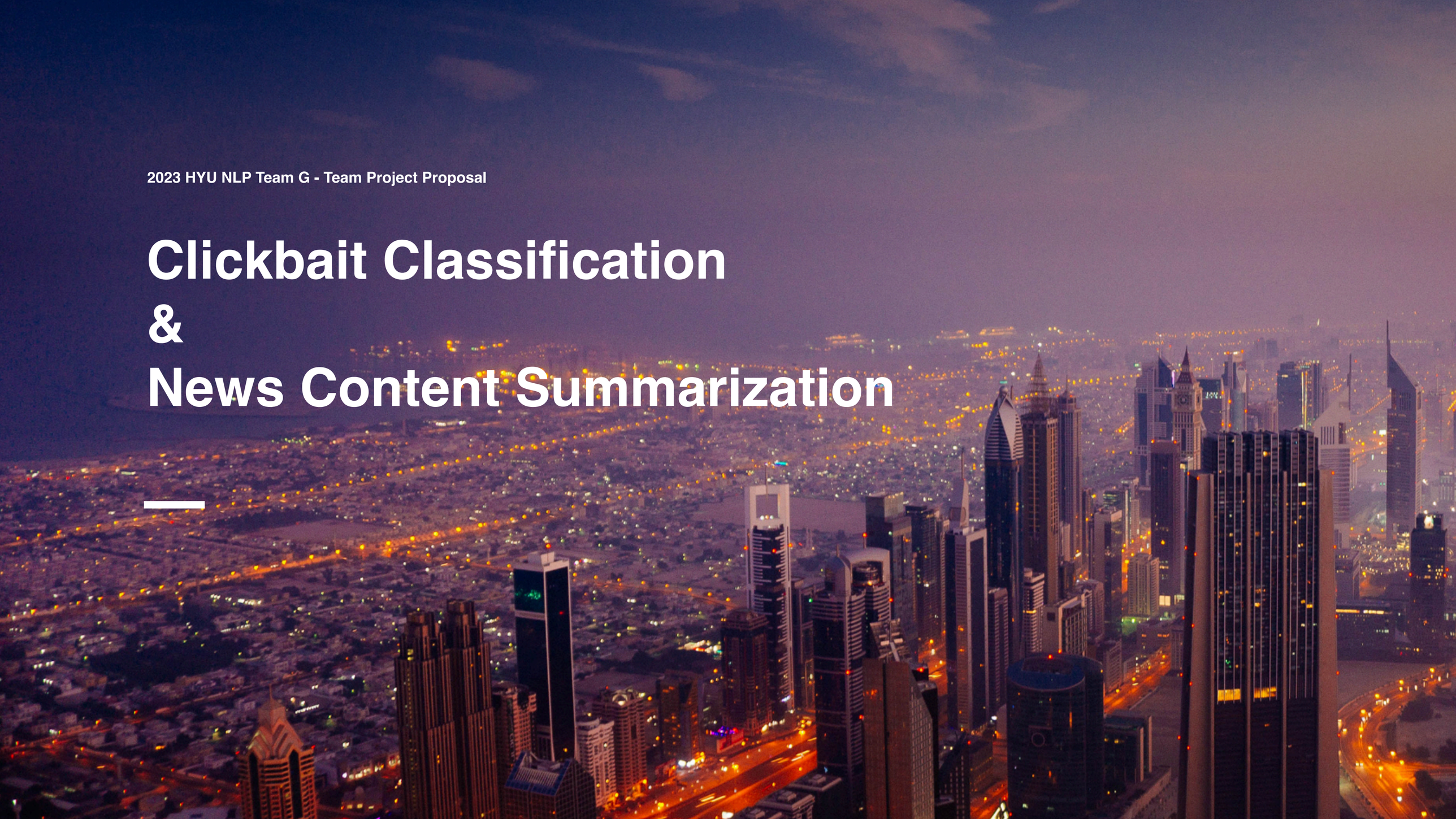
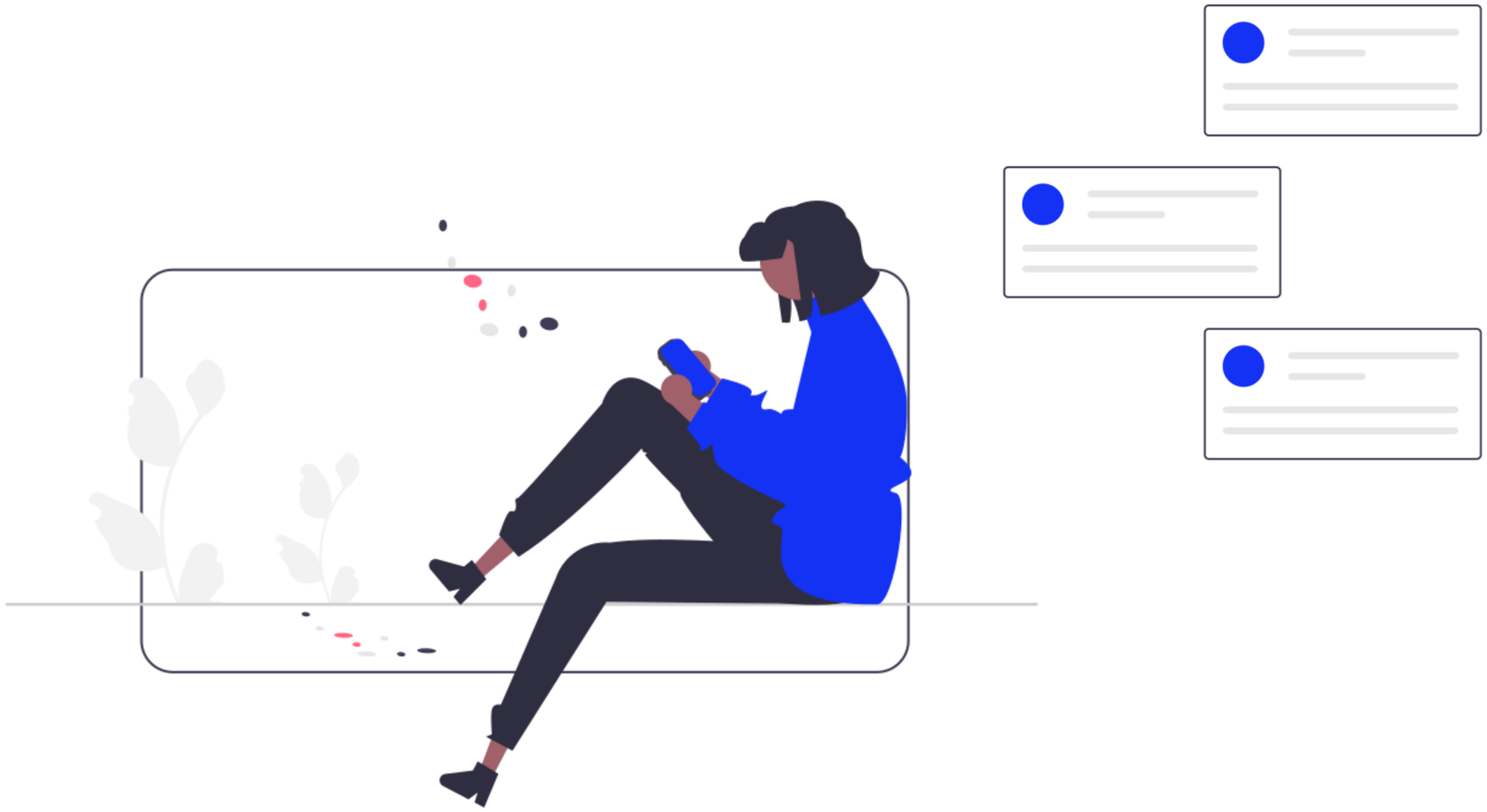
2023 HYU NLP Team G - Team Project Proposal

Clickbait Classification & News Content Summarization



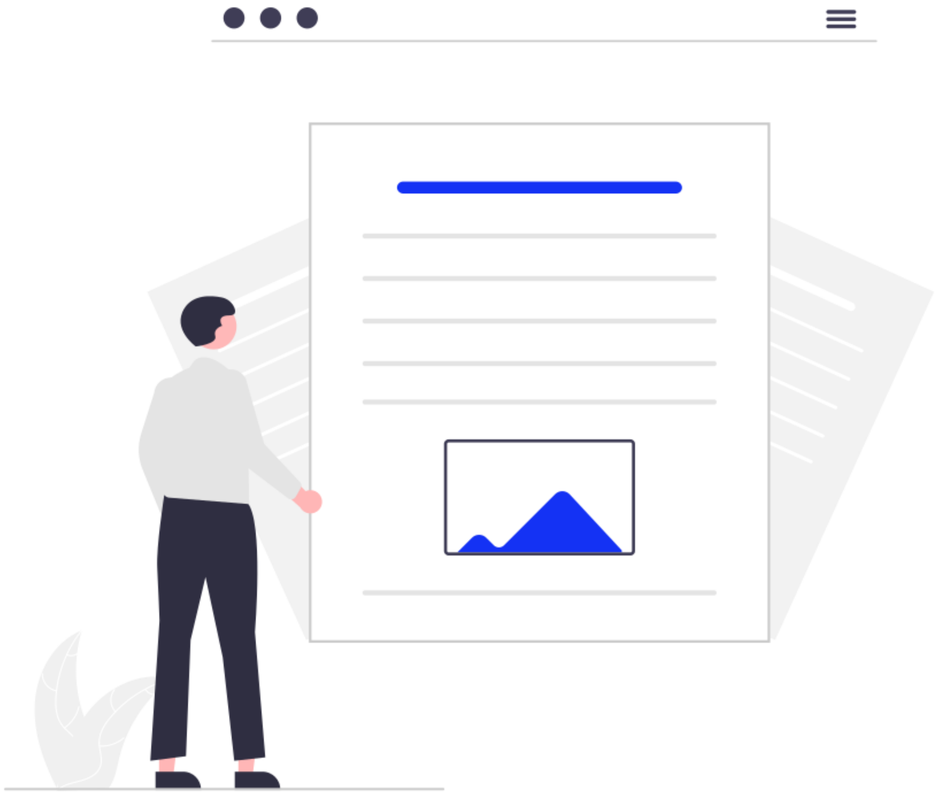

Background
of our idea

Have you ever experienced anything like this?

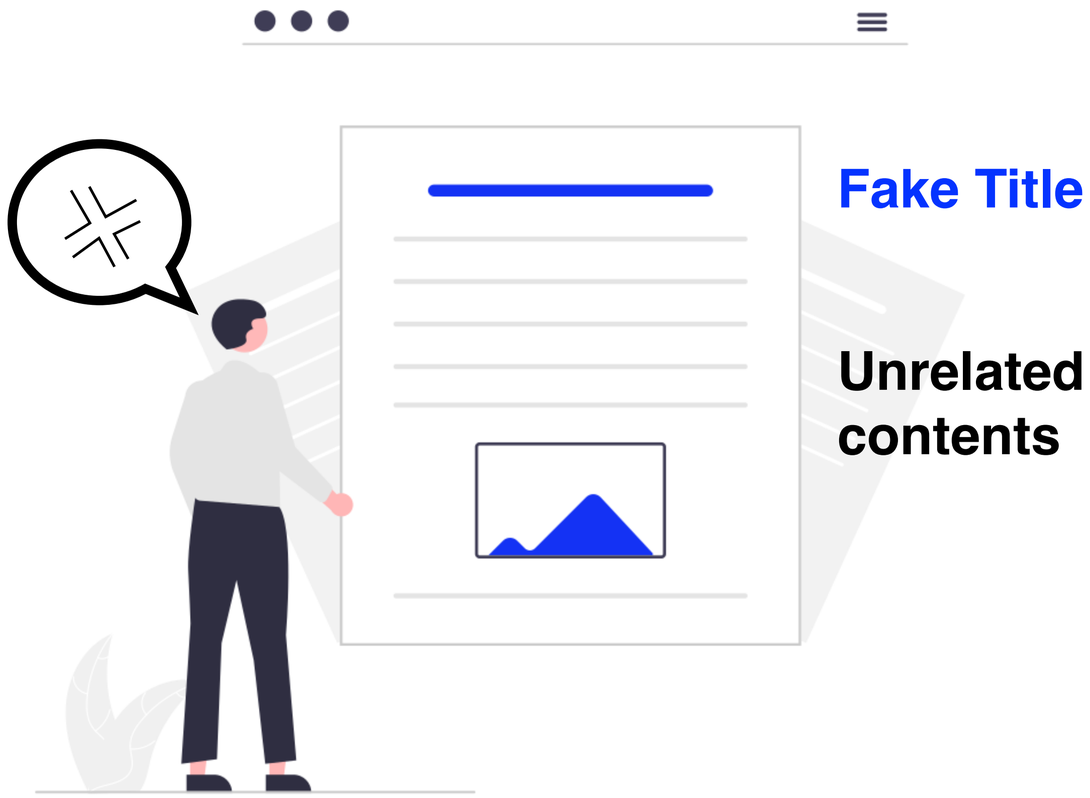


Have you ever experienced anything like this?

Have you ever experienced anything like this?



Have you ever experienced anything like this?



Fake Title

Unrelated contents

The Problem of Internet News

Fake Title

Titles that are inconsistent with the content, to induce public interest and clicks

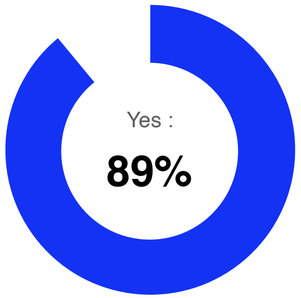
Wrong Information

Not only content that is completely different from the facts, but also cases in which facts are distorted so that they are not completely false but sufficiently misleading

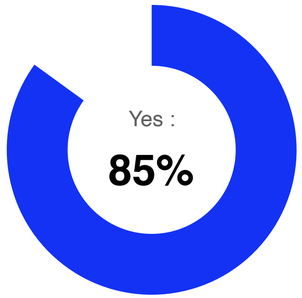
Provocative Expression

Using more provocative expressions than necessary to arouse public interest or making a narrative to cause controversy

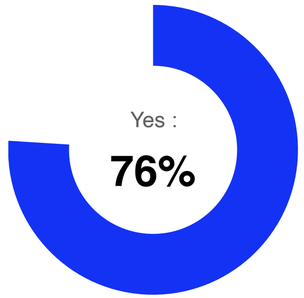
the Seriousness of Fake Articles



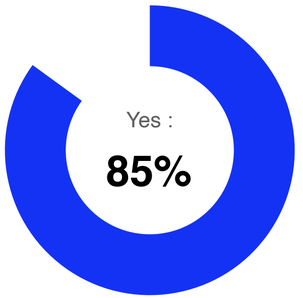
Do you feel the fake news problem is serious?



Have you ever been tricked by fake news?



Do you doubt if it's true even when you see the real news because of fake news?



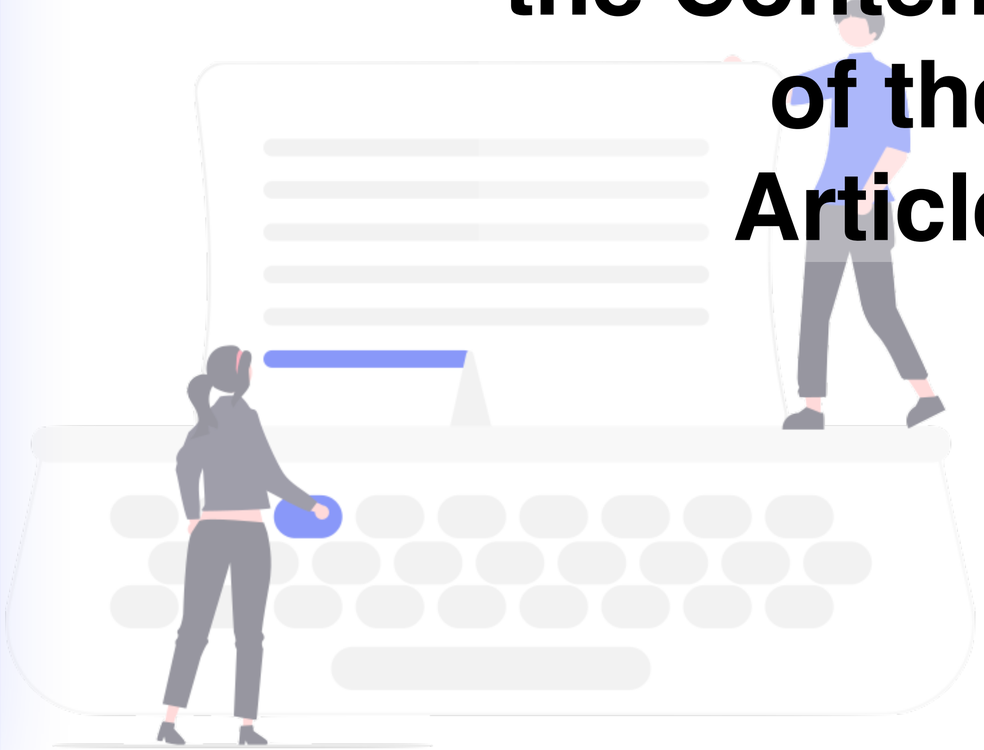
Do you feel there's too much fake news these days?

Source: 목회데이터연구소 (http://mhdata.or.kr/mailing/Numbers37th_200306_A.pdf)

Two Ideas

Filtering Articles that have Fake Title

Summarize the Content of the Article




Previous Research

ClickBait Articles

Previous Research

Feature Extraction to Detect Hoax Articles

- Utilizing **keywords** from comments -> generate large dataset
- Extract 5 classification features
- Using support **vector machine** classifier and selective **bigram model**

허성완, 손경아 - 정보과학회논문지, 2016

Multi-layered BERT-based clickbait detection model

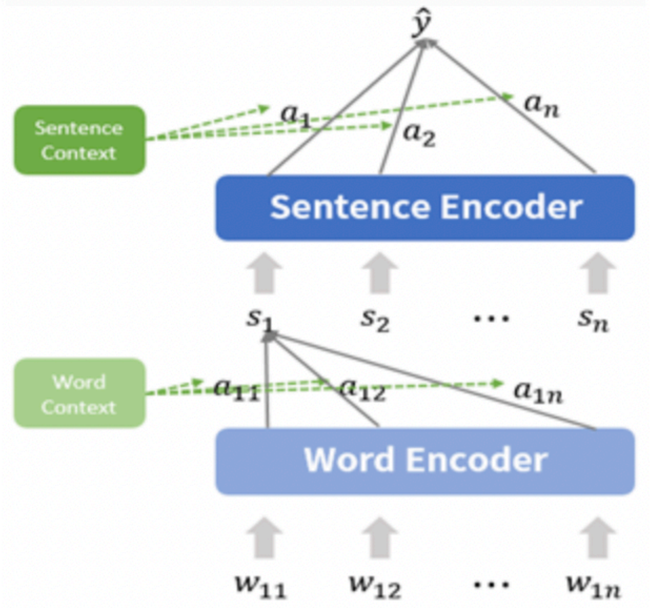
- Using pre-trained **Kobert** model
- **multi-layer BERT model** outperformed the existing phishing article classification model **HAND**
- Discuss the performance of **HAND** and the multi-layer **BERT** model

한상우, 온병원 - Proceedings of KIIT Conference, 2023

ClickBait Articles

Previous Research

HAND(Hierarchical Attention Networks for Document Classification)



- Word-level + Sentence-level

1. Word Encoder with GRU and Attention to generate word-level Hidden Representations
2. Combined to form a Sentence Vector
3. Determining matches or mismatches by passing Fully Connected Layer


Datasets

Data of Click bait article

Main data to be used for ‘Clickbait detection'

Data title: 낚시성 기사 탐지 데이터

Source: AI Hub (https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71338)

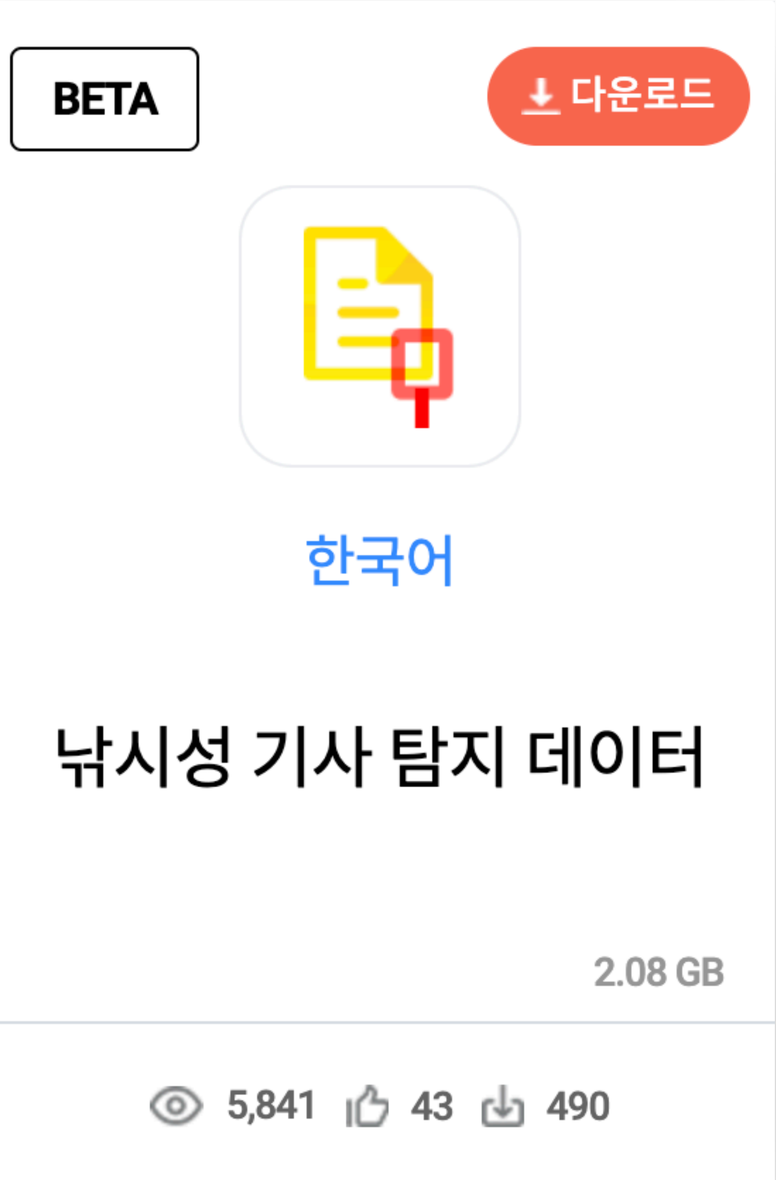
Year of Data Construction: 2022

Data Size: 733,427 articles(2.08GB)

Features: ‘newsCategory’, ‘newsTitle’, ‘newsContent’ etc

Label: clickbaitClass (0 -> clickbait / 1 -> normal news)


```
"newsTitle": "최종구 금융위원장, 임팩트금융 정착 위해 정부 차원 기금 조성",
"newsSubTitle": "",
"newsContent": "최종구 금융위원회 위원장이 \\\"'임팩트금융(사회적금융)'의 안착 차원에서 정부가 직접 기금을 만들겠다\\\"라고 강조했다.\
"clickbaitClass": 0,
```

News Summary

Main data to be used for ‘News Summary’

Data title: 문서 요약 텍스트

Source: AI Hub (https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=97)

Year of Data Construction: 2020

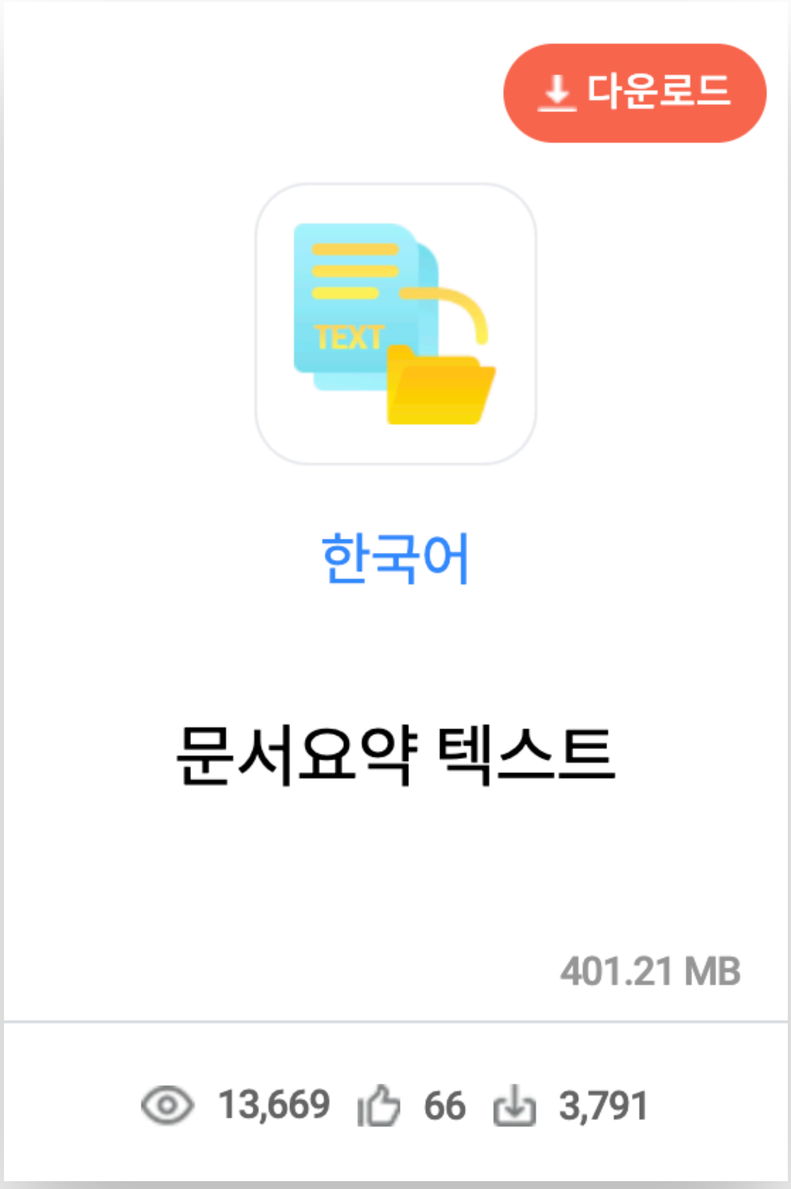
Data Size: 400,000 original data / 800,000 summary data(400,000 extraction summaries, 400,000 generation summaries)

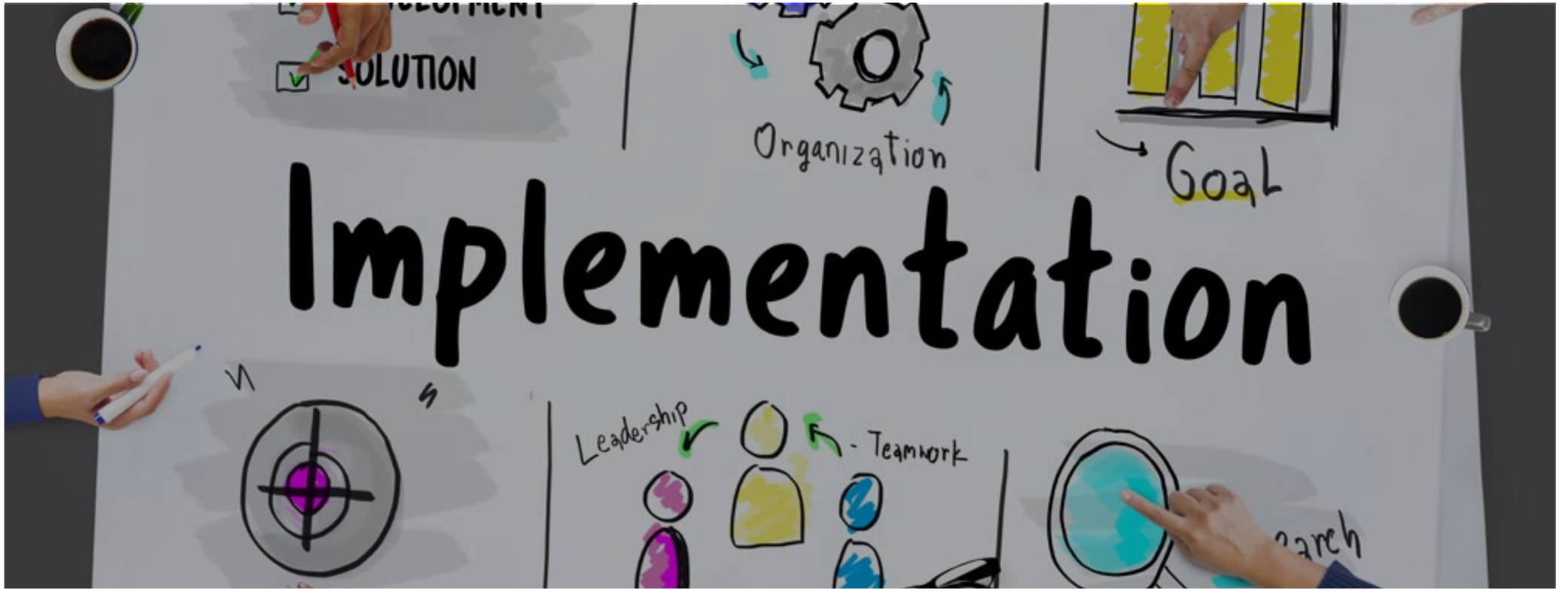
(401.21 MB)

Features: ‘category’, ‘title’, ‘text’(content) etc

Label: extractive(extraction summaries), abstractive(the generation summaries)


```
"extractive": [
    2,
    3,
    10
],
"abstractive": [
    "전라남도가 쌀 과잉문제를 근본적으로 해결하기 위해 올해부터 벼를 심었던 논에 벼 대신 사료작물이나
]
```


SOLUTION
Organization
Goal
Implementation
Leadership
Teamwork

Two main tasks.

Find Clickbait

Classify news with a consistency between title and content.
We define clickbait as the title is contradiction to it’s content.

Summarize Content

Serve a summary of the news content detected as a fake.
So user make a choice to read the news or not.

Bidirectional Encoder Representations form Transformers

BERT

Encoder only

Relatively smaller than encoder-decoder models.
So efficiently detect clickbait.

Multilingual

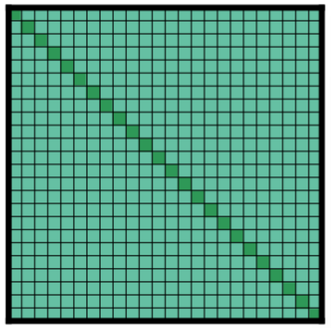
Pre-trained with millions of Korean sentences to overcome the limitations of BERT in terms of performance in Korean language.

Maximum token size

BERT has a limit on token size. It make some problems, because our input is consist of the title and content with sentence token.

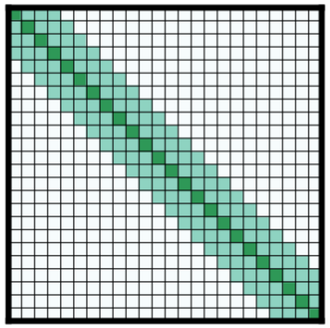
The Long Document Transformer

Longformer



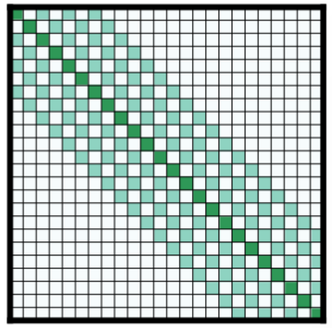
Full attention

Original self-attention. Calculate attention score with every token.



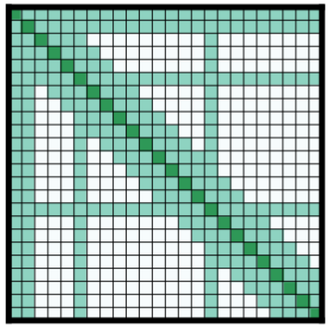
Sliding window attention

Focuses on tokens around a specific position in a sentence



Dilated sliding window

Allows attention between tokens that are farther apart, considering a wider context.



Global + sliding window

Combines both global and local attention for processing long documents.

Denoising Sequence-to-Sequence Pre-training

BART

Encoder-Decoder

BART can perform generation task like summarization, translation, and question-answering.

Denoising autoencoder

Trained to restore noisy input sentences to make effective for text generation task.

Enough max token size

It has 1024 max token size which is enough to almost news content.

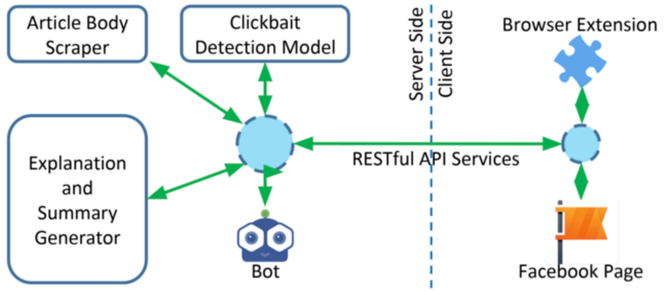
Expected Results & Evaluation

Expected Results

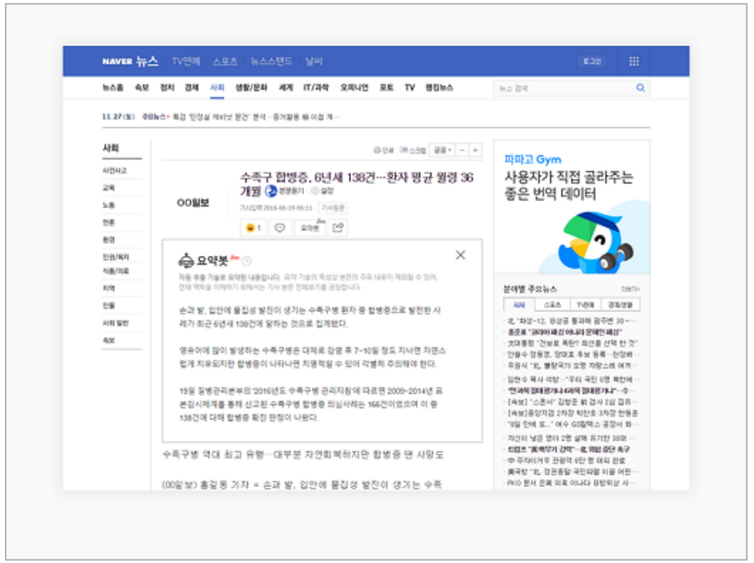
Provide a module that could:

1. **Classify input news articles as clickbait or non-clickbait**

2. **Summarize the article**



Rony, Md Main Uddin; Hassan, Naeemul; Yousuf, Mohammad; . Baitbuster: a clickbait identification framework. Thirty-Second AAAI Conference on Artificial Intelligence. 2018



네이버 뉴스 고객센터, 요약봇 서비스 소개

Result Evaluation

Potential Metrics for Evaluation

Accuracy

Checking the predicted results against results of human inspections.

F_1-Score

The harmonic sum of precision and recall of the prediction results. Often used as a metric to evaluate classification problems

ROC-AUC Score

Aggregate measure of performance across all possible classification thresholds.

ROGUE Score

Includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans.

Workload Division

Clickbait Classification

김세정
오현민

Text Summarization

송성근

Application Development

우수몽
오수영


Any Questions?


Thank you for listening :)