International Journal of Service and Knowledge Management
International Institute of Applied Informatics
2026, Vol. 10, No. 1, IJSKM908

Interpretable Document-Level Polarity Classification with Inter-sentence Attention

Shingo Kato *, Daisuke Ikeda †

Abstract

Large language models (LLMs), such as BERT and GPT-3, significantly impact our daily lives. They can derive accurate outputs for many tasks about natural languages, but they do not explain the reason for the outputs. In this sense, LLMs work as a black box and have the problem of *interpretability*. This paper is devoted to considering polarity classification of documents. Compared to simple sentence-level polarity classification, document-level one makes it more difficult to create a high-performance model because we have to consider relationships between sentences. To tackle this change, we use inter-sentence attention, which can capture the relationship between sentences: the higher an inter-sentence attention score is, the more relevant the corresponding sentences are to each other. We use inter-sentence attention scores to capture the context of sentences and propose a model whose classification is more similar to human judgment. To validate the proposed model, we conduct three types of experiments: one is to compare classification performance with prior models; the second one is to compare interpretability with prior models; and the last one is to show the ability of inter-sentence attention whether it could capture the relationship between sentences. From the first two types, we found that our model is more accurate on two real datasets. In the second type of experiment 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 and found 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. From the third type of experiment, we show that our model can capture the contexts of sentences.

Keywords: Polarity Classification, Interpretability, Inter-Sentence Attention, SCC, STAS

1 Introduction

The task to judge if a given text is positive or negative is called polarity classification, and it has been well-studied for short texts, such as review texts. Document-level polarity

* Kyushu University, Fukuoka, Japan. The current affiliation of the first author is NTT Docomo, Inc.
† Kyushu University, Fukuoka, Japan.

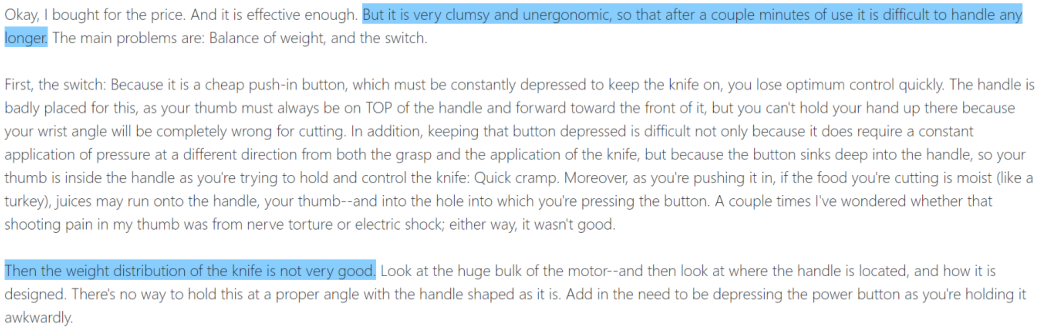

Figure 1: An example of an interpretation method in a text classification model [7]

classification becomes more and more important from the practical viewpoint. For example, it was shown that there is a relationship between the polarity of the text of Management Discussion and Analysis (MD&A) and stock prices [1].

From simple methods using polar word dictionaries to new methods, such as deep learning, many methods are used for polarity classification, and the accuracy of it has been improving year by year. In recent years, large language models (LLMs for short), such as BERT, have dramatically improved their accuracy [2]. LLMs can capture context, and thus they enable polarity classification models to correctly predict the polarity even when a given document contains a mixture of negative and positive sentences.

However, these models based on LLMs or deep learning models suffer from the issue of interpretability, and they can not explain the reason for their prediction even if prediction is correct. There exists a trade-off in general between the classification performance and interpretability [3]. Interpretability is as important as model performance, especially in some fields, such as medicine and finance, where there exists a great deal of responsibility for the predictive results. Therefore, research to improve model interpretability in these fields has been active in recent years [4].

In case of general text classification, including polar classification, many methods to increase interpretability by extracting and visualizing the features that contributed to the model prediction, such as SHAP [5] and and LIME [6]. As shown in Fig. 1, 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 [8]. Yan et al. also pointed out that because LLMs analyze the meaning of features hierarchically, interpretation by a single feature may be insufficient to explain the model's predictions [9]. 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.

Furthermore, the document-level polarity classification task addressed in this study requires more complex interpretation methods. Luo et al. point out that real-world documents often have mixed polarity and that classifying overall polarity at the document level is not suitable for real-world applications [10]. Traditional interpretation methods in text classification extract features that contribute to model predictions, so interpreting results limited to one polarity at the document level may not be practical. Even when polarity is mixed in a document, visualization of the areas that contributed to both polarities may provide a better

understanding of the overall document content and the final predictions of the model.

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 [10]. 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 use inter-sentence attention, which captures the relationship between sentences. Incorporating inter-sentence attention into an existing model, we propose a more practical and interpretable classification model. This paper is an extended version of a conference paper published in [11], where mainly additional figures, tables, and examples are added.

This paper is organized as follows: after we review related works, we'll show our proposed model. Then we evaluate the proposed model with real datasets in Section 4.

2 Related Work

In this section, we present three previous studies related to interpretable document level polarity classification models as described in Section 1. We first present a study that addresses capturing changes in textual meaning depending on context. Next, we introduce a text classification model that incorporates a mechanism to capture relationships among sentences and focus on important sentences. Finally, we present a model that can classify polarity with high accuracy and interpretability. After describing the challenges of the model, we explain how it can be applied to this study.

2.1 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 [12]. 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.

Yang et al. have improved the accuracy of implicit polarity classification, which targets sentences that do not contain explicit emotional expressions, and have proposed a new model called GACNN (Graph Attention Convolutional Neural Network) [13]. GACNN uses a graph convolutional neural network to capture relationships between words and propagate semantic information, and an attention mechanism with special constraints to direct more attention to specific essential words. Implicit polar sentences are difficult to classify because they are composed of neutral words and generate different emotions depending on the context, but GACNN achieves classification with higher accuracy than conventional models due to the aforementioned architecture.

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 of the prediction process.

2.2 The Interpretation Using Inter-Sentence Attention

Lu et al. proposed a model that uses a hierarchical Transformer to capture the relationship between sentences and extracts important sentences as interpretations based on inter-sentence attention [7]. A Transformer is a model with attention that captures the relationship between tokens [14]. 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. In addition, many studies have discussed the suitability of attention as an interpretation of the text classification model [15, 16].

2.3 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 [17]. 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 2.2.

3 Model

In this section, we explain SCC and then how to construct our model, which outputs inter-sentence attention, capturing the relationship between sentences. Since HBM is trained on

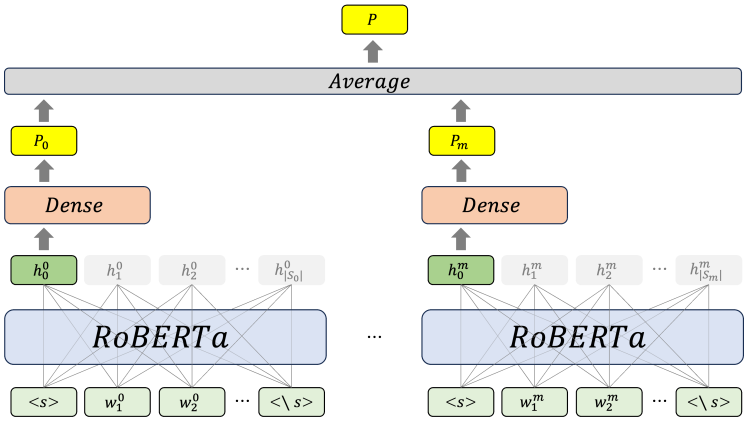

Figure 2: The architecture of SCC

an easy classification task, we will try to obtain 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.

3.1 SCC

SCC averages the polarity probabilities of each sentence to determine the polarity of the entire document. The architecture of SCC is shown in Fig. 2.

3.1.1 Model Description

SCC calculates the polarity probability of each sentence by using RoBERTa [18]. Let $D = (S_1, \ldots, S_{|D|})$ be a document, where each sentence S_i is represented by $S_i = (w_0^i, \ldots, w_{|S_i|+1}^i)$ of tokens w_j^i, and w_0^i and $w_{|S_i|+1}^i$ represent special tokens for the beginning and end of each sentence, respectively. Each tokenized sentence is independently given 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 obtained, where $\mathbf{P}_i$ is a vector representing the probability that the sentence is negative, neutral, or positive. 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.

3.1.2 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.

3.2 STAS

Now we introduce STAS [19] and explain inter-sentence attention as the key feature of STAS, and its training method.

3.2.1 Model Description

STAS is a model for unsupervised extractive summarization tasks. It aims to capture the hierarchical structure of documents, and, 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 ($Trans_S$). The architecture of STAS is shown in Fig. 3.

As well as SCC, a document D is devided into tokens $(S_1, \ldots, S_{|D|})$ and they are given to STAS. STAS differs from SCC in that tokenized documents are entered together. Tokenized documents are given 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 given to $Trans_S$. The architecture of $Trans_S$ is the same as BERT, it stacks multiple encoder layers, each of which 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

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

$Trans_S$ 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.

When STAS performs document summarization, it uses this inter-sentence attention and the score calculated by MSP, which will be explained in the next section, to rank each sentence and extract important sentences.

3.2.2 Pre-training

STAS is pre-trained with two tasks to capture relationships between sentences. The first task, called Masked Sentence Prediction, MSP for short, is to estimate the original masked sentence. This is similar to the task used in pre-training of BERT, 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 for short, is to estimate 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 [20], the original order is estimated. By training with MSP and SS, STAS learns to capture word connections across sentences and the document structure. Since related sentences tend to show higher attention, inter-sentence attention may be useful for capturing the relationship between sentences.

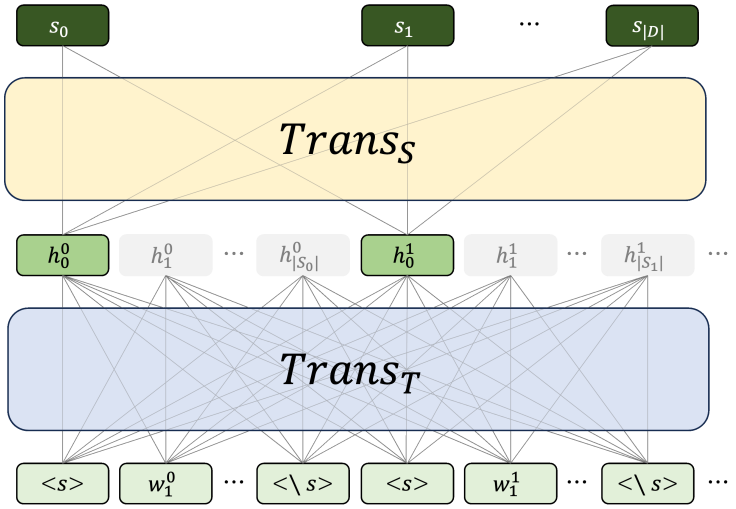

Figure 3: The architecture of STAS

3.3 Inter-sentence Attention

Here we show examples of inter-sentence attention obtained by STAS. Given the document in Table 1, we obtain inter-sentence attention from the last layer of Encoder in $Trans_S$. Table 2 shows relatively high/low inter-sentence attention.

From the table, we see that the two sentences with relatively high inter-sentence attention are those with greater contextual connection. In particular, although sentences (7) and (1) are located apart, they show a higher inter-sentence attention score maybe because they have "German perspective" appearing in common. On the other hand, sentences with lower inter-sentence attention scores make it difficult to conclude that they have a direct contextual connection between them. This tendency to show contextual connections between sentences assigned high attention suggests that inter-sentence attention derived from STAS may be useful for capturing document structure.

Table 1: Example of a document

No.	documents
(1)	I found the movie very real just as much as Saving Private Ryan and gave an upfront account from the German perspective.
(2)	The graphic action and individual characters gave this movie much to keep me interested through the whole movie.
(3)	Stalingrad was very brutal battle and the scenes gave this movie that stark reality of this battle.
(4)	I recommend this for anyone who is interested in military history.
(5)	This was my second viewing of this movie in a few years and was captivated by it's realism again.
(6)	The weapons, uniforms, and hearing it German while reading the subtitles gave it much credit for a good military movie.
(7)	In fact, this being a German perspective of the battle, I was rooting for the Germans, even though they were the agressor and ultimate loser of the war.

Table 2: An example of inter-sentence attention

(a) relatively high inter-sentence attention

direction of attention	score
(4) ⇒ (3)	0.19318
(3) ⇒ (5)	0.16124
(7) ⇒ (1)	0.15871

(b) relatively low inter-sentence attention

direction of attention	score
(2) ⇒ (4)	0.11564
(7) ⇒ (4)	0.12688
(7) ⇒ (6)	0.12888

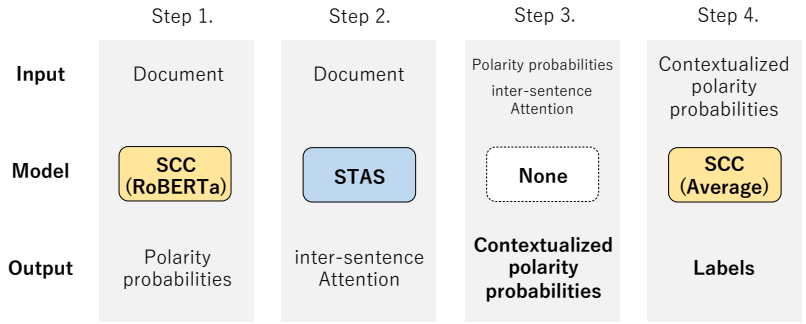

Figure 4: Our model outputs polarity labels according to these steps.

3.4 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 2.3, 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. In other words, we use inter-sentence attention scores as weights to obtain magnified polarity probabilities.

3.4.1 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. 4.

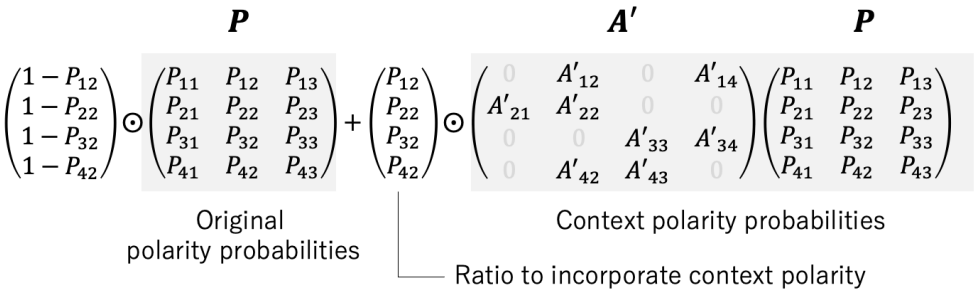

Figure 5: An example of the calculation method when $|D| = 4$ and $top\_k = 2$

3.4.2 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. 5 shows an example of the calculation method when $|D| = 4$ and $top\_k = 2$.

4 Experiments

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.

4.1 Classification Experiment

We use two polarity-labeled datasets, both created from movie review texts. The first one is the IMDB dataset[1][21]. 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][22]. It consists of a total of 2,000 review sentences, 1,000 negative review and 1,000 positive review each. Table 3 shows statistics for two datasets.

Table 3: Dataset statistics

Dataset	Label	reviews	avg. # of words	avg. # of sentences
IMDB	negative	25,000	227.2	12.6
	positive	25,000	230.7	12.0
Movie Review	negative	1,000	704.7	32.1
	positive	1,000	786.2	33.3

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) [23], 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 modified it to output inter-sentence attention instead of as a document summarization model. STAS is implemented using the fairseq [24] 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 4 shows the environement for our experiments. We use different versions of Python and PyTorch for STAS and RoBERTa.

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 in Section 3.4.2. For parameter tuning, the dataset is split into a training set and test set in

[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

Table 4: Environment for our experiments

Type	Description
OS	Ubuntu 18.04 LTS
CPU	AMD EPYC 7542
GPU	GeForce RTX 3080 10GB
CUDA Ver.	11.7
Python Ver.	3.6.12 (STAS), 3.8.10 (RoBERTa)
PyTorch Ver.	1.7.0+cu110 (STAS), 1.13.0+cu117 (RoBERTa)

Table 5: 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**

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 5). 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.

4.2 Interpretability Experiment

In Fig. 6, a true positive example document is shown, where there exist many sentences with high 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]

4.2.1 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 chose three important sentences in each document. They were told the polarity of the document in advance. Table 6 shows some annotated sentences. We also use it in this study.

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

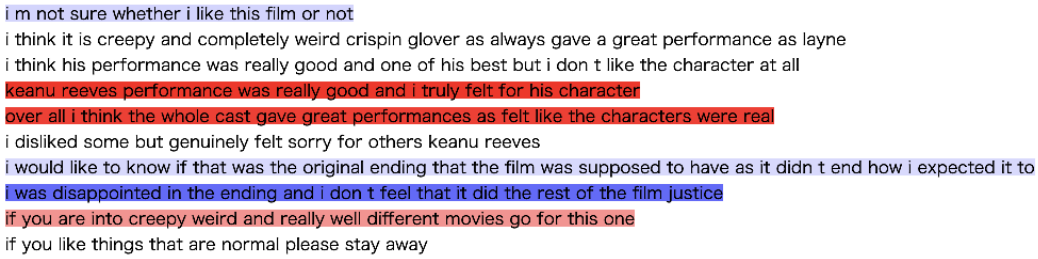

Figure 6: A true positive example with high polarity in proposed model: Blue (resp. red) indicates negative (resp. positive) statements, with the intensity indicating the polarity probability.

Table 6: Example of annotated documents made by Luca et al.

No.	documents	Ann_1	Ann_2	Ann_3	Ann_4
(1)	I found the movie very real just as much as Saving Private Ryan and gave an upfront account from the Germ an perspective.		3		3
(2)	The graphic action and individual characters gave this movie much to keep me interested through the whole movie.	2	2	2	2
(3)	Stalingrad was very brutal battle and the scenes gave this movie that stark reality of this battle.				
(4)	I recommend this for anyone who is interested in military history.	1	1	1	1
(5)	This was my second viewing of this movie in a few years and was captivated by it's realism again.	3		3	
(6)	The weapons, uniforms, and hearing it German while reading the subtitles gave it much credit for a good military movie.				
(7)	In fact, this being a German perspective of the battle, I was rooting for the Germans, even though they were the agressor and ultimate loser of the war.				

This study uses the same evaluation metric as Bacco et al. to analyze the overlap between the interpretive sentences extracted from the model and those selected by annotators. 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 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.

Table 7: 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**

4.2.2 Results

The hyper parameters determined for IMDB in Section 4.1 are used. We calculate the value of *Precision* for the 100 data available for testing, and Table 7 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).

4.2.3 Discussion

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

Table 8 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", and "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 9). 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.

4.3 Interpretability Experiment

As we showed above, there tends to be a strong contextual connection between sentences that show high attention. Therefore, the simultaneous extraction of interpreted sentences and other sentences in which the interpreted sentences incorporate polarity, those with the top top_k of attention, would capture the hierarchical structure of the document. Therefore, it could lead to an understanding of the importance of the interpreted sentences. In this section, we will discuss the applicability of this method to the aforementioned hierarchical interpretation by presenting specific examples.

Table 8: 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 9: 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.

In Fig. 7a, "I think that this movie is definitely worth watching, especially if you've lived in Japan or are interested in it" was extracted as an interpreted sentence, while "A lot of the comments seem to treat this film as a baseball movie, but I feel this is only secondary. It's really about living in Japan, and it really succeeds." was also extracted. Given the reviewer's opinion that the film is not a baseball movie but a film about the realities of life in Japan, it is possible to understand the reason for the recommendation in the interpretive text to those who have lived in or are interested in Japan, which may promote a deeper understanding of the document's content.

In Fig. 7b, while sentences praising the three aspects of plot, cinematography, and casting were selected as interpretive sentences, sentences explaining the plot and specifically stating praise were extracted at the same time. While the interpretive sentences by themselves are somewhat abstract in their praise of the film, the reviewer's evaluation of the film's plot is made concrete by following the sentences about the plot.

These examples suggest that capturing document structure by inter-sentence attention and extracting interpretive sentences and sentences related to interpretive sentences at the same time may lead to understanding the importance of interpretive sentences in a document and understanding the content of the document itself.

A lot of the comments seem to treat this film as a baseball movie, but I feel this is only secondary.
It's really about living in Japan, and it really succeeds.
I spent a few years living in Japan, and I suppose the reason that this movie didn't do too well is that you sort of have to have experienced Japan to get it.
I was watching this with a well-travelled friend who's never been to Japan, and he noted that many of the events in the movie were so ludicrous that they destroyed the suspension of disbelief.
My reply was that those events were the absolute unvarnished truth about life in Japan!
I think that this movie is definitely worth watching, especially if you've lived in Japan or are interested in it.

(a) Example 1

'A Tale of Two Sisters', or 'Janghwa, Hongryeon', is a true masterpiece.
Brilliant psychological thriller, heart-wrenching drama, and gripping horror all wrapped up in one beautifully orchestrated package.
From the intricate plot, to the beautiful cinematography, to the absolutely perfect casting, every aspect of this film is extraordinary.
For fear of revealing too much concerning the plot, I will just say it is very satisfying.
While it may appear to be a little difficult to understand at first, it does a good job of explaining things in the end.
And whether you prefer psychological thriller, drama, or horror, I promise you will not be disappointed.
From a technical standpoint, its nearly flawless.
The set, the cinematography, lighting, and especially the soundtrack, all are captivating.
The waltz seemed an odd choice at first, but proved to be an ingenious choice.
As for the casting, we're talking absolute perfection.
I'm Su-jeong is totally convincing as the defiant, yet troubled Su-mi.
Mun Keun- yeong is equally convincing as her emotionally traumatized sister Su-yeon.
These two girls were magical on the screen.
I will certainly be looking into their other films.
Yeom Jeong-ah is deceitfully cheerful and hauntingly evil as the stepmother.

(b) Example 1

Figure 7: Examples of hierarchical interpretation, where blue letters are interpreted sentences and bold letters are sentences that influenced the polarity of the interpreted sentence (towards high attention).

5 Conclusion

We have proposed a model which improves the interpretability of the polarity classification at the document level. The main idea is that we use context dependency when we compute polarity scores to achieve a natural prediction process and improve classification performance and interpretability, while SCC, which achieved better classification performance and interpretive properties, computes the polarity of each sentence independently. To incorporate the relationship between sentences, we use the document summarization model STAS and inter-sentence attention.

The proposed model using contextualized polarity shows better classification performance than SCC. 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, suggesting the possibility of improved interpretability.

There are three important future works. First, as mentioned in Section 4.2.3, 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. Second, this study uses the trained STAS model without additional training, but this model was trained on news articles. Since both datasets used here contain reviews for movies, 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. Third, since the proposed model could not deal with well some sentences including sarcasm expression, it is important to incorporate abilities for sarcasm detection, such as [25, 26], into the model when computing polarity probability.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23K28149.

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*, 15(4):915–953, 2010.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[3] 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*, 58:82–115, 2020.

[4] X.-Q. Chen, C.-Q. Ma, Y.-S. Ren, Y.-T. Lei, N. Q. A. Huynh, and S. Narayan. Explainable artificial intelligence in finance: A bibliometric review. *Finance Research Letters*, 56:104145, 2023.

[5] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[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*, pages 231–241, 2021.

[8] 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*, pages 10–16, 2022.

[9] H. Yan, L. Gui, and Y. He. Hierarchical interpretation of neural text classification. *Comput. Linguistics*, 48(4):987–1020, 2022.

[10] 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*, pages 4244–4250, 2018.

[11] S. Kato and D. Ikeda. Improving Interpretability in Document-Level Polarity Classification by Applying Attention. In *Proceedings of 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 21–26, 2024.

[12] T. Ito, K. Tsubouchi, H. Sakaji, T. Yamashita, and K. Izumi. Contextual Sentiment Neural Network for Document Sentiment Analysis. *Data Sci. Eng.*, 5(2):180–192, 2020.

[13] S. Yang, L. Xing, Y. Li, and Z. Chang. Implicit sentiment analysis based on graph attention neural network. *Engineering Reports*, 4(1):e12452, 2022.

[14] 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*, pages 5998–6008, 2017.

[15] S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics, 2019.

[16] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui. Attention interpretability across NLP tasks. *arXiv:1909.11218*, 2019.

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

[18] 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.

[19] 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*, pages 1784–1795, 2020.

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

[21] 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*, pages 142–150, 2011.

[22] 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*, pages 271–278, 2004.

[23] 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*, pages 1631–1642, 2013.

[24] 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.

[25] M. Jia, C. Xie, and L. Jing. Debiasing Multimodal Sarcasm Detection with Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18354–18362, 2024.

[26] L. Yang and D. Ikeda. The Influence of Linguistic Attribute Differences in Multilingual Datasets on Sarcasm Detection. *International Journal of Service and Knowledge Management*, 8(2), 2024.