Przedstawienie tematu

1. **SpeechMultiEmo** - rozpoznawanie emocji na podstawie
    a. Próbki głosu rozmówcy
    b. Nagrania video rozmówcy

2. Temat jest ciekawy ponieważ realizujemy go jako dodatek do projektu naukowo wdrożeniowego (DubAI - generowanie polskiego dubbingu do gier z angielską ścieżką), którego tematyka pozostaje w kręgu naszych zainteresowań. Odpowiednia detekcja emocji jest niezbędnym elementem potrzebnym do wygenerowania dubbingu adekwatnego do sytuacji. Dzięki poprawnemu transferowi nacechowania emocjonalnego wypowiedź jest przyjemniejsza w odsłuchu.

3. Chcielibyśmy się nauczyć metod analizy oraz przetwarzania danych w tematach związanych z informatyką afektywną, ze szczególnym naciskiem na dane audio. Chcielibyśmy również sprawdzić wpływ dodatkowej modalności w postaci nagrania video na skuteczność odczytywania emocji w wypowiedzi, co mogłoby wskazać, czy warto rozważyć poszerzenie PNW o nagrania video.

Opis

1. Zbiory
    - CREMA-D: multimodalny zestaw danych emocjonalnych, zawierający 7 442 klipy nagrane przez 91 aktorów (48 mężczyzn i 43 kobiety) w wieku 20–74 lat, reprezentujących różne grupy etniczne. Aktorzy wypowiadali 12 zdań, wyrażając sześć różnych emocji: złość, obrzydzenie, strach, radość, neutralność i smutek. Określone są również stopnie zadanych emocji.
        - video
        - audio
    - IEMOCAP (*) zbiór danych zawiera 302 filmy z dialogami, w których uczestniczą dwaj mówcy. Każdy segment jest oznaczony pod kątem obecności 9 emocji (złość, ekscytacja, strach, smutek, zaskoczenie, frustracja, radość, rozczarowanie, neutralność), a także walencji, pobudzenia i dominacji, a nagrania pochodzą z 5 sesji z 5 parami mówców.
      * - jeżeli uda nam się go ZDOBYĆ (wymaga złożenia podania o uzyskanie)

2. Rozwiązywany problem: określenie, czy dodanie dodatkowej modalności poprawia skuteczność rozpoznawania emocji w porównaniu do modeli wykorzystujących tylko jedną modalność.

3. Planowane podejście do rozwiązania problemu:
    - Planowane analizy:

Porównanie działania rozpoznawania emocji ze względu na rozważane modalności, w tym: tylko audio, tylko video oraz połączenie z wczesną i późną fuzją

- Modele:
  - https://huggingface.co/dima806/facial_emotions_image_detection
  - https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition
  - https://huggingface.co/trpakov/vit-face-expression

- Strategia walidacji:
  Planujemy przeprowadzić walidację krzyżową typu k-fold (z podziałem mówców) z uwagi na charakter zbioru (CREMA-D) w celu uniknięcia wycieku danych.
- Metryki:
  - F1-score
  - Accuracy
  - Recall

4. Planowane porównania: porównanie zaproponowanych przez nas rozwiązań dla różnych kombinacji modalności między sobą

Harmonogram prac

1. **Przegląd literatury** (obecnie – 31.12.24)
   Pierwszym zadaniem będzie przegląd literatury i publikacji naukowych w celu zapoznania się z tematyką detekcji emocji na danych audio i video. Chcielibyśmy zrobić research modeli SOTA do detekcji emocji, technik fuzji obu modalności, oraz metryk do ewaluacji. Efektem końcowym będzie stworzenie zestawienia modeli i algorytmów, które będą podstawą do stworzenia ich benchmarku.
2. **Analiza eksploracyjna zbioru danych** (01.01.25 – 13.01.25)
   Dla lepszego zrozumienia charakteru posiadanego zbioru danych i możliwości dogłębniejszej interpretacji przyszłych wyników przeprowadzimy analizę posiadanego zbioru danych. W ramach analizy policzymy rozkłady zadanych emocji i ich intensywności oraz zwizualizować reprezentacje modalności.
3. **Implementacja / wykorzystanie gotowych modeli. Implementacja mechanizmów fuzji, oraz metryk** (01.01.25 – 20.01.25)
   Kolejnym etapem naszego projektu będzie implementacja modeli wybranych po przeglądzie literatury, bądź wykorzystanie gotowych rozwiązań (np. na platformie *Huggingface*). Podobnie w przypadku mechanizmów fuzji, oraz metryk do ewaluacji. W przypadku implementacji modeli od podstaw niezbędne będzie ich wytrenowanie na wcześniej zdefiniowanych zbiorach danych. W przypadku wykorzystania gotowych implementacji nie będzie to konieczne, jednak nie wykluczamy opcji ich finetuningu.
4. **Walidacja, oraz ewaluacja modeli** (20.01.25 – 03.02.25)

Przetestowanie jakie wyniki osiągają obie modalności osobno, jak i razem z wyszczególnieniem wczesnej i późnej fuzji.

5. **Wnioski** (01.02.25 – 03.02.25)
   Przygotowanie wniosków podsumowujących wszystkie etapy projektu – eksploracyjnej analizy danych, oraz porównanie wyników walidacji i ewaluacji poszczególnych wariantów modeli.
6. **Obrona projektu** (04.02.25)

Kamienie milowe

1. **Eksploracyjna analiza danych** (13.01.25) – efektem będzie przygotowany plik (jupyter notebook) zawierający EDA. Szerszy opis zakresu EDA jest zawarty w punkcie 2 harmonogramu prac.
2. **Ukończenie pipeline'a** (01.02.25) – rezultatem będzie jednolity pipeline realizujący cel naszego projektu. Chcemy, aby się składał z następujących części
   a. Załadowanie danych
   b. Preprocessing danych
   c. Walidacja krzyżowa modeli
   d. Końcowa ewaluacja modeli
3. **Prezentacja wyników końcowych** (04.02.25) – wynikiem końcowym będzie prezentacja rezultatów naszego projektu na ostatnich zajęciach.