Narrative2Music: Generating Emotion-Aligned Music for Sentences

Mohammad Shokri[1], Alexandra C. Salem[1], Gabriel Levine[1], Johanna Devaney[1,2], Sarah Ita Levitan[1,3]
[1]CUNY Graduate Center, [2]Brooklyn College, [3]Hunter College

INTRODUCTION

- We introduce **Narrative2MIDI**, a sequence-to-sequence **Transformer-based model for generating emotion-aligned music from a piece of text**
- We also created Narrative2MIDI, a **dataset** of emotion-aligned music for narratives

DATASETS

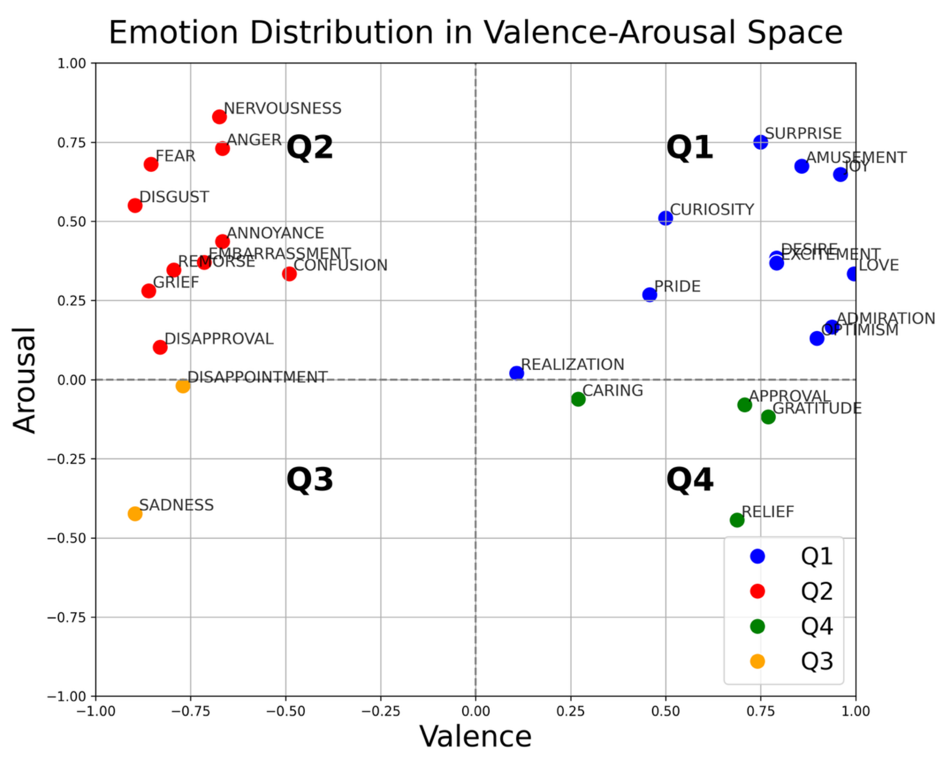
Narrative2MIDI dataset

- GoEmotionsAV
  - GoEmotions [1] is 58k English Reddit comments with 27 emotion categories or neutral
  - Recoded categories into Arousal-Valence (AV) plane (Fig. 1)
- EMOPIA [2]
  - 1,087 MIDI files with Arousal-Valence quadrants
- End up with **1,087 <Narrative, MIDI file> pairs**

GiantMIDI-Piano dataset [3]

- 10k classical MIDI files for piano, used for pre-training

Figure 1. GoEmotions Mapped to Arousal-Valence Space



TRANSFORMER-BASED MODEL

We use an **encoder-decoder Transformer** [4], shown in Figure 2

- Encoder
  - **Contrastive training** to fine-tune **RoBERTa-large** to represent the narratives
  - 500 epochs training on final four layers
- Decoder
  - **Pre-training** on **GiantMIDI-Piano**
    - 250 epochs training
  - **Fine-tuned** on **Narrative2MIDI**
    - 300 epochs training on final layer

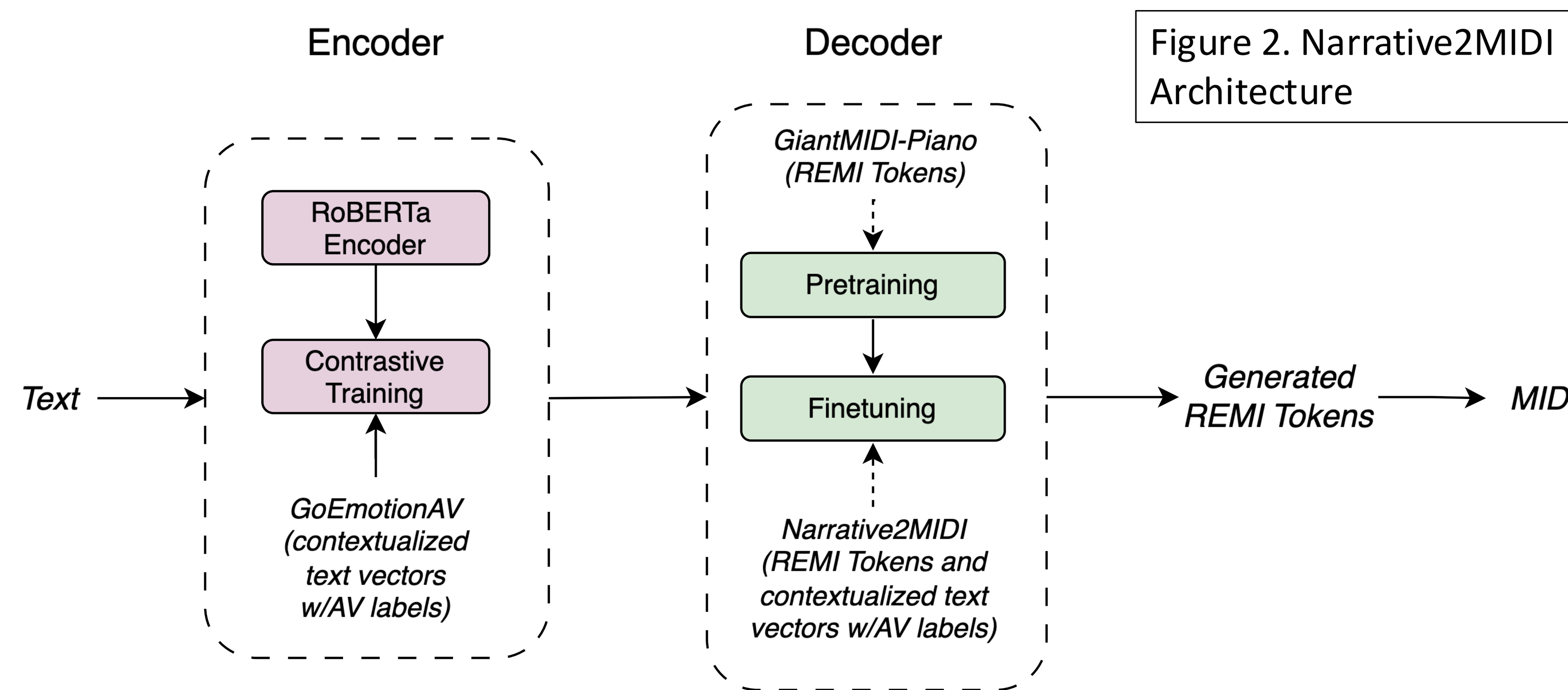


Figure 2. Narrative2MIDI Architecture

EVALUATION & RESULTS

Objective (Figure 3)

- **Valence**-Related Metric: **Major Key Ratio**
  - Prior work: **major key = positive affect (Q1, Q4), minor key = negative (Q2, Q3)**
  - Calculated key for each generated file using the Krumhansl-Kessler algorithm
  - Calculated the ratio of major key to minor key in each quadrant
  - Independent-samples t-test for Q1, Q4 vs Q2, Q3 were **significant** ($p = 0.026$)
- **Arousal**-related Metrics: **Average Note Length**
  - Expected to be **higher in low arousal (Q3, Q4)**
  - Independent-samples t-test for Q3, Q4 vs Q1, Q1 were **significant** ($p < 0.01$)
- **Arousal**-related Metrics: **Average Note Velocity**
  - Expected to be **higher in high arousal (Q1, Q2)**
  - Independent-samples t-test between Q3, Q4 vs Q1, Q1 were **not significant**

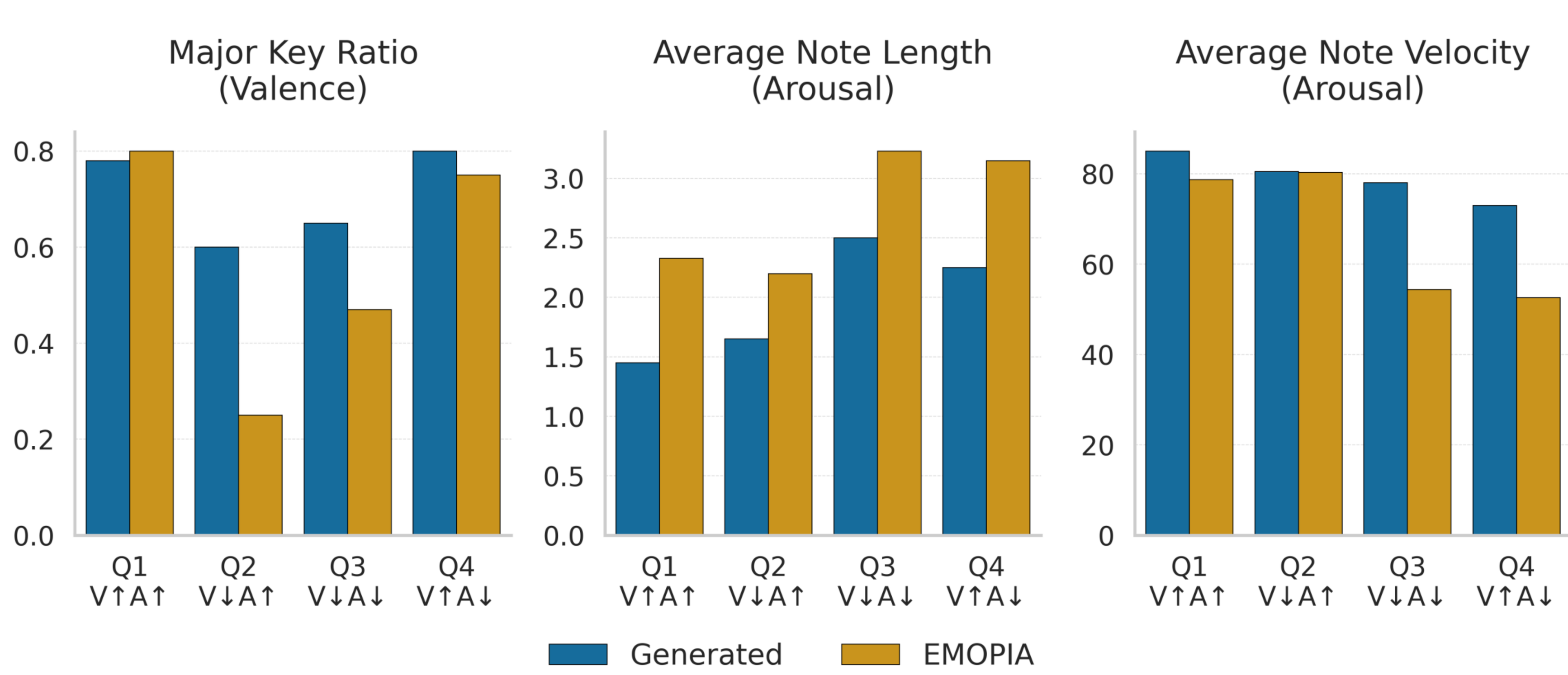


Figure 3. Valence and Arousal Metrics of Generated Files and EMOPIA

Subjective (Table 1)

- We conducted a **preliminary listening study** with 3 participants
  - Participant was given **two generated music clips**
    - One clip was generated for that narrative with Narrative2MIDI
    - One decoy clip
  - Participant chose **which clip matches the narrative**

Dimension	Accuracy
Valence	0.53
Arousal	0.70
Valence+Arousal	0.40

Table 1. Preliminary Listening Study Results

CONCLUSIONS

- **Objective evaluation** of the generated music showed that it **matched the note length** characteristics of EMOPIA more than the valence characteristics or note velocity.
- A small-scale qualitative evaluation confirmed that the model is **better at capturing arousal** than valence
- Future work will include an expanded listening study

KEY REFERENCES

[1] Dorottya Demszky et al., “Goemotions: A dataset of fine-grained emotions,” arXiv preprint arXiv:2005.00547, 2020.

[2] Hsiao-Tzu Hung et al, “Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation,” arXiv preprint arXiv:2108.01374, 2021

[3] Qiuqiang Kong et al., “Giantmidi-piano: A large-scale midi dataset for classical piano music,” arXiv preprint arXiv:2010.07061, 2020.

[4] Ashish Vaswani et al., “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) award 2228910.