기만탐지모델

1. 분석 주제

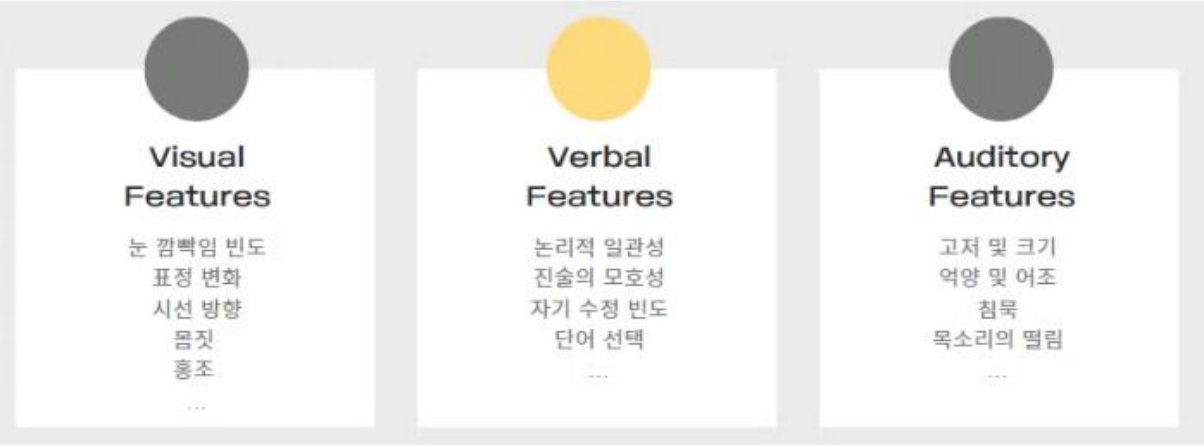
발언에서 기만을 탐지할 수 있는 모델을 고안하여 국민의 의사 결정 과정에 미치는 피해를 줄이고자 하며, 더 나아가 보이스 피싱, 전세 사기, 노인 대상 사기, 보험 사기와 같은 다양한 사기 수법에도 활용하고자 함

2. 프로젝트 배경

- 정치인들의 발언은 사회에 많은 영향을 미치며, 실제 그들이 사회적으로 중요한 주제에서 사람들을 속이려 한다면 이는 이후 국민들의 의사 결정 과정에 개입해 많은 피해를 낳음
- 따라 정치인들의 정치적 발언에 대한 기만을 탐지할 수 있는 모델을 고안해보고자 함
- **기만이란?**

  "기만은 의도적으로 진실을 감추거나 왜곡하여 다른 사람을 속이는 행위"

- **1) 사실에 대한 진위 여부 :** 기만 여부는 그 사실의 진실 여부와는 다르다. ex) 어떤 공약을 하겠다고 했지만 이후에 실현되지 못했다.
- **2) 사실의 진위 여부에 대한 확신 :** 만약 화자가 사실이 진실인지 잘 알지 못함에도 이를 진실처럼 말한다면 이 역시 기만 행위에 해당



비언어적 요인 : 눈 깜빡임 빈도, 표정 변화, 시선 방향, 몸짓, 홍조 등

→ 비언어적 요인은 전통적인 심리학 연구에서 증명된 기만의 생리적, 행동적 특징이 정리되어 있기 때문에 비교적 쉽게 접근할 수 있음

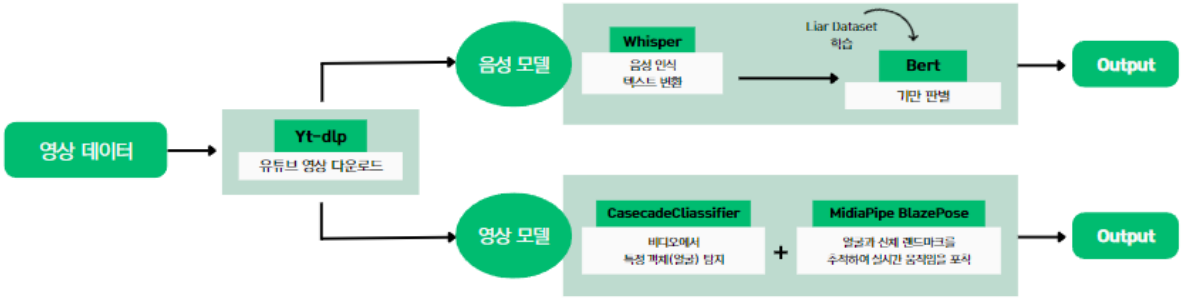
언어적 요인 : 논리적 일관성, 진술의 모호성, 자기 수정 빈도, 단어 선택 등

→ 언어적 요인은 텍스트 기반의 기만 탐지 모델과 AI 음성 인식 기술로 활용할 수 있음

반언어적 요인 : 목소리의 고저 및 크기, 억양과 어조, 침묵의 길이와 빈도 등

→ 반언어적 요인에 대한 학습이 불가능한 상황에서 다른 변수와 달리 이를 보완할 수 있는 방법론이 없었기에 반언어적 변수는 제외하기로 결정

3. 프로젝트 진행 과정



- **유튜브 영상 다운로드**
- **음성 인식 및 텍스트 변환**
- **음성 모델 구축**
- **영상 모델 구축**

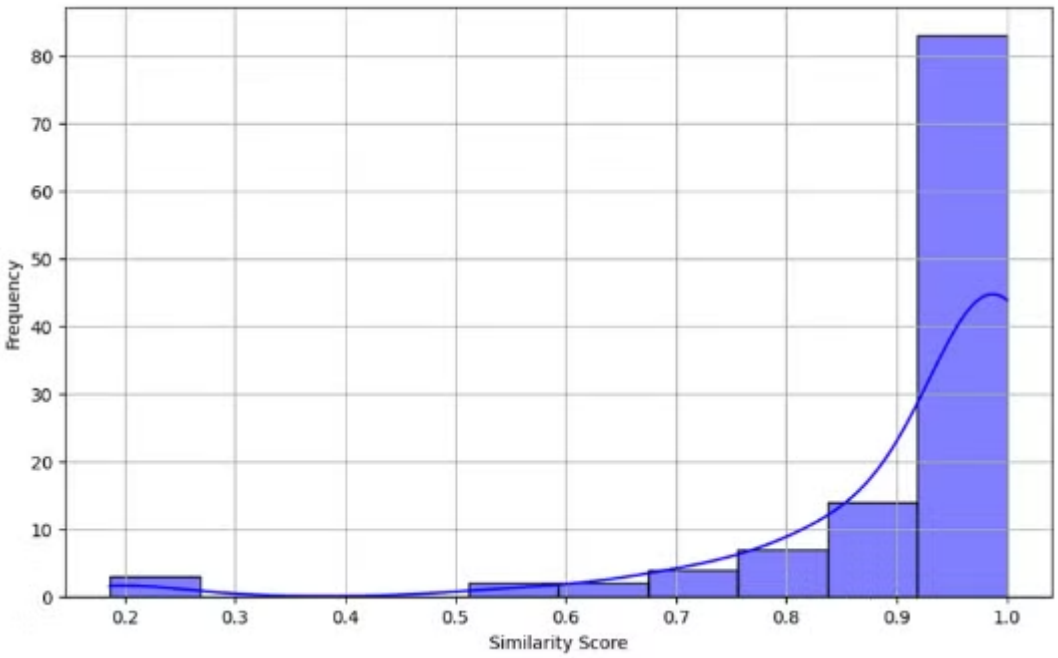
4. 분석 과정

1) 유튜브 영상 다운로드

- **Yt-dlp 를 이용해 유튜브 영상 다운로드**

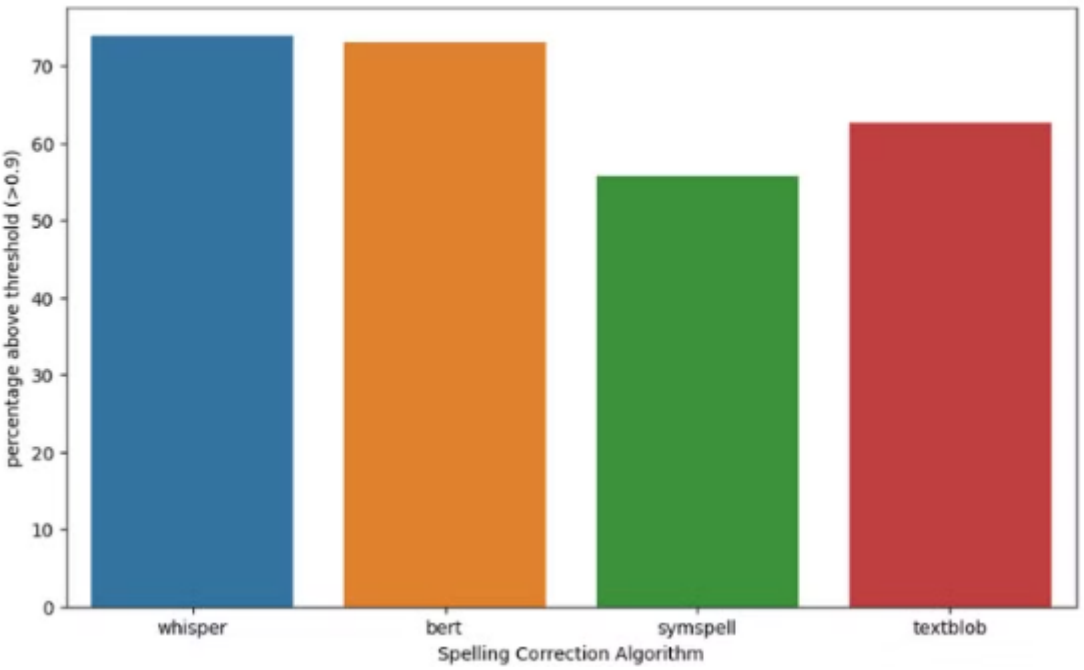
2) 음성 인식 및 텍스트 변환

- **Whisper API 를 이용해 음성 텍스트 변환**



[원 문장과 whisper 로 변환한 문장 간의 코사인 유사도]

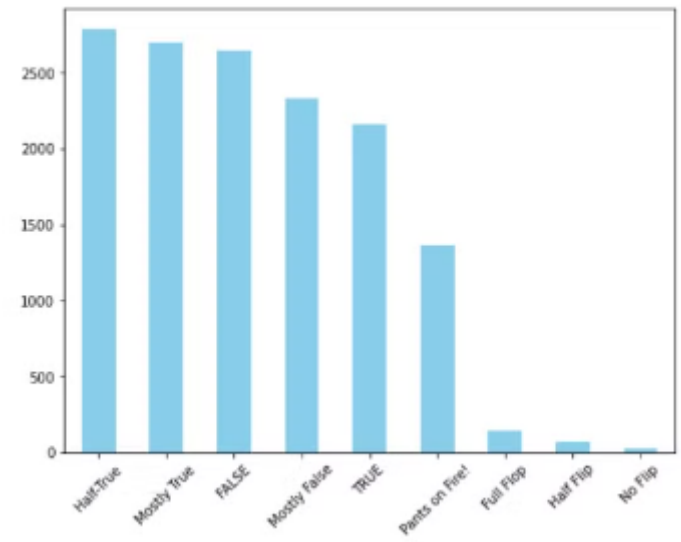
- 인식을 위한 딥러닝 기반 모델 Whisper 를 사용하여 음성을 텍스트로 변환하였으며, 음성학적 특징과 문맥을 고려하기 때문에 오차가 발생할 수 있지만, 코사인 유사도가 0.9 이상인 문장의 비율이 73.91%로 높은 정확도를 보임 (텍스트 마이닝 의미 비교 연구에서 임계값으로 제시하는 것은 0.9 이상의 코사인 유사도)
- 오타 교정 알고리즘 적용(BERT, SymSpell, TextBlob)

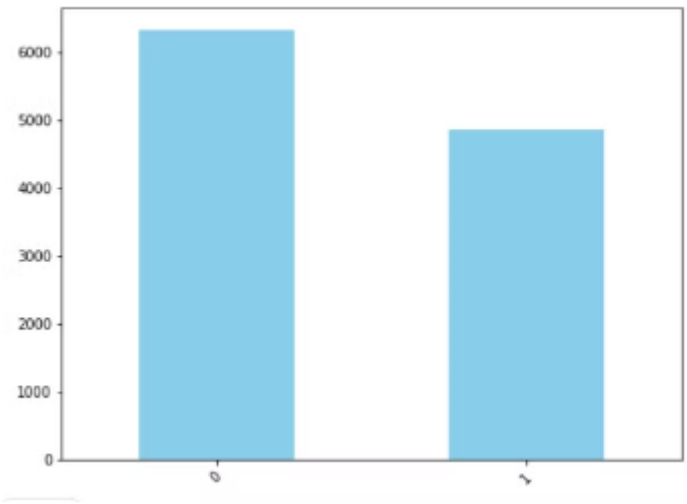


- 모델을 사용했으나, 오히려 잘 변환된 텍스트를 다른 내용으로 바꾸는 특징을 보임

3) 음성 모델 구축

- **BERT 모델**





- Flip 제외
- Half TRUE - Mostly True → 진실로 라벨링 (0)
- False - Mostly False - Pants on Fire → 거짓으로 라벨링 (1)
- BERT 모델과 LSTM 을 결합하는 모델을 생성하여 분석 진행해보았지만, 모델이 더 복잡해지기 때문에 과적합 문제를 해결할 수 없었음
- LIAR 데이터 셋을 학습시켜 발견한 기만 데이터의 7 가지 특성
- 과장된 주장 : 너무 좋은 조건, 비현실적인 약속 등이 포함된 경우.
    - "100% 보장", "위험 없음", "즉시 성공" 같은 과장된 표현.

- **모호한 표현** : 구체적인 정보 없이 추상적이거나 모호한 용어를 사용하는 경우.
    - "비밀의 방법", "독점적인 기회".
- **긴급성 강조** : 시간에 대한 압박을 느끼게 하는 문구.
    - "지금 당장 행동해야 합니다", "한정된 시간 동안만 제공". 사기와 관련된 단어나 문구가 비정상적으로 자주 사용되는지 분석.
- **유사도 분석** : 사기성 문구가 기존에 알려진 사기 사례와 유사한지 비교.
- **공포 또는 불안 조성** : 특정 행동을 촉구하기 위해 두려움이나 불안을 자극하는 경우.
    - "이 기회를 놓치면 큰 손해를 볼 수 있습니다".
- **사회적 증거** : 다른 사람들도 이 기회를 이용하고 있다는 듯한 표현.
    - "수천 명이 이미 이 기회를 잡았습니다".

sentences	label
The last thing I want to do is be involved in a primary.	LABEL_0
Okay?	LABEL_0
I could be sitting home right now, getting to watch some of the games tomorrow, getting ready, right?	LABEL_0
But it's why we need each and every one of you to get a friend, go out, get a family member, get the whole family, and bring them out to vote for Big Luther Strange, get a vote for him.	LABEL_0
We're spending a tremendous amount of money in Alabama, and you are doing great, and as long as I'm president, you are going to have so much money spent here.	LABEL_1
And you know what?	LABEL_0
Lots of other states also.	LABEL_1
I don't want to speak negatively.	LABEL_0
I said though that if I lose a selection, maybe I'll end up moving to Alabama or Kentucky or like some states.	LABEL_0
I mean, nice to go where people love you and where you love them because it's special.	LABEL_0
And by the way, Rocket Man should have been handled a long time ago.	LABEL_0
If Crookhead Hillary got elected, you would not have a second amendment, believe me.	LABEL_0
You'd be handing in your rifles.	LABEL_0
You'd be saying here, here they are.	LABEL_0
You go like, you'd be turning over your rifles.	LABEL_0
You got to speak to Jeff Sessions about that.	LABEL_0
When you love to see one of these NFL owners, when somebody disrespects our flag, to say get that son of a f*** off the field right now, he's fired.	LABEL_0
He's fired.	LABEL_0
The media, the fake news I call it.	LABEL_0
Look at the crowd.	LABEL_0
I'd love to have them show the crowd, but they don't show the crowd.	LABEL_0
They show me the whole night.	LABEL_0
I go home, I say, Melania, by the way, she's become very popular, hasn't she?	LABEL_0
So you would have people working in the VA who were sadists, who would abuse our great, great people, our great veterans.	LABEL_0
By the way, 25 years before, they would have their ass kicked by the same person that they're abusive.	LABEL_1
And by the way, folks, just in case you're like curious, no, Russia did not help me.	LABEL_0
Okay?	LABEL_0
Russia.	LABEL_1
And so on Tuesday, vote for your country, vote for your family, vote for your victory, vote for Luther Strange.	LABEL_0
She's a phenomenal person.	LABEL_0

[텍스트 문장 별 기만 여부 확인, LABEL_0 : 진실/LABEL_1 : 거짓]

1. 트럼프 대통령은 대규모 인프라 투자를 공약으로 내세웠고, 1 조 달러 규모의 인프라 계획을 발표했지만, 의회를 통한 법제화는 이뤄지지 않았음. 예를 들어, 오바마 행정부는 경제 위기 이후 경기 부양책의 일환으로 인프라 투자를 증대시킴
2. 오히려 트럼프 행정부는 일부 사회복지 프로그램의 예산을 삭감하거나 조건을 강화하려는 움직임을 보임
3. 경제 정책과 일자리 창출: 트럼프 행정부의 경제 정책: 트럼프 행정부는 대규모 감세 정책을 시행하여 기업과 고소득층에 혜택을 주었지만, 이로 인해 연방 예산 적자가 크게 늘어남. 감세 정책은 일자리 창출을 목표로 했으나, 이는 실패로 끝났고, 다른 예산의 투입도 없었음.

- 결론적으로, 트럼프 행정부 시기의 인프라와 삶의 질 관련 예산 정책은 대규모 투자 계획이 있었지만 실제로 집행된 프로젝트는 제한적이었고, 사회복지 프로그램에 대한 접근성은 일부 제한되었음

- 다른 대통령들과 비교했을 때, 트럼프 행정부는 인프라와 삶의 질 개선에 있어 대규모 계획을 세웠지만 실행 면에서 다른 행정부들에 비해 덜 적극적인 모습을 보이는 것을 확인함
- 그러나, 현실적으로 언어적 요소를 이용한 기만 탐지 모델의 예측력 판단은 불가하다는 한계가 있음
- 그렇기 때문에 비언어적 요소를 이용한 기만 탐지 모델의 결과와 비교하면서 유사한 패턴을 보이는 부분을 탐색하고 이를 시각화하는 방향을 선택

4) 영상 모델 구축

- **Truthsayer 모델 및 Media Pipe Blaze Pose 모델 구축**
- 여러 간단한 분류기를 결합해 AdaBoost 로 특징을 선택하여 강력한 객체 탐지 모델을 만들며 실시간 어플리케이션에 적합하고 훈련이 상대적으로 쉬운 Cascade Classifier 분석 방법론을 적용하고자 함
- 변수를 임의로 선택할 수 없었기 때문에 기존의 얼굴 이미지 분석 모델을 탐색하여, 심박수, 눈 깜빡임 감지, 시선 방향 변화 감지, 손과 얼굴 접촉 빈도, 입술 압박, 이목구비의 움직임을 통한 감정 분석이 가능한 Truthsayer 모델을 사용하여 얼굴 이미지 분석을 진행함
- 얼굴과 몸의 랜드마크를 실시간으로 추적할 수 있는 Media Pipe Blaze Pose 모델을 사용해 Cascade Classifier 모델과 결합하여 분석을 진행함

```
MAX_FRAMES = 120
RECENT_FRAMES = int(MAX_FRAMES / 10)
EYE_BLINK_HEIGHT = 0.5
SIGNIFICANT_BPM_CHANGE = 8
LIP_COMPRESSION_RATIO = 0.35
TELL_MAX_TTL = 30
TEXT_HEIGHT = 30
FACEMESH_FACE_OVAL = [10, 338, 297, 332, 284, 251, 389, 356, 454, 323, 361, 288, 397, 365, 379, 378, 400, 377, 152, 148, 176, 149, 150, 136, 172, 58,
EPOCH = datetime.timestamp(datetime.now())
recording = None
tells = dict()
blinks = [False] * MAX_FRAMES
blinks2 = [False] * MAX_FRAMES
hand_on_face = [False] * MAX_FRAMES
hand_on_face2 = [False] * MAX_FRAMES
face_area_size = 0
hr_times = list(range(0, MAX_FRAMES))
hr_values = [400] * MAX_FRAMES
avg_bpms = [0] * MAX_FRAMES
gaze_values = [0] * MAX_FRAMES
emotion_detector = FER(mtcnn=True)
meter = cv2.imread('meter.png')
fig = None
ax = None
line = None
peakpts = None
blink_times = []
blink_counts = []
```

[변수 임계값 설정]

- 임계값에 대한 설명
    - 눈 깜빡임 감지 EYE_BLINK_HEIGHT = 0.5

눈의 높이와 너비의 비율 (Eye Aspect Ratio, EAR)이 0.5 보다 작아질 때 깜빡임이 발생한 것으로 간주

RECENT_FRAMES = 12 (MAX_FRAMES / 10) : 최근 12 개의 프레임을 분석하여 눈 깜빡임 빈도를 계산

- 심박수 변화 감지

  SIGNIFICANT_BPM_CHANGE = 8

  평균 심박수에서 ± 8 bpm 이상 변화하면 스트레스나 긴장 상태를 나타낼 수 있음

- 입술 압박 비율 (Lip Compression Ratio)

  LIP_COMPRESSION_RATIO = 0.35 : 입술의 압박 정도를 측정하는 임계값

  입술의 비율이 0.35 이하일 때 입술이 압박된 상태로 간주하여 기만 가능성을 평가

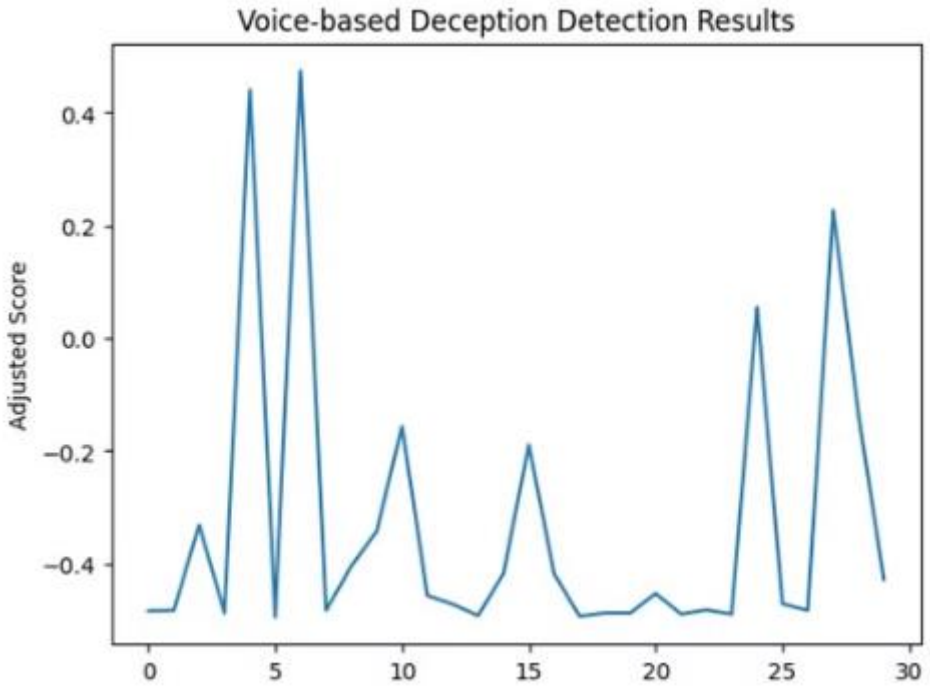
- 손이 얼굴에 닿는지 여부 감지 (Hand-on-Face Detection) 얼굴 윤곽과 손 위치를 비교하여 얼굴 윤곽 안에 손가락의 위치가 있는지를 통해 손이 얼굴에 닿는지 감지

- **얼굴 이미지 인식의 한계점**

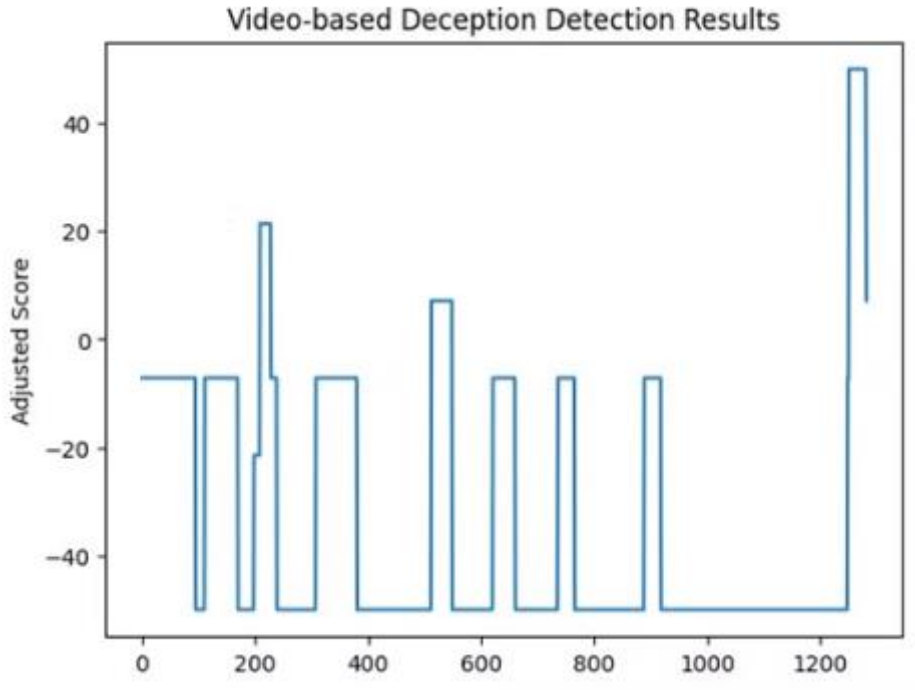
- 심작 박동 같은 경우에는 볼의 붉기와 움직임으로 카운트를 하기 때문에 실제 심장 박동과 오차가 있을 수 있음, 실제 카메라로 워치를 통한 비교를 했을 때 한 15~20 정도 차이가 있었음
- 눈 깜박임 같은 경우에는 눈을 감았다가 떴을 때 카운트하는 기능으로 구현 했지만 사람마다 눈의 크기가 다르기 때문에 눈이 작은 사람 같은 경우에는 눈을 조금만 크게 뜨기만 해도 카운트가 됨

5. 분석 결과

- 결과를 수치로 제시하는 경우, 어느 수치까지 기만으로 판단해야 하는지, 주관적인 판단이 들어가기 때문에 임계점을 설정하기 어렵다는 문제가 있음
- 따라서 두 모델이 예측하는 기만 정도를 선 그래프로 나타내어 공통적으로 그 점수가 높은 부분들에 주의를 기울이도록 하는 것이 효과적이라는 결론을 냄



[음성기반 모델 분석 결과]



[영상기반 모델 분석 결과]

6. 프로젝트를 통해 얻은 인사이트

1) 실용성

- 현재 기만 탐지 모형에 대한 연구는 동일한 데이터 셋 내에서의 분류 정확도를 높이는 것을 목표로 발전
- 새롭게 입력된 데이터에 대한 예측 정확도는 평균 70%대에 머무르고 있으므로 100% 기만을 단정하는 것은 불가능
- 데이터 오용으로 인해 알고리즘 편향이 문제를 발생시킬 여지 존재
- 그러나, 여전히 기만 행위를 경계할 수 있도록 경고하는 역할을 고려하면 성능 개선이 필요

2) 정확성

- 현재 기만 탐지 모형에 대한 연구는 언어적, 비언어적, 반언어적 요소들을 개별적으로 학습하여 앙상블 모형 적합
- 모든 변수를 동시에 종합적으로 학습하는 모델을 구축할 수 있는 새로운 딥러닝 모형의 개발이 필요
- 또한, 기만 예측의 정확도를 개선하기 위해서는 일반적 특성에 더해 개인적인 특성 역시 반영되어야 함

- 영상과 함께 대상의 온라인 상 기록을 함께 입력해 미리 학습시킨 후 모델을 적용한다면 정확도 향상을 기대할 수 있음

3) 활용성

- 보이스 피싱
- 전세 사기
- 노인 대상 사기
- 보험 사기

7. 활용 데이터

- https://paperswithcode.com/dataset/liar
- Truthsayer : Make a remote lie detector and become irresistible (https://www.youtube.com/watch?v=6esc_HD7b-A&t=0s)
- Voting-based Multimodal Automatic Deception Detection (https://arxiv.org/abs/2307.07516)
- Detecting deception using machine learning with facial expressions and pulse rate (https://link.springer.com/article/10.1007/s10015-023-00869-9)
- Scientific Validity of Polygraph Testing: A Research Review and Evaluation (https://sgp.fas.org/othergov/polygraph/ota/conc.html)
- The Definitive Guide to Reading Microexpressions (Facial Expressions) (https://www.scienceofpeople.com/microexpressions)
- Building a Better Lie Detector with BERT : The Difference Between Truth and Lies (https://ieeexplore.ieee.org/document/9206937)
- Cues to Deception (https://smg.media.mit.edu/library/DePauloEtAl.Cues to Deception.pdf)