بسم الله الرحمن الرحیم

# گزارش سوم پروژه "استخراج کلمات کلیدی از منابع صوتی و ویدئویی با استفاده از هوش مصنوعی"

فاطمه وحیدیونسی

مرکز نوآوری علوم و فناوری های شناختی

دانشگاه علم و صنعت ایران

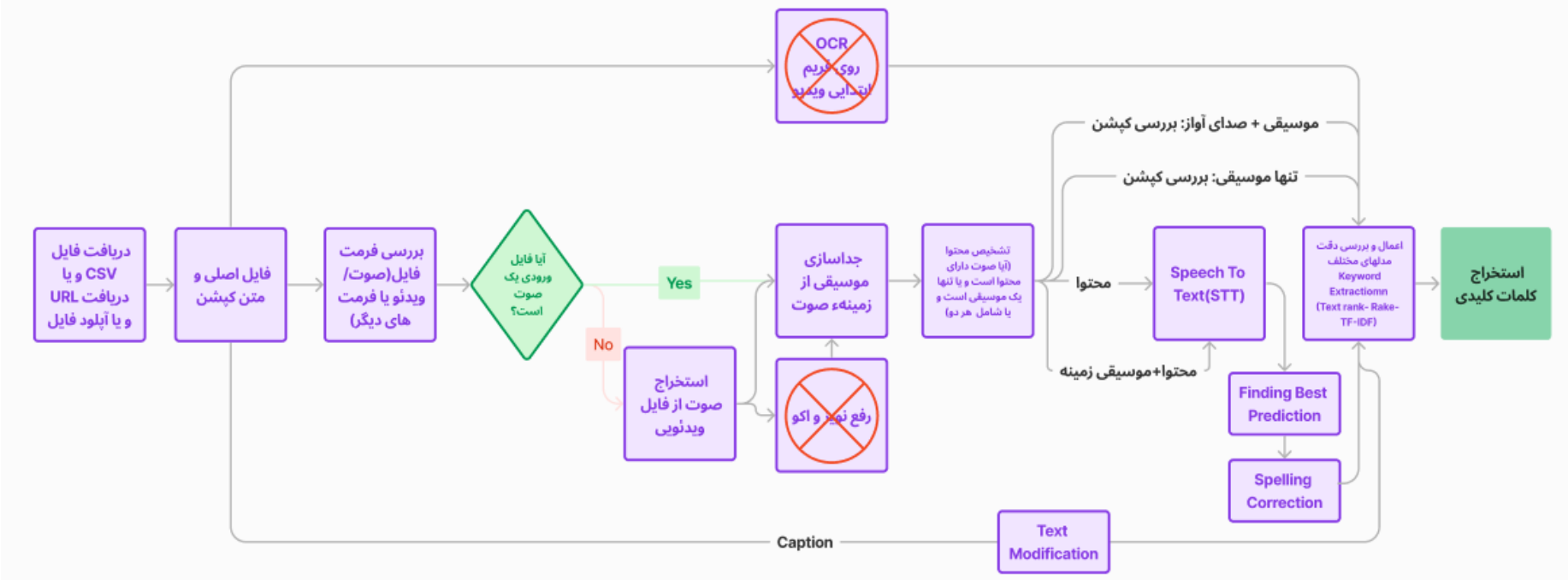
تاریخ تدوین: 1402/11/24

## 1- هدف پروژه:

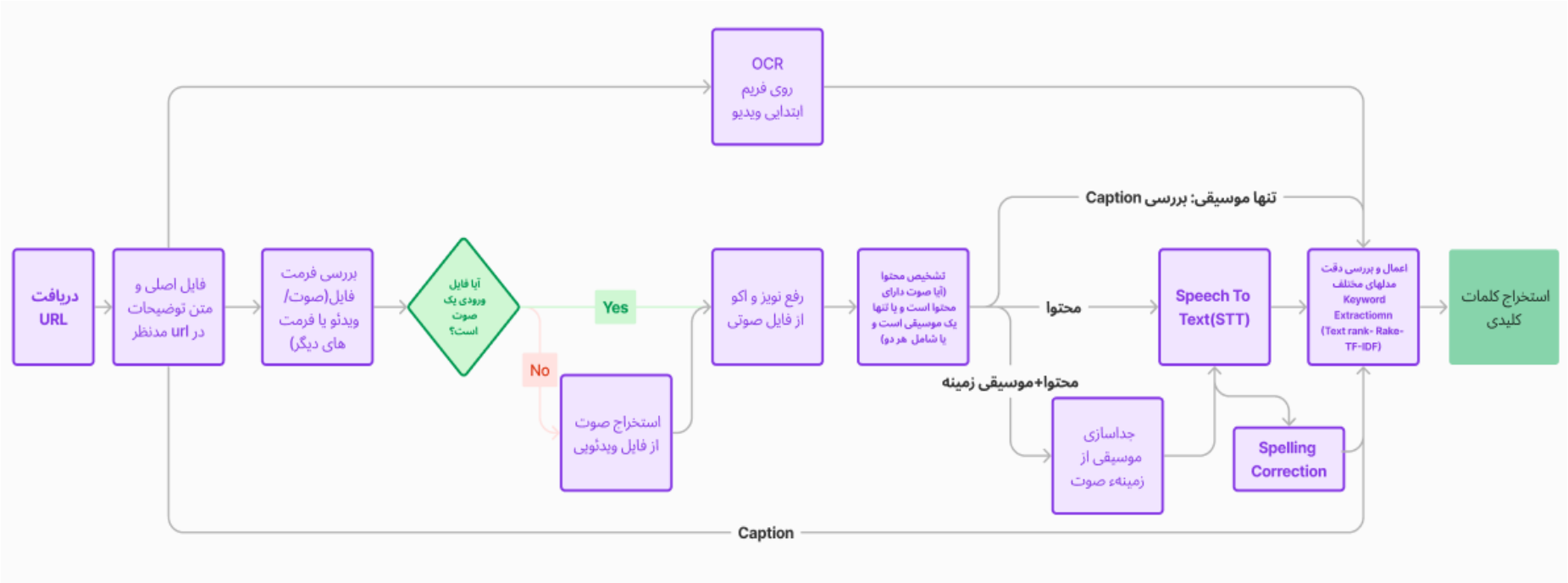
همانگونه که در گزارش های اول و دوم بیان شد، هدف از این پروژه ساخت یک فریمورک از ماژول های مختلف جهت رسیدن به مقصود "استخراج کلمات کلیدی از فایل های صوتی و ویدئویی" است. می توان شبکه های اجتماعی ای مانند اینستاگرام را یکی از این منابع به حساب آورد. توضیحات زیر جهت تقریب به ذهن برای پست های این شبکه اجتماعی بسط داده می شود.

## 2- فرآیند نهایی پروژه:

فلوچارت نهایی(**مرتبط با گزارش سوم**) قرار داده شده در زیر، مراحل انجام پروژه را نشان می دهد:




*نمودار1- فلوچارت نهایی مراحل پروژه*




*نمودار2- فلوچارت سابق مراحل پروژه **(مربوط به گزارش دوم)***

همانگونه که در مقایسه دو نمودار بالا دیده می شود، نسبت به گزارش قبلی، چند مرحله به فرآیند پروژه جهت افزایش دقت خروجی، اضافه شد. از جمله، در مهمترین مرحله در این فرآیند، مرحله STT، از ادغام دو مدل جهت افزایش دقت تبدیل صوت به متن استفاده شد. مدل های استفاده شده، مدل های Medium و Large از پروژه Whisper شرکت Google هستند. مدل های دیگر توسعه داده شده مانند مدل Hezar که مدل Fine-Tuned از مدل Whisper است و همچنین مدل های Vosk و Deep Speech دارای دقت های کافی برای عملیات STT در زبان فارسی نیستند. مدل Large از پروژه Whisper دارای دقت بیشتری نسبت به

مدل Medium است اما در برخی فایل های صوتی، به عنوان مثال فایل های دارای اکو، مدل Medium خروجی دقیق تری می دهد. در نتیجه تابعی طراحی شد تا بهترین پیش بینی STT را برای هر زوج زمانی مشابه، از میان این دو مدل شناسایی نموده و به عنوان خروجی تحویل دهد.

چالش مهم این پروژه، به خصوص در استفاده از منابعی مانند شبکه های اجتماعی از جمله اینستاگرام، وجود موزیک در اغلب ویدئوها در Post هاست. وجود موزیک در پس زمینهء صدای راوی در یک فایل صوتی می تواند بر روی عملکرد عملیات STT تاثیر گذار باشد. برای رفع این مسئله از مدل Demucs استفاده شد. اما چالش مهمتر از جداسازی موزیک از بکگراند فایل صوتی، تشخیص محتوای مفید یا غیرمفید در این فایل هاست.

فایل هایی که تنها دارای موزیک بی کلام و یا موزیک همراه با آواز هستند دارای محتوای مفیدی جهت استخراج کلمات کلیدی نبوده و از متن کپشن در این فایل ها می بایستی استفاده کرد. در مقابل، فایل هایی که فاقد موزیک بوده و یا دارای موزیک زمینه به همراه صدای یک راوی می باشند، فایل هایی هستند که محتوای آنها باید مفید تشخیص داده شده و جهت استخراج کلمات کلیدی مفید ارسال شوند. بنابراین در اینجا چالش اصلی، تشخیص فایل های دارای محتوای مفید از فایل های غیرضروری بود. طبق تحقیقات انجام شده در Literature و مقالات مرتبط، مدلی جهت تشخیص "آواز+آهنگ" از "صدای راوی+آهنگ زمینه" تولید نشده است. حتی مدلی جهت تشخیص یک شعر از یک نثر در زبان فارسی وجود ندارد. بنابراین جهت مواجهه با این چالش دو راه حل پیش رو بود:

اول، اینکه از ابتدا یک مدل Classification با استفاده از شبکه های عصبی بر روی نثر و شعر انجام گیرد.

دوم، که راه حلی با سرعت بالاتر و دقت بهتر است و نیاز به دیتا و سخت افزار، آنچنان که در راه اول نیاز است، ندارد، اینکه ویژگی های اکوستیکی متفاوت در این دو نوع محتوا، شناسایی شود و سپس با استفاده از این ویژگی ها یک فایل دارای محتوا، از یک آواز، تمیز داده شود.

در این پروژه، همانطور که گفته شد، به علت سرعت بالاتر در خروجی و دقت مناسب، از راه دوم استفاده شد.

جهت استخراج کلمات کلیدی، دقت نتایج حاصل از API معرفی شده، با دو مدل(Topic Rank و Text Rank) که شخصا پیاده سازی شد، بررسی شد. نتیجه اینکه دقت مدل Topic Rank بالاتر از دقت API شناسایی شده و از این مدل استفاده می شود.

تا به اینجا در رابطه با استخراج کلمات کلیدی از سورس صوتی صحبت شد. یکی دیگر از سورس های در اختیار، در ورودی های ویدئویی، متن های موجود در تصویر است. از مدل های Persian-OCR ، Hezar، video_ocr و Tesseract جهت تشخیص متن در فریم اول هر ویدیو(که معمولا شامل متن یا عنوان است) استفاده شد. نتیجه اینکه دقت پایین تشخیص این مدل ها در زبان فارسی، در کنار وجود عبارات انگلیسی(مانند آیدی و آدرس پیج) و متن های غیرمرتبط با موضوع ویدئو کیفیت کلمات کلیدی شناخته شده را پایین می آورد. بنابراین حداقل درمورد ویدئوهای موجود در شبکه های اجتماعی استفاده از مدل های OCR پیشنهاد نمی شود. اما درمورد ویدئوهای آموزشی(مانند ویدئوهای موجود در سایت هایی مانند مکتب خونه، فرانش، فرادرس یا سایت های مشابه خارجی) استفاده از OCR در تبدیل متن اسلایدهای آموزشی بسیار کمک کننده است.

در نهایت، مراحل نهایی پروژه، به ترتیب، همراه با جزئیات در زیر لیست شده اند. جهت درک بهتر، این مراحل، برای نمونهء استخراج کلمات کلیدی از پست های اینستاگرام تشریح شده اند(مراحل، متناسب با فلوچارت نهایی، تغییر نموده اند):

1- فایل اصلی پست شده به همراه توضیحات (Caption) موجود در پست دریافت می شود. در این پروژه از هر دو سورس(متن توضیحات کپشن و متن ناشی از تبدیل صوت(STT) جهت استخراج دقیق تر کلمات کلیدی استفاده خواهد شد).
2- فایل ورودی از جهت فرمت، بررسی شده، در صورتی که فایل ورودی ویدئویی داشته باشد، مرحلهء استخراج صوت از ویدئو اجرا می شود.
3- استفاده از مدل Demucs جهت جداسازی موزیک از زمینهء فایل صوتی. ذخیره محتوای اصلی جهت آماده سازی برای مرحلهء STT .
4- تشخیص "محتوای" فایل صوتی: درصورتیکه فایل تنها یک موسیقی است و فاقد محتواست، وارد مرحله استخراج کلمات کلیدی از سورس دیگر، متن caption ، می شویم. درصورتیکه صدای یک گوینده به همراه موسیقی بی کلام در بکگراند است، عملیات مرحله 5 ام آغاز می شود.
5- استفاده از متد های مختلف STT جهت تبدیل صدای نهایی به متن، ادغام متن حاصل از پیاده سازی دو مدل Whisper
6- اصلاح خطاهای املایی حاصل از STT
7- اصلاح متن کپشن از نظر حذف ایموجی ها، آیدی و علامت هشتگ
8- استفاده از متد Topic Rankجهت Keyword Extraction از ادغام متن Caption و متن حاصل از عملیات STT بر روی صوت

9- دررابطه با فایل هایی که دارای دو پارت موسیقیایی و پارت محتوایی، در زمان هایی متفاوت، هستند: تشخیص این پارت ها از یکدیگر، زمان بندی هر یک از آنها و تشخیص بازه زمانی که فایل صوتی دارای محتواست.

در زیر، جدول **وضعیت اجرای مراحل** فوق قرار داده شده است:

| مرحله | وضعیت |
|---|---|
| **1- دریافت فایل اصلی پست شده به همراه توضیحات (Caption) موجود در پست** | UI آپلود فایل انجام شده<br>همچنین امکان ورودی دادن فایل CSV یا آیدی اینستاگرام وجود دارد. |
| **2- بررسی فرمت فایل ورودی، در صورتی که فایل ورودی ویدئویی داشته باشد، استخراج صوت** | انجام شده |
| **کاهش نویز، اکو و صداهای زمینه** | انجام شده(بدلیل کاهش دقت خروجی از مراحل نهایی حذف شد)<br>در این مرحله با استفاده از ماژول voicefixer نویز و reverb در زمینه برخی فایل های ورودی حذف شد. |
| **3- درصورتیکه در پس زمینهء صدا، موسیقی وجود دارد، حذف موسیقی از بکگراند و ذخیره محتوای اصلی جهت STT** | انجام شده<br>این عملیات در دو مرحله انجام می پذیرد:<br>ابتدا با استفاده از مدل Htdemucs نت های موسیقیایی از زمینه صدای اصلی(vocal) حذف شده و سپس از طریق تابع طراحی شده، محتوای فایل ویدیویی/صوتی بررسی می شود. خروجی این فرآیند تشخیص 4 حالت در صداست:<br>1-صوت فاقد هرگونه موسیقی است و تماما دارای محتواست.<br>2- صوت دارای موسیقی در زمینه و محتوای گفتاری یک راوی است.<br>3- صوت تماما موسیقی و فاقد محتوای گفتار انسان است.<br>4- صوت یک آهنگ با صدای یک خواننده است.<br>این مرحله در ادامهء مراحل پروژه بسیار تعیین کننده است چرا که می توان اینگونه تعریف کرد که درصورتیکه صوت یکی از حالات 3 یا 4 را داشته باشد، فاقد محتوا تشخیص داده شود و نیازی به طی مراحل STT و اصلاح نگارش و استخراج کلمات کلیدی از |

| | |
|---|---|
| | صوت نباشد و یا محتوای آن دارای ارزش پایین تری به نسبت محتوای Caption باشد. |
| **4- تشخیص "محتوای" فایل صوتی(موسیقی باکلام است یا موسیقی بی کلام در زمینه به همراه محتوای اصلی توسط یک گوینده است.)** | انجام شده<br>بررسی منابع و مدل ها، در زمینه تشخیص محتوای موسیقیایی از محتوای عادی انجام شد، مدلی برای تشخیص یک "آهنگ" از یک "موسیقی بی کلام که صداگذاری" شده است تاکنون طراحی نشده است. کدی توسعه داده شد که ازطریق ویژگی های صوتی موجود در فایل های ورودی، با دقت نزدیک به 90٪ می تواند نوع محتوای فایل را شناسایی کند. |
| **5- استفاده از متد Whisper STT جهت تبدیل صدای نهایی به متن** | انجام شده<br>مدل های STT موجود در زبان فارسی با استفاده از کتابخانه های Whisper ، Vosk، Deep speech و Hezar(مدل Whisper Fine-tuned) بررسی شد، مدل whisper medium در برخی متون و مدل whisper large نیز در برخی متون نتایج بهتری را خروجی می دهند. به همین سبب کدی برنامه نویسی شد که بهترین خروجی را به ازای هر بازه زمانی، از هر دو مدل دریافت کرده و بهترین آن را به عنوان خروجی نشان دهد. |
| **6- اصلاح خطاهای املایی حاصل از STT** | کتابخانه های Parsivar، Kenlm و Faspell جهت ارزیابی خطاهای احتمالی موجود در خروجی مدل STT بررسی شد. در نهایت با استفاده از کتابخانه Parsivar متن نهایی آماده ورود به مرحلهء استخراج کلمات کلیدی می گردد. |
| **7- اصلاح متن کپشن از نظر حذف ایموجی ها، آیدی و علامت هشتگ** | انجام شده<br>تابع pure_caption جهت آماده سازی متن کپشن برای استخراج کلمات کلیدی طراحی شد. |
| **8- استفاده از متد Topic Rank جهت Keyword Extraction از متن ادغامی حاصل از Caption و متن حاصل از STT** | API استخراج کلمات کلیدی بررسی شد. چند متد نیز( Text-Rank, Topic Rank) شخصا پیاده سازی شد. دقت خروجی متد Topic Rank بیشتر است و از این روش استفاده می شود. |
| **9- دررابطه با فایل هایی که دارای دو پارت موسیقیایی و پارت محتوایی، در زمان هایی متفاوت، هستند: تشخیص این پارت ها از یکدیگر، زمان بندی هر یک از آنها و تشخیص بازه زمانی که فایل صوتی دارای محتواست.** | در دست انجام |

## 3-نتیجه گیری و توسعه های آتی:

همانطور که در جدول فوق نشان داده شد، تا تاریخ 24 ام بهمن، یک نمونه کامل از صفر تا صد مراحل پروژه پیاده سازی شده است. تنها مرحله نهم که شناسایی موزیک در "بخشی" از فایل های صوتی است، باقی مانده است که درصورت نیاز، فعالیت بر روی آن، می تواند به افزایش دقت شناسایی محتوای موسیقیایی فایل های صوتی/ویدئویی و در نهایت هدایت به بخش استخراج کلمات کلیدی کمک کننده باشد. درصورت نیاز به ارتقای بیشتر در مراحل پروژه، می توان بر روی این مرحله، توسعه انجام داد.

همچنین این موضوع، یعنی " شناسایی محتوای موسیقیایی فایل های صوتی/ویدئویی" می تواند به طور جداگانه ذیل یک پروژه معرفی شده و به عنوان یک API در کنار سایر سرویس ها در پروژه های مرتبط، مورد استفاده قرار گیرد.