Problem Statement:

The main issue our project aims to target is emergency and disaster situations, where multiple distressed individuals often are calling at once, making it difficult for responders to quickly assess who is in the most critical condition. Current 911 and emergency response systems lack automated methods to analyze tone and semantic meaning of speech to quantify human distress in real time. With our project, we aim to address this challenge by analyzing various features in a dataset on the tone and emotions of the person speaking (if they are screaming, crying, or talking very rapidly) and the semantic meaning of the words they are saying (classifying specific words such as "murder" or "shooting" with higher priority). This system in place can help emergency services identify the most critical situations much faster, improving both response times and potentially saving numerous lives. Beyond this, first responders can better allocate limited resources, make informed triage decisions, and reduce the cognitive burden on human dispatchers.

Strategic Aspects/Relation to Lectures:

There are a few strategic aspects involving our project, in terms of storing a large amount of audio files, parsing through each of them to analyze tone and semantic meaning, and training a machine learning algorithm based on it.

-Storage: There are over 700+ audio files in the original dataset that need to be trained on. In order to efficiently manage and process these files, we would have to use PostgreSQL database for metadata such as the file_id, duration, and emotional characteristics of the audio. An emotional severity score will be computed by classifying emotional severity of a call audio file, which will be done with machine learning.

-Speech Tone and Emotional Speech Recognition: In order to properly classify complex emotion and tone from speech, we are using the Kaggle RAVDESS emotional speech audio dataset which contains files with different emotional states such as calm, happy, sad, angry, fearful, surprise, and disgust expressions. This project will strategically use this dataset to train a model that will be able to accurately identify emotional intensity on the dataset with 911 calls. We will supplement this model with call transcription analysis as well, to evaluate the severity of the emergency being described.

-Ethical and Privacy Consideration: There are many ethical concerns involved in autonomously giving a distress severity score to an individual call, such as potentially disregarding an emergency because the model misclassified the call with a lower score. The project must avoid bias towards certain accents or speech patterns and it clearly communicates the model limitations as more of a suggestion for the first responder rather than used for a final decision.

Relevance to this Course

The aspects above relate to this course in that we are working with large datasets and applying preprocessing and machine learning techniques to create predictions that drive decision making. We also implement concepts similar to sentiment analysis, in that we evaluate the content of calls as transcripts, and identify the severity of the situation that is being described.

Novelty and Importance

There are many cases that exist in which a first responder was not able to get to an emergency on time, especially during natural disasters, violent incidents, or large-scale emergencies where they receive an overwhelming amount of distress calls. Manually identifying which calls are in the most immediate danger can be too time consuming and is subject to human error or cognitive distress at that moment. By incorporating our automated system as an assistant for 911 dispatchers and first responders, it will facilitate final decision-making by analyzing both emotion and tone of the caller's voice along with semantic meaning of their words to create a final distress severity score.

As this is a critical and contemporary issue, we are interested in applying machine learning and data management techniques to develop a system that can automatically detect and evaluate various levels of human distress in emergency situations. By combining natural language processing, audio analysis, and structured data integration, we aim to use what we have learned in the class so far to create a reliable model that supports faster decision making. Regarding existing issues in current data science practices, machine learning models that involve supervised learning involve labeled data, however labeling human emotions or distress levels can be very subjective and vague. Therefore clustering or unsupervised learning is commonly used, however this can create too many or too little groups, leading to inconsistencies in how data is categorized. Additionally, many datasets do not include diverse voices, accents, and languages, so the model may perform better for some groups than others. By using public emotional datasets (RAVDESS) that cover multiple voices and emotions, we allow our model to be receptive to a wider variety of calls and interactions. Finally, there are prior related works such as a research paper published at Cornell University which utilized end-to-end deep learning to train in the wild datasets versus acted data on identifying certain human emotions such as anger or relief (Deshamps-Berger et al, 2021). Unlike this prior work, which focused solely on recognizing emotions from single-speaker emergency calls, our project also incorporates semantic features and event metadata to produce a continuous distress severity score, for real emergency call data.

Plan

Data Utilized:

- We will utilize the RAVDESS dataset, which we sourced from Kaggle.

Our plan is to build our models and functionalities into a web application as outlined below:

Web Application Stack:

1. React + Next.js frontend (JavaScript, HTML, CSS)
2. AWS S3 bucket for audio files
3. PostgreSQL database hosted on AWS RDS for audio features
4. FastAPI backend (Python) to serve models and query databases

We define the model, implementation steps, and evaluation process below.

Audio Pre-Processing Pipeline:

1. Process RAVDESS Kaggle dataset → volume normalization, denoising (make voices clearer), audio feature extraction for Random Forests

2. Extract numeric emotion-related information from filenames (ex. 03-01-01-01-01-02-01.wav)
3. Combine filename features into a single categorical column “audio category” using defined bins
4. Python packages: Librosa, NumPy, Pandas, SciPy, pydub

Model Training & Evaluation:

1. Train Random Forest and/or CNN using “audio category” column as a predictor
2. Utilize 80-20 train-test split in data
3. k-fold cross-validation using 5 folds to verify model is not overfit
4. Evaluate with ROC AUC and other metrics
5. Will utilize CNNs trained on audio spectrograms if Random Forest ROC AUC < 0.85 (more performant but more computationally expensive → Needs more time to train)
6. Map classifications to numerical distress/urgency scores

Transcript and Context Analysis:

1. Transcribe incoming audios using OpenAI Whisper Python package
2. Utilize ChatGPT-4o powered AI agent to extract key emergency information
3. Feed into open-source conflict severity package for contextual severity

Severity Score:

1. Combine vocal emotional-severity score and contextual severity score into a unified severity score
2. Used for comparing and prioritizing emergencies

Frontend-Backend Interaction/APIs:

1. FastAPI APIs for retrieving current emergencies and processing new call recordings
2. Store and query data using SQLAlchemy with PostgreSQL tables

Deployment:

1. Deploy application using Vercel

Note:

Within this project, we only formally evaluate performance of the model we trained (Random Forest/CNN), as the conflict severity package we utilize is already evaluated to be accurate. We will, however, informally test how functional the overall application is by utilizing an additional Kaggle dataset of 911 call snippets, as well as our own recordings of sample calls. We will evaluate whether the scores calculated for these calls reflect the actual content and audible emotion in audio, by seeing if it can accurately rank the calls based on overall emergency severity.

Citation

- Deschamps-Berger, Théo, et al. “End-To-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings.” ArXiv.org, 2021, arxiv.org/abs/2110.14957.