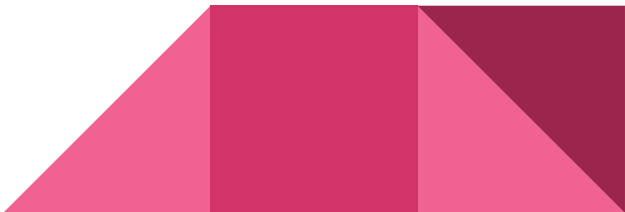


# Fake News Detection

박진영 조교  
김선명 소경민 이세희 정윤수 함영인  
발표자 : 김선명

# 목차

- ❑ 데이터 개요
  - ❑ 프로젝트 소개
  - ❑ 모델 설계
    - LIME(설명가능 AI)
    - XGBoost
    - Stacked GRU
  - ❑ 모델 평가
- 

## 가짜뉴스 사례

여성들이 남성들보다 더 많은 시간을 일한다.

1600명의 근로자들을 대상으로 한 전국 조사에 따르면 여성들은 평균 주당 **34시간**을 일하는 것으로 나타났다. 이는 5년 전의 30.4시간에 비해 반나절 더 늘어난 수치이다.

동일한 5년 동안 남성 평균 근무 시간은 주당 45.5시간에서 **44.8시간**으로 감소했다.

Privacy Policy | Feedback  Like 14.8M

# MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Australia **Femail** | Health | Science | Money | Vi

Latest Headlines | **Femail** | Fashion Finder | Food | Beauty | Gardening | Blogs | Baby Blog | Discounts

## Women work longer hours than men

by MATTHEW HICKLEY, Daily Mail

Women are working longer hours while men are putting in less time for their money.

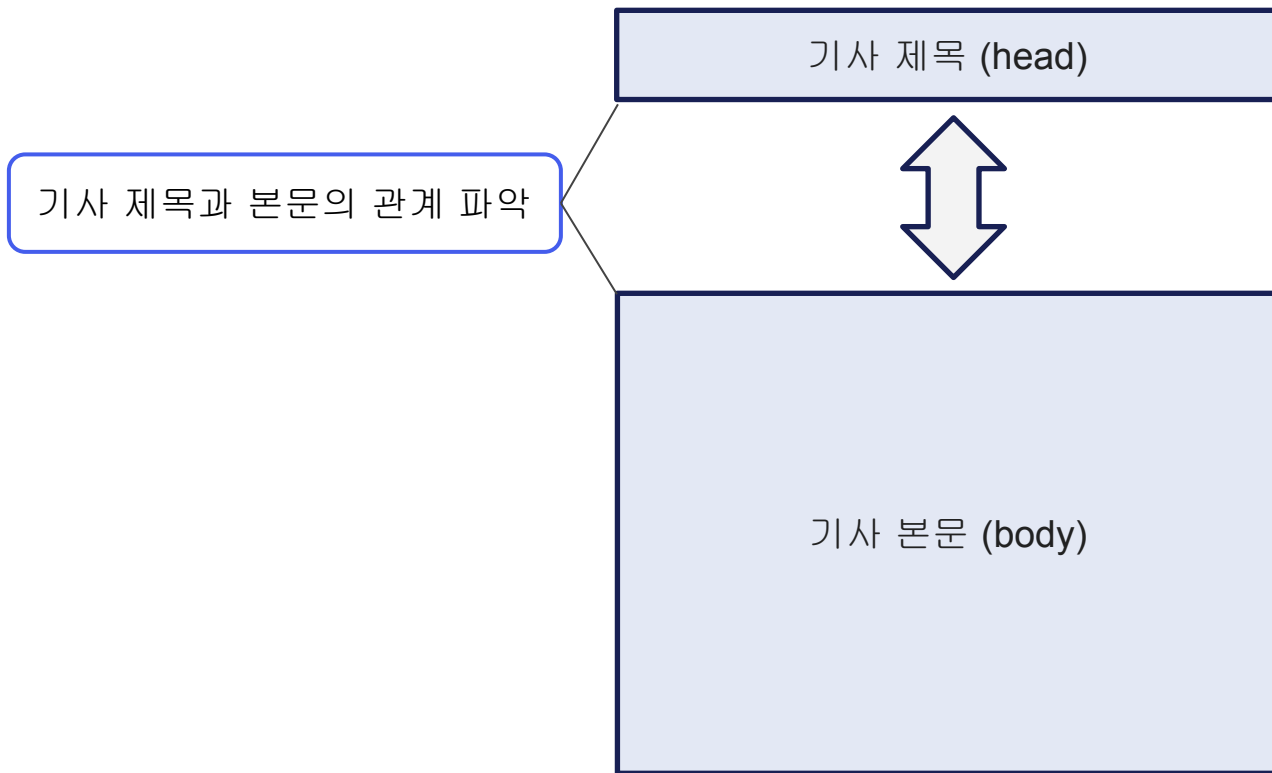
And yet the pay gap between the sexes has widened, research revealed yesterday.

A nationwide survey of 1,600 employees found that women now work almost 34 hours a week on average - half a day longer than the figure of 30.4 hours five years ago.

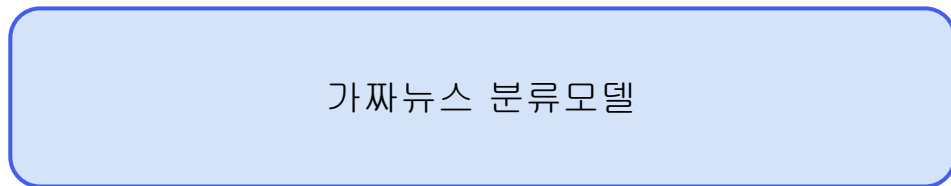
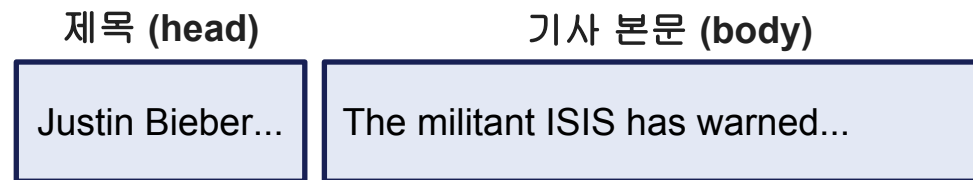
According to the Chartered Institute of Personnel and Development (CIPD) which carried out the research, the shift reflects the growing number of women in more demanding management roles and professional jobs.

Over the same five-year period, men's average work hours fell from 45.5 hours per week to 44.8 hours, although among the hardestworking males the number putting in more than 49 hours a week has passed the three million mark.

## 프로젝트 이해



# 프로젝트 설계



## 데이터 개요

Training set : 49,972 쌍

Test set : 25,413 쌍

제목 (head)	기사 본문 (body)	스탠스 (label)
Justin Bieber failed...	The militant <b>ISIS</b> has warned...	unrelated
Justin Bieber failed...	<b>Bieber failed</b> his concert in India...	agree
Justin Bieber failed...	Bieber had his first concert <b>successfully</b> ...	disagree
Justin Bieber failed...	<b>Several officials say</b> , Justin Bieber messed up....	discuss

⋮

## 가짜뉴스 사례

여성들이 남성들보다 더 많은 시간을 일한다.

1600명의 근로자들을 대상으로 한 전국 조사에 따르면 여성들은 평균 주당 **34시간**을 일하는 것으로 나타났다. 이는 5년 전의 30.4시간에 비해 반나절 더 늘어난 수치이다.

동일한 5년 동안 남성 평균 근무 시간은 주당 45.5시간에서 **44.8시간**으로 감소했다.

Privacy Policy | Feedback  Like 14.8M

# MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Australia **Femail** | Health | Science | Money | Vi

Latest Headlines | **Femail** | Fashion Finder | Food | Beauty | Gardening | Blogs | Baby Blog | Discounts

## Women work longer hours than men

by MATTHEW HICKLEY, Daily Mail

Women are working longer hours while men are putting in less time for their money.

And yet the pay gap between the sexes has widened, research revealed yesterday.

A nationwide survey of 1,600 employees found that women now work almost 34 hours a week on average - half a day longer than the figure of 30.4 hours five years ago.

According to the Chartered Institute of Personnel and Development (CIPD) which carried out the research, the shift reflects the growing number of women in more demanding management roles and professional jobs.

Over the same five-year period, men's average work hours fell from 45.5 hours per week to 44.8 hours, although among the hardestworking males the number putting in more than 49 hours a week has passed the three million mark.

## 가짜뉴스 사례

여성들이 남성들보다 더 많은 시간을 일한다.

1600명의 근로자들을 대상으로 한 전국 조사에 따르면 여성들은 평균 주당 **34시간**을 일하는 것으로 나타났다. 이는 5년 전의 30.4시간에 비해 반나절 더 늘어난 수치이다.

동일한 5년 동안 남성 평균 근무 시간은 주당 45.5시간에서 **44.8시간**으로 감소했다.

Privacy Policy | Feedback  Like 14.8M

# MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Australia | **Femail** | Health | Science | Money | Vi

Latest Headlines | **Femail** | Fashion Finder | Food | Beauty | Gardening | Blogs | Baby Blog | Discounts

## Women work longer hours than men

by MATTHEW HICKLEY, Daily Mail

Women are working longer hours while men are putting in less time for their money.

And yet the pay gap between the sexes reveals that men are still earning more than women for the same work.

**Disagree**

According to a survey by personnel and development consultants, which carried out the research, the shift reflects the growing number of women in more demanding management roles and professional jobs.

Over the same five-year period, men's average work hours fell from 45.5 hours per week to 44.8 hours, although among the hardestworking males the number putting in more than 49 hours a week has passed the three million mark.




## 데이터 개요

### □ 클래스 별 분포

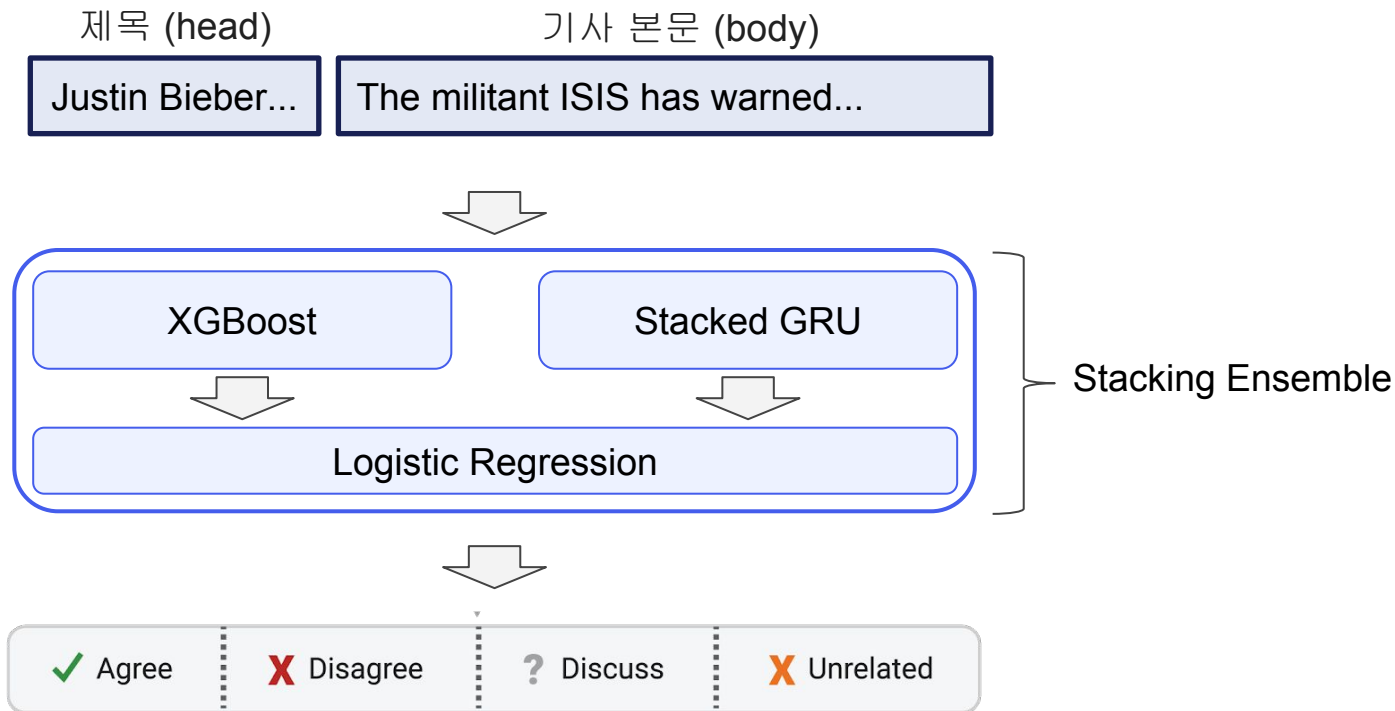
Agree	Disagree	Discuss	Unrelated	Total
3,678 (7.37%)	840 (1.7%)	8,909 (17.8%)	36,345 (72.8%)	49,924

- 영문으로 된 신문 기사 데이터
- Headline, body, stance label로 구성

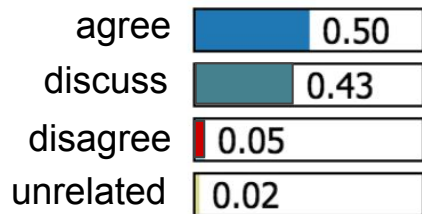
## 프로젝트 목표

- ❑ XGBoost / RNN / Stacking Ensemble 모델을 이용한 가짜뉴스 탐지 성능 개선
  - ❑ 설명가능 인공지능(Explainable AI)을 적용하여 가짜뉴스 판단 근거 제시
- 

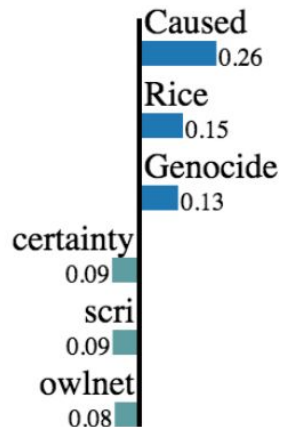
# 모델 개요



# Explainable AI

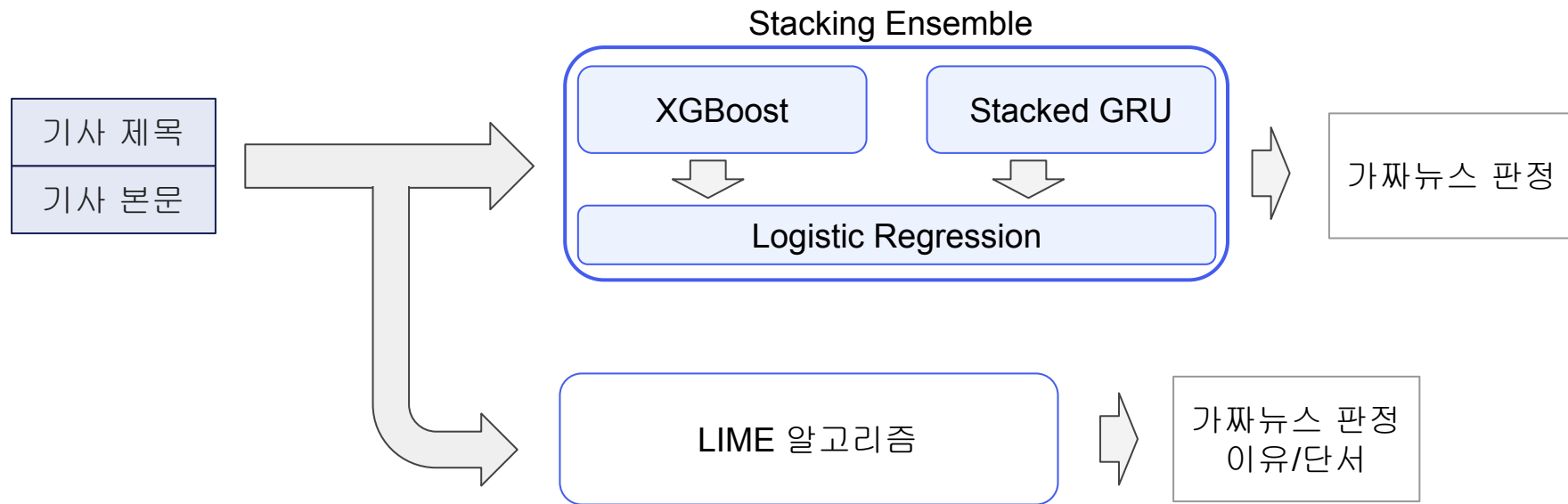


discuss | agree

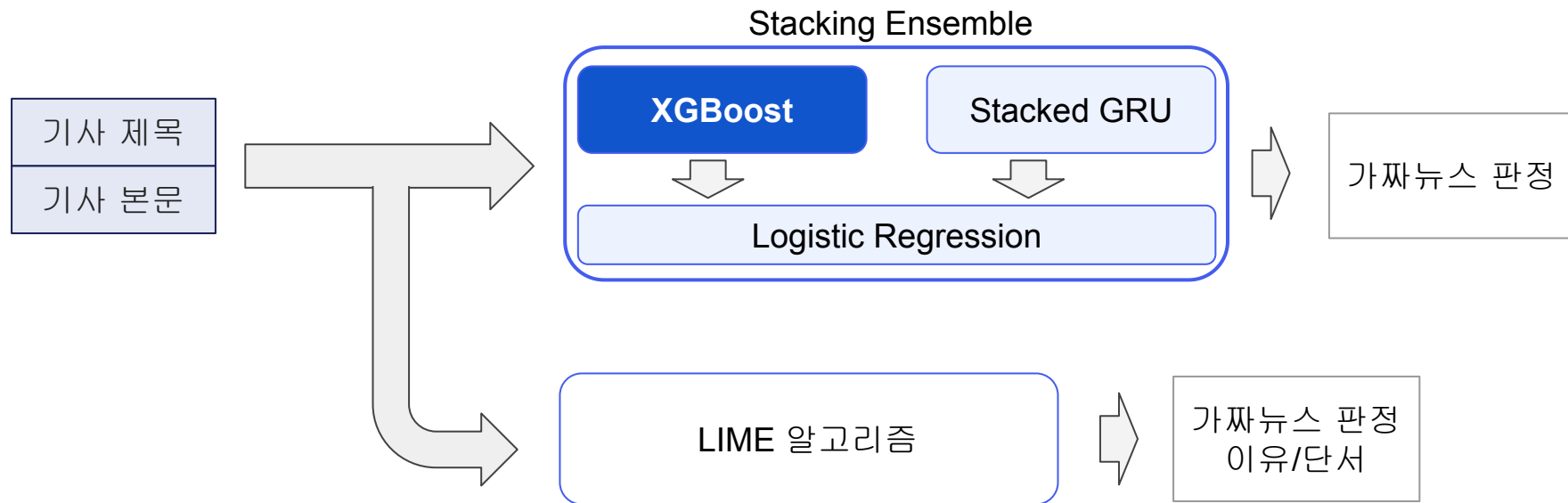


- ❑ LIME(Local Interpretable Model-agnostic Explanations) 알고리즘 적용
- ❑ 가짜뉴스를 판별했을 때, 어떤 항목을 토대로 판정했는지 근거 제시

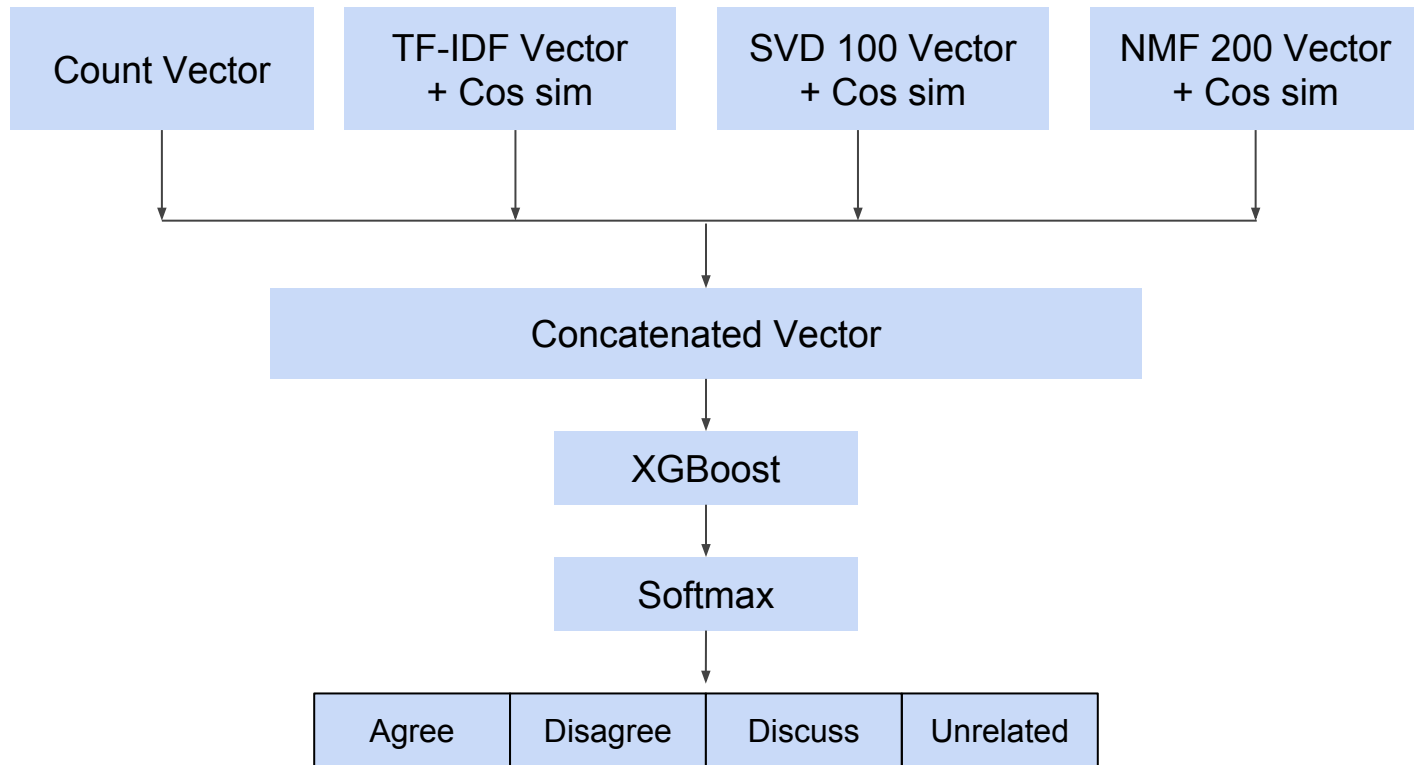
## 모델 설계



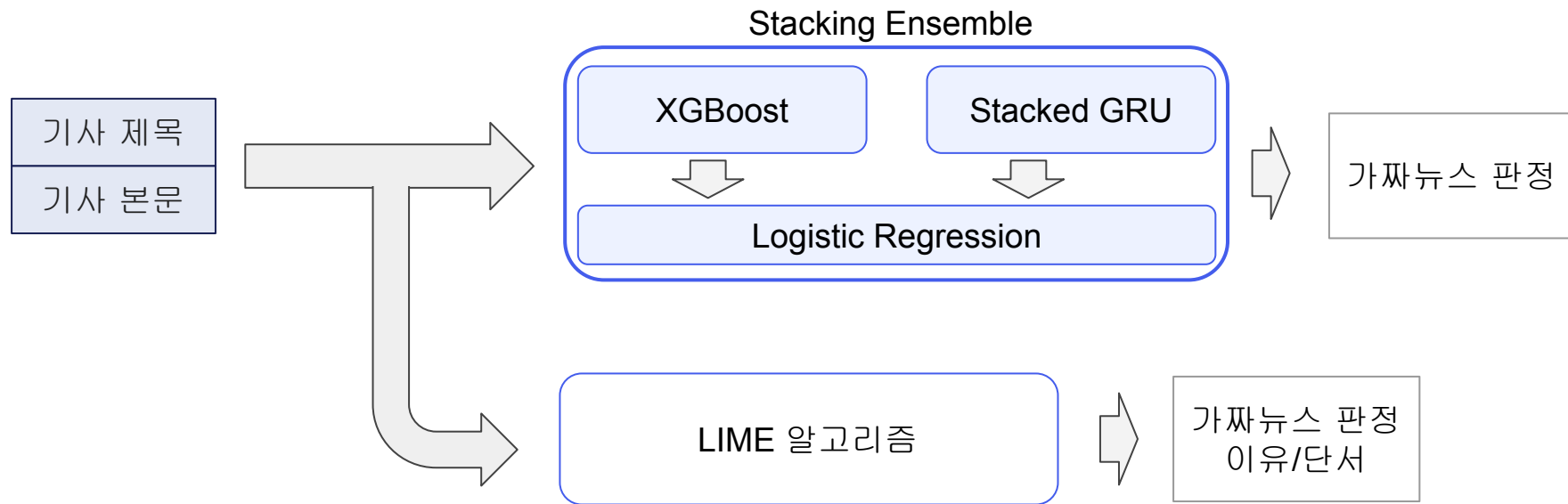
## 모델 설계



# XGBoost

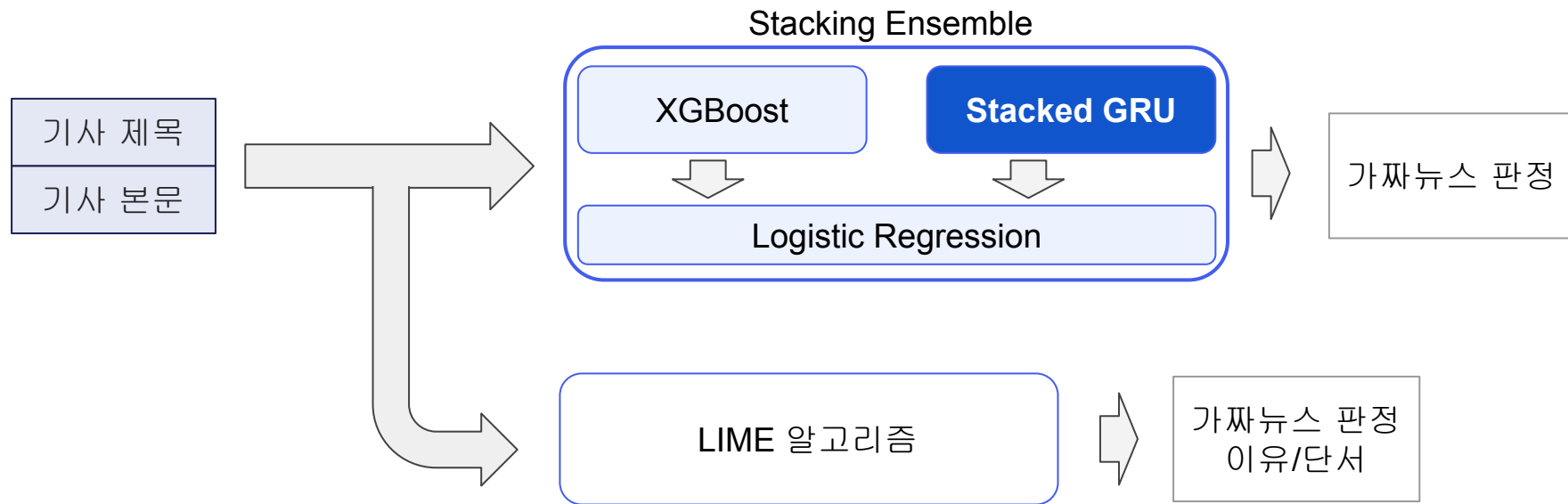


## 모델 설계

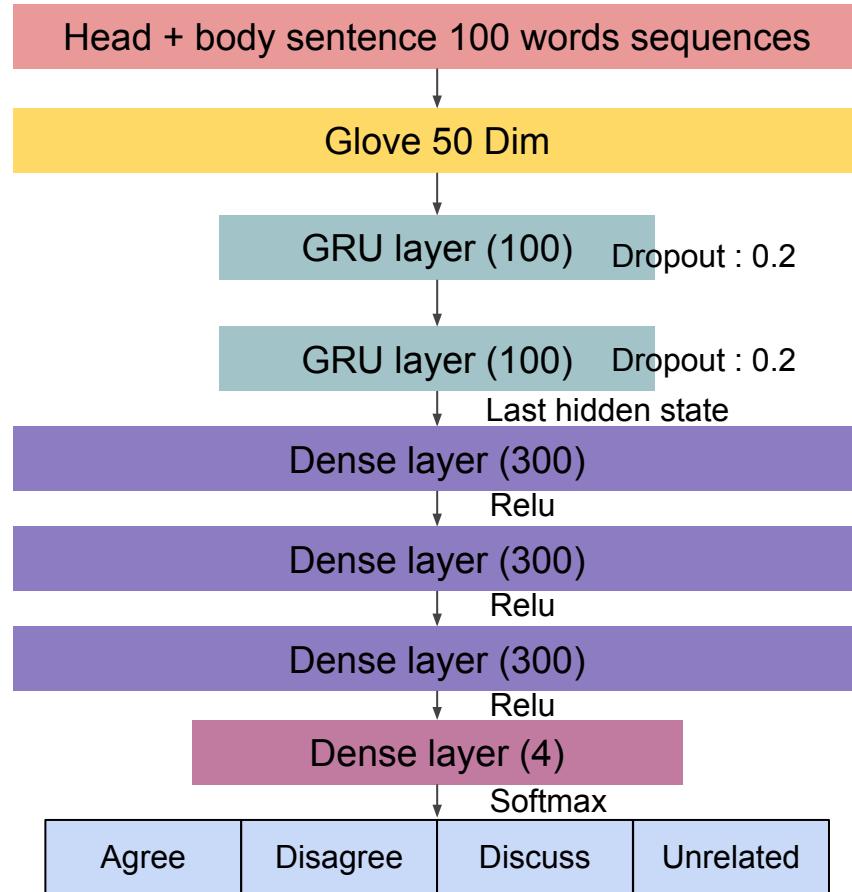




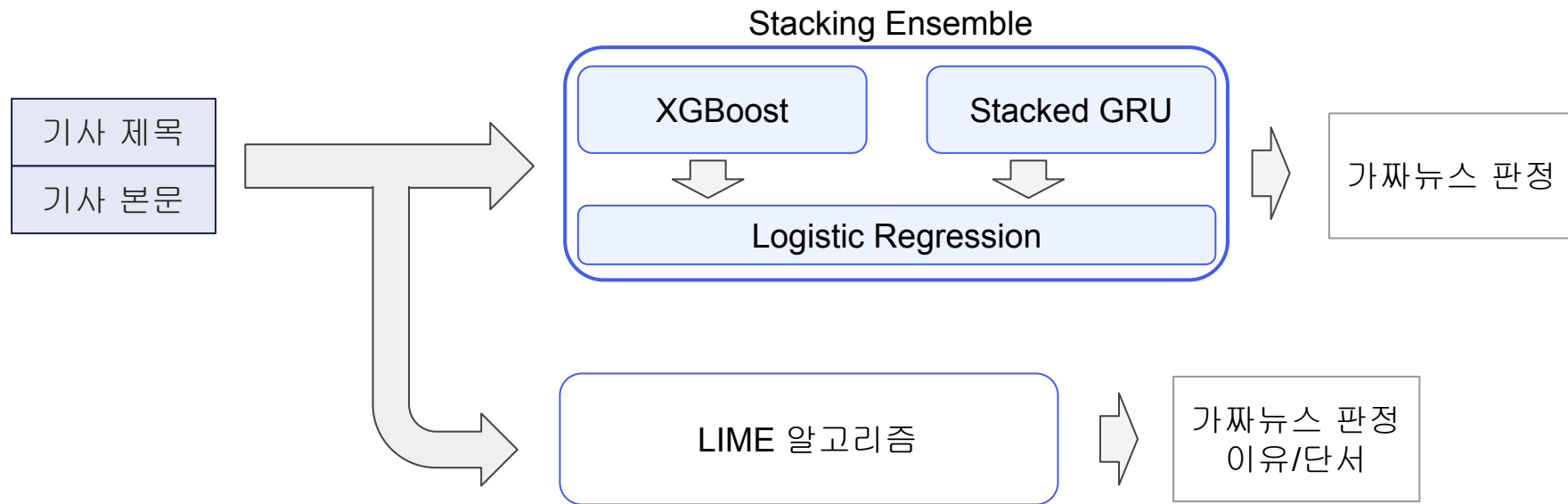
## 모델 설계



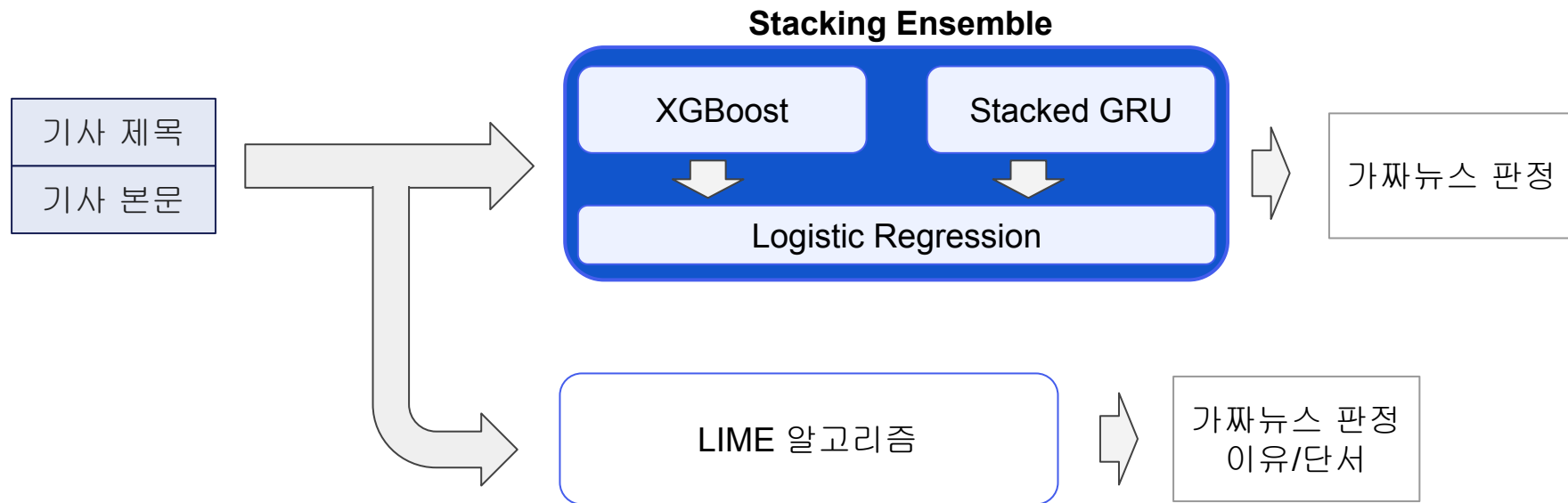
# Stacked GRU



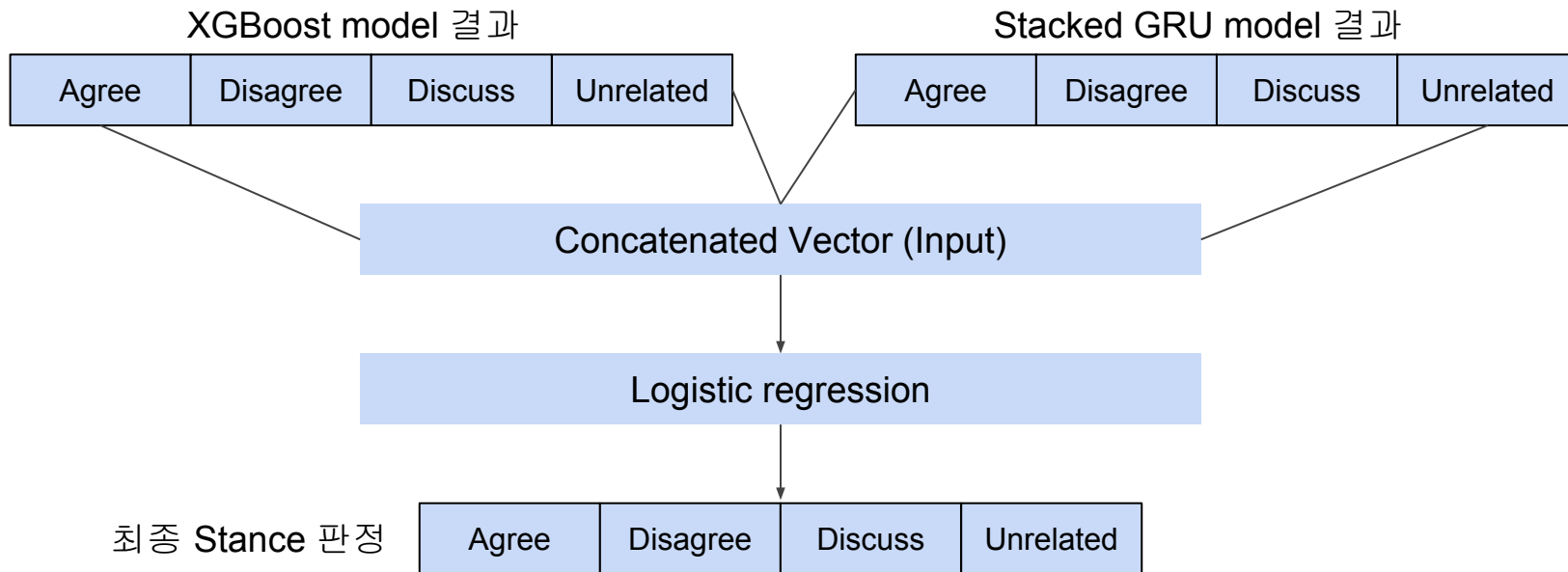
## 모델 설계



## 모델 설계




# Stacking Ensemble



# 모델 평가 환경

## Task environment

Naver Cloud Platform Cloud 이용 (Thanks to  NAVER CLOUD PLATFORM )

## 평가 measures

- FNC 점수 = (agree, disagree, discuss 맞춘 수)  $\times$  0.75 + (unrelated 맞춘 수)  $\times$  0.25
- F-score

## 우승 팀 모델 분석

- 1등 팀 : CNN + XGBoost
- 2등 팀 : 7 layer MLP ensemble model
- 3등 팀 : Single MLP model

## 모델 평가

	f1 - macro	Agree	Disagree	Discuss	Unrelated	FNC 점수	accuracy
1등 모델	58.2%	53.9%	3.5%	76.0%	99.4%	82.02%	89.1%
2등 모델	60.4%	48.7%	15.1%	78.0%	99.6%	81.97%	89.5%
3등 모델	58.3%	47.9%	11.4%	74.7%	98.9%	81.7%	88.5%
Stacked GRU	37.42%	24.70%	11.18%	41.97%	71.84%	57.44%	57.69%
XGBoost	49.3%	25.7%	0.00%	74.0%	97.6%	81.1%	87.6%
<b>Stacking Ensemble</b>	53.34%	44.15%	<b>14.82%</b>	58.42%	95.96%	72.50%	87.11%

□ 각 클래스 별 평가는 F1 score 값으로 측정

## 모델 한계점 분석

### Disagree인데 Agree로 잘못 판단한 기사 예시

제목 : Sorry, Argentina's President Didn't Actually Adopt a Jewish Werewolf

본문 : The President of Argentina, Cristina Fernandez de Kirchner, has adopted a Jewish godson – to prevent him from becoming a werewolf. Although this sounds like something straight out of a fantasy novel, the President last week.... (하락)



기사 제목과 본문 초반의  
단어 유사도가 높음

#### □ 원인 분석

- GRU 모델은 제목과 기사 본문을 붙인 것의 초반 100단어를 feature로 사용하기 때문에 본문 전체를 반영하지 못한 것으로 추정



# Future Work

## GRU 모델 성능 개선

- ❑ **GRU 모델 input feature 시퀀스 길이를 조정**
  - 본문 feature input 길이를 늘려 본문이 모델이 반영되는 비율 확대
- ❑ **Sentence embedding을 통해 다른 feature로 이용**
  - 본문을 sentence단위로 끊어 각각 임베딩
  - 임베딩 된 제목과 임베딩 된 본문 sentence를 cos similarity와 같은 measure를 이용해 추가적인 input feature로 사용

# Future Work

## 메타 데이터 활용

### □ 기사 페이지 분석

- 기사 웹페이지 내의 광고 갯수 통계 활용

### □ 언론사 통계 활용

- 언론사 사이트별 월간 방문자 수 통계 활용
- 언론사 별 재무 정보 비교

### □ 뉴스 이용자 반응 분석

- 뉴스 기사 댓글 분석 (댓글 갯수, 내용 등)
- 트윗 전파 속도, 리트윗 횟수 통계 활용

# 데모 구현

Headline \*

Bill Aims to Ban Korea's Wild Animal Cafes

Body \*

A lawmaker has proposed a bill to ban wild animal cafes, which have sprung up across Korea in recent years.

Rep. Lee Yong-deuk of the governing Democratic Party of Korea proposed a bill last week to ban displays of wild animals such as raccoons at cafes and restaurants.

The number of these cafes has increased in Seoul and other big cities over the past few years, attracting customers who want to enjoy their drinks in the company of unique animals such as meerkats, foxes, flying squirrels and even snakes.

Under the current law, cafe owners can use wild animals for commercial purposes as long as they are not endangered species recognized by the government.

If the National Assembly passes the bill, violators could face imprisonment for up to a year or a maximum fine of 10 million won (\$9,000).

3

EXPLAIN    INSPECT

Unrelated Ratio

0.22102482911077598

Agree Ratio

0.45997029770289477

Disagree Ratio

0.022408857566434476

Discuss Ratio

0.29659601561989485

# 감사합니다

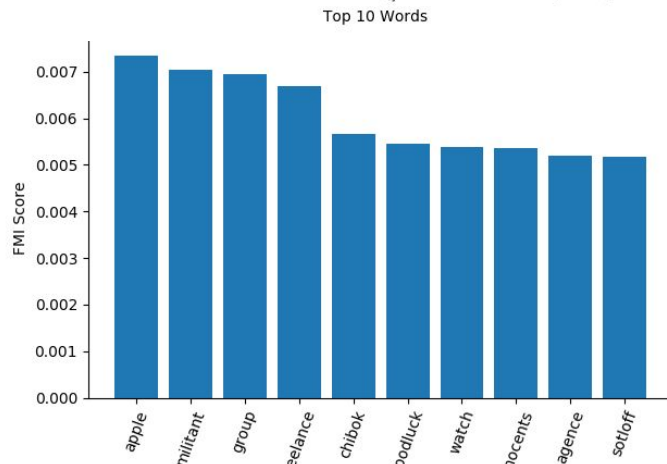
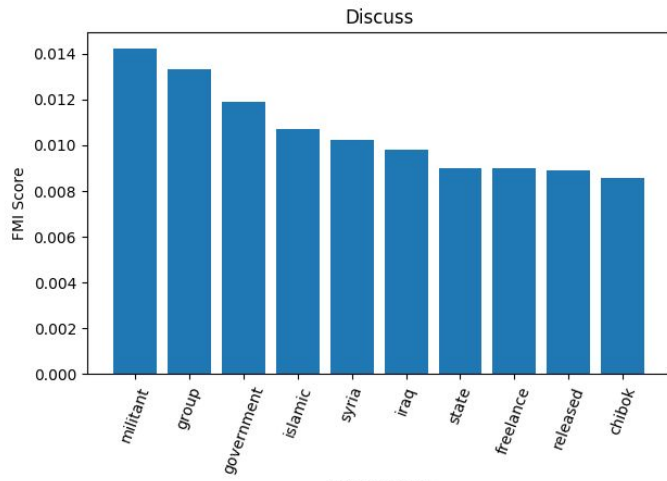
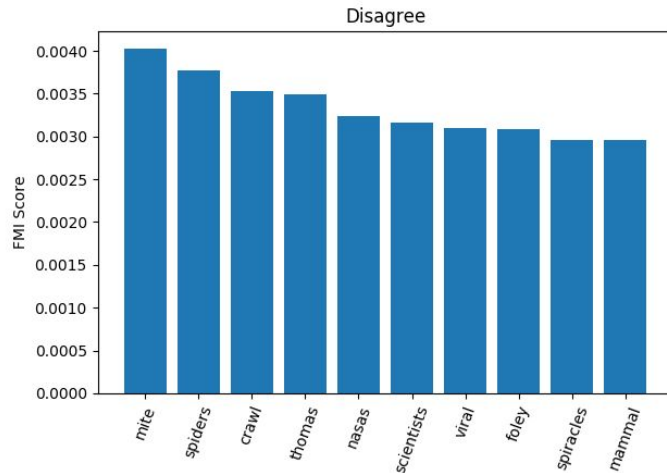
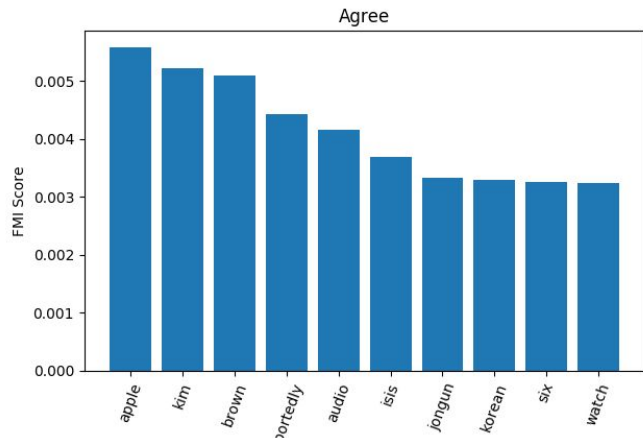
## Q&A

참고 자료 출처

- <https://www.fakenewschallenge.org>
- <https://github.com/Cisco-Talos/fnc-1> (2017 FNC winner)
- <https://arxiv.org/pdf/1806.05180.pdf>  
(A Retrospective Analysis of the Fake News Challenge Stance Detection Task)
- <https://www.kdd.org/kdd2016/subtopic/view/why-should-i-trust-you-explaining-the-predictions-of-any-classifier> (LIME)
- <https://arxiv.org/abs/1711.09784> (LIME)



# Stance별 단어 분석 결과



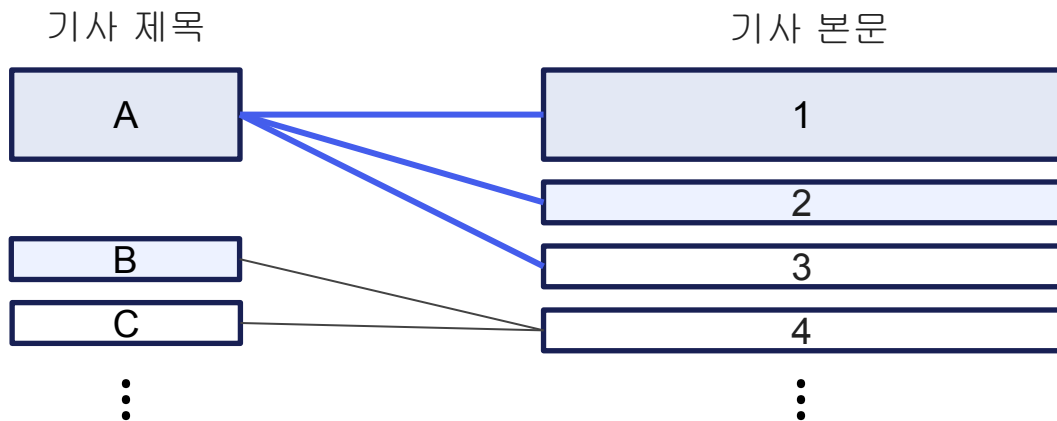
## 기사와 Stance의 상관관계 분석

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

□ **Mutual Information** 기법을 사용하여 하나의 단어가 **Stance**를 판별하는데 얼마나 영향을 미치는지 확인

1. 49,380개의 데이터를 **Stance**에 따라 **headline**과 **article body**를 합쳐서 하나의 문장으로 고려
2. 데이터에 나타나는 단어들을 모두 수집 후 각 **Stance**에서의 단어의 **Mutual Information** 계산

## 메타데이터 활용 제약 사항



Training set 데이터 : 49,972 쌍

유일한(unique) 기사 제목 : **1,643개**

한 제목 당 기사 본문 중복 매칭 횟수  
평균 : **30.41회**

유일한 기사 본문 : 1,683개

- ❑ 기사 제목이 여러 본문에 중복하여 적용되어 기사의 원본 url 파악이 어려움