathbf{s}_m$ and Transfomer decoder are used to estimate the tokens in order from the beginning. The second task, called Sentence Shuffling (SS), estimates the original order of each sentence from the shuffled documents. By using shuffled $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_{|D|})$ and Pointer Network [12], the original order is estimated. By training with MSP and SS, STAS will learn to capture word connections across sentences and document structure. Since related sentences tend to show higher attention, inter-sentence attention may be useful for capturing the relationship between sentences.

C. Proposed Model

The idea of the proposed model is to solve SCC's problem by capturing polarity information from sentences that show high values in inter-sentence attention.

As mentioned in Section II-C, SCC calculates the polarity probability of each sentence independently, so the analysis of the meaning of each sentence is not natural. It is difficult to capture implicit polarity and reverse polarity. By correctly capturing the contextual polarity of these sentences, it may be possible to improve accuracy and even extract them as interpretations. Therefore, we consider that inter-sentence attention, which tends to show higher scores between related sentences, could be applied. By using inter-sentence attention scores from surrounding sentences, we propose a method to contextualize polarity probabilities.

1) Model Description: Similar to SCC, the proposed model first divides the document D into sentences and obtains the polarity probability matrix $\mathbf{P} \in \mathbb{R}^{|D|\times 3}$ for each sentence using RoBERTa. Here, the row component of $\mathbf{P}$ corresponds to the polarity probability of each sentence in the document. Next, the document is entered into STAS and the inter-sentence attention $\mathbf{A} \in \mathbb{R}^{|D|\times|D|}$ is obtained from the sentence-level Transformer ($Trans_S$). Since the sum of the rows is always 1, $\mathbf{A}$ is a stochastic matrix. Using $\mathbf{A}$ and $\mathbf{P}$, contextualized polarity probabilities are calculated by the process explained below and averaged as in SCC to determine polarity for the entire document. These steps are shown in Fig. 1.



	Step 1.	Step 2.	Step 3.	Step 4.
Input	Document	Document	Polarity probabilities inter-sentence Attention	Contextualized polarity probabilities
Model	SCC (RoBERTa)	STAS	None	SCC (Average)
Output	Polarity probabilities	inter-sentence Attention	**Contextualized polarity probabilities**	**Labels**

Fig. 1. Our model outputs polarity labels according to these steps.

2) Contextualization of Polarity Probabilities: The idea is based on the product $\mathbf{AP} \in \mathbb{R}^{|\mathbf{D}|\times\mathbf{3}}$ of $\mathbf{A}$ and $\mathbf{P}$. First, $\mathbf{AP}$ is also a stochastic matrix since $\mathbf{A}$ and $\mathbf{P}$ are both stochastic matrices. Here the ith row of $\mathbf{AP}$ is the polarity probability of the context related to S_i, reflecting the greater polarity probability of the sentence to which S_i paid higher attention. By using $\mathbf{AP}$ and the original polarity probability matrix $\mathbf{P}$, we contextualize the polarity probabilities of each sentence.

In this study, $\mathbf{A}$ obtained from STAS is used for the the processing. First, we use $\mathbf{A}$ transposed from the results of the prior experiment, where $\mathbf{A}_{i,j}$ represents the attention from S_j to S_i. By transposing, it takes in more polarity information from sentences which pay high attention to itself, instead of sentences which itself pays high attention to. We also believe that by targeting sentences with attention of *top_k*, we can capture polarity information of only those sentences that are more relevant. Therefore, we obtain a new stochastic matrix $\mathbf{A}'$ using Softmax with temperature, leaving only *top_k* attention. Softmax with temperature can adjust the distribution of the output probability according to the value of the parameter *temperature*. By giving $temperature < 1$, the probability of the sentence with the highest attention will be more emphasized. Finally, the new context's polarity stochastic matrix $\mathbf{A}'\mathbf{P}$ and the original $\mathbf{P}$ are added together in a given ratio to obtain the contextualized polarity probabilities. This ratio adjusts how much contextual polarity information is included (how much of its own original polarity information is retained). In this study, we use "the neutral probability of each sentence" for this ratio. This is because we have considered that it may be easier to capture sentences that have implicit polarity information. Fig. 2 shows an example of the calculation method when $|D| = 4, top\_k = 2$.

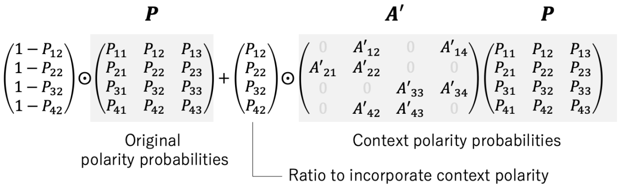

P
A′
P
Original
polarity probabilities
Context polarity probabilities
Ratio to incorporate context polarity

Fig. 2. An example of the calculation method when $|D| = 4, top\_k = 2$

IV. Experiment

This section describes experiments comparing classification performance and interpretability between the proposed model and SCC. We try to improve the problem that SCC cannot capture contextual polarity by applying inter-sentence attention, and we test its impact on accuracy and interpretability.

A. Classification Experiment

We use two polarity-labeled datasets, both created from movie review texts. The first one is the IMDB dataset[1][13]. Each data is labeled negative or positive and consists of a total of 50,000 review sentences, 25,000 for each. The second one is the Movie Review[2][14]. It consists of a total of 2,000 review sentences, 1,000 negative review and 1,000 positive review each.

Punctuation, tags, and consecutive spaces are removed and words are converted to all lowercase.

Both SCC and the proposed model should calculate the polarity probability of each sentence using RoBERTa. In this study, we use RoBERTa, which is fine-tuned with a polarity classification task available on HuggingFace[3]. For domain adaptation, we additionally have fine-tuned RoBERTa using SST-5 (Stanford Sentiment Treebank with 5 labels) [15], in which single sentences extracted from movie reviews are labeled with five labels, and "negative" and "somewhat negative" are re-labeled as "negative" and "somewhat positive" and "positive" are re-labeled as "positive".

Xu et al. published a trained model of STAS on GitHub[4], and we use it modified to output inter-sentence attention instead of as a document summarization model. STAS is implemented using the fairseq [16] toolkit for natural language processing. $Trans_T$ in STAS is initialized with $RoBERTa_{BASE}$ parameters: the number of encoder layers $L = 12$, one of heads $A = 12$, and the dimension of embedding $H = 768$. Similarly, $L = 6$, $A = 12$, and $H = 768$ are used for $Trans_S$. The inter-sentence attention can be obtained by the number of heads in each encoder layer of $Trans_S$. We average for each layer and use the inter-sentence attention of the final layer. In STAS, the maximum number of sentences per document is 30, and if the number of tokens exceeds 512, subsequent tokens are truncated. When calculating the polarity probability of each sentence using RoBERTa, only those sentences that could be entered into STAS in the document are used.

TABLE I
Results of classification experiments

Dataset	Model	Accuracy	F1 Score
IMDB	SCC	88.11	87.44
	Proposed Model	**88.35**	**87.89**
Movie Review	SCC	81.20	78.64
	Proposed Model	**83.80**	**82.04**

The dataset used in this study has two labels (negative and positive), but the polarity probabilities are calculated for three labels, including neutral. Therefore, when averaging the polarity probabilities of the sentences and determining the polarity label of the document, the label with the higher value among negative and positive is used.

The hyperparameters of the proposed model are the *top_k* and the *temperature* described. For parameter tuning, the dataset is split into training set and test set in a 3:1 ratio, and further 10-fold cross-validation is conducted using the data for training. The parameters that showed the best performance in 10-fold cross-validation are used to evaluate the test data. We use accuracy and F1 scores as our evaluation metrics.

After parmeter tuning, we obtain $top\_k = 3$, $temperature = 0.01$ for IMDB and top_k =3, $temperature$ =0.05 for Movie Review. As a result, our method improves the accuracy for both IMDB and Movie Review (see Table I). We found some examples in SCC's misclassified data, where the most of such documents consist of sentences with the polarity other than the document's polarity. It is remarkable that the proposed method allows us to correctly classify some of these difficult documents by using contextualized polarity probabilities.

B. Interpretability Experiment

In Fig. 3, a true positive example document is shown, where there exist many sentences with hight positive scores. Both models can correctly classify such "easy" cases. In such cases, sentences with high probabilities of predictive labels can be extracted as interpretations. If this interpretive sentence is truly important to the polarity of the document, the model is considered highly interpretable. Therefore, we examine the interpretation performance by quantitatively analyzing the differences in interpretive sentences obtained from SCC and the proposed model using the annotated dataset.[5]

1) Evaluation Method: To evaluate the interpretability of SCC, Bacco et al. examined the overlap between interpretive sentences extracted with SCC and those humans judged as important. They used 150 randomly selected documents from IMDB, and four annotators choose three important sentences in each document. We also use it in this study.

For the jth document, let $N_{j,k}$ represent the number of sentences considered important by k or more annotators. Next, the jth document is input to both models, and sentences are ranked

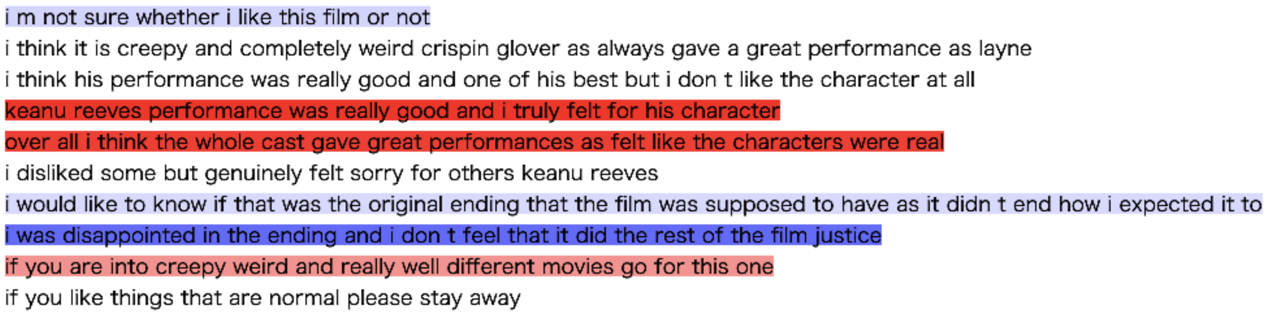
[1] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[2] https://www.cs.cornell.edu/people/pabo/movie-review-data/

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[4] https://github.com/xssstory/STAS

[5] https://github.com/lbacco/ExS4ExSA/tree/main



i m not sure whether i like this film or not

i think it is creepy and completely weird crispin glover as always gave a great performance as layne

i think his performance was really good and one of his best but i don t like the character at all

keanu reeves performance was really good and i truly felt for his character

over all i think the whole cast gave great performances as felt like the characters were real

i disliked some but genuinely felt sorry for others keanu reeves

i would like to know if that was the original ending that the film was supposed to have as it didn t end how i expected it to

i was disappointed in the ending and i don t feel that it did the rest of the film justice

if you are into creepy weird and really well different movies go for this one

if you like things that are normal please stay away

Fig. 3. A true positive example with high polarity in proposed model: Blue (resp. red) indicates negative (resp. positive) statements, with the intensity indicating the polarity probability.

based on the predicted labels and the polarity probability of each sentence to obtain the top $N_{j,k}$ interpretive sentences. Let $TP_{j,k}$ denote the number of interpretive sentences annotated by k or more persons. Then $TP_{j,k}/N_{j,k}$ denotes the overlap per document. Since there are cases $N_{j,k} = 0$, especially when $k = 3$ or more, we only include documents for which $N_{j,k} \geq 1$ and the model correctly classifies polarity. The evaluation index $Prec_k$ when the number of documents is D_k is calculated as follows:

$$Prec_k = \frac{1}{D_k} \sum_{j}^{D_k} \frac{TP_{j,k}}{N_{j,k}}.$$

We compare the value of $Prec_k$ between SCC and the proposed model at $k = 1 \sim 4$ to test which interpretive sentence is closer to a human judgment.

2) Result: The hyper parameters determined for IMDB in Section IV-A are used. We calculate the value of $Precision$ for the 100 data available for testing, and Table II shows the results. The proposed method improves the accuracy and the value of $Prec_k$ from SCC except for $k = 2$. It suggests that the interpretive sentences obtained by contextualizing the polarity probabilities may be closer to human annotations (i.e., sentences that humans consider important naturally are more likely to contribute significantly to the model's predictions).

3) Discussion: We analyze the relationship between changes in polarity probability by the proposed method and annotation with examples.

Table III shows an example, where the proposed method captures sentences that implicitly contain polarity information. The "No." in the table is the index of each sentence in the document, and the polarity probabilities are "negative", "neutral", "positive" in that order. First, more than three annotators considered sentences (2), (4), and (5) important. It is clear from reading the subsequent sentence that (2) is a negative sentence, but it does not contain direct negative expressions. Therefore, although the polarity probabilities are somewhat biased toward the negative, the probability of neutrality is also relatively high, and (2) is not an interpretive sentence in SCC. Here, inter-sentence attention within the proposed model shows high attention to (2) from (4) and (5), which have strong negative meaning. By taking in the polarity information of sentences with high attention scores, the contextual polarity probability in (2) increases the negative value, and it is chosen as the interpretation in the proposed model. In this example, it can be said to capture a sentence that implicitly contains polarity whose meaning changes with context. On the other hand, we observe some cases, where unintended changes in polarity (Table IV). The negative sentence (8) is followed by the positive meaning (9), which is like an "ironic" sentence structure that emphasizes the negative meaning of (8). Sentences (8) and (9) pay high attention to each other, and the proposed model has been able to capture this relationship by inter-sentence attention. However, as a result of the interaction on the polarity probability, the value of the negative in (8) is decreased. It is suggested that the proposed method may not be robust against sentence structures like this example.

V. Conclusion

In this study, we have addressed improving the interpretability of the polarity classification model at the document level. Although SCC has better classification performance and interpretive properties, the prediction process is not natural in that the polarity of each sentence is calculated independently. The main idea is that the polarity contextualization would achieve a natural prediction process, and improve classification performance and interpretability. To capture the relationship between sentences, the document summarization model STAS and inter-sentence attention are used. The proposed model using contextualized polarity shows better classification performance than SCC for the two datasets. In addition, we have examined the overlap between the interpretive sentences obtained from both models and the sentences judged by humans to be important for the polarity of the document, and have found that our model shows greater overlap. By applying inter-sentence attention and contextualizing the polarity of each sentence, it has become easier to capture natural changes in the meaning of sentences and extract sentences that resemble human judgments as interpretations.

There are two important future works. As mentioned, unintended changes in polarity could occur depending on the document structure. We need to examine other similar cases and devise more robust methods of contextualizing polarity probabilities. In addition, this study uses the trained STAS

TABLE II
RESULTS OF INTERPRETABILITY EXPERIMENTS

Model	Accuracy	F1 Score	$Prec_1$	$Prec_2$	$Prec_3$	$Prec_4$
SCC	91.00	90.32	73.07	**66.04**	57.59	44.87
Proposed Model	**92.00**	**91.49**	**73.44**	65.33	**58.79**	**44.94**

TABLE III
AN EXAMPLE OF CAPTURING A SENTENCE WITH IMPLICIT POLARITY, WHERE ANN., PROB. AND CONTEXTUALIZED PROB. STAND FOR ANNOTATION, POLARITY PROBABILITY, AND CONTEXTUALIZED POLARITY PROBABILITY, RESPECTIVELY.

No.	Ann.	SCC	Proposed	Prob.	Contextualized Prob.
(2)	✓		✓	(0.57,0.41,0.02)	(0.68,0.31,0.01)
(3)		✓		(0.63,0.35,0.02)	(0.63,0.36,0.01)
(4)	✓	✓	✓	(0.95,0.05,0.00)	(0.91,0.06,0.03)
(5)	✓	✓	✓	(0.95,0.05,0.00)	(0.94,0.06,0.00)

No.	Sentences (excerpt)
(2)	Problem was that I spent most of the time trying to keep my finger away from the fast forward button.
(3)	It sure would have sped up the film's slow pacing, but then again I wouldn't know about too much that was going on, which was reasonably hard to figure out or keep interest in the first place.
(4)	The performances ranged from too melodramatic or just plain dull, and that's probably because these characters are unconvincing, stale and coma inducing.
(5)	The actual back-story of the old bed and the spirits is incredibly boring and messily put together, with too much focus on a flimsy romance, being laughable when it shouldn't be and overall it's constructed in an ordinary manner that just lacks the oomph or conviction to carry the film.

TABLE IV
AN EXAMPLE OF A WRONG POLARITY CHANGE
(8) AND (9) PAY HIGH ATTENTION TO EACH OTHER AND INTERACT IN POLARITY PROBABILITY.

No.	Ann.	Prob.	Contextualized prob.
(8)	✓	(0.85, 0.14, 0.01)	(0.78, 0.17, 0.05)
(9)		(0.00, 0.12, 0.88)	(0.10, 0.12, 0.78)

No.	Sentences (excerpt)
(8)	Do yourself a favor and avoid this movie at all costs.
(9)	You'll be glad you did.

model without additional training, but this model was trained on news articles. Since both datasets used in this study consist of movie reviews, additional pre-training of STAS in the same domain may help inter-sentence attention more accurately capture the relationship between sentences and improve the performance of the proposed model.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23K28149.

REFERENCES

[1] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal, "Management's tone change, post earnings announcement drift and accruals," *Review of Accounting Studies*, vol. 15, no. 4, pp. 915–953, 2010.

[2] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[3] E. Mosca, D. Demirtürk, L. Mülln, F. Raffagnato, and G. Groh, "GrammarSHAP: An Efficient Model-Agnostic and Structure-Aware NLP Explainer," in *Proceedings of the First Workshop on Learning with Natural Language Supervision*, 2022, pp. 10–16.

[4] H. Yan, L. Gui, and Y. He, "Hierarchical interpretation of neural text classification," *Comput. Linguistics*, vol. 48, no. 4, pp. 987–1020, 2022.

[5] L. Luo, X. Ao, F. Pan, J. Wang, T. Zhao, N. Yu, and Q. He, "Beyond Polarity: Interpretable Financial Sentiment Analysis with Hierarchical Query-driven Attention," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 4244–4250.

[6] T. Ito, K. Tsubouchi, H. Sakaji, T. Yamashita, and K. Izumi, "Contextual Sentiment Neural Network for Document Sentiment Analysis," *Data Sci. Eng.*, vol. 5, no. 2, pp. 180–192, 2020.

[7] J. Lu, M. Henchion, I. Bacher, and B. M. Namee, "A Sentence-Level Hierarchical BERT Model for Document Classification with Limited Labelled Data," in *Proceedings of 24th International Conference on Discovery Science*, 2021, pp. 231–241.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[9] L. Bacco, A. Cimino, F. Dell'Orletta, and M. Merone, "Explainable Sentiment Analysis: A Hierarchical Transformer-Based Extractive Summarization Approach," *Electronics*, vol. 10, p. 2195, 2021.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.

[11] S. Xu, X. Zhang, Y. Wu, F. Wei, and M. Zhou, "Unsupervised Extractive Summarization by Pre-training Hierarchical Transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 1784–1795.

[12] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer Networks," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2692–2700.

[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 142–150.

[14] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, pp. 271–278.

[15] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.

[16] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A Fast, Extensible Toolkit for Sequence Modeling," in *Proceedings of NAACL-HLT 2019*, 2019.