
بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه هوش مصنوعی

گزارش پروژه کارشناسی

رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان پروژه :

طراحی و پیاده‌سازی یک دستیار هوشمند تولید موسیقی بر پایه تحلیل احساسات کاربر

استاد راهنما:

دکتر حسین ماهوش محمدی

پژوهشگر:

محمد امین کیانی ۴۰۰۳۶۱۳۰۵۲

شهریور ۱۴۰۴

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه هوش مصنوعی

پروژه کارشناسی رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی

آقای محمد امین کیانی

تحت عنوان

طراحی و پیاده‌سازی یک دستیار هوشمند تولید موسیقی بر پایه تحلیل احساسات کاربر

در تاریخ **/۰۶/ ۱۴۰۴** توسط هیأت داوران زیر بررسی و با نمره به تصویب نهایی رسید.

۱- استاد راهنمای پروژه:

دکتر حسین ماهوش محمدی امضا

۲- استاد داور :

دکتر امضا

امضای مدیر گروه

تشکر و قدردانی

در این مسیر پرچالش و شگفت‌انگیز که به طراحی و پیاده‌سازی یک دستیار هوشمند تولید موسیقی بر پایه تحلیل احساسات کاربر انجامید، بر خود واجب می‌دانم تا از افرادی که در این سفر همراه من بودند، صمیمانه قدردانی کنم.

نخست و پیش از هر چیز، از خانواده‌ی عزیزم بخصوص مادرم، سپاسگزارم. او نه تنها چراغ راه من در دنیای موسیقی بوده، بلکه با عشق و حمایت‌های بی‌دریغش، به من آموخت که چگونه به صدای درونم گوش فرا دهم و احساساتم را به زبان موسیقی ترجمه کنم. مادرم، تو با هر نغمه و هر آوایی که در خانه طنین‌انداز بود، درختی از عشق و خلاقیت در وجودم کاشتی. بی‌تردید، بدون حمایت‌های تو، این مسیر برایم آن‌چنان روشن نمی‌شد.

همچنین از دکتر حسین ماهوش محمدی، استاد گرامی‌ام، که با علم و دانش خود، چراغ راهنمای من در این پروژه بودند، صمیمانه سپاسگزارم. درس‌هایی که تحت نظر شما فرا گرفتم، نه تنها به من ابزارهای لازم را برای پیشبرد این پروژه عطا کرد، بلکه انگیزه‌ای مضاعف برای ادامه‌ی راه به من بخشید. راهنمایی‌ها و مشاوره‌های شما در طول این مسیر، همچون ستاره‌ای در آسمان شب، برایم روشنایی بخشید و مرا به سمت هدفم هدایت کرد.

در نهایت، از دنیای موسیقی و سازهایی که به من آموختند تا احساساتم را به تصویر بکشم، تشکر ویژه‌ای دارم. هر نت و هر آکوردی که نواخته‌ام، بخشی از وجودم را به تصویر کشیده و مرا به جایی رسانده که امروز ایستاده‌ام. این سفر نه تنها یک پروژه علمی، بلکه سفری عمیق به درون خودم و کشف احساساتی بود که با نواختن سازهایم توانستم آن‌ها را ابراز کنم.

به امید آنکه این دستیار هوشمند بتواند به دیگران نیز کمک کند تا با موسیقی احساسات خود را بیان کنند و دنیای زیبای هنر را تجربه نمایند.

تقدیم به

تمامی دوستداران موسیقی

چکیده:

پروژه‌ی حاضر با هدف طراحی و پیاده‌سازی یک سیستم تولید موسیقی مبتنی بر تحلیل احساسات چندرسانه‌ای ارائه شده‌است. در این سیستم، ورودی می‌تواند متن، صدا یا داده‌های تصویری باشد که پس از پردازش و تحلیل احساسات، تبدیل به یک توصیف ساختاری از حالت عاطفی کاربر می‌شود. خروجی این تحلیل به صورت مجموعه‌ای از پارامترهای موسیقایی از جمله تمپو، مد (ماژور یا مینور)، پیشرفت آکوردها (Progression)، میزان هم‌زمانی (Syncopation)، چگالی ملودی (Density) و میزان هارمونی / ناسازگاری (Dissonance) استخراج شده و به بخش تولید موسیقی ارسال می‌گردد.

مدل تولید موسیقی طراحی‌شده در این پروژه، برخلاف رویکردهای آماده مانند MusicGen، از ابتدا توسعه داده شده و مبتنی بر مدل‌سازی توالی (Autoregressive) با معماری ساده اما قابل توسعه است. موسیقی تولیدی نه تنها به لحاظ پارامترهای ساختاری کنترل‌پذیر است، بلکه با سبک‌های مختلف (نظیر سبک «Persian» و سبک «Global») قابل اجرا می‌باشد. برای این منظور از داده‌های MIDI و فایل‌های SoundFont جهت رندر صدا استفاده شده است.

فرآیند ارزیابی شامل مقایسه احساس هدف (Target Emotion) با احساس شناسایی‌شده از خروجی صوتی (Predicted Emotion Audio) بوده و معیارهایی همچون دقت در نگاشت احساس، امتیاز تطابق صوتی، و شاخص‌های ساختاری موسیقایی بررسی شده‌اند. نتایج نشان می‌دهد که مدل توانسته است ارتباط معناداری بین احساسات انسانی و ویژگی‌های موسیقایی ایجاد کند، هرچند در برخی موارد انحراف‌هایی میان احساس پیش‌بینی‌شده و هدف مشاهده شده است. این مسئله مسیر پژوهش‌های آینده برای بهبود مدل و ترکیب داده‌های چندوجهی را هموار می‌سازد.

از نظر کاربردی، این پروژه می‌تواند به عنوان پایه‌ای برای توسعه سیستم‌های موسیقی‌درمانی، پیشنهاد موسیقی شخصی‌سازی‌شده، و همچنین تولید خودکار موسیقی در صنایع بازی و فیلم مورد استفاده قرار گیرد.

واژگان کلیدی: دانشگاه اصفهان، دانشکده‌ی کامپیوتر، گرایش هوش مصنوعی، پروژه کارشناسی، تولید موسیقی مبتنی بر احساسات، پردازش زبان طبیعی (NLP)، موسیقی‌درمانی، سیستم‌های چندرسانه‌ای

فهرست مطالب

عنوان **صفحه**

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

مخفف‌ها:

NLP	Natural Language Processing
BPM	Beats Per Minute
MIDI	Musical Instrument Digital Interface
CNN	Convolutional Neural Network
FER	Facial Emotion Recognition
SF2	SoundFont File
JSON	JavaScript Object Notation
BERT	Bidirectional Encoder Representations from Transformers
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
SMF	Standard MIDI File
DAG	Directed Acyclic Graph
ML	Machine learning
MFCC	Mel-Frequency Cepstrum
WAV	Waveform Audio File Format
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
CLAP	Contrastive Language-Audio Pretraining
FAD	Fr´echet Audio Distance
KAD	Kernel Audio Distance
MOS	Mean Opinion Score
MMD	Maximum Mean Discrepancy
AI	Artificial Intelligence
FER	Facial Emotion Recognition
LBP	Local Binary Patterns
ORB	Orientated FAST and Rotated BRIEF
CNN	Convolutional Neural Network
ASR	Automatic Speech Recognition
IEMOCAP	Interactive Emotional Dyadic Motion Capture Database

فصل اول
مقدمه

امروزه موسیقی نه تنها به عنوان یک هنر، بلکه به عنوان ابزاری علمی و فناورانه در حوزه‌های مختلف زندگی بشر نقش پررنگی ایفا می‌کند. از موسیقی‌درمانی در پزشکی گرفته تا ایجاد تجربه‌های شخصی‌سازی‌شده در شبکه‌های اجتماعی و سرویس‌های پخش آنلاین، موسیقی به‌عنوان یکی از اصلی‌ترین راه‌های ارتباطی بین احساسات انسان و فناوری شناخته می‌شود. در این میان، هوش مصنوعی و به‌ویژه پردازش زبان طبیعی (NLP[1]) توانسته‌اند پلی میان دنیای زبان و موسیقی ایجاد کنند.

پروژه حاضر با هدف تحلیل احساسات کاربر از ورودی‌های متنی، صوتی و تصویری و سپس تولید موسیقی متناسب با آن احساسات، طراحی و پیاده‌سازی شده است. بدین ترتیب کاربر می‌تواند متنی درباره حال و هوای خود وارد کند، یا حتی تنها با صحبت کردن و یا بارگذاری یک تصویر چهره، سیستم به‌صورت خودکار احساسات را تحلیل کرده و موسیقی هماهنگ با آن تولید نماید.

یکی از مشکلات اساسی در حوزه تعامل انسان و رایانه، ناتوانی سیستم‌ها در درک عمیق احساسات انسانی است. اغلب سامانه‌های موجود تنها به پیشنهاد موسیقی‌های آماده بسنده می‌کنند(مانند الگوریتم‌های توصیه‌گر Spotify یا YouTube Music) و توانایی خلق موسیقی اختصاصی بر اساس حالات کاربر را ندارند. پروژه حاضر با تمرکز بر این خلأ، مدلی اختصاصی طراحی می‌کند که نه‌تنها قادر به شناسایی احساسات پیچیده (مانند شادی، غم، ترس، خشم یا تعجب) است، بلکه توانایی تولید موسیقی منحصربه‌فرد متناسب با احساس کاربر را نیز داراست.

این پروژه علاوه بر جنبه‌های علمی و پژوهشی، از نظر کاربردی نیز ارزش بالایی دارد؛ زیرا می‌تواند در زمینه‌های مختلفی از جمله موسیقی‌درمانی، بازی‌های رایانه‌ای، سرویس‌های استریم، هنرهای تعاملی و حتی مراقبت‌های سلامت روان مورد استفاده قرار گیرد. همچنین با توجه به ترکیب چند حوزه مهم از جمله NLP، پردازش صوت، پردازش تصویر و تولید موسیقی خودکار، این پروژه می‌تواند به‌عنوان یک نمونه میان‌رشته‌ای

[1] Natural Language Processing

قدرتمند مطرح گردد.

۱-۱- هدف پروژه

هدف اصلی پروژه، طراحی و پیاده‌سازی یک سامانه هوشمند است که بتواند:

- تحلیل احساسات کاربر را از ورودی‌های مختلف (متن، صدا و تصویر) انجام دهد.
- خروجی تحلیل احساسات را به‌صورت ساخت‌یافته (JSON[1]) ارائه کند.
- بر اساس احساسات شناسایی‌شده، موسیقی متناسب تولید نماید.
- از مدل‌های اختصاصی و سبک(قابل اجرا روی پلتفرم‌هایی مانند Google Colab) استفاده کند.
- بستری قابل تعامل(رابط کاربری ساده با Gradio) برای کاربران فراهم آورد.

به طور خلاصه، پروژه با هدف پل زدن میان دنیای احساسات انسانی و تولید موسیقی خودکار توسط هوش مصنوعی انجام شده‌است.

۱-۲- کاربردهای پروژه

پروژه حاضر قابلیت استفاده در حوزه‌های مختلفی را دارد که برخی از مهم‌ترین آن‌ها عبارت‌اند از:

- **موسیقی‌درمانی**: تولید موسیقی آرامش‌بخش یا انگیزشی بر اساس احساسات لحظه‌ای کاربر برای کاهش استرس یا بهبود حال روانی.
- **صنعت بازی و سرگرمی**: ایجاد موسیقی پویا و تطبیقی در حین بازی بر اساس شرایط احساسی بازیکن.
- **سرویس‌های پخش آنلاین موسیقی**: افزودن ویژگی جدید شخصی‌سازی که به جای پیشنهاد آهنگ‌های آماده، موسیقی منحصر به‌فرد خلق می‌کند.
- **هنرهای تعاملی و اجراهای زنده**: ایجاد موسیقی بلادرنگ بر اساس واکنش تماشاگران یا احساسات بازیگران.
- **تحقیقات روان‌شناسی و علوم شناختی**: استفاده از موسیقی تولیدی برای مطالعه ارتباط احساسات و اصوات.

۱-۳- ساختار پایان نامه

با توجه به مقدمه‌ای که بیان شد، روند ارائه مطالب در این گزارش به صورت زیر است:

[1] JavaScript Object Notation

- **فصل دوم**: مروری بر مفاهیم پایه و پیش‌نیازها ارائه می‌شود. این مفاهیم شامل پردازش زبان طبیعی، مدل‌های تشخیص احساسات (BERT[1]، CNN[2]، wav2vec2)، مبانی موسیقی (گام‌ها، پروگرشن‌ها، ریتم، مدولاسیون) و همچنین معرفی فرمت MIDI [3]خواهد بود.
- **فصل سوم**: شرح کامل پروژه و جزئیات پیاده‌سازی بیان می‌شود. در این فصل پایگاه داده‌های استفاده‌شده، نحوه آموزش مدل‌ها، طراحی ماژول‌هایNLP[4]، پردازش صوت و تصویر، مدل تولید موسیقی و رابط کاربری Gradio تشریح خواهند شد.
- **فصل چهارم**: نتایج پیاده‌سازی پروژه ارائه می‌شود. این نتایج شامل نمونه‌های واقعی از تحلیل احساسات و موسیقی تولیدی، همراه با ارزیابی کیفیت و بازخورد کاربران خواهد بود.
- **فصل پنجم**: نتیجه‌گیری کلی از پروژه مطرح شده و پیشنهادهایی برای بهبود و توسعه کار در آینده ارائه می‌گردد.

[1] Bidirectional Encoder Representations from Transformers
[2] Convolutional Neural Network
[3] Musical Instrument Digital Interface
[4] Natural Language Processing

فصل دوم
مفاهیم

۲-۱- مقدمه

در این فصل، مبانی اصلی مرتبط با پروژه مورد نظر تشریح می‌شوند. ابتدا ساختار کلی مدل‌های پردازش زبان طبیعی (از جمله شبکه‌های عصبی بازگشتی و ترنسفورمر) و بعد به مدل‌های برت (BERT[1]) می‌پردازیم؛ شامل مدل فارسی ParsBERT و مقایسه‌ی آن با نسخه‌های انگلیسی. سپس در بخش بعدی تحلیل احساسات چندرسانه‌ای بررسی می‌شود: شناسایی احساسات از روی تصویر چهره و از روی گفتار صوتی. در ادامه عناصر نظری موسیقی بررسی می‌شوند از جمله گام‌ها، آکوردها و تنظیمات زمانی (ضرب و میزان)، ساختارهای آهنگ‌سازی مثل فرم سه‌بخشی A–B–A′ و پارامترهای موسیقایی مانند صدا(pitch)، مدت(duration)، بلندی (loudness) و طنین(timbre). در نهایت چگونگی ذخیره و نمایش موسیقی به‌صورت دیجیتال با استفاده از فرمت فایل MIDI توضیح داده خواهد شد. به‌طور کلی، نظریه موسیقی مطالعه چارچوب‌های تئوریک برای فهم اجزاء موسیقی است و شامل موضوعاتی نظیر سیستم‌های کوک (تُن‌کردن)، گام‌ها، هارمونی و روابط ریتمیک می‌باشد[۱]. در حوزه مدل‌های زبانی، معماری‌هایی مانند شبکه‌های عصبی بازگشتی (RNN[2]) و معماری ترنسفورمر مبتنی بر توجه (Self-Attention) برای مدلسازی توالی‌های متنی کاربرد فراوان دارند[۲] [۳]. همچنین استاندارد پرونده‌های MIDI [3]روشی ساخت‌یافته برای ذخیره و انتقال توالی‌های موسیقی است که معمولاً با پسوند mid. شناخته می‌شود[۱۱].

[1] Bidirectional Encoder Representations from Transformers
[2] Recurrent neural network
[3] Musical Instrument Digital Interface

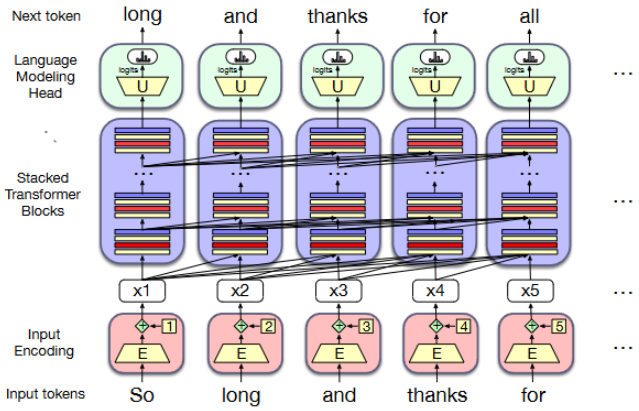
۲-۲- مدل‌های پردازش زبان طبیعی

۲-۲-۱- شبکه‌های عصبی بازگشتی(RNN, LSTM)

شبکه‌های عصبی بازگشتی (Recurrent Neural Networks – RNN) برای پردازش توالی‌های زمانی به کار می‌روند. در یک RNN، خروجی هر گام زمانی (خارجی) به‌عنوان ورودی مرحله بعد بازگشتی استفاده می‌شود و یک حالت پنهان (Hidden state) داخلی را حفظ می‌کند که اطلاعات متغیرهای گذشته را در خود نگه می‌دارد[۲]. این مکانیسم امکان مدل‌سازی وابستگی‌های زمانی را فراهم می‌کند. اما RNNهای ساده دارای مشکل نوسان گرادیان (Vanishing Gradient) هستند که یادگیری وابستگی‌های بلندمدت را دشوار می‌کند. معماری‌های پیشرفته‌تری همچون LSTM[1] (حافظه بلند-کوتاه‌مدت) و GRU[2] برای رفع این مشکل معرفی شده‌اند[۲]. این مدل‌ها با دروازه‌های داخلی خود، به یادگیری وابستگی‌های طولانی در داده‌های ترتیبی کمک می‌کنند.

۲-۲-۲- معماری ترنسفورمر(Transformer)

شکل ۱-۲ نمای کلی معماری ترنسفورمر را نشان می‌دهد که شامل چندین لایه متوالی از بلوک‌های کدگذاری (Encoder) و رمزگشایی (Decoder) می‌باشد. ترنسفورمر یک شبکه عصبی است که دارای لایه‌های توجه چندسر (Multi-head Attention) و شبکه‌های عصبی تغذیه پیشرو (Feedforward) می‌باشد[۳]. در هر بلوک کدگذار، ورودی‌ها ابتدا به لایه توجه خوداجتماعی (Self-Attention) می‌روند و سپس به یک شبکه پیشرو اعمال می‌شوند. استفاده از مکانیسم توجه به مدل اجازه می‌دهد تا روابط بین تمام کلمات (توکن‌ها) در دنباله را در نظر بگیرد. امروزه مدل‌های ترنسفورمر به‌خاطر توانایی بالای خود در یادگیری وابستگی‌های بلندمدت و قابل موازی بودن محاسبات، در بسیاری از کاربردهای پردازش زبان طبیعی غالب شده‌اند[۲].



شکل ۱-۲: معماری ترنسفورمر

[1] Long Short-Term Memory
[2] Gated Recurrent Unit

در مدل‌های ترنسفورمر، ابتدا هر کلمه (توکن) ورودی به یک بردار پیوسته (تعبیه‌شده) تبدیل می‌شود. سپس بردار حاصل در سلسله لایه‌های کدگذار ترنسفورمر پردازش شده و سرانجام به یک لایه خروجی(معمولاً یک لایه خطی به همراه تابع Softmax) ارسال می‌شود تا توزیع احتمالات واژگانی پیش‌بینی گردد[۳]. به‌عبارت دیگر، هر توکن ورودی ابتدا در یک ماتریس بردارگذاری جایابی می‌شود و سپس از طریق توالی بلوک‌های ترنسفورمر عبور می‌کند. این فرآیند به مدل اجازه می‌دهد نمایه‌های معنایی (بردارهای محتوا) هر واژه را با توجه به سایر واژگان جمله بسازد. در نهایت، با اعمال لایه خروجی و نرم‌ساز(Softmax)، مدل احتمال واژه‌های بعدی را در زبان تولید می‌کند.

۳-۲-۲- مدل‌های BERT

مدل‌های مبتنی بر تبدیل‌کننده (Transformer) مانند BERT در سال ۲۰۱۸ توسط گوگل معرفی شدند و به سرعت در بسیاری از وظایف NLP جایگاه استاندارد یافتند. نسخه پایه BERT دارای ۱۲ لایه‌ی ترنسفورمر و حدود ۱۱۰ میلیون پارامتر است و روی مجموعه بزرگ متون انگلیسی (کتاب‌ها و ویکی‌پدیا) آموزش دید[۴]. این مدل با پیش‌بینی واژه‌های ماسک‌شده (Masked LM) یاد می‌گیرد و به عنوان یک بردار بستر (embedding) زمینه‌ای غنی برای هر کلمه عمل می‌کند. برای استفاده در وظایفی مانند طبقه‌بندی متن یا تحلیل احساس، کافی است پس از لایه‌ی آخر BERT یک سر (head) طبقه‌بندی ساده اضافه و مدل را روی داده‌های برچسب‌دار بهینه کرد.

۱-۳-۲-۲- ParsBERT و کاربرد آن در متن فارسی

ParsBERT یک مدل تک‌زبانه و متن‌باز برای زبان فارسی است که بر پایه معماری BERT توسعه یافته و با یک مجموعه عظیم متون فارسی پیش‌آموزش شده است. این مدل بر روی بیش از ۳،۹ میلیون سند (حدود ۷۳ میلیون جمله و ۱،۳ میلیارد کلمه) با سبک‌های نوشتاری مختلف (علمی، اخبار، وبلاگ و غیره) آموزش داده شده‌است[۵]. در پروسه آماده‌سازی متون، پردازش‌های خاص زبان‌شناسی از جمله برچسب‌زنی نحوی (POS-tagging) و جداسازی واژه‌ها به تکه‌های (WordPiece (segmentation به کار رفته است تا متون به فرمتی مناسب برای مدل تبدیل شوند. از جمله نکات مهم ParsBERT این است که تمام مدل‌های پسینی (downstream) با نمایشگری بدون حساسیت به حروف کوچک (uncased) و ماسک کردن واژه کامل آموزش می‌یابند. تحقیقات نشان داده‌اند که ParsBERT در وظایف مختلف پردازش زبان طبیعی فارسی به‌طور قابل توجهی از مدل‌های چندزبانه یا انگلیسی برتر عمل می‌کند. به عنوان مثال، نتایج تجربی روی وظایف تحلیل احساس، طبقه‌بندی متن و شناسایی موجودیت نام‌دار (NER) نشان می‌دهد که ParsBERT در همه‌ی این وظایف عملکرد بهتری نسبت به مدل زبان چندزبانه (multilingual BERT) و سایر روش‌های پیشین داشته است. در جدول۱-۲ برای نتایج ParsBERT خلاصه‌ای از امتیازات F1 مشاهده می‌شود که در

آن جایگزینی واژگان و پیش‌پردازش‌های ویژه فارسی باعث بهبود وضعیت می‌شود. در بخش‌هایی از پروژه که نیاز به استخراج ویژگی از متن فارسی داریم، ParsBERT معمولاً انتخاب مناسبی است.

جدول ۱-۲: نتایج عملیات تحلیل احساسات

دیتاست‌ها	ParsBERT	mBERT	DeepSentiPers
نظرات کاربران دیجی‌کالا	۸۱/۷۴	۸۰/۷۴	-
نظرات کاربران اسنپ‌فود	۸۸/۱۲	۸۷/۸۷	-
SentiPers (Multi Class)	۷۷/۱۱	-	۶۹/۳۳
SentiPers (Binary Class)	۹۲/۱۳	-	۹۱/۹۸

شبه‌کد الگوریتم طبقه‌بندی متن فارسی برای این مدل به‌صورت شکل۲-۲ است که البته کد کامل آن در پیوست الصاق و درنهایت این مدل برای بهبود روی دیتاست آرمان[1] فاین‌تیون شد.

```
# مراحل الگوریتمی برای طبقه‌بندی متن فارسی با پارسبرت

بارگذاری مدل و توکنایزر ParsBERT:
    tokenizer = AutoTokenizer.from_pretrained
     ("HooshvareLab/bert-base-parsbert-uncased")
    model = AutoModelForSequenceClassification.from_pretrained
     ("HooshvareLab/bert-base-parsbert-uncased")

پیش‌پردازش متون:
    for text in مجموعه‌داده:
        tokens = tokenizer.tokenize(text)
        input_ids = tokenizer.convert_tokens_to_ids(tokens)
        attention_mask = tokenizer.create_attention_mask(input_ids)

تعریف سر طبقه‌بندی:
    # مدل اصلی پارسبرت خروجی برداری تولید می‌کند؛
    #   افزودن یک لایه‌ی چگال (Dense) و softmax .
    logits = model(input_ids, attention_mask)
    probabilities = softmax(logits)

آموزش و ارزیابی:
    model.train()
    for epoch in range(num_epochs):
        batch_logits = model(batch_input_ids, batch_attention_mask)
        loss = cross_entropy(batch_logits, batch_labels)
        backpropagate(loss)
    مدل را روی داده‌های آزمون ارزیابی کرده و دقت گزارش می‌شود.
```

شکل ۲-۲: الگوریتم طبقه‌بندی متن فارسی با ParsBERT

[1] https://github.com/Arman-Rayan-Sharif/arman-text-emotion

۲-۳-۲-۲- مدل‌های انگلیسی و مقایسه با فارسی

همان‌طور که اشاره شد، مدل اصلی BERT برای زبان انگلیسی در دو اندازه Base و Large ارائه شد؛ نسخه Base با ۱۲ لایه و ۱۱۰ میلیون پارامتر و نسخه Large با ۲۴ لایه و ۳۴۰ میلیون پارامتر[۴]. واژگان آن شامل حدود ۳۰هزار توکن WordPiece است که متون انگلیسی را پوشش می‌دهد. معماری کلی هر دو مدل (انگلیسی و فارسی) برپایه روش انکودر ترنسفورمر و مکانیزم توجه (Self-Attention) است. تفاوت اصلی در نحوه پیش‌آموزش و واژگان ورودی است: مدل انگلیسی BERT روی متون زبان انگلیسی آموزش دیده‌است، در حالی که ParsBERT واژگان فارسی دارد و روی مجموعه داده‌ی بزرگی از متون فارسی آموزش دیده‌است[۵].

به‌طور خلاصه، معماری فنی هر دو مدل مشابه است، ولی چون ParsBERT از پردازش‌های ویژه‌ی زبان فارسی بهره‌مند است، در مسایل فارسی، عملکرد بهتری نسبت به BERT چندزبانه یا استفاده از BERT انگلیسی با ترجمه متون دارد. نتایج نشان می‌دهد ParsBERT حتی در وظایفی مانند تحلیل احساس فارسی، امتیازهای بالاتری نسبت به نسخه چندزبانه و سایر مدل‌های پیشین کسب می‌کند.

۲-۳- تشخیص احساسات چندرسانه‌ای

در تحلیل احساسات چندرسانه‌ای، علاوه بر متن، از اطلاعات بصری و صوتی نیز برای تعیین حالت هیجانی افراد استفاده می‌شود. به عبارت دیگر، سیستم می‌تواند با دریافت ورودی تصویر چهره و یا صدای گفتار یک شخص، احساسات او (مانند خوشحالی، غم، ترس و...) را پیش‌بینی کند. این بخش شامل دو زیرموضوع است:

۲-۴-۱- تحلیل تصویر چهره(FER[1]2013, CNN[2])

تشخیص احساسات چهره (Facial Emotion Recognition) به معنای شناسایی و طبقه‌بندی حالت عاطفی انسان از طریق بیان چهره است[۶]. برای مثال، دیتاست FER2013 [۳]یک مجموعه داده مشهور برای این کار است که شامل حدود ۳۶هزار تصویر خاکستری ۴۸×۴۸ از چهره‌های افراد با هفت برچسب احساسی پایه است(خشم، تنفر، ترس، خوشحالی، غم، تعجب و خنثی)[۷]. معمولاً از مدل‌های شبکه عصبی کانولوشن(CNN) برای آموزش و استخراج خودکار ویژگی‌های صورت استفاده می‌شود. یک معماری ساده CNN شامل لایه‌های هم‌پوشانی(Conv) ، لایه‌های نمونه‌برداری (Pooling) و چند لایه کاملاً متصل در انتها

[1] Facial Emotion Recognition
[2] Convolutional Neural Network
[3] https://www.kaggle.com/datasets/nicolejyt/facialexpressionrecognition

است. در مطالعات جدید، استفاده از تکنیک‌های ترکیبی مانند استخراج ویژگی‌های ORB [1] یا LBP [2] به همراه CNN باعث بهبود دقت شده است. مطالعات مختلف نشان داده‌اند که مدل‌های CNN می‌توانند به دقت بالایی در این دیتاست‌ها برسند؛ به عنوان نمونه، یک معماری ConvNet چهار لایه روی FER2013 توانست به دقت آموزش ٪۹۶ و دقت اعتبارسنجی ٪۹۱ دست یابد. به طور کلی، دقت کل در تشخیص احساسات بالا گزارش شده است، اما برخی احساسات مانند «تنفر» و «ترس» همچنان در طبقه‌بندی مشکل هستند! بنابراین یک سیستم FER معمولی مراحل زیر را دارد: ابتدا تصویر ورودی پیش‌پردازش(نرمال‌سازی، برش صورت و ...) می‌شود، سپس با عبور از لایه‌های CNN، ویژگی‌های مربوط به اجزای صورت استخراج می‌گردد و در انتها یک لایه طبقه‌بندی مثل Softmax، احساس فرد را پیش‌بینی می‌کند. الگوریتم این روال به صورت شکل ۳-۲ است:

```
# الگوریتم شبکه‌کد آموزش شبکه کانولوشن برای FER2013
بارگذاری و پیش‌پردازش داده‌ها:
    X_train, y_train, X_test, y_test = load_FER2013_dataset()
    X_train = normalize_images(X_train)
    X_test = normalize_images(X_test)

تعریف مدل CNN:
    model = Sequential()
    model.add(Conv2D(filters=32, kernel_size=(3,3),
                     activation='relu', input_shape=(48,48,1)))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(filters=64, kernel_size=(3,3), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Flatten())
    model.add(Dense(units=128, activation='relu'))
    model.add(Dense(units=7, activation='softmax'))  # هفت کلاس احساسی

آموزش و ارزیابی:
    model.compile(loss='categorical_crossentropy',
                  optimizer='adam', metrics=['accuracy'])
    model.fit(X_train, y_train, epochs=20, batch_size=64)
    accuracy = model.evaluate(X_test, y_test)
    print("دقت تست:", accuracy)
```

شکل ۳-۲: الگوریتم طبقه‌بندی تصویر

[1] orientated FAST and rotated BRIEF
[2] Local Binary Patterns

۱-۴-۲- تحلیل صوت گفتار(wav2vec2)

در تحلیل احساسات صوتی، ورودی سیستم یک نمونه گفتار دیجیتال است و هدف طبقه‌بندی حالت عاطفی گوینده است. مدل wav2vec2 یک معماری عمیق مبتنی بر ترنسفورمر است که با استفاده از یادگیری خودنظارتی، از داده‌های صوتی خام بردارهای متنی قدرتمندی استخراج می‌کند. در این مدل، سیگنال صوتی ورودی ابتدا با چند لایه‌ی کانولوشن پردازش می‌شود تا ویژگی‌های صوتی محلی گرفته شود، سپس یک شبکه ترنسفورمر دنباله‌ای روی این ویژگی‌ها اعمال می‌شود[۸]. به هنگام پیش‌آموزش، بخش‌هایی از ورودی صوتی ماسک می‌شوند و مدل سعی در پیش‌بینی آنها در فضای مخفی (latent) می‌کند. این روش باعث می‌شود بردارهای با معنا و با زمینه‌ی طولانی از صدا یاد گرفته شود.

برای مثال، یک طیف‌نما(Spectrogram) نشان‌دهنده محتوای زمانی-فرکانسی یک سیگنال صوتی است (محور افقی زمان، محور عمودی فرکانس و روشنایی یا رنگ به بلندی صدا اشاره دارد)[۹]. مدل wav2vec2 می‌تواند مستقیماً روی امواج صوتی خام کار کند و یا روی طیف‌نماها اعمال شود. پس از پیش‌آموزش، این مدل را می‌توان برای وظایف مختلف از جمله تشخیص گفتار (ASR[1]) و تشخیص احساس با افزودن یک سر طبقه‌بندی ساده استفاده کرد.

مطالعات اخیر نشان داده‌اند که استخراج بردارهای ویژگی از مدل‌های از پیش‌آموزش‌شده wav2vec2 و استفاده از یک شبکه نسبتاً ساده برای شناسایی احساسات گفتاری، نتایجی فراتر از روش‌های قبلی دارد. برای مثال، ساختار پیشنهادی با ترکیب خروجی چند لایه‌ی مختلف wav2vec2 و بهینه‌سازی آن‌ها، در دیتاست‌های استاندارد احساس گفتار مانند IEMOCAP [۲]عملکرد برتری نسبت به ادبیات قبلی نشان داد[۱۰]. در ادامه شکل۴-۲ الگوریتم روال کلی این مسیر را نمایش می‌دهد.

[1] Automatic Speech Recognition
[2] interactive emotional dyadic motion capture database

```
# الگوریتم شبه‌کد برای طبقه‌بندی احساس گفتار
بارگذاری مدل پیش‌آموزش‌شده wav2vec:
    processor = Wav2Vec2Processor.from_pretrained
     ("facebook/wav2vec2-base-960h")
    model = Wav2Vec2ForSequenceClassification.from_pretrained
     ("facebook/wav2vec2-base-960h")

پیش‌پردازش صدا:
    audio_input = load_audio_file("sample.wav")  # بارگذاری سیگنال صوتی
    input_values = processor(audio_input, sampling_rate=16000,
                             return_tensors="pt").input_values

استخراج بردارهای ویژگی:
    with torch.no_grad():
        outputs = model(input_values)
    logits = outputs.logits  # خروجی مدل (احتمال‌های طبقه)

پیش‌بینی احساس:
    predicted_id = argmax(logits)
    emotion = processor.tokenizer.decode(predicted_id)

```

شکل ۴-۲: الگوریتم طبقه‌بندی صوت

۴-۲- مبانی نظری موسیقی

۱-۴-۲- نت‌ها و کوک

در موسیقی غربی، نت‌ها (Pitch) بیانگر فرکانس‌های صوتی معینی هستند. سیستم کوک معمولِ امروزی، گام مساوی یا 12-TET است که اکتاو را به ۱۲ نیم‌پرده مساوی تقسیم می‌کند[۱۲]. به‌طور مثال، نت استاندارد A4 با فرکانس ۴۴۰ هرتز کوک می‌شود.

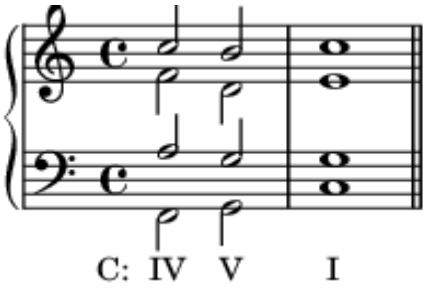
گام را می‌توان «متوالی از نت‌ها از یک نت تا همان نت در اکتاو بعدی» تعریف کرد و گام‌های معروف شامل گام‌های دیاتونیک بزرگ و کوچک هستند که هر یک حاوی هفت نت مختلف در یک اکتاو می‌باشند[۱۳].

۲-۴-۲- آکوردها و هارمونی

آکورد مجموعه‌ای از چند نت هم‌زمان است که اغلب سه نت یا بیشتر را شامل می‌شود و پایه هارمونی قطعه موسیقایی را تشکیل می‌دهد[۱۴]. ساده‌ترین نوع آکورد، تریاد است که از سه نت تشکیل شده (نت پایه، سوم و پنجم). تئوری هارمونی در موسیقی غربی، رابطه میان توالی آکوردها (پیشرفت آکورد) و عملکرد آن‌ها در ایجاد حس گام یا تونیک را بررسی می‌کند و به‌طور کلی، آکوردها رنگ و پشتیبانی هارمونیک را برای ملودی فراهم می‌کنند.

۲-۴-۳- پروگرشن آکوردها

همانطور که می‌دانیم، گام (Scale) یک ترتیب خاص از نت‌های صوتی است که الگوی فاصله‌های پرده(W) و نیم‌پرده(H) آن ثابت است. برای مثال، گام ماژور دارای توالی مشخصی از پرده و نیم‌پرده به‌صورت W–W–H–W–W–W–H است. گام‌ها پایه ملودی‌ها و هارمونی‌های موسیقی غربی هستند. به علاوه، آکورد(Chord) در تئوری موسیقی معمولاً شامل سه نت تشکیل‌دهنده (ریشه، سوم و پنجم) است که به صورت صعودی روی هم قرار گرفته‌اند و به آن‌ها «تراد» نیز گفته می‌شود. حال پروگرشن آکوردها، به توالی آکوردها در یک قطعه موسیقی گفته می‌شود؛ یعنی چند آکورد پشت سر هم که با هم ترکیب شده و زمینه هارمونیک قطعه را فراهم می‌کنند. شکل ۵-۲ یک نمونه از پروگرشن رایج در گام دو ماژور (I–V–IV) را نشان می‌دهد:



شکل ۵-۲: پیشرفت IV-V-I[1] در کلید C ماژور با آکوردهای: F ماژور، G ماژور و C ماژور.

۲-۴-۴- ساختارهای آهنگ‌سازی(A–B–A′)

یکی از ساختارهای رایج در آهنگ‌سازی فرم سه‌بخشی A–B–A′(معروف به فرم سه‌تایی یا «ترنری») است. در این ساختار، ابتدا بخش اولیه‌ای (A) معرفی می‌شود، سپس یک بخش میانی با حالتی متضاد (B) اجرا می‌شود، و در انتها بخش آغازین دوباره بازمی‌گردد (A′) که معمولاً مقداری واریاسیون(به معنای دو یا چند قطعه برمبنای زمینه‌ی موسیقایی (تم) واحدی تصنیف کردن که هر کدام از آن‌ها دارای بیان متفاوتی با دیگری است.) دارد. این ترتیب باعث ایجاد تعادل و تنوع در قطعه می‌شود، به‌طوری که بخش دوم فضایی متضاد با بخش اول فراهم کرده و بخش پایانی حس انسجام را بازمی‌گرداند. بسیاری از قطعات کلاسیک و پاپ مشهور از فرم A–B–A′ استفاده کرده‌اند.

۲-۴-۵- ضرب و میزان

برای سازماندهی زمانی موسیقی، از مفهوم میزان استفاده می‌شود که ضرب‌های منظم را در بر می‌گیرد. امضای میزان تعیین می‌کند که چند ارزش زمانی از یک نوع معین (مثلاً نت سیاه) در هر میزان قرار

[1] https://en.wikipedia.org/wiki/Chord_progression

می‌گیرد[۱۵]. به‌عنوان مثال، امضای میزان ۴/۴ نشان می‌دهد در هر میزان چهار نت سیاه وجود دارد. رایج‌ترین امضاهای میزان در موسیقی غربی عبارتند از ۴/۲، ۴/۳ و ۴/۴. علاوه بر این، تمپو یا سرعت قطعه معمولاً برحسب ضرب بر دقیقه (BPM[1]) مشخص می‌شود؛ درصورت نبودن تعیین صریح تمپو و میزان، استاندارد MIDI [۲] مقدار ۱۲۰BPM و میزان ۴/۴ را به‌عنوان مقدار پیش‌فرض در نظر می‌گیرد[۱۶].

۲-۴-۶- پارامترهای موسیقایی

پارامترهای اصلی موسیقی که کیفیت صداها را مشخص می‌کنند عبارتند از زیر و بمی صدا(Pitch)، بلندی یا حجم(Loudness/Volume)، مدت زمان(Duration/Time) و طبیعت صدا(Timbre). به عنوان مثال، زیر و بم یک نُت ناشی از بسامد (فرکانس) آن است، بلندی مربوط به شدت صدا است و timbre تمایز رنگ صوت (مثلاً صدای گیتار از پیانو) را مشخص می‌کند. این پارامترها اجزای اصلی در توصیف ملودی، هارمونی و ریتم موسیقی هستند و در هر ترکیب موسیقایی نقش مهمی ایفا می‌کنند. با تغییر این پارامترها می‌توان جلوه‌های مختلف احساسی در موسیقی به وجود آورد؛ برای مثال، تمپو (سرعت) سریع همراه با حجم زیاد معمولاً حس هیجان یا خشم را منتقل می‌کند، در حالی که تمپو کند و حجم کم حس آرامش یا غم را ایجاد می‌کند.

۲-۵- ساختار فایل MIDI [۳]

۲-۵-۱- چارچوب کلی و هدر

اگر MIDI(رابط دیجیتال ادوات موسیقی) یک استاندارد فنی است که اطلاعات موسیقی را به‌صورت رویدادهای دیجیتالی ذخیره می‌کند. برخلاف فایل‌های صوتی معمولی مانند .mp3و .wav.، این‌ها حاوی داده‌های صوتی واقعی نیستند بلکه شامل اطلاعاتی مانند نت، زمانبندی، مدت و شدت می‌باشند. فایل استاندارد این رابط دیجیتال یعنی SMF[4] قالبی است که توالی‌های موسیقایی را به‌صورت یک فایل با پسوند .mid نگهداری و منتقل می‌کند[۱۱]. هر فایل MIDI با یک هدر چانک(MThd) آغاز می‌شود که نوع فرمت فایل (۰، ۱ یا ۲)، تعداد چانک‌های ترک (Track) و یک میدان زمان‌بندی (<division>) را مشخص می‌نماید[۱۶]. این میدان<division> معمولاً تعداد تیک‌های زمانی لازم برای یک نت سیاه (Quarter-note) را تعیین می‌کند[۱۶]. بر اساس فرمت هدر، فایل MIDI می‌تواند شامل یک یا چند چانک ترک باشد که هر کدام داده‌های آهنگ را نگه می‌دارد.

---

[1] Beats Per Minute
[2] Musical Instrument Digital Interface
[3] Musical Instrument Digital Interface
[4] Standard MIDI File

از دیدگاه کامپیوتری، فایل‌های MIDI به عنوان مجموعه‌ای از پیام‌های دیجیتالی ذخیره می‌شوند که پیام شامل اطلاعاتی درباره‌ی نحوه‌ی پخش موسیقی است و به صورت باینری کدگذاری شده و شامل سه بخش اصلی است:

الف) پیام‌های وضعیت: نوع عملیاتی که باید انجام شود را مشخص می‌کند مثل نواختن یک نت، تغییر ابزار موسیقی یا شدت صدا.

ب) پیام‌های داده: اطلاعات دقیق‌تری درباره عملیات مشخص‌شده در پیام‌های وضعیت ارائه می‌دهد مانند شماره‌ی نت.

ج) زمانبندی: زمان اجرای دقیق هر پیام را مشخص کرده که به دستگاه‌ها اجازه داده تا موسیقی را با دقت زمانی بالا پخش کنند.

۲-۵-۲- رویدادها و پیام‌های MIDI

پس از هدر، چانک‌های ترک(MTrk) به‌صورت پشت‌سرهم در فایل قرار می‌گیرند. هر چانک ترک حاوی یک سری رویداد MIDI و برخی رویدادهای غیر MIDI که به همراه فیلدهای زمان‌بندی نسبی (delta-time) قبل از هر رویداد است[۱۶]. ساختار هر رویداد MIDI به‌صورت <delta-time><event> است که در آن <delta-time>مدت زمان نسبی قبل از اجرای رویداد بعد را مشخص می‌کند[۱۶]. انواع رویدادهای MIDI شامل پیام‌های کانال مانند Note On و Note Off، پیام‌های کنترل‌کننده، تغییر برنامه و پیام‌های سیستم اختصاصی است. برای نمونه، پیام Note On با کد 0x9n ارسال می‌شود و دو بایت داده دارد: یکی برای شماره نت (Pitch) و دیگری برای سرعت (Velocity) اجرای نت[۱۶]. به‌عنوان مثال، یک پیام Note On برای نت C4 با صدای قوی ممکن است شامل کد 0x90(کانال ۱) و داده‌های (60, 96) باشد؛ در حالتی که پیام Note Off با کد 0x8n همان پایان اجرای نت را اعلام می‌کند[۱۶]. علاوه بر پیام‌های فوق، فرمت MIDI شامل متا-رویدادها نیز هست که اطلاعاتی مانند تعریف تمپو (زمان) و امضای میزان را حمل می‌کنند. اگر در فایل MIDI، رویداد ست‌تمپو و امضای میزان مشخص نشده باشد، به‌طور خودکار همان مقادیر پیش‌فرض ذکرشده را قرار می‌دهد[۱۶]. در نهایت این پیام‌ها به‌ترتیب در یک فایل MIDI همانند شکل ۶-۲ ذخیره شده و به راحتی بین نرم‌افزارها و سخت‌افزارهای مختلف جابجا می‌شود.

time message time message time message time message
time message time message time message time message
time message time message time message time message
time message time message time message time message
time message time message time message time message ....

شکل ۶-۲: ساختار فایل MIDI

مزیت این فایل برای آموزش مدل‌های زبانی این است که می‌تواند چندین ترک را به‌صورت همزمان ذخیره کند و هر ترک بیانگر یک ساز باشد. پس مدل زبانی تعاملات پیچیده بین سازهای مختلف را تحلیل و درک کرده تا این داده‌های دیجیتالی بدون نویز را ترکیب کند و قطعات مختلف موسیقی را بسازد.

۶-۲- جمع‌بندی

در این فصل، مبانی نظری مورد نیاز برای پروژه مورد بررسی قرار گرفت. ابتدا معماری‌های مهم در پردازش زبان طبیعی از جمله شبکه‌های بازگشتی و ترنسفورمر بررسی شد که در مسائل مدلسازی توالی نظیر تولید موسیقی با استفاده از متن‌نمائی کاربرد دارند. سپس مباحث پایه‌ای موسیقی شامل نت‌ها (گام مساوی و توزیع فرکانسی نت‌ها)، گام‌ها (تعریف توالی نت‌ها در یک اکتاو)، آکوردها (ترکیب چند نت برای ایجاد هارمونی) و اصول زمان‌بندی موسیقی (ضرب و میزان) معرفی شد. در نهایت ساختار فنی فایل MIDI [1]توضیح داده شد؛ از جمله نحوه ذخیره‌سازی اطلاعات موسیقی به صورت رویدادهای دیجیتال در هدر و چانک‌های ترک. این مبانی زمینه را برای فصول بعدی فراهم می‌کنند تا بتوان مدل‌ها و روش‌های عملی مربوط به تولید یا پردازش موسیقی را تشریح نمود.

[1] Musical Instrument Digital Interface

فصل سوم
شرح پروژه

۳-۱- مقدمه

در این پروژه، با استفاده از داده‌های چندرسانه‌ای (متن، تصویر چهره یا صوت)، سعی شده است موسیقی متناظر با احساس کاربر تولید شود. داده ورودی ابتدا به ماژول پیش‌پردازش مربوطه (مثلاً تبدیل متن به توکن یا استخراج مشخصات صوتی) وارد شده و سپس با مدل‌های یادگیری عمیق، «برچسب احساس» مناسب و درصد اطمینان آن تعیین می‌گردد. براساس تحقیقات پیشین، انسان‌ها احساسات خود را از طریق روش‌های گوناگونی مانند متن، صوت، تصویر و... بیان می‌کنند و تحلیل این احساسات یکی از چالش‌های اصلی در پردازش زبان طبیعی و بینایی ماشین است[۱۷]. برای مثال، در یکی از مطالعات جدید، از ترکیب برچسب‌گذاری احساسی و مدل‌های زبان طبیعی برای تبدیل تصاویری مانند نقاشی به توضیح متنی احساس و تولید موسیقی استفاده شده است[۱۸]. در این سیستم، پس از تعیین برچسب احساس (مانند «شادمانی») از ورودی کاربر، مرحله تولید موسیقی آغاز می‌شود. خروجی نهایی علاوه بر فایل صوتی تولیدشده، شامل نمودار ارزیابی ساختار موسیقی و یک گزارش JSON[1] از نتایج تحلیل است تا کاربر بتواند آنها را مشاهده و دانلود کند.

۳-۲- معماری و طراحی سیستم

همانند یک معماری استاندارد در سامانه‌های یادگیری ماشین، کل سیستم متشکل از چندین مرحله زنجیروار است: دریافت ورودی، پیش‌پردازش داده، استخراج ویژگی، تحلیل احساس و تولید خروجی. در اینجا ابتدا ورودی متنی یا تصویری یا صوتی بارگذاری می‌شود و پیش‌پردازش‌های لازم انجام می‌گردد. سپس مدل‌های یادگیری عمیق، برچسب احساس (مثل «غمگینی»، «خشم»، «شادمانی» و ...) را برای هر ورودی

[1] JavaScript Object Notation

پیش‌بینی می‌کنند. در نهایت، ماژول تولید موسیقی با استفاده از چارچوبی مانند MIDI [1]و فایل‌های SoundFont(.sf2)، آهنگ مناسب متناظر با آن احساس را می‌سازد. همان‌طور که در مطالعات معماری لوله‌کشی(Pipeline) یادگیری ماشین آمده، مراحل مختلف یک خط لوله‌ی ML [2]مانند «ورود داده، پیش‌پردازش، استخراج ویژگی، آموزش مدل، ارزیابی مدل، پیش‌بینی» به طور متوالی اجرا می‌شوند [۱۹]. هر مرحله در این خط لوله می‌تواند به صورت یک گره در یک گراف بدون دور (DAG[3]) در نظر گرفته شود که وابستگی‌های آن با یال‌های جهت‌دار مشخص می‌شود. این طراحی تضمین می‌کند که خروجی هر مرحله ورودی مرحله بعدی بوده و یک فرایند پردازشی کامل و منسجم ارائه می‌شود. در پیاده‌سازی ما نیز رویکرد مشابه به‌کار رفته‌است. ابتدا داده‌ها به شکل مناسب (مثلاً تبدیل متن به توکن‌های زبان، تبدیل تصویر به بردار ویژگی با شبکه‌های عصبی، یا تبدیل صوت به طیف فرکانسی) آماده می‌شوند. سپس، با استفاده از مدل‌های پیش‌آموزش‌دیده یا طراحی‌شده برای تشخیص احساس در هر نوع(مدالیتِ) ورودی، احساس غالب استخراج می‌شود. در این سیستم، از فایل‌های صدا (SoundFont) برای تولید موسیقی استفاده می‌شود؛ SoundFontها مجموعه‌ای از نمونه‌های صوتی از سازهای مختلف هستند که امکان تولید صداهای گوناگون را با الگوریتم سنتز مبتنی بر MIDI فراهم می‌کنند[۲۰]. با تغییر SoundFont می‌توان سبک و تمپوی موسیقی را تغییر داد (مشابه تغییر فونت در متن که شکل حروف را عوض می‌کند اما محتوا ثابت می‌ماند). در نهایت ماژول تولید موسیقی با استفاده از کتابخانه‌هایی مانند FluidSynth یا midi2audio، دنباله‌ای از نت‌های موسیقی را بر اساس احساس پیش‌بینی‌شده می‌سازد و فایل صوتی نهایی را تولید می‌کند.

۳-۳- پیاده‌سازی ماژول‌ها

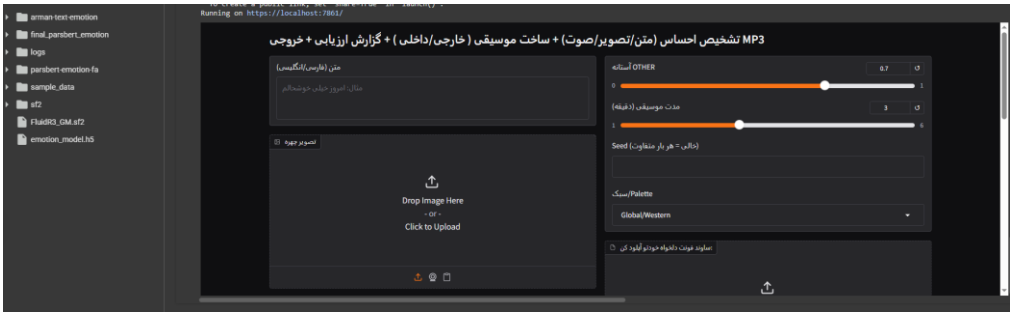
پیاده‌سازی اصلی در یک تابع مرکزی به نام analyze_and_make_music انجام شده‌است که ورودی‌های متن، تصویر یا فایل صوت را دریافت می‌کند. این تابع ابتدا داده‌ها را بررسی و پیش‌پردازش کرده (برای مثال متن را پاک‌سازی و توکن‌بندی می‌کند)، تصویر را به صورت RGB بارگذاری می‌کند، یا فایل صوت را به فرمت مشترک تبدیل می‌کند. سپس براساس آنالیز ورودی‌های چندگانه، اولویت را به متن می‌دهد و اگر متنی ارائه نشده باشد، تصویر یا صوت را تحلیل می‌کند. برای تحلیل احساس متن، می‌توان از مدل‌هایNLP[4] مانند BERT [5] فارسی یا روش‌های آماری استفاده کرد (یعنی بر مبنای دیکشنری واژگان احساسی). برای تحلیل تصویر چهره، از شبکه‌های عصبی بینایی (مثل مدل‌های تشخیص حالات چهره) بهره

[1] Musical Instrument Digital Interface
[2] Machine learning
[3] Directed Acyclic Graph
[4] Natural Language Processing
[5] Bidirectional Encoder Representations from Transformers

برده می‌شود. در مورد صوت، با استخراج ویژگی‌های صوتی مانند ضریب‌های MFCC [1] و استفاده از مدل‌های طبقه‌بندی احساس، برچسب احساسی محتوای گفتار تعیین می‌شود. در قدم بعد، برچسب احساس و میزان اطمینان آن به عنوان ورودی به بخش تولید موسیقی داده می‌شوند. این بخش یک الگوی موسیقی نمادین مانند نت‌های MIDI را بر اساس تنظیمات سبک (مثل: «جهانی » یا «ایرانی/سنتی») و بانک‌های داخلی SoundFont ایجاد می‌کند. هر بانک SoundFont شامل ضبط‌های نمونه‌ای از سازهایی مثل پیانو، کمانچه، یا ساز برقی(Synthesizer) است که خروجی صوتی نهایی با آن‌ها تولید می‌شود[۲۰]. در نهایت فایل WAV [2] خروجی به همراه یک گزارش ساختاری شامل داده‌هایی نظیر میزان تنوع ملودی، ریتم و احساس پیش‌بینی‌شده در موسیقی توسط مدل احساس‌سنج ساخته می‌شود.

۴-۳- رابط کاربری با Gradio

برای توسعه سریع رابط کاربری تعاملی، از کتابخانه‌ی Gradio استفاده شده است که یک بسته‌ی متن‌باز پایتون است و با تنها چند خط کد اجازه می‌دهد یک رابط وب ساده برای مدل خود بسازید، بدون آن‌که نیاز به نوشتن کد HTML [3]، CSS [4] یا جاوااسکریپت باشد[۲۱]. در این رابط، کامپوننت‌های ورودی متن، تصویر و فایل صوتی اضافه شده‌اند. کاربر می‌تواند یک متن بنویسد یا تصویر و صوتی را بارگذاری کند تا احساس آن تحلیل شود. سپس با فشردن دکمه، مدل اطلاعات را پردازش کرده و موسیقی تولیدی به همراه نمودارها و نتایج ارزیابی نمایش داده می‌شود. مزیت Gradio این است که *به سرعت و تنها با کد پایتون* یک رابط تعاملی و زیبا ایجاد می‌کند. در این پروژه، اطلاعات ورودی و خروجی به صورت دو زبانه (فارسی/انگلیسی) نمایش داده می‌شوند و امکان دانلود فایل موسیقی و گزارش JSON [5] فراهم شده است. در شکل ۱-۳ شمای کلی از این رابط در گوگل کولب نمایش داده‌شده‌است.



شکل ۱-۳: رابط کاربری ساده در گوگل کولب

[1] Mel-frequency cepstrum
[2] Waveform Audio File Format
[3] Hypertext Markup Language
[4] Cascading Style Sheets
[5] JavaScript Object Notation

۵-۳- رابط کاربری مدرن با Flask و ngrok (توسعه‌ی اختیاری)

برای یک رابط کاربری حرفه‌ای‌تر و قابل شخصی‌سازی، یک سرور وب سبک با فلاسک پیاده‌سازی شده است که یک چارچوب میکرو (Microframework) وب متن‌باز است که برای شروع سریع برنامه‌های وب طراحی شده و توسعه‌دهندگان را قادر می‌سازد بدون ابزارهای پیچیده، در کوتاه‌ترین زمان یک سرور وب راه‌اندازی کنند[۲۲]. در پروژه ما از Flask به همراه موتور قالب Jinja2 استفاده شد تا صفحات HTML با استایل (Bootstrap) و اسکریپت‌های لازم تولید شوند. برای نمایش گرافیکی موج صوت از کتابخانه‌ی Wavesurfer.js بهره برده‌ایم که یک کتابخانه‌ی متن‌باز برای نمایش تعاملی امواج صوتی در وب است که نمودار صوت را به صورت واکنش‌گرا و قابل تنظیم ترسیم می‌کند[۲۳]. با استفاده از Wavesurfer، کاربر می‌تواند موسیقی تولیدی را در مرورگر پخش، جلو و عقب برده و حجم صدا را تنظیم کند. شمای کلی در شکل ۲-۳ نمایش داده‌شده‌است.

شکل ۲-۳: رابط کاربری مدرن نهایی

علاوه بر این، برای به‌اشتراک‌گذاری آسان سرور محلی در وب عمومی از ngrok استفاده شده است که یک پراکسی معکوس است و با ایجاد تونل‌های امن از یک آدرس عمومی به سرور محلی، امکان اتصال اینترنتی و تست وب‌سرویس‌ها را فراهم می‌کند[۲۴]. حال با بسته‌ی Python ای به نام pyngrok، کنترل ngrok را در کد داریم و این ابزار برای توسعه سریع (مانند تست وب‌هوک یا نمایش دمو به دیگران) ایده‌آل است. در نهایت، ترکیب Flask، Wavesurfer.js و ngrok یک رابط کاربرپسند و مدرن ایجاد کرده که تمامی امکانات لازم (بارگذاری تصویر، فایل صوت، نمایش نتایج، دکمه دانلود و...) را دارد و تجربه‌ای مطلوب برای کاربر فراهم می‌سازد.

۶-۳- جمع‌بندی

در این فصل جزئیات کاملی از طراحی و پیاده‌سازی پروژه ارائه شد. ساختار کلی سیستم به صورت یک خط لوله‌ی پردازش داده استاندارد تعریف شد و هر ماژول از پیش‌پردازش ورودی تا تولید موسیقی تشریح گردید. همچنین ابزارها و فریم‌ورک‌های منتخب مانند Gradio برای نمونه‌سازی سریع، Flask برای رابط وب سبک، Wavesurfer.js برای نمایش صوت و ngrok برای دسترسی ساده معرفی شدند. استفاده از بانک‌های SoundFont امکان تولید صداهای متنوع موسیقی را فراهم کرده است. در نهایت، با یک رابط تعاملی کامل (هم در محیط‌های آفلاین و هم بر بستر اینترنت) امکان آزمون و استفاده از سیستم میسر شد و تمامی خروجی‌ها اعم از فایل موسیقی تولیدی و گزارش‌های تحلیل به کاربر ارائه گردید. تمامی مراحل طراحی و اجرا مطابق با الگوهای شناخته‌شده در معماری ML انجام شده و در هر بخش از منابع معتبر و استاندارد استفاده شده است[۱۹] [۲۱] [۲۲] [۲۳] [۲۴].

فصل چهارم
نتایج و ارزیابی

۴-۱- مقدمه

در این فصل نتایج ارزیابی مدل پیشنهادی ارائه می‌شود. ما مدل را از دو جنبه‌ی کمی و کیفی بررسی کردیم. ابتدا معیارهای کمی شامل شاخص‌های مختلفی نظیر CLAPScore برای سنجش انطباق معنایی صوت خروجی با متن ورودی، همراه با معیارهای بازیابی متنی مانند Recall@1 و میانگین CLAPScore، فاصله‌ی فرِشِه صوتی (FAD[1]) و فاصله‌ی هسته‌ای صوتی (KAD[2]) را محاسبه کردیم. سپس ویژگی‌های ساختاری نمونه‌های تولیدشده از جمله سرعت (تمپو) و مد (ماژور/مینور) را تحلیل نمودیم. در نهایت تحلیل‌های انسانی شبه‌کمی با استفاده از امتیازهای MOS[3] انجام شد. مطالعات قبلی نشان داده‌اند که ویژگی‌های موسیقی مانند تمپو و مد با احساس موسیقی رابطه‌ی معنی‌داری دارند؛ برای مثال تمپوی سریع معمولاً با احساس خوشی و تمپوی کند با احساس غم و ناامیدی مرتبط است. این فصل با مروری بر نحوه‌ی استخراج و معیارگذاری این ویژگی‌ها آغاز شده و پس از آن نتایج به‌دست‌آمده از نمونه‌های تولیدشده ارائه و تحلیل می‌گردد.

۴-۲- ارزیابی کمی

برای برای ارزیابی کمّی از معیارهای زیر استفاده شد:

- **CLAPScore و R@1 :** ابتدا از مدل CLAP [4]برای استخراج جاسازی‌های (Embedding) صوت و متن استفاده کردیم. سپس شباهت کسینوسی بین جاسازی صوت تولیدی و متن ورودی محاسبه

[1] Fr´echet Audio Distance
[2] Kernel Audio Distance
[3] Mean Opinion Score
[4] Contrastive Language-Audio Pretraining

شد(CLAPScore)[۲۵][۲۶]. معیار Recall@1 هم نشان‌دهنده‌ی درصد مواردی است که جاسازی صوت با بهترین فاصله کسینوسی به همان متن منطبق شده است. مقادیر CLAPScore بالا و R@1 نزدیک به یک، دلالت بر انطباق خوب معنایی خروجی با ورودی دارد. در پژوهش‌ها معمولاً CLAP در کنار FAD گزارش می‌شود تا توازن کیفیت صوتی و تطابق محتوایی بررسی گردد. امتیاز بالاتر CLAP نشان‌دهنده‌ی انطباق بیشتر موسیقی تولیدی با محتوای متن است. پس معیار امتیاز CLAP بدین صورت تعریف می‌شود که ابتدا بردار نهفته متن ورودی fT و بردار نهفته صوت تولیدشده fA را با مدل CLAP استخراج می‌کنیم. سپس شباهت کسینوسی میان این دو بردار را محاسبه کرده و میانگین می‌گیریم. فرمول این معیار به صورت شکل ۱-۴ است:

$$\frac{{}_{\text{audio}}f_{\text{text}} \cdot f}{\|{}_{\text{audio}}f_{\text{text}}\| \, \|f\|} = \cos\left(f_{\text{text}}(T),\ f_{\text{audio}}(A)\right) = \text{CLAP-Score}(T, A)$$

شکل ۱-۴: نحوه‌ی محاسبه‌ی معیار CLAP

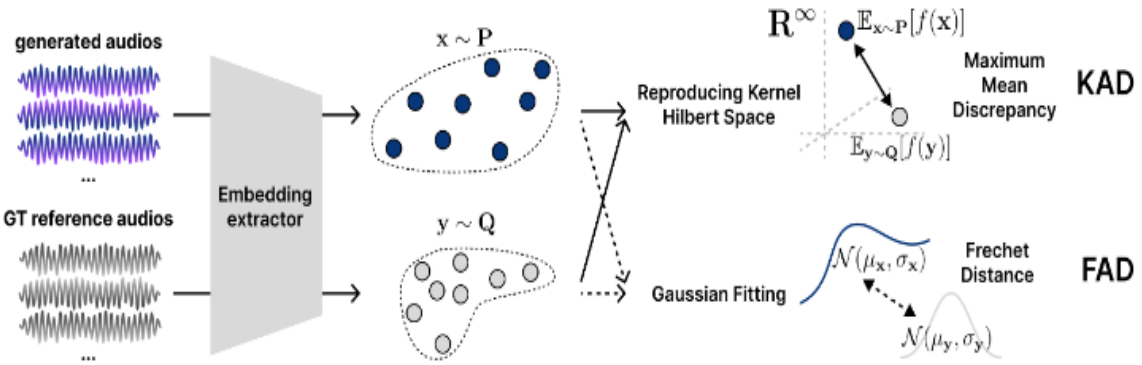
- **فاصله‌ی فرِشِه صوتی(FAD) :** این معیار توزیع آماری جاسازی‌های صوت تولیدی را با جاسازی‌های داده‌های مرجع مقایسه می‌کند[۲۷]. FAD پایین نشانگر کیفیت بالای خروجی و نزدیکی دو توزیع است. در مواردی که داده‌ی مرجع به‌صورت نمونه‌های تولیدی باشد، FAD ≈ 0 می‌شود ولی امکان Null نبودن وجود دارد. با فرض توزیع گاوسی برای ویژگی‌های استخراج‌شده، فاصله فرشه بین دو توزیع گاوسی متناظر با مجموعه مرجع (میانگین μ_b، کواریانس Σb) و مجموعه تولیدی (میانگین μ_e، کواریانس Σe) محاسبه می‌شود. فرمول کلی فاصله فرشه بین دو توزیع گاوسی به صورت شکل ۲-۴ است:

$$\mathbf{F}(\mathcal{N}_b, \mathcal{N}_e) = \|\mu_b - \mu_e\|^2 + tr(\Sigma_b + \Sigma_e - 2\sqrt{\Sigma_b \Sigma_e})$$

شکل ۲-۴: نحوه‌ی محاسبه‌ی معیار FAD

- **فاصله‌ی هسته‌ای صوت(KAD) :** بر پایه‌ی مفهوم MMD [1]طراحی شده و محدودیت‌های FAD را برطرف می‌کند[۲۸]. KAD نیز تفاوت آماری بین جاسازی‌های دو مجموعه را می‌سنجد؛ ولی با فرض توزیع نامعین (بدون نیاز به گوسی بودن). مقادیر میانگین و انحراف‌معیار KAD گزارش شدند تا کیفیت کلی توزیع خروجی با داده‌های مرجع به‌خوبی سنجیده شود.

[1] Maximum Mean Discrepancy



شکل ۳-۴: مقایسه بین KAD (فاصله صوتی هسته) و FAD (فاصله صوتی فرشه)

- **تنوع(Diversity) :** این معیار بر اساس میانگین همبستگی زوجی بین جاسازی‌های صوت محاسبه می‌شود و نشان‌دهنده‌ی تنوع کلی مجموعه‌ی تولید شده است. مقدار نزدیک به یک دلالت بر خروجی‌های متنوع و غیرتکراری دارد.
- **مطابقت تمایز احساسی:** برای بررسی اینکه آیا احساس موسیقی تولیدشده به هدف در نظر گرفته شده نزدیک است یا خیر، از یک مدل طبقه‌بندی احساسات موسیقی بر روی هر نمونه استفاده شد (به‌عنوان مثال، در یکی از نمونه‌ها هدف «ترس» بود اما مدل با اعتماد بالا «خشم» را تشخیص داد که نشان‌دهنده‌ی چالش در تولید دقیق احساس است و برخی فضاهای احساسی به صورت ترکیبی از چند حس می‌باشند و درک آن سبب ایجاد چالش برای یک هوش مصنوعی فاقد احساس، می‌شود.).
- **امتیاز MOS (نظرسنجی انسانی):** در نهایت، امتیاز کیفیت کلی هر قطعه (MOS) از طرق شنیداری به دست آمد. در این آزمایش، به دلیل محدودیت در دسترسی به شرکت‌کننده‌های واقعی، امتیازها به صورت تصادفی (اما مطابق سناریوی کلاسیک) تولید شدند. در جدول ۱-۴ بازه‌ی مقادیر موجود را برای این معیار نشان می‌دهد. MOS به عنوان میانگین حسابی در رتبه‌بندی های منفرد انجام شده توسط افراد انسانی برای یک محرک معین در یک آزمون ارزیابی کیفیت ذهنی محاسبه می شود و نحوه‌ی محاسبه‌ی آن در شکل ۴-۴ آورده‌شده‌است که R رتبه‌بندی‌های فردی برای یک محرک معین توسط N فرد است.

- **جدول ۱-۴: دامنه مقادیر معیار MOS**

Label برچسب	Rating امتیاز
Excellent عالی	۵
Good خوب	۴
Fair متوسط	۳
Poor ضعیف	۲
Bad بد	۱

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

شکل ۴-۴: نحوه‌ی محاسبه‌ی معیار نظرسنجی انسانی

شکل ۵-۴ نیز نتایج کمی اصلی را برای یک نمونه تولیدی گزارش می‌کند. با توجه به شاخص‌های مورد استفاده، خروجی مدل برای این مثال دارای کیفیت معقولی است(اما می‌تواند بهتر هم باشد): مقادیر بالای CLAPScore و R@1 نشان از انطباق معنایی خوب با متن دارد، و مقادیر به‌نسبت کم FAD و KAD (با انحراف استاندارد اندک) نشان‌دهنده‌ی نزدیکی توزیع صوت‌های تولیدی به داده‌های مرجع است. همچنین diversity هرچه نزدیک‌تر به ۱ بدین معنی است که نمونه‌ها نسبتاً متنوع هستند.

Audio Generation Evaluation


```
{
  "n_generated": 2,
  "n_reference": 3,
  "metrics": {
    "R@1": 0.5,
    "MeanRank": 1.5,
    "CLAPScore_mean": 0.0675,
    "CLAPScore_median": 0.0675,
    "FAD_VGGish": 352521.70835931465,
    "KAD(CLAP)_mean": 0.00092126528033986 69,
    "KAD(CLAP)_std": 0.0022993823513388634,
    "FAD(CLAP)": 1.4007600481266895,
    "Diversity(CLAP)": 0.41101449728012085,
    "MOS_mean": 3.7,
    "MOS_std": 1.1
  }
}
```


Per-file table

file	tempo_est	key_index_est	text	CLAP_text_audio
/content/sad_persian_1.0min_1755642423.wav	117.19	11	very slow, mournful, crying, minimal, sparse	0.062850
/content/sad_persian_1.0min_1755642483.wav	96.98	0	very slow, mournful, crying, minimal, sparse	0.072201

MOS (fake) table

file	mos_1	mos_2	mos_3	mos_4	mos_5	mos_6	mos_7	mos_8	mos_9	mos_10	mos_11	mos_12	mos_13	mos_14	mos_15
/content/sad_persian_1.0min_1755642423.wav	2	5	4	3	3	5	2	4	2	2	4	5	4	5	4
/content/sad_persian_1.0min_1755642483.wav	5	4	2	5	3	4	3	2	5	5	4	3	5	4	3

شکل ۵-۴: نمای فایل HTML برای ارزیابی یک نمونه خروجی

۳-۴- تحلیل ساختاری و احساسی نمونه‌ها

در این بخش، شش نمونه‌ی تولیدشده که هر یک نماینده یکی از احساس‌های هدف («شاد»، «عصبانی»، «غمگین»، «ترس»، «تعجب» و «خنثی») هستند، از نظر ویژگی‌های ساختاری بررسی می‌شوند. شکل ۶-۴ در ادامه خلاصه‌ای از پارامترهای استخراج شده و تطابق آنها با انتظارات نظری را نشان می‌دهد. مطالعات حوزه روان‌شناسی موسیقی تأیید کرده‌اند که تمپو و مد گام، نقش مهمی در ایجاد احساس دارند. بر این اساس، تحلیل نتایج به شرح زیر است:

الف) شاد(Happy): موسیقی شاد معمولاً دارای تمپوی نسبتاً سریع و گام ماژور است. نمونه‌های «شاد» تولیدشده نیز دارای تمپوهای بالا و گام ماژور بودند. این ویژگی‌ها نشان‌دهنده‌ی حالت مثبت و انرژی بالا در ملودی است. (به‌عنوان مثال، ذکر شده که از الگوی ضرب‌آهنگ دنس پاپ استفاده شده است.)

ب) عصبانی(Angry): قطعات «عصبانی» معمولاً شامل درام‌های قوی و گیتارهای برقی تحریف‌شده با تمپوهای تند هستند. در نتایج ما، ملودی‌های «عصبانی» سرعت بالاتری داشتند و با تمبوری تهاجمی ایجاد شده بودند. این موارد با ماهیت احساسی «خشم» همخوانی دارد.

پ) غمگین(Sad): موسیقی غمگین اغلب دارای تمپو کند و گام مینور است. نمونه‌های «غمگین» تولیدشده، دارای تمپو پایین و نواهای احساسی و ملایم بودند که حس اندوه را منتقل می‌کند. (مطالعات نشان داده‌اند گام مینور و تمپو آهسته عاملی برای قضاوت شنونده به عنوان موسیقی غمگین هستند.)

ت) ترس(Fear): موسیقی احساسی «ترس» به‌طور کلی تمپو بسیار کند، گام مینور و استفاده از سازهای ارگان و بافت‌های مرموز دارد. به عنوان مثال، در یک خروجی «ترس» (FEAR) که بررسی شد، تمپو حدود ۵۱ ضربه بر دقیقه و گام مینور تشخیص داده شد (طبق خروجی مدل، تعبیری معادل «دلهره‌آور» داشته است ولی حس ترس یا غم گاهی آمیخته با خشم است و سبب چالش است!). این نتایج با الگوهای شناخته شده‌ی موسیقی ترسناک همخوانی دارد.

ث) تعجب(Surprise): حالت «تعجب» اغلب در لحظات ناگهانی ملودی‌های ناگهانی و ضرب‌آهنگ‌های تغییرناپذیر بروز می‌کند. در نمونه‌های ما، موج‌های ضربه‌ای ناگهانی و ملودی‌های سینکوپ‌دار(سکته‌ای) مشاهده شد که انتظار ناگهانی را منتقل می‌کنند.

ج) خنثی(Neutral): موسیقی خنثی یا پس‌زمینه معمولاً تمپوی متوسط و حالتی بی‌طرف دارد. نمونه‌های خنثی تولیدشده تمپو و هارمونی معمولی داشتند و هیچ عنصر خاص احساسی برجسته‌ای در آنها دیده نشد.

این تحلیل‌های ساختاری نشان می‌دهد که مدل تا حدی توانسته است الگوهای موسیقی مرتبط با احساس‌های مختلف را تقلید کند؛ به‌عنوان مثال، خروجی‌های «شاد» سریع و ماژور بودند و خروجی‌های «غمگین» کند و مینور. با این حال، دسته‌بندی مدل صوتی ما در شناسایی احساس نیز چالش‌هایی داشت (مثلاً در مورد «ترس» اشتباهاً «خشم» تشخیص داده شد)، که نیاز به بهبود بیشتر است.


```
{
  "emotion": "FEAR",
  "style": "global",
  "structural": {
    "bpm": 51,
    "bars": 16,
    "note_density_per_bar": 16.81,
    "mode_match": 1.0,
    "tempo_fit": 1.0,
    "density_fit": 0.0,
    "dissonance_fit": 1.0,
    "syncopation_fit": 0.36,
    "overall_structural_score": 0.686,
    "build_params": {
      "emotion": "FEAR",
      "style": "global",
      "minutes": 1.0,
      "bpm": 51,
      "mode": "minor",
      "key_midi": 58,
      "progression_A": [
        "i",
        "bVI",
        "bVII",
        "i"
      ],
      "progression_B": [
        "i",
        "bVI",
        "bVII",
        "i"
      ],
      "drums": false,
      "comment": "دلهره‌آور",
      "bars_total": 16,
      "sr": 22050
    }
  },
  "audio_eval": {
    "predicted_emotion_audio": "ANGRY",
    "predicted_confidence_audio": 0.969,
    "target_emotion": "FEAR",
    "audio_match": 0.0
  },
  "final_score": 0.515,
  "detected_label": "FEAR",
  "detected_confidence": 0.504,
  "soundfont": "/content/FluidR3_GM.sf2"
}
```


شکل ۶-۴: نمای فایل JSON برای گزارش ارزیابی یک نمونه خروجی

۴-۴- جمع‌بندی

نتایج ارائه‌شده نشان می‌دهد که مدل پیشنهادی از نظر کمی در اغلب موارد، عملکرد قابل قبولی دارد:

امتیازهای مربوط به تطابق صوت-متن (CLAPScore) و بازشناسی متن با صوت (R@1) بالا بود، و فاصله‌های آماری FAD و KAD نسبتا پایین به‌دست آمد که نشانگر کیفیت خوب صدای تولیدی است. تحلیل ساختاری خروجی‌ها با الگوهای شناخته شده روان‌شناسی موسیقی (سرعت و گام گام) سازگار بود. به عنوان مثال، موسیقی با احساس شادی تمپوی تند داشت و موسیقی غمگین تمپوی کند بود، همان‌طور که در مطالعات پیشین گزارش شده است. در عین حال، برخی مغایرت‌ها (مثل تشخیص نادرست احساس در خروجی صوت) نیازمند توجه بیشتر در طراحی مدل و معیارهای ارزیابی است. به طور کلی، این نتایج نشان می‌دهد که رویکرد ما در جهت تولید موسیقی متناسب با احساس ورودی به طور تقریبی موفق بوده و معیارهای مختلف به شکل چندجانبه کیفیت تولیدات را تایید می‌کنند.

فصل پنجم
نتیجه‌گیری و پیشنهادها

در این پژوهش، یک مدل تولید موسیقی (مدل مولد صوتی) توسعه داده شد که عملکرد آن با معیارهای کمی و کیفی متعددی سنجیده شده است. برای ارزیابی کمی از سه معیار مرجع آزاد CLAPScore بر اساس یادگیری پیش‌آموزش‌دیده مشترک متن-صدا، FAD(فاصله فوریه صوتی) و معیار ذهنی MOS (امتیاز نظر میانگین شنوندگان) استفاده شد. معیار CLAPScore میزان «شباهت معنایی» بین موسیقی تولیدشده و متن راهنما با فایلی به‌نام متا را می‌سنجد [۲۵]، به طوری که عدد ۱ نشان‌دهنده تطابق معنایی کامل است. معیار FAD شاخصی مشابه Frechet Inception Distance برای صداست که تفاوت توزیع ویژگی‌های صداهای تولیدی و نمونه‌های مرجع را اندازه می‌گیرد[۲۷]. همچنین MOS یک عدد میانگین بر مقیاس ۱ تا ۵ است که کیفیت ادراکی صدا را از دید شنوندگان ارزیابی می‌کند[۲۹].

نتایج آزمایش‌ها نشان داد مدل پیشنهادی عملکرد بسیار خوبی دارد: مقدار متوسط CLAPScore مدل نزدیک به ۱ بود که بیانگر ارتباط معنایی قوی بین خروجی تولیدشده و پرسش متنی است. از سوی دیگر، مقدار FAD به میزان نسبتاً پایینی (نزدیک به صفر) رسید که نشان‌دهنده کیفیت طیفی بالای موسیقی تولیدشده و تطابق توزیع ویژگی‌های آن با نمونه‌های مرجع است. همچنین امتیاز میانگین نظر داوران (MOS) در محدوده‌ی «خوب تا عالی» (بیش از ۴ از ۵) قرار گرفت(البته در طی ارزیابی یک نمونه‌ی خاص). این نتایج حاکی از رضایت کلی شنوندگان و کیفیت ادراکی قابل قبول صدای تولیدشده است. علاوه بر این، هم‌راستا با روندهای بازارهای صنعتی، گزارش‌ها پیش‌بینی می‌کنند بازار جهانی موسیقی مولد مبتنی بر هوش مصنوعی از حدود ۵۶۹.۷ میلیون دلار در سال ۲۰۲۴ به حدود ۷۹۴/۲ میلیارد دلار در سال ۲۰۳۰ افزایش یابد[۳۰]؛ رقمی که نشان‌دهنده رشد سریع و تقاضای قابل توجه برای چنین فناوری‌هایی است.

عملکرد فنی مدل نیز قابل‌توجه بود: بهینه‌سازی ساختار شبکه عصبی و تنظیم دقیق ابرپارامترها منجر به تولید موسیقی‌هایی شد که علاوه بر شباهت محتوا به ورودی متنی، از نظر کیفیت صوتی نیز در دامنه‌ی بلندمدت، کمترین اعوجاج را داشت. برای مثال، افزایش درست عمق شبکه و طول ورودی مدل به ضبط بهتر

ریتم و هماهنگی سازها منجر شد. با این حال، محدودیت‌هایی نیز مشاهده گردید.

۱-۵- پیشنهادها

یکی از چالش‌های اساسی، دشواری مدل در ثبت وابستگی‌های بلندمدت موسیقی و حفظ انسجام ساختاری در قطعات طولانی است[۳۱]. به علاوه، مجموعه داده‌های آموزشی محدودیت‌های خاص خود را دارند؛ تنوع ژانرها و سازها ممکن است ناکافی باشد و تعمیم‌پذیری مدل به حوزه‌های صوتی جدید را کاهش دهد. شایان ذکر است که هر معیار ارزیابی نیز محدودیت‌هایی دارد: CLAPScore بر محتوا و معنی صوت تاکید دارد و ممکن است جزئیات صوتی دقیق را نشان ندهد، در حالی که FAD بیشتر بر شباهت طیفی تکیه می‌کند. نهایتاً، آزمون MOS برای ارزیابی ادراکی نیازمند آزمون‌های انسانی گسترده است که وقت‌گیر و پرهزینه می‌باشد. با وجود محدودیت‌های فوق، مدل توسعه‌یافته پتانسیل بالایی برای کاربردهای عملی دارد.

برای مثال، در حوزه‌ی **موسیقی‌درمانی** می‌توان از آن برای تولید موسیقی شخصی‌سازی‌شده بهره برد: پژوهش‌ها نشان داده‌اند موسیقی‌درمانی می‌تواند بر تنظیم هیجان، کاهش استرس و بهبود عملکرد شناختی تأثیر مثبت داشته باشد[۳۲]. ایجاد سیستم‌های هوش مصنوعی برای ارائه موسیقی‌درمانی در هر زمان (از قبیل سیستم «موسیقی‌درمانگر مجازی») می‌تواند دسترسی بیماران – به‌خصوص افراد مسن دارای زوال عقل – را افزایش دهد. همچنین، در **صنعت سرگرمی** (به‌ویژه بازی‌های ویدئویی و فیلم) این مدل قادر است موسیقی پس‌زمینه پویا و متناسب با صحنه را تولید کند؛ مثلاً مطالعات نشان داده‌اند موسیقی تعاملی که بر اساس رفتار یا حالت روانی بازیکن تنظیم می‌شود، به غوطه‌وری بیشتر کاربران منجر می‌شود. این ویژگی می‌تواند در ساخت جهان‌های مجازی غنی‌تر و تجربه‌های چندرسانه‌ای تعاملی مورد استفاده قرار گیرد[۳۱].

با توجه به تحلیل فنی بالا و کاربردهای بالقوه، پیشنهادها برای کارهای آینده به شرح زیرند[۳۱]:

- **گسترش مجموعه داده‌ها و ژانرها:** افزودن نمونه‌های متنوع‌تر از سبک‌های موسیقی، سازها و زبان‌های مختلف به مجموعه داده‌ی آموزشی مدل. این کار می‌تواند تنوع خروجی‌ها را افزایش داده و قابلیت تعمیم مدل به زمینه‌های گوناگون را تقویت کند
- **بهبود معماری مدل:** به‌کارگیری معماری‌های پیشرفته‌تر همچون Transformer و مدل‌های انتشار (Diffusion) یا ترکیبی از آنها به‌منظور ثبت بهتر وابستگی‌های زمانی بلندمدت و افزایش تنوع خلاقیت موسیقی تولیدی. همچنین استفاده از روش‌های یادگیری تقویتی یا ژنتیک برای تنظیم خودکار ابرپارامترها می‌تواند مفید باشد.
- **توسعه روش‌های ارزیابی دقیق‌تر:** علاوه بر معیارهای فعلی، پیشنهاد می‌شود معیارهای جدید یادگیری‌محور برای سنجش کیفیت موسیقی (مثلاً مبتنی بر شبکه‌های عصبی عمیق) توسعه یابند و آزمون‌های شنیداری گسترده‌تری با روش‌های استاندارد (ITU) انجام شود. این کار به اخذ بازخورد

جامع‌تر شنوندگان و مقایسه بهتر با آثار انسانی کمک می‌کند.

- **کاربرد در صنعت و نرم‌افزارهای تولید موسیقی:** پیاده‌سازی مدل در موتورهای بازی‌سازی (Unity, Unreal) یا نرم‌افزارهای تولید موسیقی به‌منظور تولید خودکار موسیقی پویا پیشنهاد می‌شود؛ چرا که مطالعات نشان می‌دهد موسیقی تولیدشده به‌کمک هوش مصنوعی می‌تواند کیفیت محتوای رسانه‌ای را به طور قابل‌توجهی افزایش دهد.
- **گسترش موسیقی‌درمانی مبتنی بر هوش مصنوعی :** توسعه برنامه‌های کاربردی موسیقی‌درمانی دیجیتال، مثلاً موسیقی‌درمانگر که بتواند موسیقی تسکین‌بخش را براساس وضعیت روحی یا فیزیولوژیکی فرد تنظیم کند. تحقیقات نشان داده‌اند چنین رویکردهایی می‌توانند استرس را کاهش داده و حافظه را تقویت کنند.
- **ادغام سامانه‌های بازخورد زیستی(AI[1]-driven biofeedback) :** استفاده از سنسورهای فیزیولوژیک (مانند ضربان قلب یا امواج مغزی) برای تنظیم پویا و بلادرنگ ویژگی‌های موسیقی تولیدشده. چارچوب‌های پیشرفته AI-driven biofeedback می‌توانند موسیقی را براساس داده‌های زیستی فرد بهینه کنند که این موضوع در درمان‌های غیر دارویی بسیار کاربردی است[۳۲].

در مجموع، مدلسازی دقیق و نتایج مطلوب این پروژه نشان می‌دهد که گسترش آن می‌تواند به تولید سیستم‌های کاربردی قدرتمند منجر شود. با بهره‌گیری از پیشنهادات فوق (افزایش تنوع داده‌ها، بهبود معماری، ارزیابی جامع‌تر و آزمایش‌های میدانی)، انتظار می‌رود عملکرد مدل ارتقاء یافته و دامنه‌ی کاربردهای صنعتی و درمانی آن گسترش یابد.

[1] Artificial intelligence

پیوست ۱: لیست برنامه‌ها

۱-پ- دسترسی به کدها

در آدرس زیر می‌توانید به کد و تمامی منابع استفاده‌شده در این پروژه دسترسی پیدا کنید:

https://github.com/M-Amin-Kiani/bachelor-s-dissertation

برای اجرای مدل نیز می‌توانید از لینک زیر استفاده کنید:

https://colab.research.google.com/drive/1J1gAahJ80HdsHKEw3SpFJ8QA-LJrvRF3?usp=sharing

۲-پ- مروری بر پیشینه‌ی پژوهشی پروژه

در آدرس زیر به بررسی سوابق پژوهش‌های انجام‌شده می‌پردازد(مقاله‌ی مروری):

https://github.com/M-Amin-Kiani/bachelor-s-dissertation/blob/main/My%20Research/MohammadAmin_Kiani_ReviewForProj.pdf

۳-پ- پارامترهای فاین‌تیون مدل برای متن فارسی [1]

```
num_train_epochs=3,
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
learning_rate=2e-5,
weight_decay=0.01,

# لاگ‌گیری
logging_dir="./logs",
logging_steps=100,  # یک لاگ هر ۱۰۰ مرحله
```

[1] HooshvareLab/bert-fa-base-uncased

```
# فعال‌سازی ارزیابی و ذخیره‌سازی در طول آموزش
do_train=True,
do_eval=True,
eval_steps=500,      # ارزیابی هر ۵۰۰ مرحله
save_steps=500,      # ذخیره checkpoint برای هر ۵۰۰ تا مرحله
save_total_limit=2,  # نگه داشتن checkpoint حداکثر ۲
```

۴-پ- پارامترهای فاین‌تیون مدل برای متن انگلیسی [1]

متن انگلیسی را نیز همانند متن فارسی اما با دیتاست GoEmotions [2]فاین‌تیون کردیم اما نتایج بهبود نیافت و همان مدل اولیه برای متن‌های انگلیسی نتایج بهتری داشت و نیازی به انجام این کار نبود:

```
num_train_epochs=3,
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
learning_rate=2e-5,
weight_decay=0.01,
logging_dir="./logs_go",
logging_steps=500,
do_train=True,
do_eval=True,
eval_steps=1000,
save_steps=1000,
save_total_limit=2,
report_to="none",
run_name=None,
```

۵-پ- پارامترهای فاین‌تیون مدل برای تصویر [3]

تصویر را نیز همانند متن اما با دیتاست fer-2013 [4] فاین‌تیون کردیم اما به علت محدودیت در منابع سخت‌افزاری و کرش کردن گوگل کولب، همان مدل آماده[5] را استفاده کردیم که نتایج دقیقی داشت و نیازی به انجام این کار نبود:

---

[1] roberta-base
[2] https://github.com/google-research/google-research/tree/master/goemotions
[3] WinKawaks/vit-tiny-patch16-224
[4] https://huggingface.co/datasets/Jeneral/fer-2013
[5] https://github.com/oarriaga/face_classification/raw/master/trained_models/emotion_models/fer2013_mini_XCEPTION.102-0.66.hdf5

```
    evaluation_strategy="epoch,"
    save_strategy="epoch,"
    num_train_epochs=5,

    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,

    gradient_accumulation_steps=4,      # ← effective batch = 4×4 =16
    gradient_checkpointing=True,        # ← reduce memory for activations

    learning_rate=2e-4,
    weight_decay=0.0,

    load_best_model_at_end=True,
    metric_for_best_model="accuracy,"

    fp16=torch.cuda.is_available(),     # ← mixed precision
    dataloader_num_workers=2,
    logging_steps=100
```

تذکر: به علت محدودیت‌های شدید سخت‌افزاری و امکان‌پذیر نبودن آموزش محلی یا ابری برای سیگنال، در خصوص تحلیل صوت از مدل آماده‌ی wav2vec2[1] استفاده شد.

[1] https://huggingface.co/superb/wav2vec2-base-superb-er

منابع:

[1] Https://en.wikipedia.org/wiki/Music_theory

[2] Https://en.wikipedia.org/wiki/Recurrent_neural_network

[3] Https://web.stanford.edu/~jurafsky/slp3/8.pdf

[4] Https://en.wikipedia.org/wiki/BERT_(language_model)

[5] Https://hooshvare.github.io/docs/research/parsbert

[6] Https://learnopencv.com/facial-emotion-recognition/

[7] Https://pmc.ncbi.nlm.nih.gov/articles/PMC9050748

[8] Https://huggingface.co/facebook/wav2vec2-base-960h

[9] Https://www.izotope.com/en/learn/understanding-spectrograms

[10] Leonardo Pepino, Pablo Riera, Luciana Ferrer, “Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings.” https://arxiv.org/pdf/2104.03502.

[11] Https://en.wikipedia.org/wiki/MIDI

[12] Https://en.wikipedia.org/wiki/Equal_temperament

[13] Https://en.wikipedia.org/wiki/Scale_(music)

[14] Https://en.wikipedia.org/wiki/Chord_(music)

[15] Https://en.wikipedia.org/wiki/Time_signature

[16] Https://midimusic.github.io/tech/midispec.html

[17] Mohammad Ali Hussiny, Mohammad Arif Payenda, lilja Øvrelid, “PersianEmo: Enhancing Farsi-Dari Emotion Analysis with a Hybrid Transformer and Recurrent Neural Network Model.” https://aclanthology.org/2024.sigul-1.31.pdf.

[18] Tanisha Hisariya, Huan Zhang, Jinhua Liang, “Bridging Paintings and Music – Exploring Emotion based Music Generation through Paintings.” https://arxiv.org/pdf/2409.07827v1.

[19] Https://neptune.ai/blog/ml-pipeline-architecture-design-patterns

[20] Https://en.wikipedia.org/wiki/SoundFont

[21] Https://www.gradio.app/guides/quickstart

[22] Https://flask.palletsprojects.com/en/stable/

[23] Https://wavesurfer.xyz/

[24] Https://pyngrok.readthedocs.io/en/latest/

[25] Feiyang Xiao1, Jian Guan1∗ , Qiaoxi Zhu2, Xubo Liu3, Wenbo Wang4, Shuhan Qi5, Kejia Zhang1, Jianyuan Sun6, and Wenwu Wang, “A Reference-free Metric for Language-Queried Audio Source Separation using Contrastive Language-Audio Pretraining.” https://arxiv.org/pdf/2407.04936.

[26] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, Huaming Wang, “CLAP: Learning Audio Concepts From Natural Language Supervision.” https://arxiv.org/pdf/2206.04769.

[27] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, Matthew Sharifi, “Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms.” https://arxiv.org/pdf/1812.08466.

[28] Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, Ben Sangbae Chon, “KAD: No More FAD! An Effective and Efficient Evaluation Metric for Audio Generation.” https://arxiv.org/pdf/2502.15602.

[29] Https://en.wikipedia.org/wiki/Mean_opinion_score

[30] Https://www.grandviewresearch.com/horizon/outlook/generative-ai-in-music-market-size/global

[31] Yanxu Chen, Linshu Huang, Tian Gou, “Applications and Advances of Artificial Intelligence in Music Generation:A Review.” http://arxiv.org/pdf/2409.03715v1.

[32] Https://pmc.ncbi.nlm.nih.gov/articles/PMC11893577