Towards Generating Robust, Fair, and Emotion-Aware Explanations for Recommender Systems

Bingbing Wen
University of Washington
Seattle, WA, US
bingbw@uw.edu

Yunhe Feng
University of Washington
Seattle, WA, US
yunhe@uw.edu

Yongfeng Zhang
Rutgers University
New Brunswick, NJ, US
yongfeng.zhang@rutgers.edu

Chirag Shah
University of Washington
Seattle, WA, US
chirags@uw.edu

ABSTRACT
As recommender systems become increasingly sophisticated and complex, they often suffer from lack of fairness and transparency. Providing robust and unbiased explanations for recommendations has been drawing more and more attention as it can help address these issues and improve trustworthiness and informativeness of recommender systems. However, despite the fact that such explanations are generated for humans who respond more strongly to messages with appropriate emotions, there is a lack of consideration for emotions when generating explanations for recommendations. Current explanation generation models are found to exaggerate certain emotions without accurately capturing the underlying tone or the meaning. In this paper, we propose a novel method based on a multi-head transformer, called Emotion-aware Transformer for Explainable Recommendation (EmoTER), to generate more robust, fair, and emotion-enhanced explanations. To measure the linguistic quality and emotion fairness of the generated explanations, we adopt both automatic text metrics and human perceptions for evaluation. Experiments on three widely-used benchmark datasets with multiple evaluation metrics demonstrate that EmoTER consistently outperforms the existing state-of-the-art explanation generation models in terms of text quality, explainability, and consideration for fairness to emotion distribution. Implementation of EmoTER will be released as an open-source toolkit to support further research.

KEYWORDS
Explainable Recommendation; Emotion-aware Explanation; Fairness; Robustness

1 INTRODUCTION
Recommender systems assist users in finding relevant, contextual information and making decisions in diverse domains, such as shopping, entertainment, education, and healthcare. Advances of deep learning and big data [45] have made recommender systems increasingly complex and opaque, which hurts the trustworthiness and confidence in these systems [44]. Incorporating explainability [16] has been explored as one of the reasonable ways to mitigate such concerns and increase transparency in decision-making processes. Natural language sentence explanation, as a type of explanation that is easily understandable by average users, has gained much attention in recent years. In particular, many text generation methods such as LSTM [17] and Transformers [42] have been applied to explanation generation models. The Neural-Template (NETE) generation method [24] integrates both template and generation-based approaches to make the explanation generation process more controllable. PETER [26] proposed a Personalized Transformer to generate personalized explanation sentences. These works mainly focus on the text quality [27, 33] and explainability [24, 26].

High-quality explanations should be informative, i.e., containing factual information for users to understand. In addition, the generated explanations should also sound natural and be comprehensible as well as empathetic towards users. Monotonous and dull sentences could make users feel disengaged and less trustworthy [18]. Emotion-aware text could address these issues by diversifying explanations and making them more useful and trustworthy to the users. Effects of emotion-induced messages to comprehension and trust has been well established in fields such as psychology, philosophy, and cognitive science (e.g., [13, 34, 39]. Areas where emotion-induced content has been applied include text generation [15, 19], conversations [12], and social media [20]. However, thus far web search or recommendation systems scholars have not focused on emotional aspects of explanations and their impacts on informativeness or usefulness in providing explainability for recommendations. This is a novel problem as it is challenging, because we need to pay attention to or extract emotional information while retaining the original meaning of the explanation.

To explore the emotional aspect of explanations, we selected six widely-used emotion categories: happy, angry, surprise, sad, fear, and neutral in recommendation scenarios inspired by emotion theories proposed by Plutchik [35, 36]. These are also at the intersection of common emotion categories used in the research work, such as emotion recognition [10, 18, 46] and emotion text generation [15, 19]. Table 1 shows some explanation examples corresponding to different emotion categories.

When examining datasets and generative models for recommendation explanations, we find certain emotions are over-represented and considerable emotion biases exist, leading to lack of fairness as per the disparate impact definition [41]. For instance, in the TripAdvisor dataset, one of the most commonly-used recommendation explanation datasets, the “happy” emotion dominates others significantly. To be specific, 42.8% and 59.2% explanations in TripAdvisor...
Table 1: Explanation examples on six emotion categories on the TripAdvisor dataset.

| Emotion | Explanation |
|---------|-------------|
| Happy   | on the rooftop an open sky swimming pool - fun |
| Angry   | Shame on you for your terrible customer service |
| Surprise| The big surprise is the fantastic dining room for lunch with stations for every food |
| Sad     | The restaurant staff was nice but the buffet breakfast was pretty sad for the 18 |
| Fear    | It was a miserable way to end a trip to a great country |
| Neutral | the hotel is directly opposite budapest famous chain bridge and the danube |

are classified as “happy” by emotion detector Text2Emotion [32] and GoEmotion [10], respectively (see the ground-truth distributions in Figure 1). More importantly, we also observe that state-of-the-art models (e.g., PETER [26]) amplify such emotion bias in generated explanations and fail to perform robustly regarding different emotions. As shown in Figure 1, 61.3% and 69.5% explanations generated by PETER are classified as “happy” by Text2Emotion and GoEmotion respectively, leading to less “sad”, “surprised”, “fear” explanations than the ground-truth TripAdvisor dataset, which implies that the explanation generator could be unproportionally generating pleasant explanations to attract or even cheat users into the recommendations, and this hinders the trustworthiness of the explanations and the recommender system.

In order to address the robustness and emotion bias problem in explanation generation, in this paper, we take the emotion factor into account and propose a novel method based on a multi-head transformer, called Emotion-aware Transformer for Explainable Recommendation (EmoTER). EmoTER can use explicit contextual information on emotions to generate emotion-aware explanations with high text quality, explainability, and robustness. Specifically, we extract the contextual emotion information as a guiding signal and introduce an emotion-aware module to explicitly integrate the emotion information during the encoding process. Then we add an emotion recognition head in the decoding process that enables emotion constraints. To measure the linguistic quality and emotion fairness of the generated explanations, we adopt automatic text metrics and human perceptions for evaluation.

Our main contributions in this paper are:

- [Conceptual] We first address the emotion bias issues on three benchmark datasets and current state-of-the-art models for explainable recommendation and propose an emotion-aware generation framework to generate robust and fair explanations. To the best of our knowledge, this is the first work to robust, fair, emotion-aware explanation generation in explainable recommendations.
- [Theoretical] We propose EmoTER (Emotion-aware Transformer for Explainable Recommendation), a novel multi-head Transformer architecture, which enhances the emotion of the explanations with a multi-task learning approach for personalized explanation generation.
- [Empirical] Experiments on three widely-used benchmark datasets with multiple evaluation metrics demonstrate that EmoTER consistently and significantly outperforms existing state-of-the-art explanation generation models in terms of robustness of emotion distribution, text quality, and explainability. We also demonstrate that the proposed method leads to more equitable distribution of emotions in explainable, providing enhanced fairness based on disparate impact definition [41].

2 RELATED WORK

We begin by summarizing the related literature about explainable recommendation, emotional text generation and robustness of natural language generation in this section.

**Explainable recommendation.** Explainable recommendation has two major research perspectives: machine learning (ML) [5, 8, 44, 45] and human-computer interaction (HCI) [2, 38]. Our work incorporates both of these perspectives with a new model to generate emotion-aware explanations as well as a set of methods for evaluating such explanations using various metrics, including human judgments. From the ML perspective, the methods for providing natural language explanations, which are most related to this work, can be categorized into template-based and generation-based methods. Pre-defined templates [45] are not only simple and easy to use but also intuitive. However, designing templates is labor intensive and may limit the flexibility of the explanations. Recently, we have seen techniques involving LSTM [17] and Transformers [42] making great strides in natural language explanation generation. For instance, Chen et al. [7] proposed time-aware GRU to model dynamic user preferences. The Neural-Template (NETE) method [24] integrates both template and generation-based approaches to make the explanation generation process more controllable. PETER [26] is a Personalized Transformer to generate personalized explanation sentences. A big shortcoming of these works is that they primarily focused on the text quality [27, 33] and explainability [24, 26], leaving the emotional aspects of explanations under-explored, which motivates our work.

Generating and presenting explanations is not sufficient; we also need to assess if these explanations are effective. Explanations should serve to improve the transparency, persuasiveness, effectiveness, trustworthiness, efficiency, scrutability and user satisfaction [44]. There are several methods and metrics for assessing generated explanations. An ideal way of evaluating the explainability of machine generated explanations is through an online user study. For instance, Balog et al [2] measured recommendation explanation quality by collecting users’ judgments on seven pre-designed goals. But doing such evaluations can be expensive, time-consuming, and challenging at scale. Therefore, offline evaluation is often a more suitable solution for general research scenarios. The most commonly used metrics for evaluating machine generated explanation sentences are BLEU [33], METEOR [3] and ROUGE [27]. In this paper, we adopt both automatic metrics and human evaluation to access the emotion-aware explanations.

**Emotional Text Generation.** Although research works on generating emotion-aware explanations are scant, there are several advancements made in emotion-inclusive natural language generation. For instance, Keshtkar and Inkpen [19] used syntactic rules
to generate emotional sentences. Since the corpus used for training was relatively small, this technique was only able to generate simple emotional sentences. Colombo et al. [9] used a sequence-to-sequence approach and emotion vector representation, and proposed a model for generating words with different emotions. Ghosh et al. [14] came up with Affect-LM, which is a LSTM-based model and able to generate sentences in four emotion categories including positive, anxious, sadness, and anger.

Another area where emotional text generation has been very important is conversational agents as several recent works have investigated incorporating emotions into conversational agents (e.g., [1]). Emotional Chatting Machine [46] takes sequence-to-sequence framework to create emotion-rich information in the context of dialogue scenarios. It utilizes eight emotion categories such as anger, disgust, fear, happiness, like, sadness, surprise, and unknown. Other researchers [18] have adopted transfer learning to deal with emotional generation problems. They use pre-trained language models such as BERT [11] and GPT-2 [37], which have achieved reasonably high performance. However, these models are typically very large and take a long time to train.

In contrast, our proposed model is very light-weight, utilizes no prior syntactic knowledge and no pre-trained model, but is able to generate expressive emotional explanations.

Robustness in Natural Language Generation Robustness failures sometimes result from dataset biases introduced during data collection or human labeling, affecting model generalization and model performance. For example, Lewis et al. [23] showed that if there is a considerable overlapping on the test dataset in open-domain question answering, many QA models will not get memorized from training data and exhibit much less accuracy. In natural language inference, McCoy et al. [29] demonstrated that crowdsourced datasets used for training natural language inference models might introduce specific patterns more easily detected by statistical learning. Moreover, Bras et al. [4] proposed light-weight adversarial filtering to filter dataset bias.

Apart from finding the reasons for robustness failures, we should increase the robustness of the models with better use of minority examples. For example, Yaghoobzadeh et al. [43] proposed fine-tuning the model on full data and then on minority data. DRO (Distributitional Robust Optimization) [40] addressed the training strategy on particularly hard examples. There are lots of extensions for DRO, e.g., Nam et al. [31] emphasized the model’s early-stage decision-making behaviors to train another model. Lahoti et al. [22] adopted additional model to recognize examples. Liu et al. [28] proposed to weight minority examples that have high training loss.

Inspired by previous work on robustness, we propose an emotion-aware module to better use minority examples such as training examples belonging to some under-represented emotion categories. This will also enhance disparate impact view of fairness [41], which aims to provide representation of data proportional to the underlying distribution of those categories.

3 METHODOLOGY

In this section, we formulate the problem of the emotion-aware explanation generation, describe the input representation, and present the detailed design and implementation of EmoTER.

![Figure 2: Input representation of the explanation “beautiful lobby and nice bar” generated for the item hotel, with lobby serving as the feature of item hotel.](image)

### 3.1 Problem Formulation

We aim to propose a model $M$ to generate emotion-aware explanations for recommendations. We denote the generated recommendation explanation for the user $u$ and the item $i$ as $EXP_{u,i} = e_1, ... , e_{|E_{u,i}|}$, where $e_1, ..., e_{|E_{u,i}|}$ are the explanation’s word sequence and $E_{u,i}$ denotes the number of explanation words. Item features in the ground-truth datasets are represented by $T_{u,i}$. We take the user $u$ and item $i$ as the input $S = \{u, i, f_{0}, f_{1} \ldots f_{|T_{u,i}|}, f_{emo}, e_{1} \ldots e_{|E_{u,i}|}\}$, where $f_{0}, f_{1} \ldots f_{|T_{u,i}|}$ are topic features and $f_{emo}$ is emotion tag feature. $f_{u,i}$ represents the number of topic features. The output of context encoder would be $h_{context}$.

### 3.2 Input Representation

As shown in Figure 2, the input representation consists of three parts: emotion embedding, token embedding, and feature embedding. All the three embeddings are fed into the transformer encoder layers, which encode the context information of explanations with an embedding size of 512. We use positional embedding with a length of 512, which is the same as the embedding size.

#### 3.2.1 Emotion Embedding

We construct the word emotion embedding based on the emotion category and intensity extracted from NRC Emotion Lexicon [30], which is a word dictionary containing words and their corresponding emotion. When creating the word emotion embedding, we consider six most commonly used emotions including anger, fear, trust, surprise, sad, and happy. For example, the representation of word “lucky” is expressed as a 6-dimension vector $V_{NRC}$, which equals to $\{happy:0.721, angry:0, surprise:0.539, sad:0, fear:0, neutral:0\}$. If we cannot find a word in the NRC Emotion Lexicon, then we assign a vector $\{happy:0, angry:0, surprise:0, sad:0, fear:0, neutral:1.0\}$ to it.

Once we have constructed a 6-dimension vector $V_{NRC}$, we convert it into a 512-dimension emotion embedding $V_{emo}$ using Eq. 1.

$$V_{emo} = g(V_{NRC})$$

where $g(\cdot)$ is a multi-layer perceptron with an input of 6 and an output size of 512.

#### 3.2.2 Token Embedding

We use a simple tokenization method and treat a word as a token. The vocabulary used in our study consists of 20,000 most frequent tokens. There are 10 special tokens including four functional words: (bos) (begin of sentence), (eos) (end of sentence), (pad) (padding word), (unk) (unknown word) as well as six frequently used emotion category words to denote word emotions: (happy), (angry), (surprise), (sad), (fear), (neutral).
3.2.3 Feature Embedding. We integrate item features and emotion tags to construct feature embeddings, as shown in the orange box of Figure 2. The item feature \( f_{item} \) is self-contained in the dataset, while the emotion tag \( f_{emo} \) is assigned by an external emotion classification detector. In our study, we choose a widely-used emotion classifier named Text2Emotion [21] to create emotion tags.

3.3 Model Description

EmoTER adopts two encoders and a two-head decoder based Transformer architecture and takes as input a sequence in the embedding space of emotion embeddings, token embeddings, and feature embeddings, as shown in Figure 3. To be specific, one encoder takes emotion embeddings as input and creates the emotion hidden state, while the other encoder is responsible to process the concatenated feature embeddings and token embeddings to generate the context hidden state. Next, the two hidden states are fused by vector summation. We design a two-head decoder, i.e., the emotion head and the language modeling head, to project the last layer of the merged hidden state to two emotion and language modeling representations with desired dimensions. Accordingly, the final loss is expressed as a weighted sum of the two losses based on the emotion and language modeling constraints.

EmoTER is designed based on the framework of PETER [26] by adding one more encoder and two feed-forward linear heads on decoder. For the two encoders in EmoTER, we adopt a 2-layer architecture because: (1) it is fair to compare with PETER’s single 2-layer encoder [26]; (2) a 2-layer encoder is light-weight and easy to apply to different tasks. Each Transformer encoder layer is composed of a multi-headed self-attention and a position-wise feed-forward network. In the following, we introduce these two encoders and two different heads and their corresponding loss functions.

3.3.1 Emotion Encoder. The emotion encoder serves as an emotion-aware module in EmoTER to encode the emotion information when processing the input. In particular, the emotion-aware module guides EmoTER to pay more attention to the emotional words and strengthen the weights on them, which explicitly encodes the emotional information and enables emotion-guided decoding for explanation generation.

The input sequence of the user \( u \) and the item \( i \) to emotion transformer can be expressed as:

\[
S = [u, i, emo_{t_1}, \ldots, emo_{t_{F_u,i}}, emo_{emo}, emo_{t_1}, \ldots, emo_{e_{Emo}}]
\]  

where \( emo_{t_k} \) means the emotion embedding of \( i_k \) item feature. \( F_u,i \) represents the number of item features and \( E_{ emo } \) denotes the number of explanation words. We denote the hidden state of emotion encoder as \( hidden_{emo} \).

3.3.2 Context Encoder. Input sequence to context transformer can be expressed as:

\[
S = [u, i, f_{t_1}, \ldots, f_{t_{F_u,i}}, f_{emo}, e_1, \ldots, e_{Emo}]
\]  

where \( f_{t_1}, \ldots, f_{t_{F_u,i}} \) are topic features and \( f_{emo} \) is emotion tag feature. \( e_1, \ldots, e_{Emo} \) are the explanation’s word sequence. \( F_u,i \) represents the number of item features and \( E_{ emo } \) denotes the number of explanation words. The output of context encoder would be \( hidden_{context} \).

When the emotion hidden state \( hidden_{emo} \) and the context hidden state \( hidden_{context} \) are ready, we merge them into \( hidden_{merge} \) by:

\[
hidden_{merge} = intensity \times hidden_{emo} + hidden_{context}
\]  

where the coefficient \( intensity \) works as a modifier to adjust the strength of emotion.

3.3.3 Emotion Head. Emotion head is trained to perform the emotion-constrained classification task. We add this head to constrain the model to focus on the emotion information expressed in the context.

The prediction probability of target emotion is computed as:

\[
P(emo_1|emo_1, emo_2, \ldots, emo_t) = softmax(hidden_{merge}[-1] \times M_1)
\]  

where \( emo_1 \) is the predicted emotion category \( i \), and \( emo_1, \ldots, emo_t \) are all emotion categories. \( hidden_{merge}[-1] \) is the last hidden layer of our model and \( M_1 \) is the weight to be learned during the training for the emotion classification task. Then we have the following cross-entropy loss:

\[
L_{emo}(E) = - \sum_{t=1}^{N} \log(P(emo|emo_1, emo_2, \ldots, emo_t))
\]  

3.3.4 Language Modeling Head. We add a 2-layer transformer decoder, which predicts the next token based on the merged context vector \( hidden_{merge} \). We define the explanation as \( E = e_1, e_2, \ldots, e_n \), then the conditional probability of the next token is computed as:

\[
P(e_t|e_1, e_2, \ldots, e_{t-1}) = softmax(hidden_{merge}[-1] \times M_2)
\]  

where \( hidden_{merge}[-1] \) is the last hidden layer and \( M_2 \) is the token embedding matrix. During training, \( M_2 \) updates its weights. We
Table 2: Statistics of the three datasets.

|               | TripAdvisor | Yelp       | Amazon     |
|---------------|-------------|------------|------------|
| #user         | 9765        | 27147      | 7506       |
| #item         | 6280        | 20266      | 7360       |
| #features     | 5069        | 7340       | 5399       |
| #records      | 320023      | 1293247    | 441783     |
| #records/user | 32.77       | 47.64      | 58.86      |
| #records/item | 50.96       | 63.81      | 60.02      |
| #word/explanation | 13.01 | 12.32 | 14.14 |

also have the following cross-entropy loss:

\[ L_{lm}(E) = -\sum_{i=1}^{N} \log(P(e_i|e_1, e_2, ..., e_{i-1})) \] (8)

where \( N \) is the word count in the generated explanation.

3.3.5 Multi-task Learning. Finally, we integrate the two tasks into a multi-task learning framework with the following loss function:

\[ L_{total} = c_1L_{lm} + c_2L_{emo} \] (9)

where \( c_1 \) and \( c_2 \) are two hyper-parameters to control the weights of emotion loss and language modeling loss.

4 EXPERIMENT SETUP

In this section, we introduce the datasets and metrics for evaluation. The baselines and model training are also presented.

4.1 Datasets

For experimentation, we adopted three publicly available explainable recommendation datasets [24, 25, 32]: TripAdvisor (hotel), Yelp (restaurant) and Amazon (movies, TV). Details of the datasets are shown in Table 2. Each dataset is randomly split into training, validation, and testing sets with a ratio of 8:1:1 for 5 times. Each user and each item contains at least one record in the training set. Each record is comprised of a user ID \( (u) \), an item ID \( (i) \), an explanation \( (Exp_u(i)) \), and at least one item feature \( (F_{item}) \). The explanations are extracted from user reviews [25].

4.2 Evaluation Metrics

We evaluate our model’s ability of generating explanations with expressive emotion, text quality, and efficacy of explainability on the test set. Overall, evaluations of the explanation performance could be divided into two parts: automatic metrics and human evaluations.

4.2.1 Automatic Evaluations. We use three various automatic evaluation methods to evaluate the emotion fairness, text quality, and explainability of explanations generated by our model EmoTER. Explanations with emotion fairness and robustness As reported earlier, we found there is huge bias about the emotion distribution both on three benchmark datasets and explanations generated by state-of-the-art models. Here, we equate this bias to lack of fairness, where fairness is considered to be based on disparate impact [41]. Our goal is to reduce this bias and increase fairness by having our explanations exhibit emotions that match the underlying distributions in ground truth. We use the same emotion classifiers to measure the emotion distribution of explanations generated by our model EmoTER for fair comparison. Text2Emotion [21] is a widely-used emotion detection API working on the emotional words of text. While GoEmotions [10] is the largest manually labeled dataset on English Reddit comments with high quality of 28 emotion categories including Neutral. Its BERT-based model is one of the state-of-the-art emotion recognition models. In order to show the robustness and fairness effect of our emotion-aware module, we chose these two very different models as our emotion classifiers to identify the emotion distribution of the explanations from the ground-truth dataset, generated by PETER, and generated by our model, respectively.

Text quality To assess the text quality of EmoTER-generated explanations and compare them with state-of-the-art models, we follow the widely-used automatic evaluation metrics. We adopt BLEU [33] and ROUGE [27], which are commonly used in machine translation and in text summarization respectively. Since BLEU and ROUGE often fail to detect the problem of always generating simple and repetitive sentences such as "I recommend it," we adopt USR (Unique Sentence Ratio) [24] to measure the diversity of the generated explanations.

\[ USR = \frac{\text{Set}(N)_{\text{num}}}{N_{\text{num}}} \] (10)

where \( N \) is the collection of generated explanations and \( \text{Set}(N) \) is the collection of unique explanations. We use the similarity between explanations using exact matches.

Explainability To assess the explainability of EmoTER-generated explanations, we use metrics that are widely used in previous work. We adopt three metrics related to features in the datasets. Since people care about specific features of the recommended items [6], we use three feature-based metrics: Feature Matching Ratio (FMR), Feature Coverage Ratio (FCR) and Feature Diversity (DIV), which are proposed in [24].

\[ FMR = \frac{N_f^{\text{num}}}{G_f^{\text{num}}} \] (11)
We considered three state-of-the-art explanation generation models where $F$ is ground-truth explanation with feature and $N$ is generated explanations with feature. FMR measures generated explanations in terms of feature level.

$$\text{FCR} = \frac{F_G^N}{F_G'^N}$$

where $F_G$ is the distinct features in the ground-truth and $F_N$ is the distinct features in the generated explanations.

$$\text{DIV} = F_{u,i} \cap F'_{u',i'}$$

where $F_{u,i}$ and $F'_{u',i'}$ are two sets of distinct features discussed in explanations of $u,i$ and $u',i'$ respectively. DIV measures the feature intersections between various $u,i$ pairs. Since different users may not always talk about the same feature about items, a lower DIV means better performance.

4.2.2 Human Evaluations. We conduct two different user studies to measure emotion fairness/robustness and quality of explanations from humans’ perspectives.

Explanations with emotion fairness and robustness We measure EmoTER’s ability to generate fair and robust emotional explanations by conducting an extensive user study on Amazon’s Mechanical Turk (MTurk) platform. As figure 4(a) shows, we designed a survey to ask annotators to choose which explanation is more emotionally close to the ground-truth explanation. We presented a set of explanations generated using a baseline method PETER and the corresponding explanations generated using EmoTER in a randomly-ordered pair. The Turk workers did not know which explanation was emotion-aware. We used three sets of 50 such explanation pairs from three datasets each to create a HIT (Human Intelligence Task) on MTurk. Each HIT was done by three different Turk workers. Providing ratings on 50 items in this way took about 30 minutes. Each was paid 4.00 USD. This experiment was authorized by our Institutional Review Board (IRB) after reviewing it for ethical and privacy considerations. Each task was conducted by human raters that were located in the United States and had high approval rates.

4.3 Baselines

We considered three state-of-the-art explanation generation models as baselines in our experiments.

- NETE [24] integrates template-based and generation-based approaches to make the explanation generation process more controllable.

Table 3: Emotion distribution of explanations evaluated by Text2Emotion TripAdvisor dataset, PETER and our model EmoTER.

|        | Ground Truth | PETER | EmoTER | Debiasing |
|--------|--------------|-------|--------|-----------|
| happy  | 42.8%        | 61.3% | 47.2%  | 21.9%     |
| angry  | 5.9%         | 5.4%  | 4.8%   | -10.2%    |
| surprise | 12.7%       | 5.0%  | 9.8%   | 37.8%     |
| sad    | 7.7%         | 4.6%  | 5.8%   | 10.4%     |
| fear   | 18.7%        | 11.7% | 17.4%  | 30.5%     |
| neutral| 12.2%        | 11.8% | 14.9%  | 25.4%     |

Table 4: Emotion distribution of explanations evaluated by GoEmotion on TripAdvisor dataset, PETER and our model EmoTER.

|        | Ground Truth | PETER | EmoTER | Debiasing |
|--------|--------------|-------|--------|-----------|
| happy  | 59.2%        | 69.5% | 63.6%  | 10.0%     |
| angry  | 1.5%         | 0.3%  | 0.5%   | 13.3%     |
| surprise | 1.6%        | 0.3%  | 0.4%   | 6.3%      |
| sad    | 1.8%         | 0.6%  | 1.3%   | 38.9%     |
| fear   | 0.3%         | 0.05% | 0.06%  | 3.3%      |
| Neutral| 35.6%        | 29.2% | 34.2%  | 14.0%     |

4.4 Model Training

When training models, we split one dataset into three parts – the training set to train models, the validation set to fine-tune hyperparameters, and the test set to measure the performance. The reported results are averaged on five random data splittings. Since NETE and PETER are open-sourced, we reused the published codes and followed their default settings in training and testing.

When training the proposed EmoTER, the word embedding size $d$ was set to 512, the dimension of the feed-forward network was set to 2048, and both the number of layers and attention heads of the decoder block were set as 2. The hyperparameter of language modeling task $c_1$ and the hyperparameter of emotion task $c_2$ were set to 1.0 and 1.0 respectively after performing grid search. We set the batch size to 128 and the learning rate to 1.0. The optimizer was SGD (stochastic gradient descent), and gradient clipping was adopted with a threshold of 1.0.

5 RESULT AND ANALYSIS

This section presents automatic and human evaluation results of EmoTER’s performance.

5.1 Automatic Evaluation on Explanation

We compared EmoTER with existing state-of-the-art explanation generation models from the perspectives of the robustness of emotion distribution, explainability, and text quality.

Explanations with emotion fairness and robustness To illustrate that EmoTER is capable of reducing emotion biases in recommendation explanations, we qualitatively evaluated its debiasing performance by comparing emotion distributions in explanations generated by PETER and by the proposed EmoTER on the TripAdvisor, Amazon, and Yelp datasets. We adopted two widely-used emotion detection models to recognize the emotion of generated explanations: (1) Text2emotion [32], a lightweight Python library;
Table 5: Performance comparison on three datasets TripAdvisor, Yelp and Amazon. R1 and R2 denote ROUGE-1 and ROUGE-F. P, R, F denote Precision, Recall and F1. The best-performing values are boldfaced. The second-performing values are underlined. BLEU and ROUGE are percentage values.

| Dataset | Model | Explainability | Text quality |
|---------|-------|----------------|--------------|
|         |       | FMR | FCR | DIV | USR | BLEU-1 | BLEU-4 | R1-P | R1-R | R1-F | R2-P | R2-R | R2-F |
| TripAdvisor | ACMLM | 0.07 | 0.41 | **0.78** | **0.94** | 3.45 | 0.02 | 4.86 | 3.82 | 3.72 | 0.18 | 0.20 | 0.16 |
|           | NETE  | 0.79 | 0.27 | 2.23 | 0.56 | 22.41 | 3.60 | 35.70 | 24.68 | 27.70 | 10.23 | 6.98 | 6.55 |
|           | PETER | **0.89** | 0.36 | 1.59 | 0.25 | 24.19 | 4.54 | 37.57 | 29.15 | **30.49** | 11.99 | **8.93** | **9.27** |
|           | EmoTER | 0.87 | 0.42 | 1.52 | 0.39 | **25.14** | **4.89** | **39.29** | **30.04** | **31.76** | **13.12** | **9.64** | **10.08** |
| Yelp      | ACMLM | 0.05 | 0.31 | 1.48 | 0.52 | 19.34 | 2.63 | 33.95 | 22.31 | 25.65 | 8.88 | 5.45 | 6.67 |
|           | NETE  | 0.81 | 0.29 | 1.48 | 0.52 | 19.34 | 2.63 | 33.95 | 22.31 | 25.65 | 8.88 | 5.45 | 6.67 |
|           | PETER | **0.86** | 0.36 | 1.07 | 0.30 | 20.55 | 3.40 | 35.63 | 29.15 | 27.91 | 11.99 | 8.93 | 9.27 |
|           | EmoTER | 0.87 | 0.38 | 1.00 | 0.41 | **21.61** | **3.84** | **37.41** | **27.16** | **29.41** | **11.90** | **8.17** | **9.28** |
| Amazon    | ACMLM | 0.10 | 0.31 | 2.07 | **0.96** | 9.52 | 0.22 | 11.65 | 10.39 | 9.69 | 0.71 | 0.81 | 0.64 |
|           | NETE  | 0.69 | 0.19 | 1.95 | 0.57 | 18.82 | 2.47 | 33.78 | 21.31 | 24.75 | 7.64 | 4.81 | 5.52 |
|           | PETER | **0.77** | 0.26 | 1.25 | 0.39 | 19.58 | 3.00 | 35.23 | 23.84 | 26.44 | 8.99 | 6.18 | 6.63 |
|           | EmoTER | 0.76 | 0.26 | 1.22 | 0.46 | **20.20** | **3.21** | **35.56** | **24.47** | **27.06** | **9.32** | **6.57** | **7.03** |

Table 6: Ablation study of multitasks settings performed on the TripAdvisor dataset. ↑ and ↓ indicate performance changes compared to EmoTER. BLEU and ROUGE are percentage values.

| Dataset | Model | Explainability | Text quality |
|---------|-------|----------------|--------------|
|         |       | FMR | FCR | DIV | USR | BLEU-1 | BLEU-4 | R1-P | R1-R | R1-F | R2-P | R2-R | R2-F |
| TripAdvisor | Disable $L_{emo}$ | 0.87 | 0.35 | 1.53 | 0.30 | 24.62 | 4.70 | 39.41 | 29.62 | 31.56 | 12.98 | 9.41 | 9.89 |
|           | Disable $L_{lm}$ | 0.88↑ | 0.31↑ | 1.64↑ | 0.30 | 25.06↑ | 4.76↑ | 39.15↑ | 29.87↑ | 31.58↑ | 12.71↑ | 9.44↑ | 9.78↑ |
|           | EmoTER | 0.87 | 0.39 | 1.53 | 0.42 | 25.03 | 4.80 | 39.35 | 29.95 | 31.77 | 13.04 | 9.56 | 10.03 |

Table 7: Ablation study of intensity performed on the TripAdvisor dataset. ↑ and ↓ indicate performance change compared to EmoTER. BLEU and ROUGE are percentage values.

| Dataset | Model | Explainability | Text quality |
|---------|-------|----------------|--------------|
|         |       | FMR | FCR | DIV | USR | BLEU-1 | BLEU-4 | R1-P | R1-R | R1-F | R2-P | R2-R | R2-F |
| TripAdvisor | intensity=0.5 | 0.87 | 0.39 | 1.57↓ | 0.43↑ | 25.72↑ | 4.90↑ | 38.56↑ | 30.32↑ | 31.68↑ | 12.51↑ | 9.64↑ | 9.87↑ |
|           | intensity=2 | 0.86↑ | 0.39↑ | 1.53↑ | 0.33↑ | 25.30↑ | 4.83↑ | 39.00↑ | 30.00↑ | 31.65↑ | 12.69↑ | 9.52↑ | 9.87↑ |
|           | EmoTER (intensity=1) | 0.87 | 0.39 | 1.53 | 0.42 | 25.03 | 4.80 | 39.35 | 29.95 | 31.77 | 13.04 | 9.56 | 10.03 |

Explainability In terms of explainability, EmoTER is very competitive and close to the best performing values regarding the FMR metric. EmoTER beats the other three state-of-the-art models by a large margin on FCR metric which considers whether explanations contain all the features in the ground-truth. EmoTER achieves much better performance on explainability even compared to pre-trained models (e.g., ACMLM) that learn large-scale external knowledge. Lower DIV means higher diversity. It is not surprising that ACMLM has the highest diversity. However, EmoTER also achieves a competitive score.

Text quality We also compared the text quality of generated text by EmoTER with three state-of-the-art baselines on three datasets (see the text quality columns in Table 5 for details). We used unique
We found that the explanations maintain high performance of explainability and text quality when intensity was small. If emotion intensity becomes too strong, it may cause performance of USR to drop nearly 50%. These two experiments confirm EmoTER’s effectiveness in generating emotional and high-quality explanations. In addition, we changed the intensity strength to generate emotional and diverse explanations to enhance the user engagement with various informative recommendations.

Explanations with emotion fairness and robustness

As an additional checkpoint and evaluation, we conducted an ablation study on the TripAdvisor dataset to show the effectiveness of our multi-task design shown in Table 6 and hyper-parameter of intensity shown in Table 7. First, we launched an experiment without emotion classification loss by setting $c_1$ in Equation 9 to be 0. We found that disabling the emotion classification task causes USR to drop by nearly 25%, and performances of almost all text quality metrics decrease sharply. Similarly, we disabled language modeling loss by setting $c_2$ to be 0. It not only degraded the explainability performance but also caused USR to drop nearly 50%. These two experiments confirm EmoTER’s effectiveness in generating emotional and high-quality explanations. In addition, we changed the intensity strength to generate emotional explanations with different emotion intensities shown in Table 7. We found that the explanations maintain high performance of explainability and text quality when intensity was small. If emotion intensity becomes too strong, it may cause performance of USR and DIV to drop dramatically.

5.2 Human Evaluation on Explanations

In addition to automatic metrics, we also evaluate human perceptions of emotion-aware explanations. We conducted one Amazon MTurk user study to validate that our model EmoTER increases the robustness and fairness of emotion-aware explanation.

Table 8: Qualitative examples from three datasets

|          | TripAdvisor | Yelp          | Amazon        |
|----------|--------------|---------------|---------------|
| Ground Truth | Beautiful lobby and nice bar. | The wait is long for food. | Don’t waste your time on this loser. |
| PETER     | The lobby was very nice and the rooms were very comfortable | If you are a fan of the original, you’ll love this film. |
| EmoTER    | The lobby is impressive and the rooms are spacious | Don’t worth the wait. | If you are a fan of the films, you will be disappointed. |

Table 5 shows that EmoTER generates explanations with the same negative emotion as ground-truth while PETER generates more diverse and unique vocabularies. An example of Yelp dataset shows that EmoTER generates explanations with the same negative emotion as ground-truth while PETER generates explanations as it is a template-like approach. ACMLM achieves the highest USR because it is a well-tuned model trained on a large-scale dataset. We also use BLUE-1, BLUE-4, ROUGE-1 (R1 column in Table 5), and ROUGE (R2 column in Table 5) to evaluate text quality. EmoTER consistently outperforms baselines by a large margin on BLEU and ROUGE metrics, verifying our model’s ability to generate high-quality explanations.

Figure 5 shows the human evaluation results of explanations with emotion fairness and robustness. More users think explanations generated by our model EmoTER are emotionally closer to the ground-truth than explanations generated by the current state-of-the-art model PETER. We can see that our model EmoTER consistently outperforms PETER in three benchmark datasets in terms of robustness and fairness. We perform student t-test among the human evaluation result and obtain significant result ($t = 4.90, p = 0.008$).

6 Conclusion

Providing explanations that are appropriately emotion-induced could help users of recommender systems understand and trust the recommendations better, but existing approaches have failed to incorporate a wide range of emotions in unbiased and naturalistic manner. In this paper, we proposed an emotion-aware generation framework to generate robust and fair explanations. To the best of our knowledge, this is the first work to explore robust, fair, emotion-aware explanation generation in explainable recommendations. We presented EmoTER (Emotion-aware Transformer for Explainable Recommendation), a novel multi-head Transformer architecture, which enhances the emotion of the explanations with a multi-task learning approach for personalized explanation generation. Experiments on three widely-used benchmark datasets with multiple evaluation metrics demonstrated that EmoTER consistently outperforms existing state-of-the-art explanation generation models in terms of robustness of emotion distribution, text quality, and explainability.

The next steps in this research will involve paying attention to the possible dataset biases and further exploring fairer methods to generate explanations for recommendation systems. Moreover, we will extend our framework to emotion-aware multi-modal explanations and develop our solution for explainable conversational recommendations, which significantly needs the emotional elements to enhance the user engagement with various informational systems, including conversational agents.

Figure 5: Human evaluation result about emotion robustness on generated explanations on three benchmark datasets. The t-test result demonstrated that EmoTER outperformed PETER significantly in generating robust explanations ($p < 0.05$).
