NH투자증권 NH INVESTMENT & SECURITIES

DACON DATA TO VALUE

"데이터, 문화가 되다 : League1"

AI야, 진짜 뉴스를 찾아줘!

발표 목차

CONTENTS

NH투자증권 NH INVESTMENT & SECURITIES

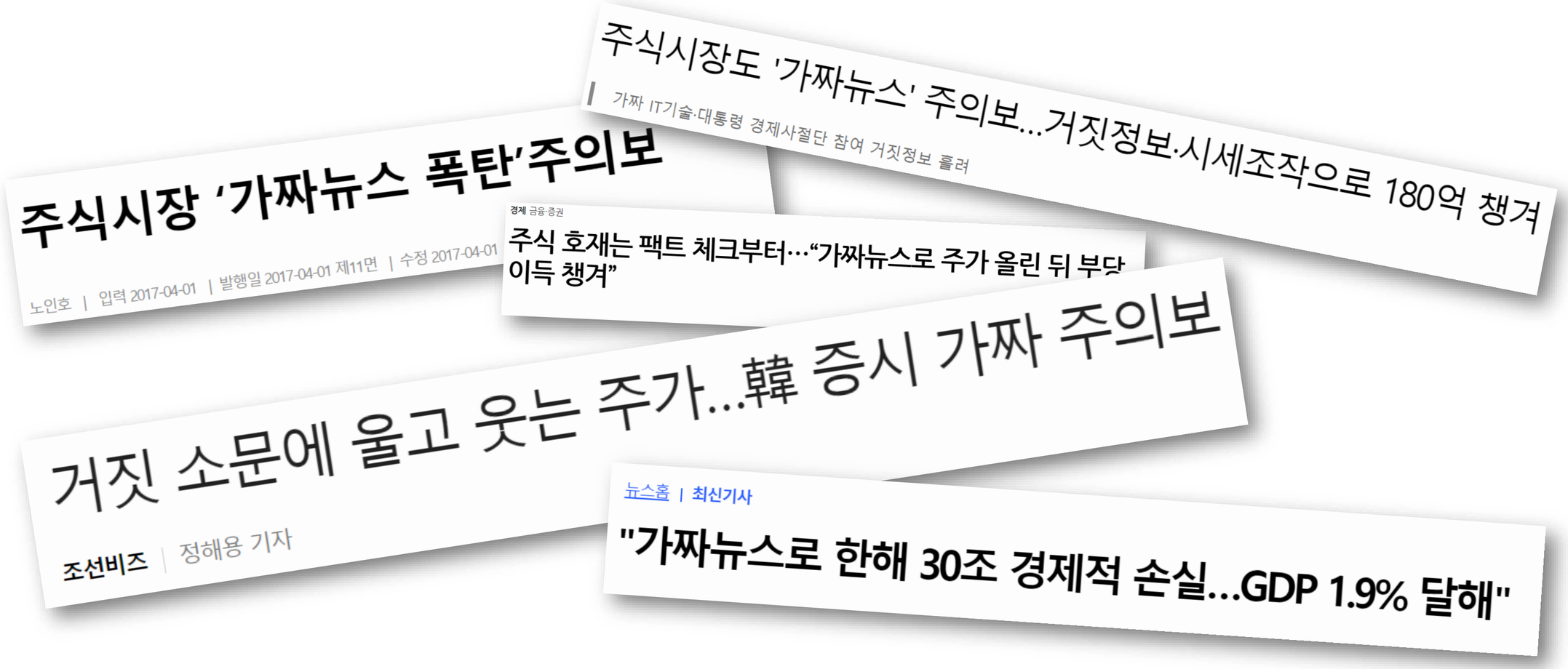
DACON DATA TO VALUE

01.

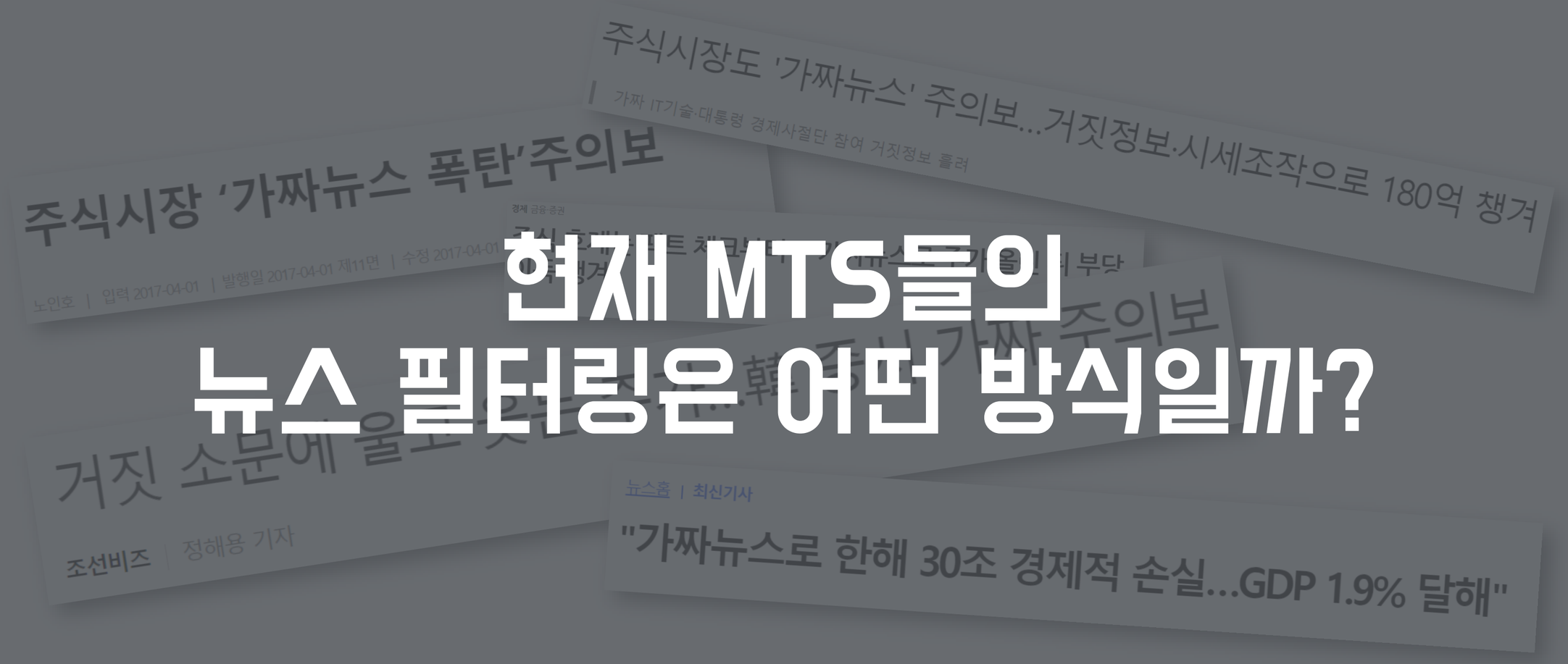
현황 분석

가짜뉴스로 인한 투자 피해와 현재 증권사들의 상황

01) 현황 분석



NH투자증권 NH INVESTMENT & SECURITIES DACON DATA TO VALUE



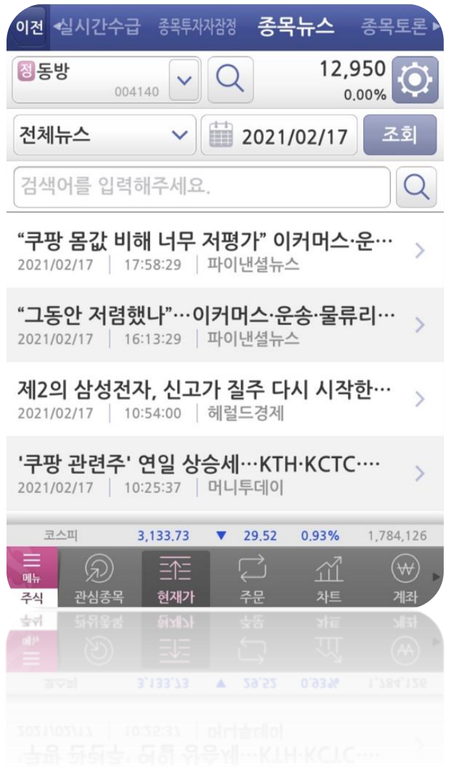
현재 MTS들의 뉴스 필터링은 어떤 방식일까?

01) 현황 분석

A 증권사

- 'AI' 기반의 스팸성 광고뉴스 필터링 제공
- 2019년 7월에 광고뉴스 필터링 특허 등록
- 단어 뿐 아니라 뉴스의 문맥 의미 해석

B 증권사



- 1차적으로 단어 필터링을 거친 뉴스들에 대해 사용자가 필터링 단어를 추가 지정하는 기능 제공
- 사용자 필터링 가능 단어의 선택지 기본 제공, 추가 단어 지정 가능

C 증권사

- 머신러닝 기반의 뉴스 필터링 제공
- 2019년 10월에 MTS 적용
- 일 평균 8천여건의 6개월 치 뉴스를 머신러닝으로 학습

NH투자증권 NH INVESTMENT & SECURITIES DACON DATA TO VALUE

01) 현황 분석

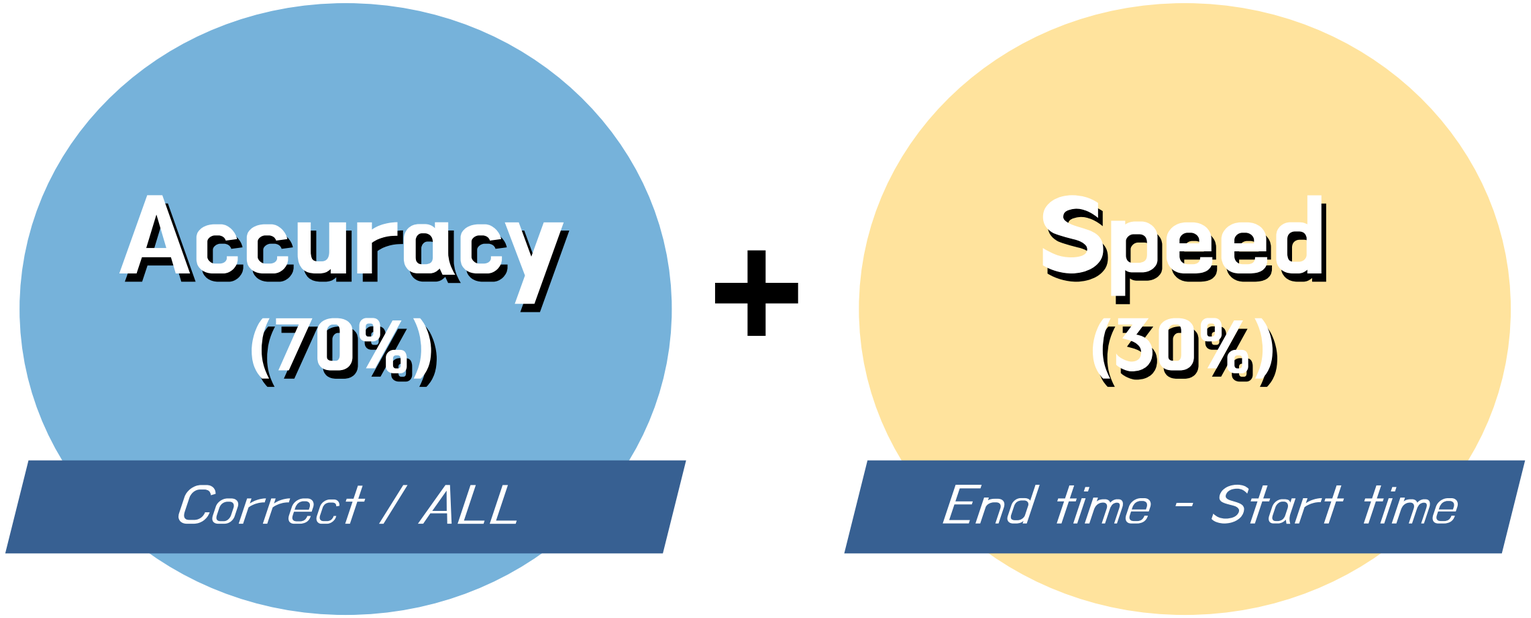
NAMUH

- 리서치 보고서 분석 및 종목 추천, 로보 어드바이저 등 AI 기반의 서비스들이 제공되고 있다.
- 뉴스 제공 서비스의 경우 HTS에는 단어 기반의 필터링이 제공되고 있지만, MTS는 필터링 기능 확인 불가
- 단어 필터링을 보완한 AI 알고리즘 기반의 뉴스 필터링 필요

02.

데이터 탐색 및 인사이트 도출

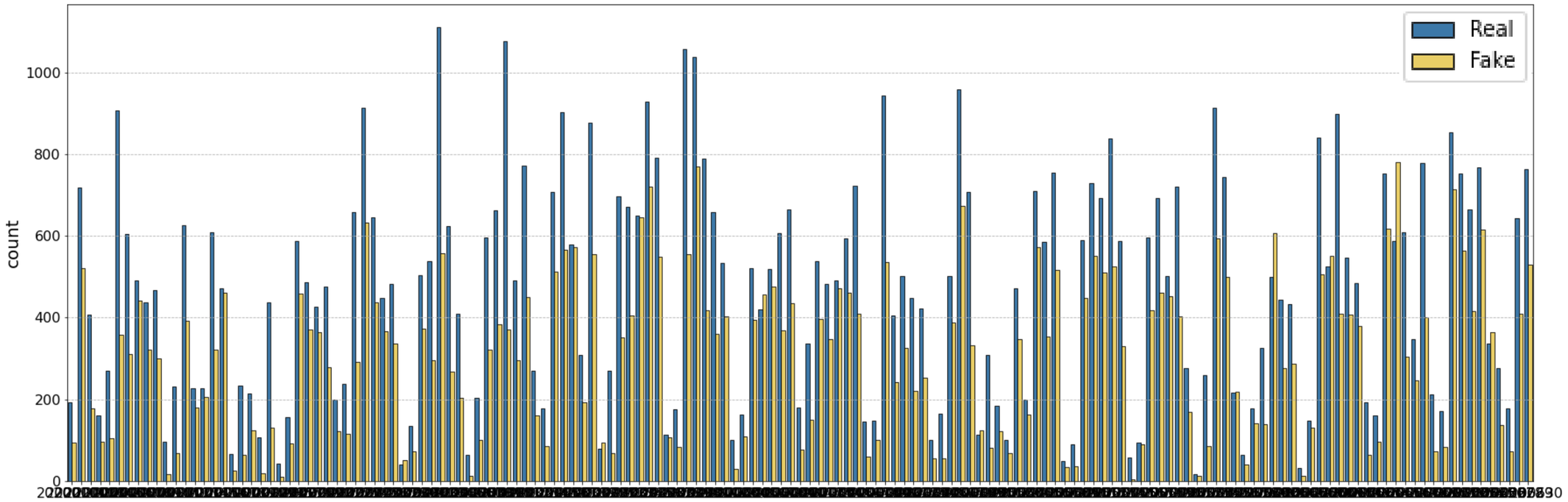
데이터 전반을 이해하기 위한 탐색결과와 인사이트 정리



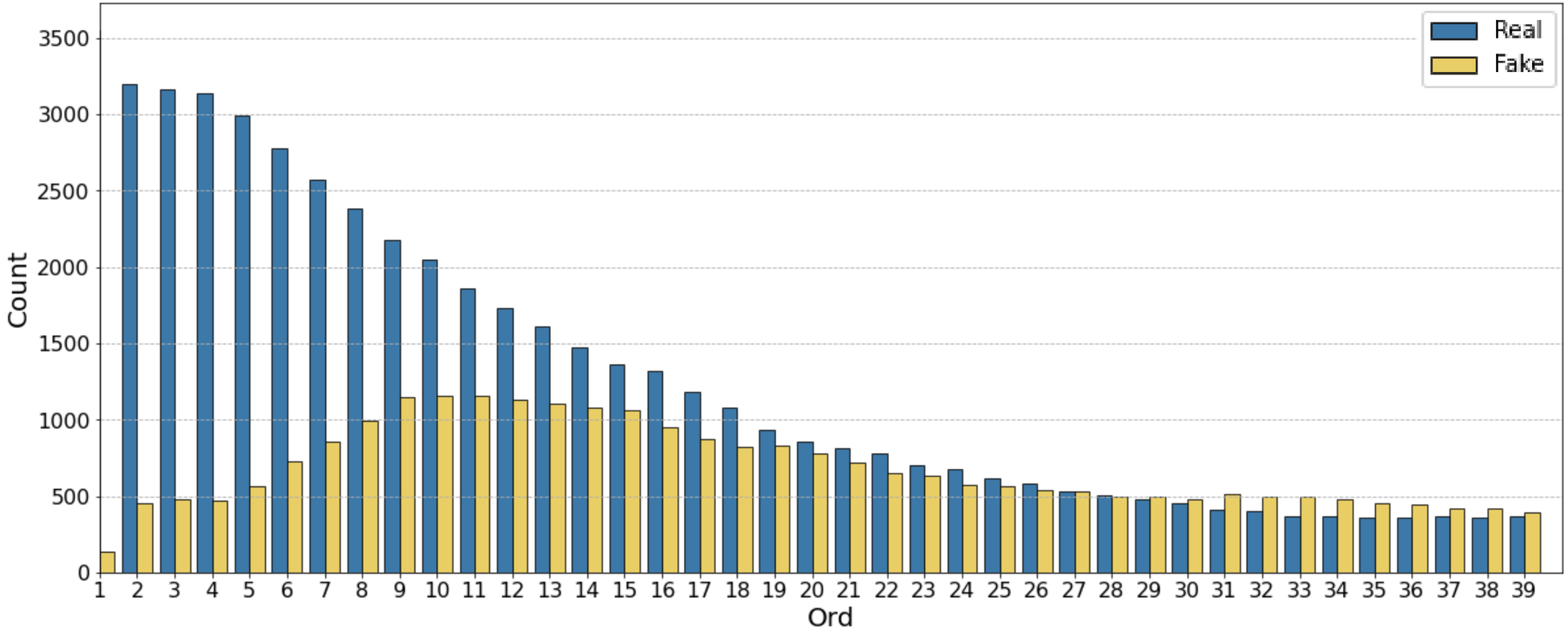
정확성이 높으면서도

빠르게 탐지할 수 있는 딥러닝 모델 개발!

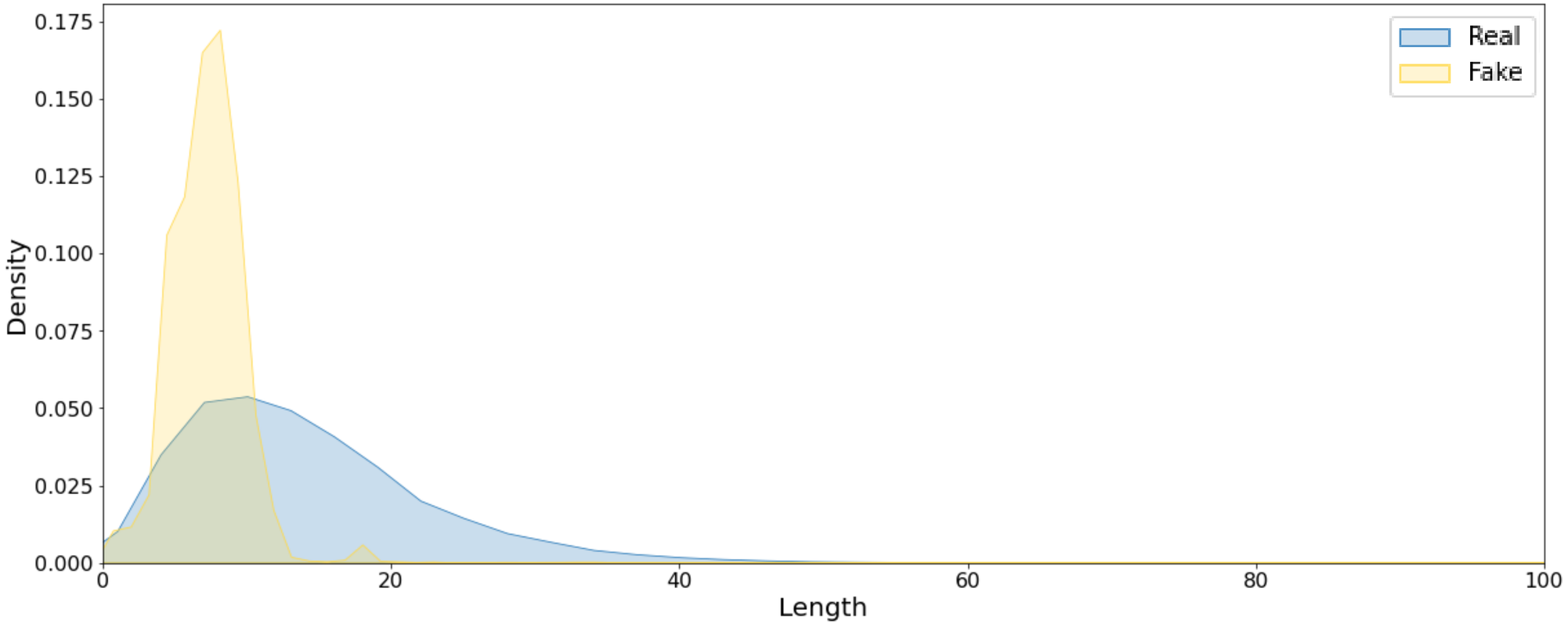
date



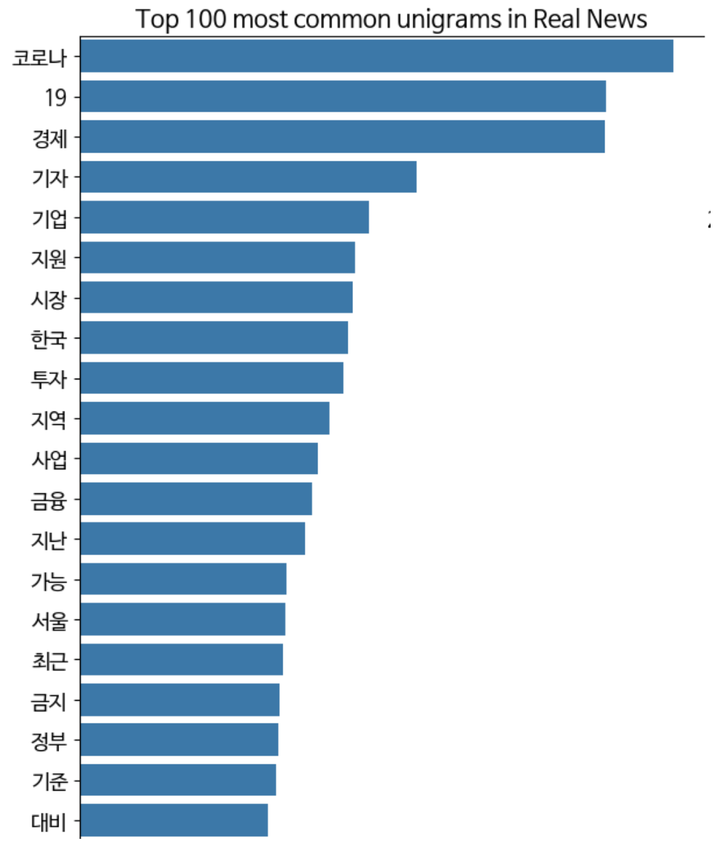
ord



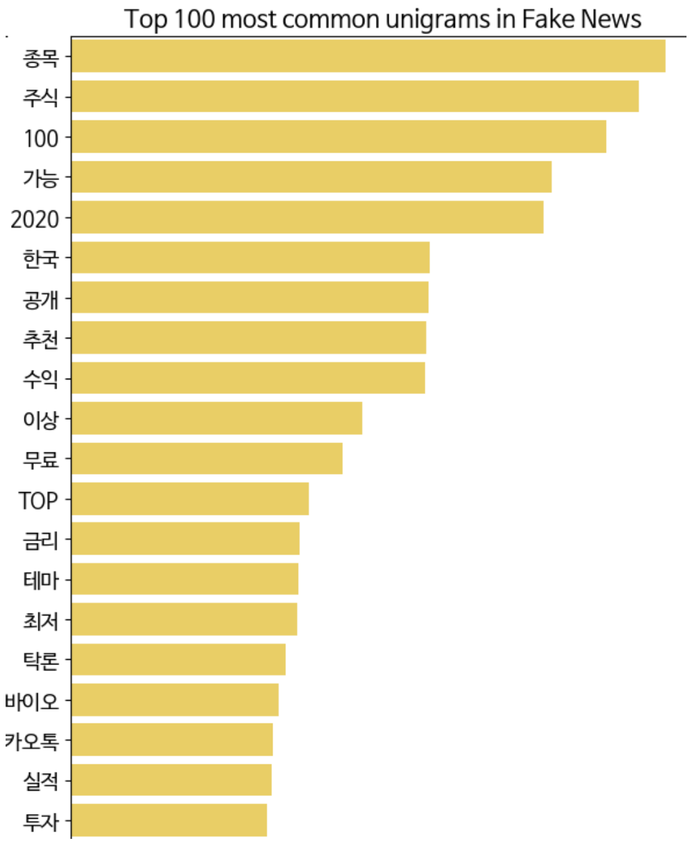
Word Level Length



단어 빈도 분석



Real News



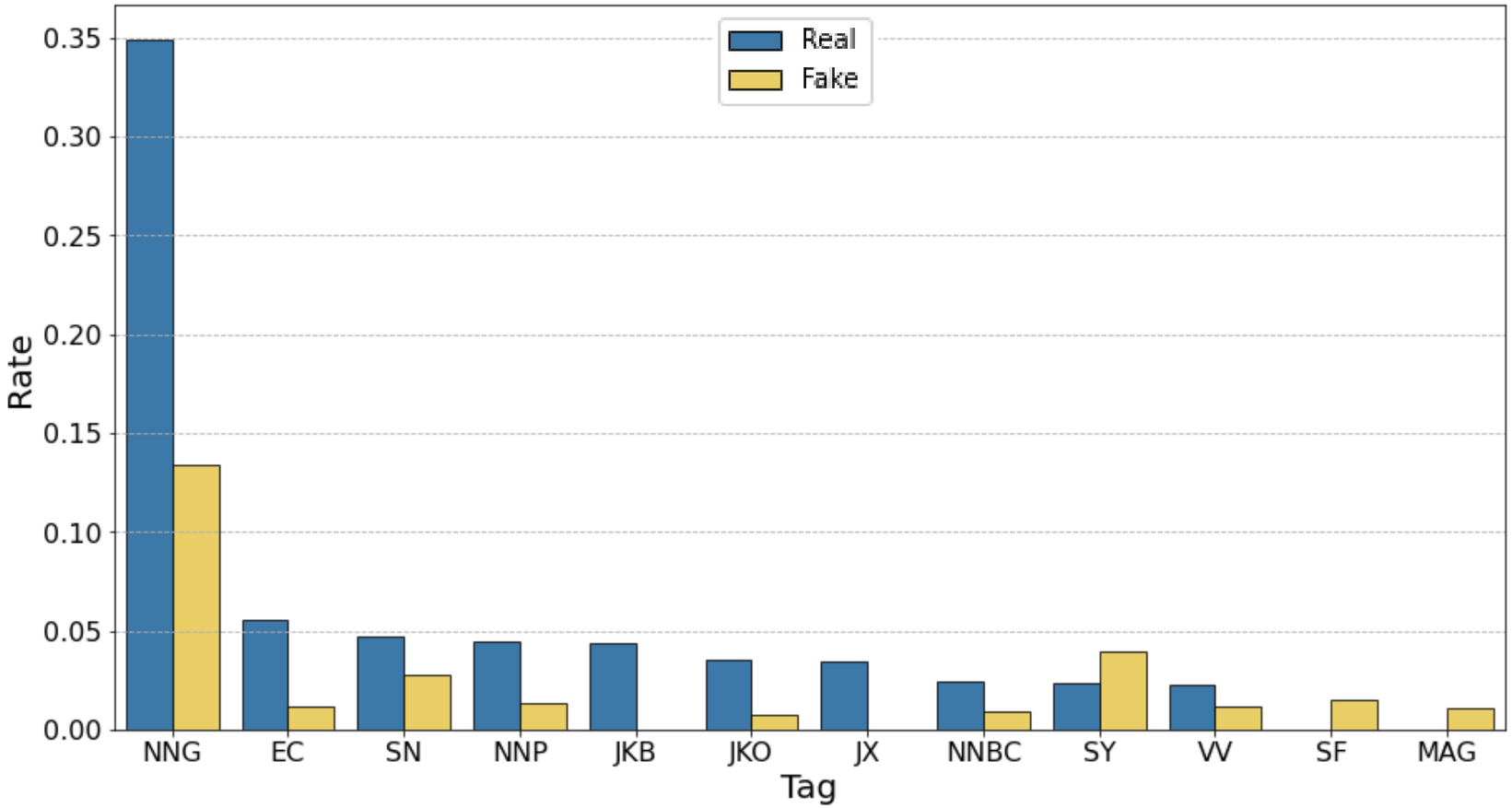
Fake News

Word Cloud

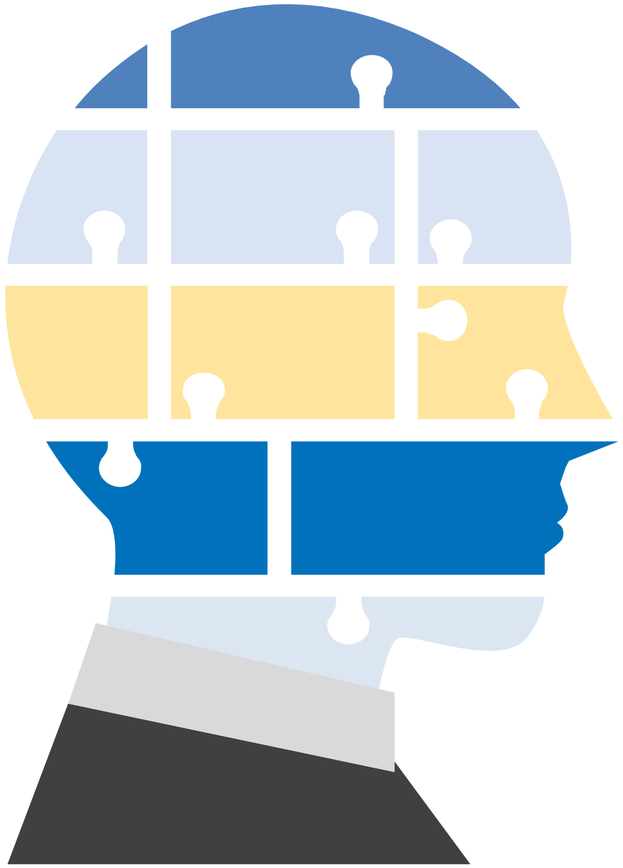
Real News

Fake News

형태소 분석



INSIGHT !



01 날짜 변수는 진짜뉴스와 가짜뉴스 판별에 큰 도움이 되지 않을 것이다

02 순서 변수는 진짜뉴스와 가짜뉴스를 판별하는데 도움을 줄 것으로 보인다

03 텍스트의 길이는 진짜뉴스와 가짜뉴스 판별에 도움을 줄 것으로 보인다.

04 진짜뉴스와 가짜뉴스는 단어 단위로도 충분히 상이한 내용을 담고 있다

05 진짜뉴스와 가짜뉴스의 단어들은 매우 다른 분위기를 가지고, 가짜뉴스는 사전에 정의되지 않은 단어가 등장한다

06 진짜뉴스와 가짜뉴스는 문법적으로 다른 문장 구조를 가지는 것으로 보인다

아이디어 도출

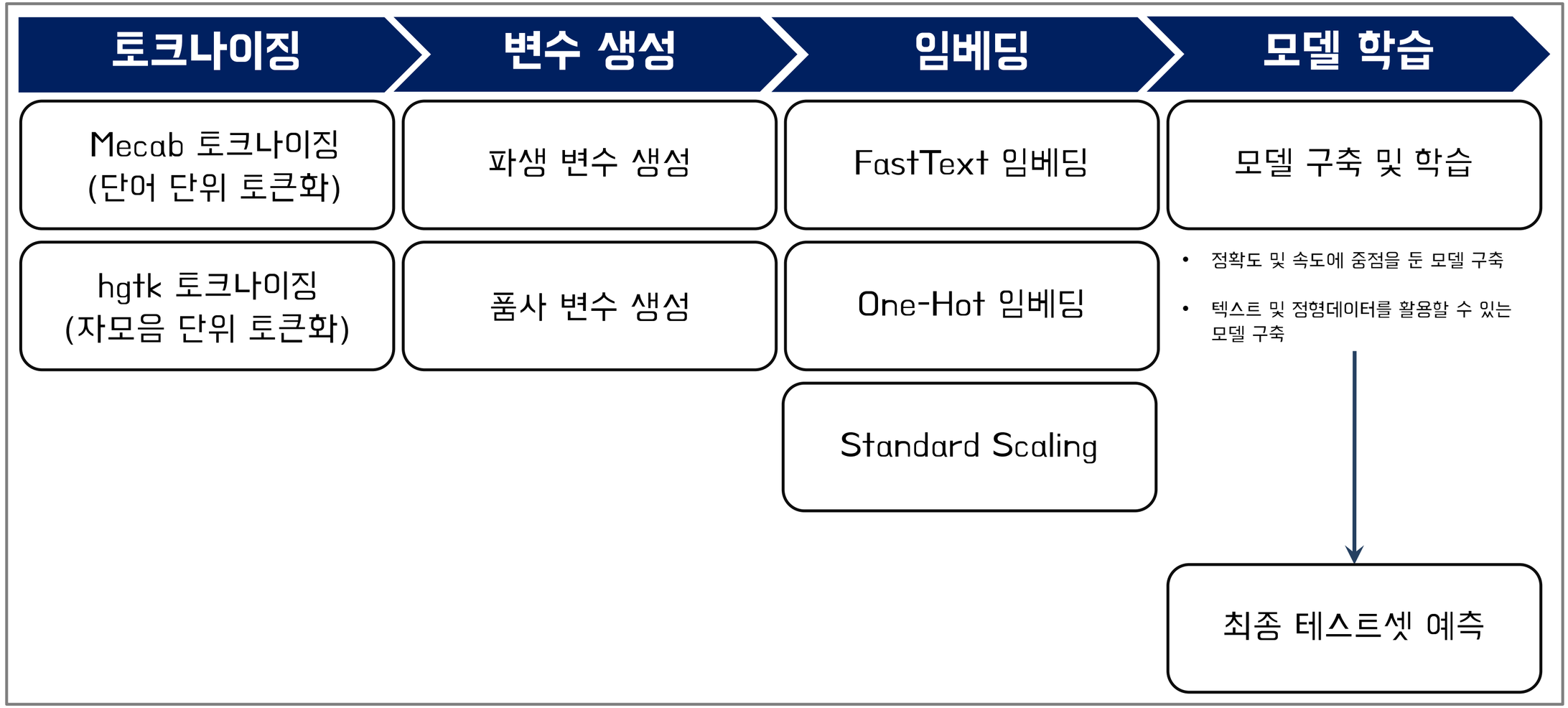
- EDA를 통해 알아낸 분류에 **도움이 될만한 변수들을 모델에 사용**해보자
- 모델복잡도와 시간을 고려하여 **단어 단위로도 좋은 성능을 내는 무겁지 않은 모델을 구축**해보자
- Mecab 형태소 분석기에 정의되어 있는 **품사를 활용하여 변수를 생성**해보자
- **신조어에 유연**하고 다른 분위기의 문장을 파악하기 위해 **특징을 잘 잡아낼 수 있는 모델**을 사용해보자

03.

변수 생성

인사이트와 그래프로 설명하는 논리성 있는 변수 생성 및 선택

워크플로우



Tokenizing:

주어진 텍스트를 토큰(token)이라 불리는 개별 의미를 가지는 단위로 나누는 작업

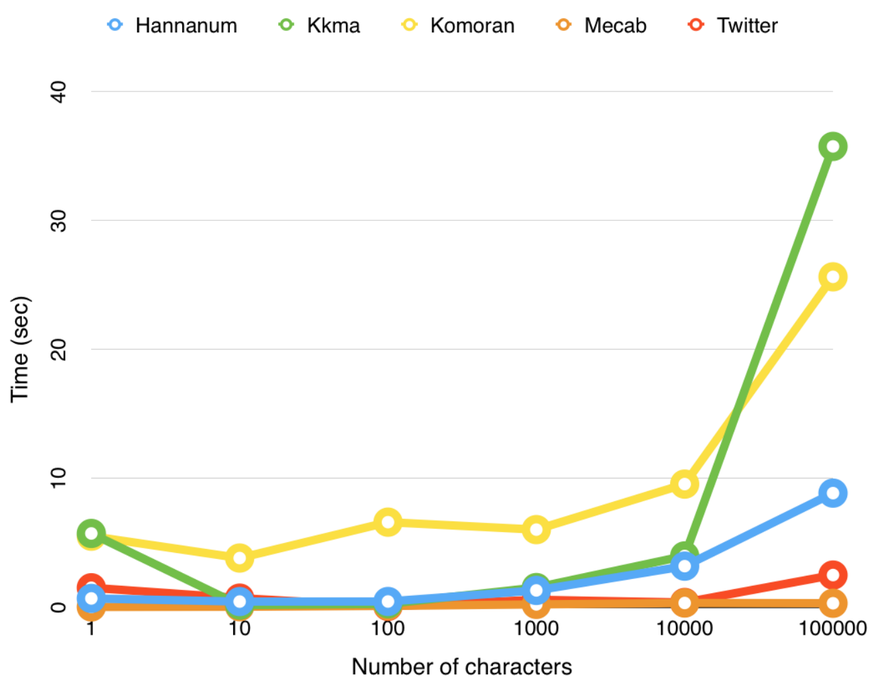
1) 단어 단위 토크나이징

2) 자음/모음 단위 토크나이징

단어 단위 토큰화

KonlPy 형태소 분석기 사용

- Mecab은 대용량 말뭉치를 사용해도 빠른 수행시간을 보여줌
- 또한, 상세한 tagset을 통해 고품질의 품사 태깅이 가능
- 처리 속도가 중요하다는 점과 품사를 통한 파생변수 생성을 위해 최종 토크나이저로 Mecab을 선정



<문자 개수에 따른 한글 형태소 분석기별 수행속도>

Mecab.tagset

: Mecab에서 제공하는 한국어 품사 태그 셋

<(표) Mecab의 품사와 태그정보>

품사	태그	설명	품사	태그	설명	태그	설명	태그	설명
체언	NNG	일반 명사	부사	MAG	일반 부사	EC	연결 어미	SSC	)]
	NNP	고유 명사		MAJ	접속 부사	EF	종결 어미	SSO	([
	NNB	의존 명사	감탄사	IC	감탄사	EP	선어말어미	SY	기타 기호
	NNBC	단위 표현 명사	조사	JKS	주격 조사	ETM	관형형 전성 어미	XPN	체언 접두사
	NR	수사		JKC	보격 조사	ETN	명사형 전성 어미	XR	어근
	NP	대명사		JKG	관형격 조사	SC	구분자, - / :	XSA	형용사 파생 접미사
용언	VV	동사		JKO	목적격 조사	SE	줄임표 …	XSN	명사 파생 접미사
	VA	형용사		JKB	부사격 조사	SF	. ? !	XSV	동사 파생 접미사
	VX	보조 용언		JKV	호격 조사	SH	한자	SN	숫자
	VCP	금정 지정사		JKQ	인용격 조사	SL	외국어		
	VCN	부정 지정사		JX	보조사				
			관형사	MM	관형사				

캐릭터 단위 토큰화

hgtk 라이브러리 사용

- 2017년에 공개된 한글 자모 분해 오픈소스 라이브러리
- 한글 자모 분해, 조합(오토마타), 조사 붙이기, 초/중/종 분해조합, 한글/한자/영문 여부 체크 등을 지원
- ᴥ 라는 문자를 자음과 모음의 구분자로 사용

ㄱ	ㄴ	ㄷ	ㄹ	ㅁ
0	4	0	3	1
3	1	4	2	2
1	2	6	2	5
6	8	5	7	10

<캐릭터 단위 매트릭스 표현 방식>

Feature Engineering:

모델이 학습하는데 도움을 줄 수 있는 여러가지 정보를 추출하여 변수로 생성하는 작업

1) 길이 및 순서에 관한 변수

2) Mecab 품사 태깅 변수

3) 제목과 내용의 유사성에 관한 변수

4) 자음/모음 관련 변수

길이 및 순서에 관한 변수

1) len

: n_id 별 content의 개수

2) ord/len

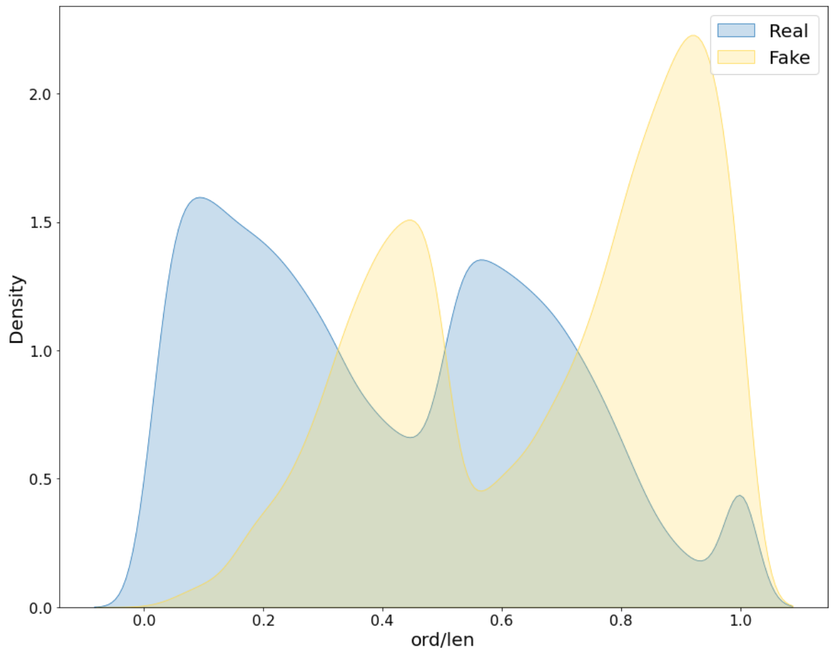
: 한 뉴스 기사 내에서 해당 content가 위치한 순서를 비율로 나타낸 것

3) content_len

: Content와 Title의 문장을 합친 후 전체 텍스트의 길이 반환

4) content_word_count

: Content와 Title의 문장을 합친 후 띄어쓰기 단위로 자른 후 단어 개수 반환



<ord/len의 밀도 그래프>

Mecab 품사 태깅 변수

1) special_char_len

: 정규표현식을 통해 텍스트 내의 특수문자 및 숫자 추출 (한자, 숫자, 기타기호 등)

2) content_singleSY

: Mecab의 SY 태그를 활용한 특수문자 count

3) about_num

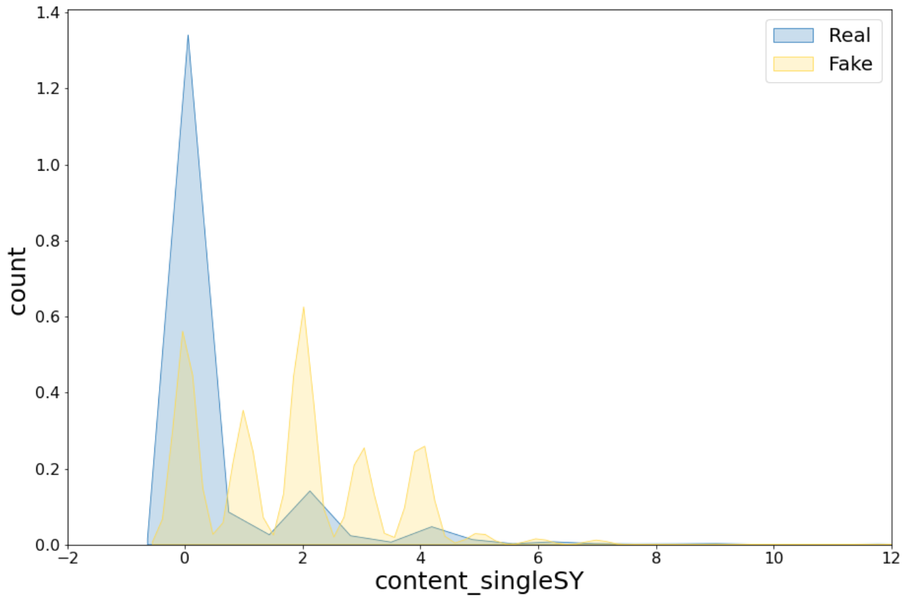
: content에서 품사가 수사이거나 숫자이면 count

4) content_variable4

: content에서 긍정, 부정, 접속부사, 접속조사의 단어 수 count

5) content_variable5

: content_variable4에서 더해준 품사에 추가로 수사를 더한 후 count



<content_singleSY의 밀도 그래프>

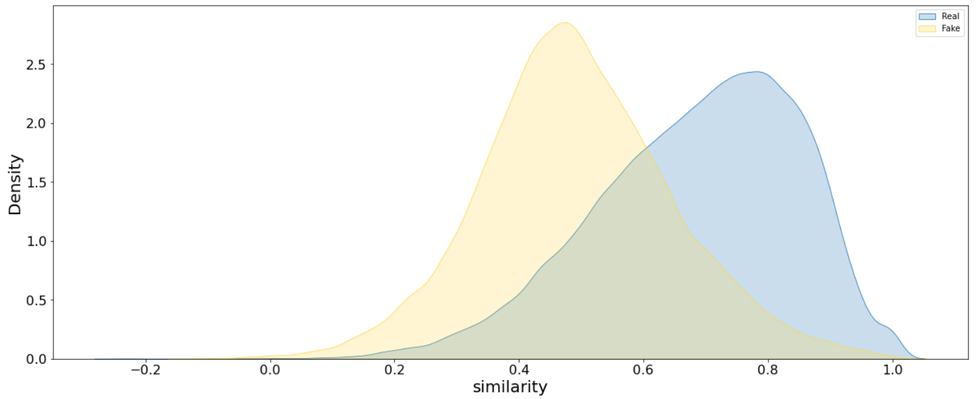
제목과 내용의 유사성에 관한 변수

1) title_in_content_noun

: Title에 있는 명사인 단어가 content의 텍스트에 포함되는 단어라면 count

2) title_content_similarity

: Title과 content간의 코사인 유사도 도출



<title_content_similarity의 밀도 그래프>

자음/모음 관련 변수

1) per_eng

: character 단위로 분리된 토큰 중 알파벳 토큰의 비율

2) per_digit

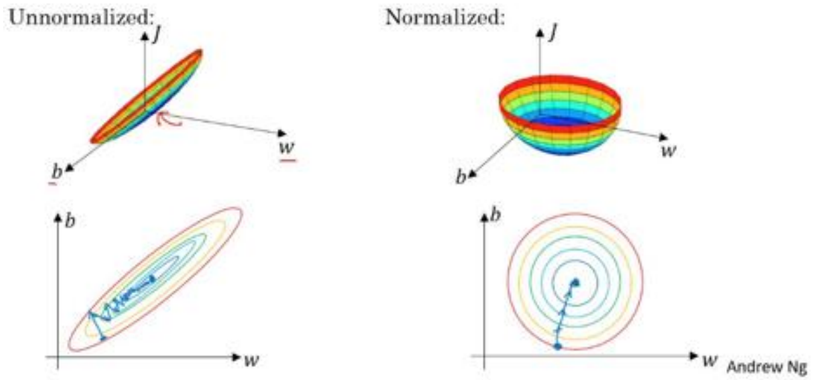
: character 단위로 분리된 토큰 중 숫자 토큰의 비율

변수 스케일링

Standard Scaling

정형 데이터 스케일링

- Scaling을 거친 데이터의 경우 쉽게 최적값에 도달할 수 있음
- 학습률을 상대적으로 높여서 사용할 수 있기 때문에 빠르게 훈련시킬 수 있음



* Scaling을 해주는 이유는?

: 신경망의 학습을 빠르게 할 수 있기 때문, Scaling을 거친 데이터의 경우 쉽게 최적값에 도달할 수 있으며, 학습률을 상대적으로 높여서 사용할 수 있기 때문에 빠르게 훈련시킬 수 있다.

Embedding:

자연어를 컴퓨터가 이해하고, 효율적으로 처리하게 하기 자연어를 적절히 변환해주는 작업

1) FastText Embedding

2) One-Hot Embedding

텍스트 임베딩

FastText Embedding

단어 토큰 임베딩

- 등장하는 주변단어 벡터는 중심단어 벡터와 가까워지게끔 학습하는 방식인 Word2Vec의 방법론을 따라가면서 부분단어(subword)의 벡터들로 표현한다는 특징을 가지는 임베딩 기법
- 빠른 연산량과 오탈자에 유연하다는 장점

One-Hot Embedding

자음/모음 토큰 임베딩

- 컬럼의 단위가 단어가 되며 각 레코드마다 그 단어가 얼마나 출연하는지에 대한 빈도수가 담겨있는 행렬
- 자음/모음 토큰의 경우 컬럼이 텍스트에 등장하는 자음과 모음을 의미한다.

최종 모델 INPUT

TEXT Title+Content	단어 단위 텍스트 (FastText 임베딩)
TEXT Title+Content	자모 단위 텍스트 (One-Hot 임베딩)
TABULAR 총 13개의 정형데이터 Feature	len ord/len content_len content_word_count special_char_len content_singleSY about_num content_variable4 content_variable5 title_in_content_noun title_content_similarity per_eng per_digit

04.

모델 선정 및 학습 결과

여러 조건에 부합하는 모델 선정, 파라미터에 따른 학습결과

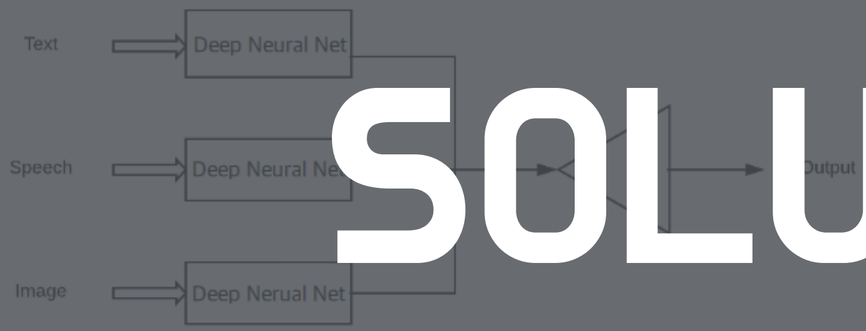
모델 선정에 대한 3가지 조건 !

1. 신조어에 민감하지 않으면서 텍스트 자체의 전체적인 분위기와 특징을 잘 파악할 수 있는 모델

2. 단어 단위로 임베딩한 텍스트를 입력값으로 받는 가벼운 모델

3. 분류에 있어서 도움을 줄 수 있는 텍스트 외 정형 데이터를 사용할 수 있는 모델

MultiModal Neural Network



- MultiModal Learning은 이미지, 음성, 텍스트, 정형 데이터등
  다양한 형태의 데이터를 학습하여 서로 다른 소스의 정보를 결합하는 모델
- MultiModal 모델을 통해 여러 소스의 특징 추출이 포함되어
  더 [illegible]로 예측할 수 있음
- 상호 보완적인 정보를 입력받아 보이지 않는 패턴 반영이 가능함

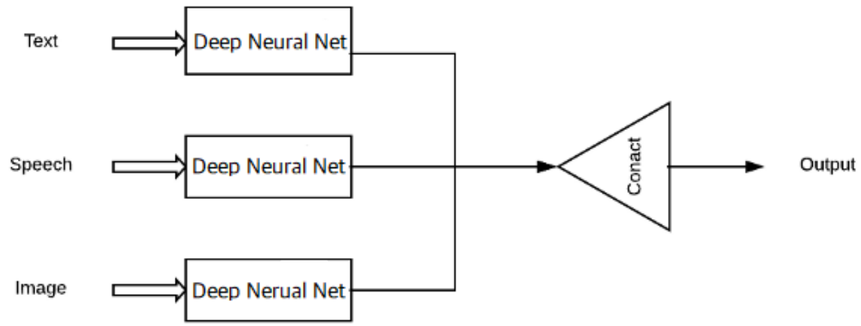
SOLUTION!

기본 사용 데이터인 텍스트를 상호 보완적인 임베딩 방식으로 입력받고,

추가적인 정보를 정형 데이터로 받을 수 있는

MultiModal 모델을 구축함으로 모델의 예측력을 향상시키고자 함

MultiModal Neural Network



- MultiModal Learning은 이미지, 음성, 텍스트, 정형 데이터등
  다양한 형태의 데이터를 학습하여 서로 다른 소스의 정보를 결합하는 모델
- MultiModal 모델을 통해 모든 소스의 특징 추출이 포함되어
  더 큰 규모로 예측하는데 기여할 수 있음
- 상호 보완적인 정보를 입력받아 보이지 않는 패턴 반영이 가능함

기본 사용 데이터인 텍스트를 상호 보완적인 임베딩 방식으로 입력받고,

추가적인 정보를 정형 데이터로 받을 수 있는

MultiModal 모델을 구축함으로 모델의 예측력을 향상시키고자 함

Character Level CNN
For Text Pattern

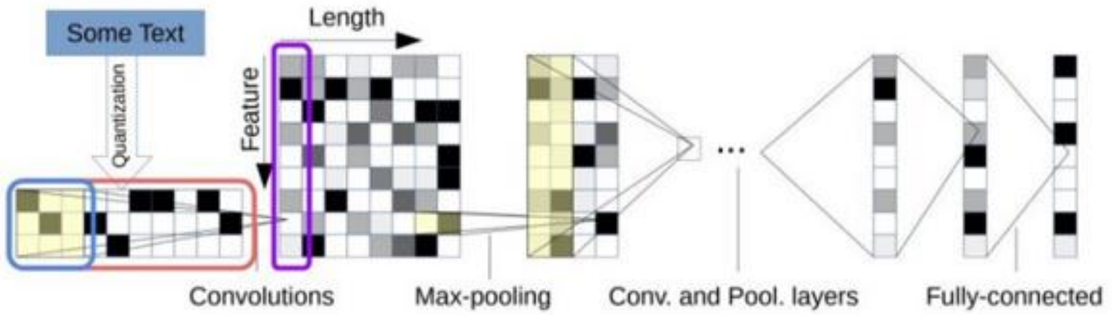
+

LSTM
For Sequential Text

+

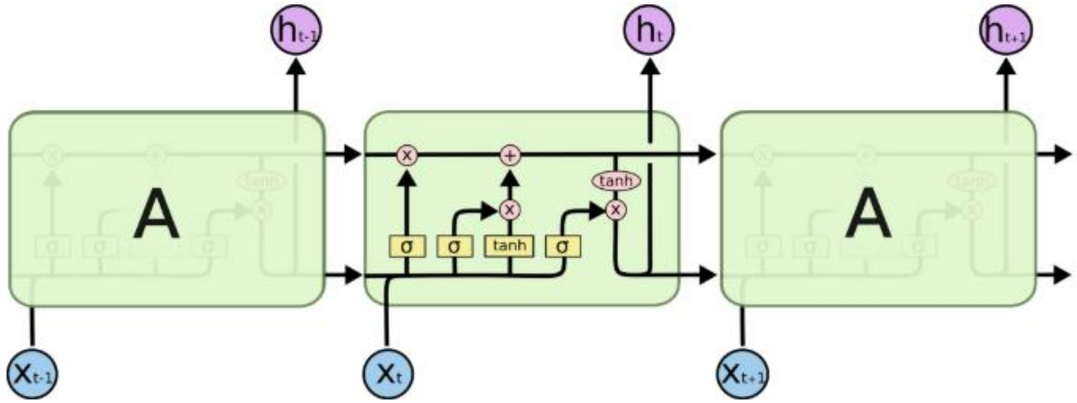
Dense Layer
For Tabular Data

Character Level CNN



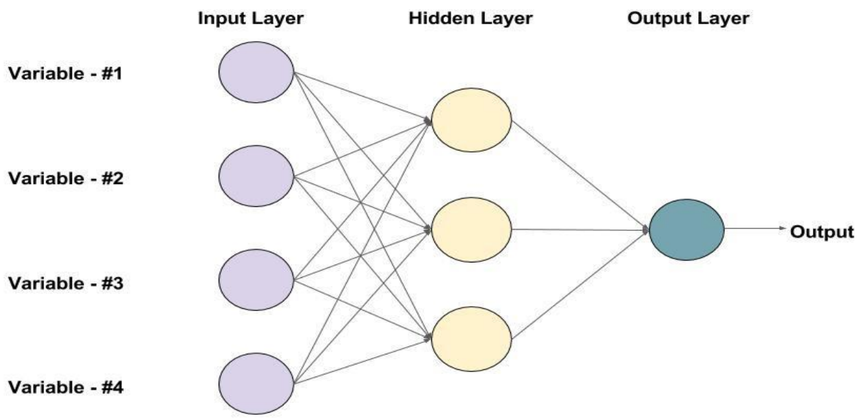
- 신조어에 민감하지 않으면서 텍스트 자체의 전체적인 분위기와 특징을 잘 파악할 수 있는 모델
- 진짜 뉴스와 가짜 뉴스에 등장하는 단어는 매우 다르며, 문장의 분위기 또한 매우 다름
- 또한 가짜뉴스에는 사전에 정의되어 있지 않은 단어들이 많이 분포함
- 한글 character 단위인 자음, 모음으로 분해한 후 임베딩을 진행
- Dilation 파라미터의 단위를 다르게 하여 적은 연산량으로 문장을 다양한 스케일로 파악할 수 있도록 함

LSTM



- 시퀀스 데이터를 처리하는 모델로, 기사 문장의 문맥 파악 가능
- 앞서 데이터 탐색 결과 뉴스 문장의 길이는 다양한 분포를 가졌음
- 시퀀스가 길더라도 전체 문맥을 파악할 수 있음
- 토큰화된 뉴스 내용 Embedding한 벡터에 convolution, pooling과 같은 layer를 적용한 후 LSTM을 적용

Dense Layer



- 분류에 도움이 될만한 기존 변수들과 함께 EDA 과정에서 얻은 인사이트를 통해 생성한 변수를 입력값으로 받음
- 정형 데이터를 신경망으로 학습시키기 위해 Dense layer를 사용
- 총 13개의 변수를 정형 데이터 입력값으로 사용

Input 1
자음/모음 One-Hot embedding

Input 2
단어 토큰 FastText embedding

Input 3
정형 데이터 Standard Scaling

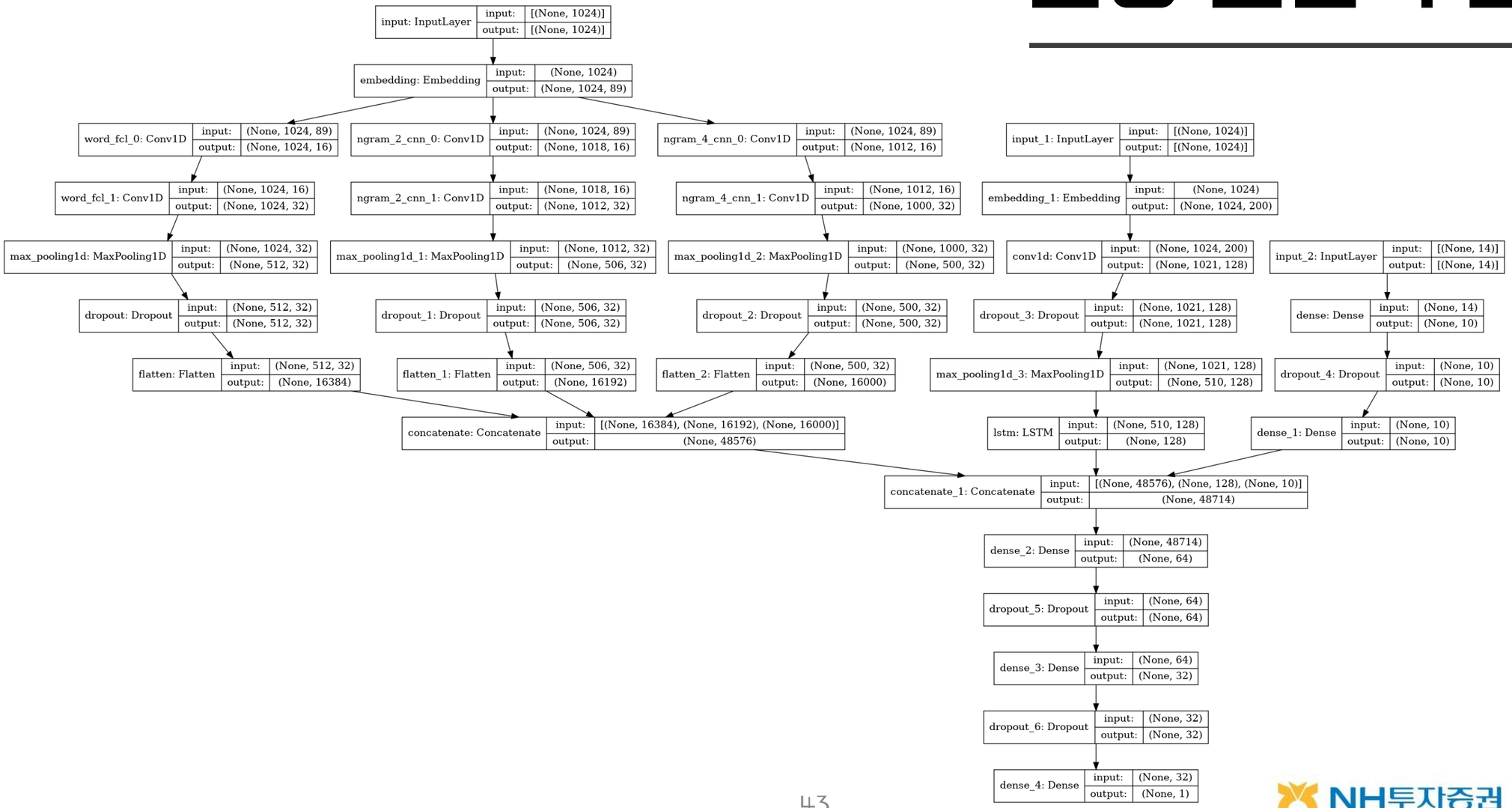
Character Level CNN
For Text Pattern

LSTM
For Sequential Text

Dense Layer
For Tabular Data

concatenate

최종 모델 구조



모델 결과

정확도 (Accuracy)

- Public Score: **99.13%**
- Private Score: **98.74%**

예측 소요 시간 (Time)

- NH투자증권 제공 142,565개 데이터:

  1m27s 소요

한 케이스 데이터 분류(17개의 n_id):

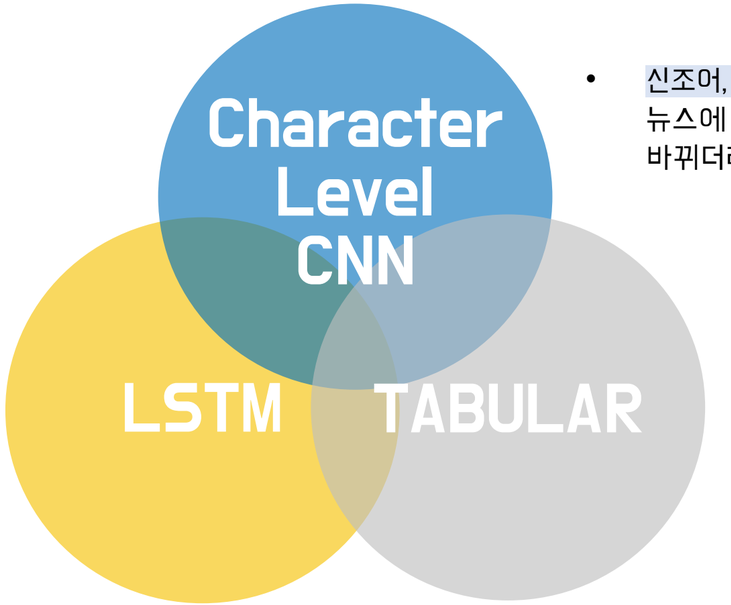
119ms. 초 단위 환산 시 0.000119s 소요

05. 결론 및 활용방안 제시

모델 활용 장점부터 실제 적용할 수 있는 서비스 방안까지

모델 활용 장점

MULTIMODAL을 통해 높은 정확도와 빠른 속도를 갖춘 알고리즘



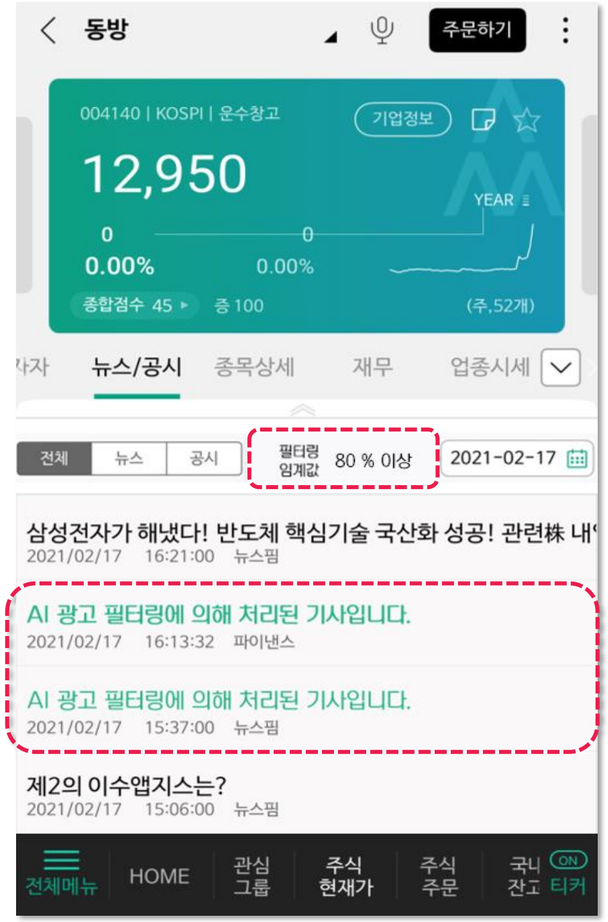
- 신조어, 오탈자, 띄어쓰기 등에 강하기 때문에 향후 뉴스에 새로운 신조어들이 많이 나타나 단어집이 바뀌더라도 유연하게 대처 가능

- 무거운 언어모델 구조보다 비교적 가벼우며, 시퀀셜한 텍스트의 문맥 파악 가능

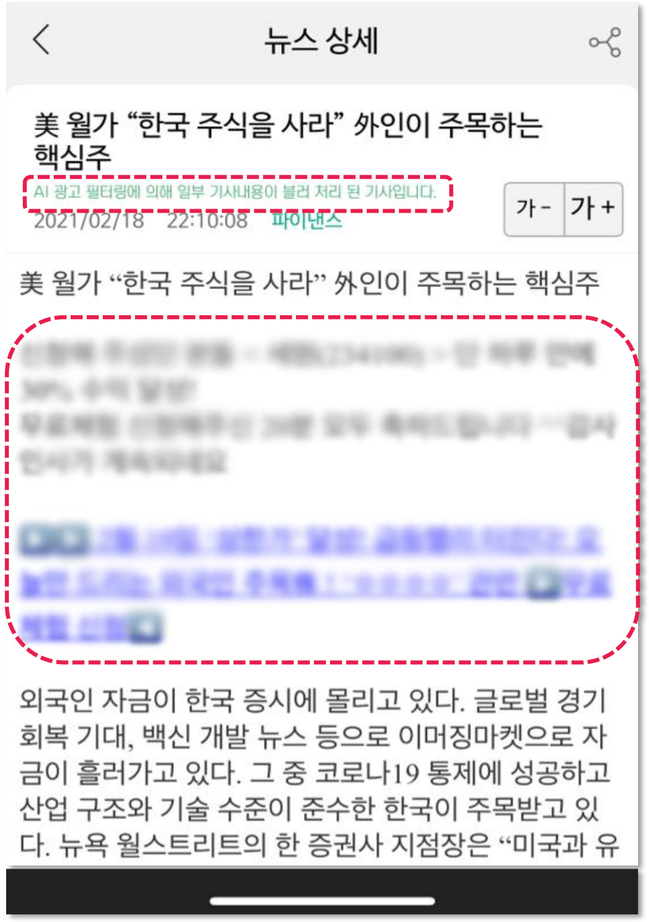
- 다양한 변수를 활용하여 단순 문맥 정보 외의 다양한 정보 반영 가능

고성능/무거운 모델을 사용하지 않고도 가볍고, 정확한 모델 사용 가능
→ 실시간 진짜/가짜 뉴스 탐지에 유리함

서비스 방안 제시



<임계값 조정 기능>



<블러 처리 기능>

뉴스 필터링 임계값 조정 기능

- 필터링 임계값을 사용하여 사용자 주관에 따른 기사 노출 선택 가능
- 사용자가 선택한 임계값보다 높은 비율은 갖는 가짜뉴스는 노출 제외

가짜 뉴스 텍스트 블러처리 기능

- AI로 분류된 가짜뉴스의 텍스트만 블러 처리하여 정확한 정보만 제공 가능
- 사용자가 옵션을 선택하여 블러 처리 여부 선택 가능

참고 기사

http://www.hani.co.kr/arti/economy/finance/864237.html

https://www.yeongnam.com/web/view.php?key=20170401.0101107334200001

http://www.bosa.co.kr/news/articleView.html?idxno=2134535

참고 논문

https://papers.nips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

https://static.googleusercontent.com/media/research.google.com/ko//pubs/archive/43905.pdf

https://papers.nips.cc/paper/2018/file/500e75a036dc2d7d2fec5da1b71d36cc-Paper.pdf

https://icml.cc/2011/papers/399_icmlpaper.pdf

https://www.sciencedirect.com/science/article/abs/pii/S0167739X21000340

Q&A

감사합니다