Explaining Multimodal Emotion Recognition: How Do AI Models Balance Content and Tone?

이지호 이상엽

연구 목표	AI 감정 인식 과정에서 “언어 정보(Semantics)”와 “음향 정보(Prosody)”가 서로 어떤 기여를 하는지 정량·정성적으로 규명하는 것
핵심 방향	• **Dual-Signal Contribution 분석** 텍스트 의미 vs 음향·억양 신호가 감정 판단에 미치는 영향 비교 • **Feature Importance 파악** Prosody(속도·톤·강세) 및 언어적 표현이 정서 분류에 기여하는 정도 분석
Target OutCome	• 감정 인식 시 어떤 신호가 어떤 상황에서 더 유효한지 설명 가능한 인사이트 도출 • 감정 AI 설계 시 모달 우선순위·설계 방향성 제안 • XAI 관점에서 판단 근거를 구조화해 해석 가능성 확보

AI는 무엇을 기준으로 판단하는가?

" 참 잘했다. "

텍스트 분석: 100% 긍정 (칭찬)

억양 분석: 100% 부정 (비꼬는 어조)

이러한 ' **불일치(Mismatch)** ' 를 AI가 어떻게 판단하고 해결할 것인가가 본 연구의 핵심 과제입니다.

감정 인식의 핵심 입력구조에 대한 실증적 이해 필요	• 현재 감정 AI는 텍스트·음성을 조합하지만 각 신호의 영향력과 기여도가 명확히 설명되지 않음 • 이를 해석하면 AI의 추론 경로를 더 신뢰성 있게 제시 가능
Prosody의 정서 전달력 주목	• 사람 간 의사소통에서 정서 정보의 70% 이상이 비언어 신호(tone·pitch·rate) • Prosody가 중요하지만 모델 관점에서 그 기여도는 덜 연구됨
Explainable AI(XAI) 적용 시 가치가 큼	• 감정 AI는 의료·온라인 상담·고객 VOC 분석 등 신뢰 기반 도메인에서 활용도 높음 → 해석력 필수
경량 모델·RAG·디바이스 온보드 등 확장성	• 텍스트/음향 기여도를 알면 → 입력 축소·모델 경량화·도메인 최적화 설계 가능 → 서비스 적용 시 운영 효율 증가

멀티모달 감정 인식 모델 구축	• 텍스트(BERT) + 음성(wav2vec2 or Mel-spectrogram CNN)
XAI 기법 적용	• SHAP → 각 모달별 특징 중요도 계산 • Grad-CAM → 음성 스펙트로그램 시각화
근거 비중 비교	• 감정별로 Text/Audio 기여도 비율 산출 (예: 분노: 70% 억양 / 슬픔: 55% 텍스트)

Data set: MELD (Multimodal EmotionLines Dataset)

-대화 기반 감정 데이터셋 (텍스트 + 오디오 + 감정 라벨)

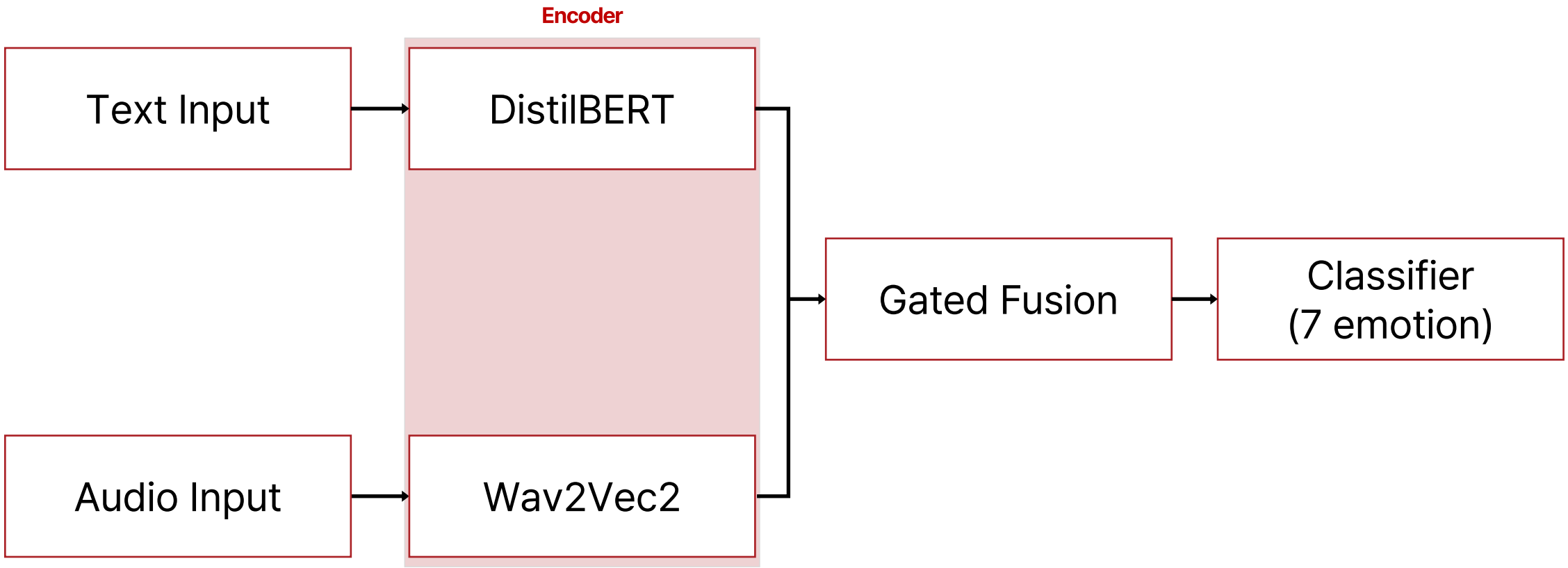
-Emotion class: 7(Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise)

Text Encoder: DistilBERT-base-uncased(hugging face)

-대사 내용을 입력을 받아 **단어 의미 기반 감정 단서** 추출

Audio Encoder:Wav2Vec2-base (Hugging Face)

-억양, 속도, 강세 등 **감정적 음성 특징** 추출


Encoder
Text Input
DistilBERT
Audio Input
Wav2Vec2
Gated Fusion
Classifier
(7 emotion)

텍스트 또는 오디오를 각각 제거(masking)했을 때 감정 예측 확률이 얼마나 변하는가?
어떤 feature가 더 중요하게 감정을 추출할 때 중요하게 적용되는가?

실험	입력 구성	의미
Text Masking	[MASK] or 빈 문장	말의 내용 제거
Audio Masking	무음 or 노이즈 대체	억양 제거
비교	예측 확률 변화(Δ)	각 모달의 중요도

예시 기대 결과

텍스트 SHAP: “정말”, “미쳤다” → 분노에 높은 기여
음성 Grad-CAM: 고주파(2–4 kHz), 급격한 pitch 상승 → 분노 신호
기여도 비교:

분노: Audio 72% / Text 28%

슬픔: Audio 55% / Text 45%

행복: Text 60% / Audio 40%

결론

AI는 감정 종류에 따라 “내용 vs 억양” 의존도가 다르다.
이는 인간의 감정 인식 패턴과 유사한지 비교 가능.

분야	기대효과
AI Explainability	블랙박스 감정 인식 모델의 내부 판단 근거를 해석
Human Alignment	인간 감정 판단 구조와 AI의 근거 비교
모델 개선	특정 감정에 대한 모달 편향(예: 음성 과의존) 보정 가능
응용 가능성	감정 챗봇, 상담 분석, 감정 TTS, 인간-로봇 상호작용 개선