On Stance Detection in Image Retrieval for Argumentation

Miriam Louise Carnot
Leipzig University and ScaDS.AI

Lorenz Heinemann
Leipzig University

Jan Braker
Leipzig University

Tobias Schreieder
Leipzig University

Johannes Kiesel
Bauhaus-Universität Weimar

Maik Fröbe
Martin-Luther-Universität Halle Wittenberg

Martin Potthast
Leipzig University and ScaDS.AI

Benno Stein
Bauhaus-Universität Weimar

Venue: SIGIR 2023

발표자: HUMANE Lab 석사과정생 이다현

2024-10-18

Can Image be Argumentative?

- 소셜 미디어에서 자신의 입장이나 주장을 설명하거나,
  글로 작성된 주장을 더 강하게 표현하기 위해 이미지가 자주 사용됨
- 이미지가 독자적으로 "**논쟁적**"일 수 있는지,
  즉 이미지만으로도 주장을 표현할 수 있는지는 논란의 여지가 있음
- Kjeldsen et al. (2014)은 이미지가 주장을 뒷받침하고, 사실을 명확히 하며,
  텍스트보다 더 효과적으로 전달할 수 있다고 주장
- 논쟁적인 주제에 관련된 이미지를 검색해주는 전용 검색 엔진은
  소셜 미디어나 다른 곳에서 자신의 입장을 지지할 이미지를 찾거나,
  다양한 의견을 한눈에 시각적으로 파악할 수 있게 하는 데 유용할 수 있음

Image Retrieval for Argumentation

- 논쟁적인 주제에 대한 텍스트 쿼리가 주어졌을 때, 해당 주제에 대한 논의를 뒷받침할 수 있는 이미지를 얼마나 잘 찾을 수 있는지에 따라 이미지를 순위화하는 작업
  - 중요한 하위 과제 중 하나는 검색된 이미지의 입장을 파악하는 것
- 2022년 CLEF Touché 연구실에서 첫 번째 Shared Task로 수행
- 주제에 대한 쟁점 또는 주장에 대한 키워드 쿼리가 주어졌을 때,
  - (1) 이를 지지하는 데 도움이 되는 이미지와
  - (2) 이를 반박하는 데 도움이 되는 이미지를 각각 순위가 매겨진 두 목록으로 검색

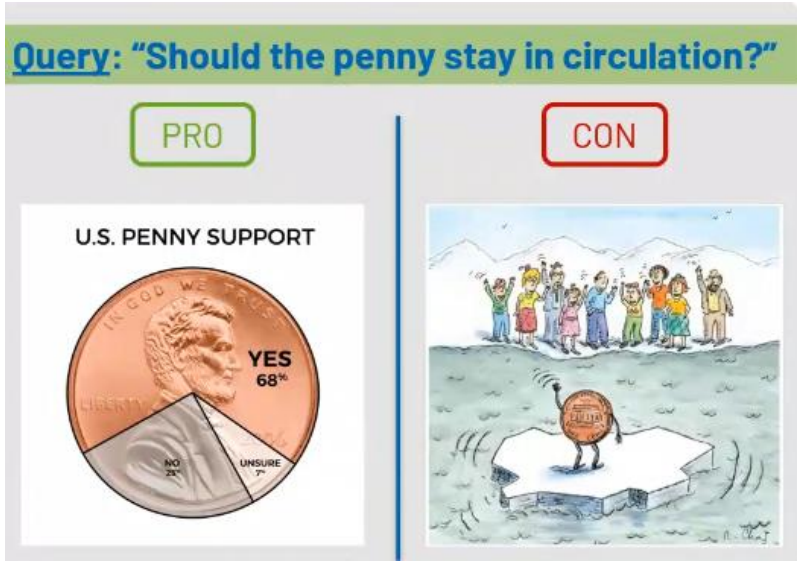
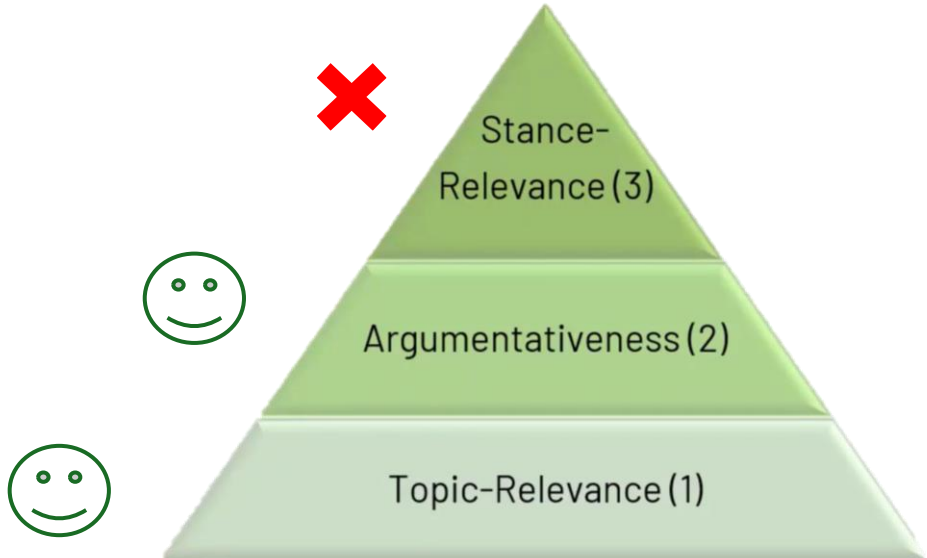


Image Retrieval for Argumentation

- Kiesel et al. (2021) 연구에서 제안된 논증을 위한 이미지 검색의 3단계 평가 방식
  - **Topic Relevance**: 이미지가 주제나 쿼리와 얼마나 잘 맞는지를 평가
  - **Argumentativeness**: 이미지가 주제에 대한 어떤 진술로 간주될 수 있는지 여부를 평가
  - **Stance Relevance**: 이미지가 찬성(지지) 또는 반대(공격) 입장 중 어느 쪽을 표현하는지를 평가
- 이 논문에서는 Touché22에서 사용한 방법들을 재현하고 확장
- 입장 감지(stance detection) 모델에 주로 관심을 두고 연구를 진행



이미지 출처 :

Related Work - CLEF Touché Dataset

- 50개의 논쟁적인 주제(쿼리)에 대한 23,841개의 이미지
  - “대체 에너지가 화석 연료를 효과적으로 대체할 수 있는가?”
  - “골프가 스포츠인가?”
  - “교육이 무료여야 하는가?”

(1) Is the image in some manner related to the topic?

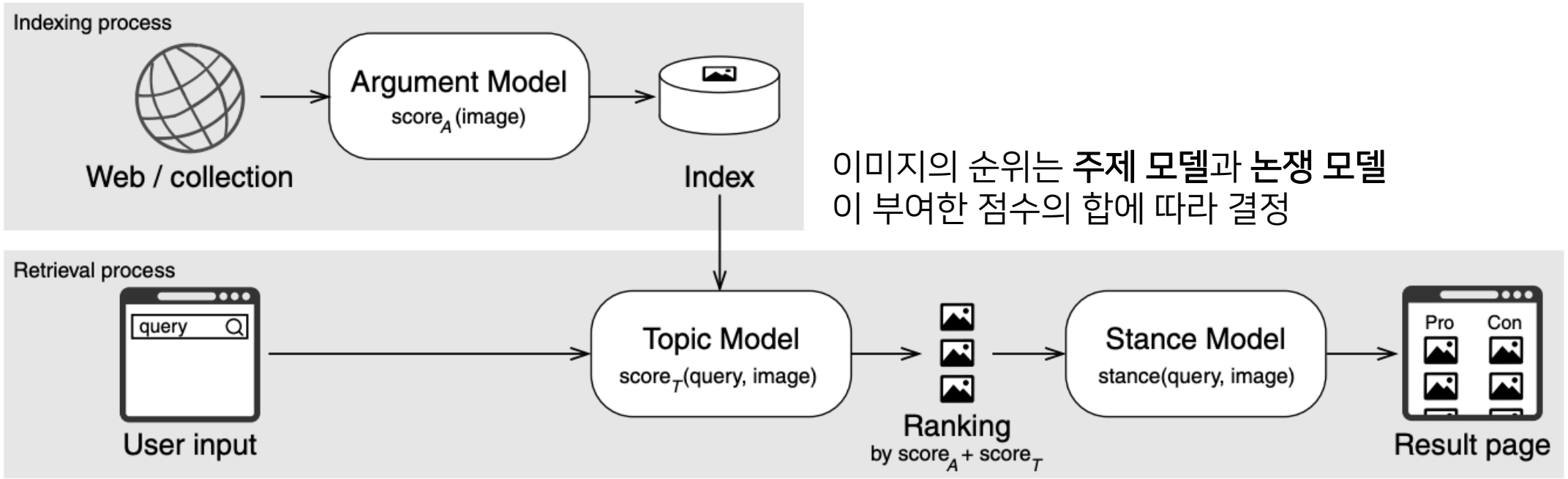
(2) Do you think most people would say that,
if someone shares this image without further comment,
they want to show they approve of the pro-side to the topic?

(3) Or do you think most people would rather say the one who shares this image does so to show they disapprove?

```
1 ONTOPIC Ib7fc7d5f8ee59d62 1
1 PRO Ib7fc7d5f8ee59d62 1
1 CON Ib7fc7d5f8ee59d62 0
```

제안 방법

- 세 가지 AI 모델을 활용한 모듈형 검색 시스템을 제안
  - 쿼리와 관련된 이미지를 식별하는 주제 모델
  - 논쟁에 적합한 이미지를 식별하는 논쟁 모델
  - 이미지를 찬성과 반대로 분류하는 입장 모델

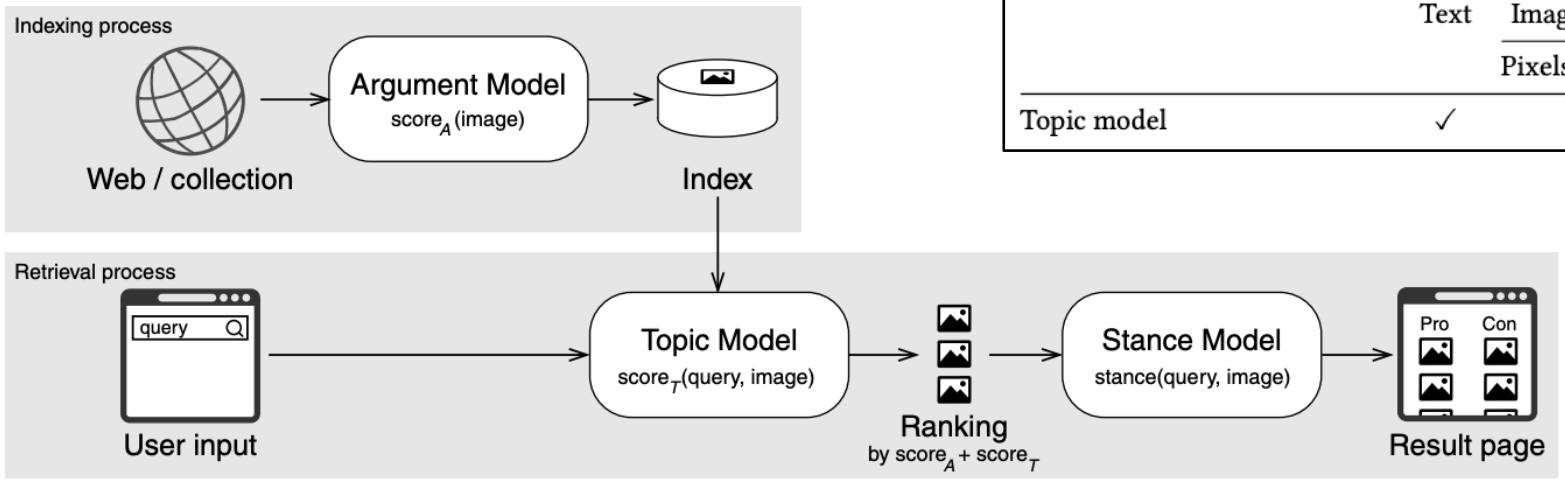


이미지의 순위는 **주제 모델**과 **논쟁 모델**이 부여한 점수의 합에 따라 결정

주제 모델

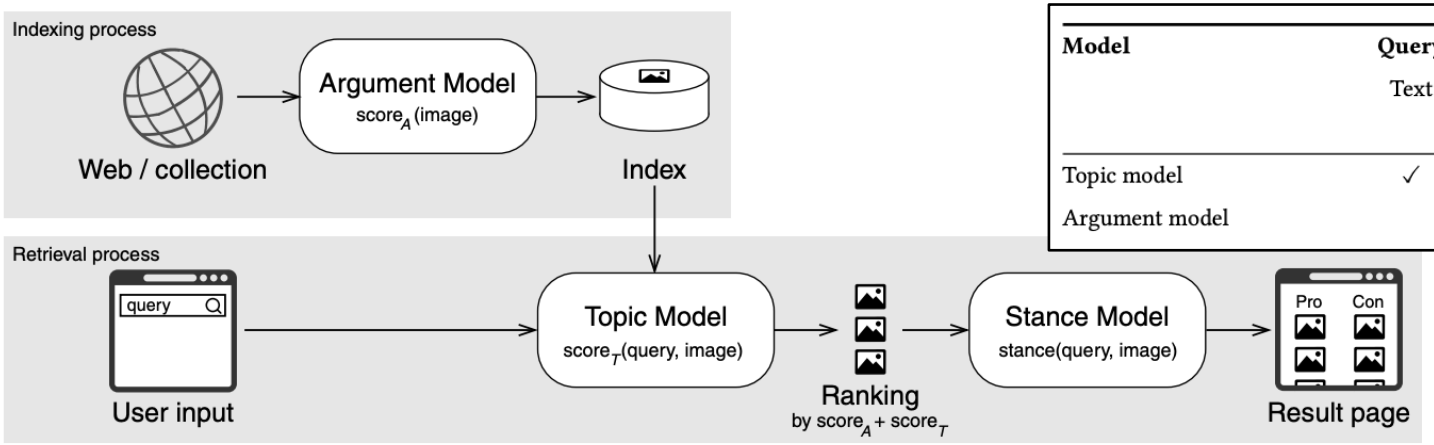
- 각 이미지가 쿼리와 얼마나 관련이 있는지 점수를 부여하여 이미지를 순위화
  - Touché'22 공유 과제에서 가장 성능이 뛰어났던 두 접근 방식, Boromir et al. (2022)와 Aramis et al.(2022)의 특징을 결합한 주제 모델을 사용
- 다음의 두 가지 특징을 활용하여 이미지가 주어진 쿼리와 얼마나 잘 맞는지 평가
  - **이미지와 관련된 웹 페이지의 텍스트**
    - 이미지가 포함된 웹 페이지의 HTML 소스 코드를 분석하여 텍스트를 추출 후 이미지와 가까운 부분은 BM25 알고리즘을 사용하여 인덱싱
  - **이미지 자체의 텍스트 활용**
    - OCR기술을 사용하여 이미지 안의 글자를 추출 후 인덱싱

Model	Query	Image features			
	Text	Image file		Web page	SERP
		Pixels	Text	Title Source	Rank
Topic model	✓		✓	✓	



논쟁 모델

- 이미지가 논증에 얼마나 적합한지에 따라 점수를 부여
- 주어진 쿼리와 독립적으로 작동하기 때문에 각 점수는 인덱싱 과정에서 활용됨
- Touché'22 공유 과제에서 Aramis et al.(2022)이 사용한 쿼리 독립적 특징을 사용
  - 이미지의 전반적인 분위기를 포착하기 위한 색상 속성
  - 이미지 유형(그래픽인지 사진인지)
  - 다이어그램 유사성
  - 텍스트의 일반적인 사용
- Aramis et al.(2022)에서 사용된 신경망 분류기로 특징들로부터 논쟁성 점수를 계산

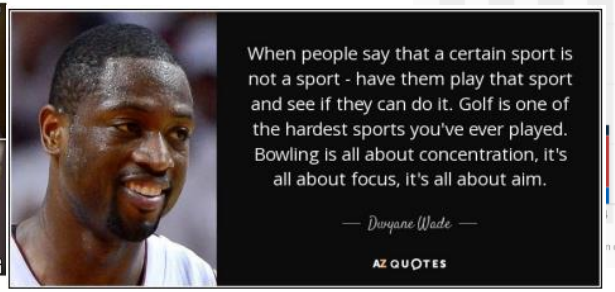


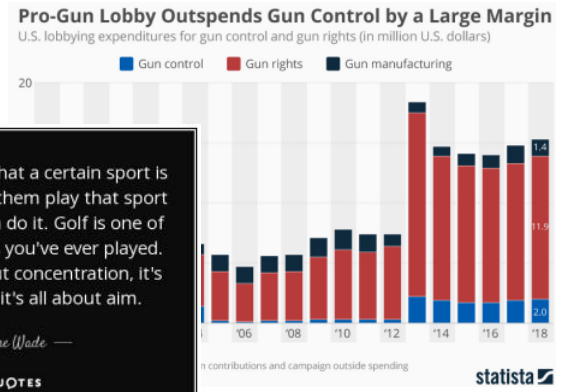
Model	Query	Image features				
	Text	Image file		Web page		SERP
		Pixels	Text	Title	Source	Rank
Topic model	✓		✓		✓	
Argument model		✓	✓			

논쟁 모델

- 이미지의 전반적인 분위기를 포착하기 위한 색상 속성
  - 이미지의 전반적인 분위기를 표현하는 색상
  - 평균 색상과 주요 색상(빨강, 초록, 파랑, 노랑)이 차지하는 비율을 RGB 값으로 계산

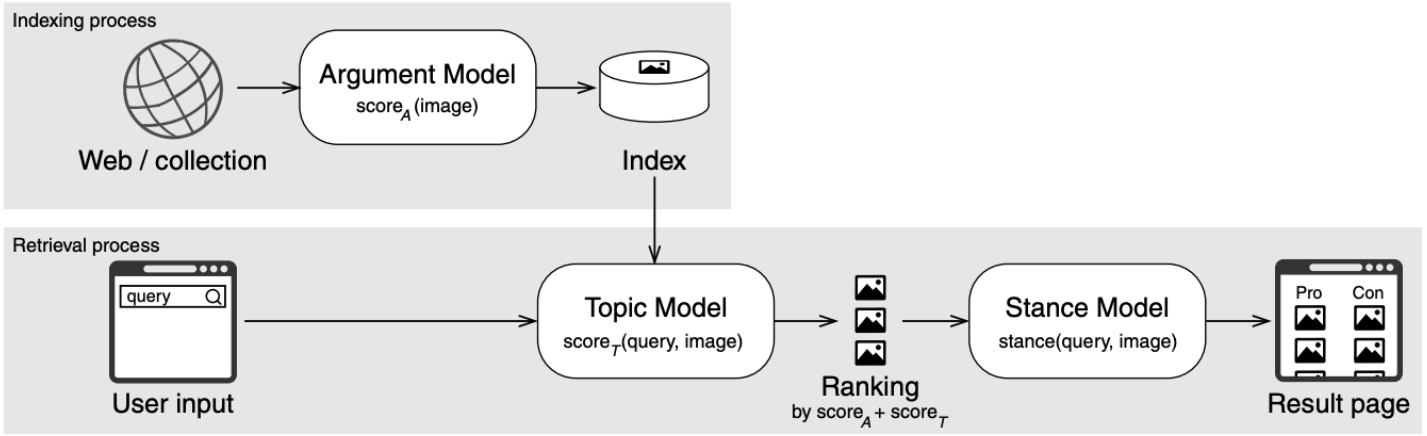
- 이미지 유형
  - 그래픽(만화, 클립아트 등)인지 또는 사진인지 분류
  - 이미지에서 가장 많이 사용된 10가지 색상이 전체 이미지의 30% 이상을 차지하면 그래픽, 아니면 사진
- 다이어그램 유사성
  - 이미지에 포함된 짧은 텍스트의 비율을 기반으로 다이어그램과 유사한지 평가
  - 수평으로 배치된 긴 텍스트는 제거, 수직으로 배치된 짧은 텍스트만 남겨 다이어그램처럼 보이는지 평가
  - 연구에서 다이어그램은 일반적으로 논쟁적 성격을 띠고 있다고 주장
- 텍스트의 일반적인 사용
  - 이미지에 포함된 텍스트의 길이
  - 텍스트가 긍정적인지, 부정적인지 감정 분석
  - 이미지에서 텍스트가 차지하는 면적 비율
  - 이미지를 8x8 그리드로 나누어, 텍스트 위치 파악





입장 모델

- 이미지를 특정 주제에 대해 **찬성(Pro)**, **반대(Con)**, **둘 다**, 또는 **둘 다 아님**으로 라벨링
  - 하나의 이미지가 찬성 입장과 반대 입장 모두에 해당할 수 있음 (Touché 정의)
- Touché'22 결과에 따르면,
  참가 모델 중 어느 것도 입장 감지에서 높은 정확도를 달성하지 못함
- 입장 감지 하위 과제에 있어 14가지 접근 방식을 비교
  - 두 가지 Baseline 접근법과 Oracle도 포함



입장 모델

Model	Query	Image features				
	Text	Image file		Web page		SERP
		Pixels	Text	Title	Source	Rank
Topic model	✓		✓		✓	
Argument model		✓	✓			
Stance models						
Oracle						
Both-sides baseline		✓	✓			
Random baseline						
Crawl query stance	✓					✓
CLIP query stance	✓	✓				
BERT title sentiment				✓		
AFINN text sentiment					✓	
Aramis Formula	✓	✓	✓		✓	
Aramis Neural	✓	✓	✓		✓	
Neural text+image 3class	✓	✓	✓			
Neural text+image 2x2class	✓	✓	✓			
Neural text 3class	✓		✓	✓		
Neural text+page 3class	✓		✓	✓	✓	

- Oracle: 실제 라벨(ground truth)
- Both-sides Baseline: 모든 이미지를 찬성과 반대 입장 모두로 분류하여 동일한 결과 목록 두 개를 생성
- Random Baseline: 이미지를 찬성 또는 반대로 무작위로 분류
- Crawl query stance: 이미지가 처음 크롤링될 때 쿼리 확장을 통해 얻어진 검색 결과 목록을 기반으로 라벨 지정
- CLIP query stance : 크롤링된 검색 결과 목록 대신 CLIP 모델을 사용
- BERT title sentiment: Boromir et al. (2022) - Large Movie Review Dataset에서 훈련된 감정 분석 BERT 모델을 사용해 웹 페이지 타이틀의 감정 분석
- AFINN text sentiment : Boromir et al. (2022) - AFINN 감정 사전을 사용해 웹 페이지 텍스트의 감정을 분석
  각 단어의 감정 점수를 AFINN 사전에서 찾아 합산한 뒤, 점수가 음수면 반대, 양수면 찬성으로 라벨링
- Aramis Formula: Aramis et al.(2022)이 개발한 13개의 특징을 기반으로 하는 수식을 사용
- Aramis Neural : Aramis et al.(2022)이 개발한 신경망 모델로, Aramis Formula 에서 사용된 것과 동일한 특징들을 사용/찬성, 중립, 반대로 분류
- Neural text+image 3class: 이미지를 256x256 픽셀로 리사이즈한 뒤, 쿼리 텍스트와 이미지에서 인식된 텍스트를 입력으로 사용
  신경망은 BERT 모델과 ResNet50V2를 결합 → 찬성, 중립, 반대 3개의 출력
- Neural text+image 2x2class: 위와 동일한 신경망 아키텍처를 사용하지만, 찬성과 반대를 독립적으로 판단
  입장에 맞는 점수를 계산하고, "현재 쿼리에 대해 모든 이미지 중에서 가장 높은 점수"의 절반을 넘으면 해당 입장으로 분류
- Neural text 3class: Neural text+image 3class와 동일하지만, 이미지 픽셀 대신 웹 페이지의 타이틀을 입력으로 사용
  쿼리와 이미지에서 인식된 텍스트도 함께 사용
- Neural text+page 3class: Neural text+image 3class와 동일하지만, 여기에 이미지 주변 HTML 텍스트도 추가로 사용

실험 결과

- 검색 대상 데이터: 이미 주제 관련성, 논쟁성, 입장에 대한 평가가 되어 있는 Touché'22 데이터셋의 6607개 이미지
- 기계 학습 기반 접근법에 대해서는 5-fold 교차 검증을 사용
- Touché22 에서는 precision@10만 사용
  - 사용자가 한 페이지의 결과 이미지들을 보는 상황과 가장 유사하기 때문
- 이 연구는 NDCG@10도 사용하여 평가를 확장
  - NDCG: 검색 결과의 순위에 따라 가중치를 부여하여 평가

Stance Model	**Precision@10**									**NDCG@10**								
	Topic-relevance			Argumentativeness			Stance-relevance			Topic-relevance			Argumentativeness			Stance-relevance		
	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both
Oracle	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.802	0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.929	0.964
Neural text+image 2x2class	0.924	0.822	0.873	0.830	0.766	0.798	0.660	**0.310**	**0.485**	0.928	0.847	0.887	0.831	0.789	0.810	0.657	**0.341**	**0.499**
BERT title sentiment	0.892	0.872	0.882	0.806	0.802	0.804	0.674	0.250	0.462	0.909	0.885	0.897	0.813	0.814	0.814	0.673	0.266	0.470
CLIP query stance	**0.932**	**0.932**	**0.932**	0.836	0.824	0.830	0.662	0.256	0.459	**0.937**	**0.934**	**0.935**	**0.843**	0.830	**0.836**	0.667	0.267	0.467
Aramis Formula	0.920	0.814	0.867	**0.838**	0.742	0.790	**0.690**	0.216	0.453	0.920	0.837	0.878	0.835	0.757	0.796	**0.685**	0.239	0.462
Both-sides baseline	0.926	0.926	0.926	0.832	**0.832**	**0.832**	0.662	0.232	0.447	0.928	0.928	0.928	0.831	**0.831**	0.831	0.658	0.246	0.452
Neural text+image 3class	0.924	0.866	0.895	0.830	0.800	0.815	0.660	0.226	0.443	0.928	0.878	0.903	0.831	0.805	0.818	0.657	0.234	0.446
Random baseline	0.894	0.888	0.891	0.816	0.812	0.814	0.664	0.222	0.443	0.908	0.895	0.901	0.823	0.815	0.819	0.654	0.239	0.447
Aramis Neural	0.694	0.676	0.685	0.668	0.640	0.654	0.588	0.278	0.433	0.733	0.708	0.721	0.703	0.668	0.686	0.602	0.303	0.453
Best of Touché'22 (Boromir)	0.884	0.872	0.878	0.782	0.754	0.768	0.594	0.256	0.425	0.895	0.877	0.886	0.787	0.746	0.767	0.609	0.260	0.435
Crawl query stance	0.830	0.728	0.779	0.744	0.694	0.719	0.610	0.214	0.412	0.842	0.761	0.801	0.761	0.720	0.740	0.612	0.227	0.420
AFINN text sentiment	0.766	0.908	0.837	0.708	0.814	0.761	0.564	0.222	0.393	0.797	0.904	0.851	0.735	0.809	0.772	0.587	0.241	0.414
Neural text+page 3class	0.644	0.616	0.630	0.598	0.560	0.579	0.504	0.154	0.329	0.691	0.675	0.683	0.649	0.611	0.630	0.541	0.176	0.358
Neural text 3class	0.668	0.668	0.668	0.602	0.602	0.602	0.458	0.190	0.324	0.704	0.704	0.704	0.632	0.632	0.632	0.469	0.219	0.344

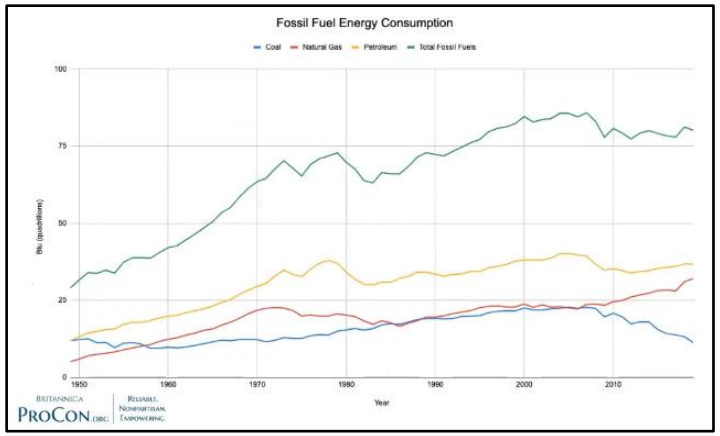
실험 결과

- Both-sides Baseline이 Touché22에서 경쟁한 다른 모든 방법들보다 뛰어난 성능을 보임
- 주제 관련성에서 92.6%, 논증성에서 83.2%의 정밀도(precision@10)를 달성
- 가장 좋은 결과는 Neural text+image 2x2class 모델이 48.5%의 정밀도를 달성
  - 찬성 입장의 이미지는 비교적 잘 분류됨(최대 69%)
  - 그러나 반대 입장의 이미지 분류는 매우 낮은 성능을 보임(최대 31%)
- Student's t-test와 Bonferroni 보정을 한 결과, 오직 oracle만이 기준선을 유의미하게 능가

Stance Model	Precision@10									NDCG@10								
	Topic-relevance			Argumentativeness			Stance-relevance			Topic-relevance			Argumentativeness			Stance-relevance		
	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both
Oracle	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.802	0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.929	0.964
Neural text+image 2x2class	0.924	0.822	0.873	0.830	0.766	0.798	0.660	**0.310**	**0.485**	0.928	0.847	0.887	0.831	0.789	0.810	0.657	**0.341**	**0.499**
BERT title sentiment	0.892	0.872	0.882	0.806	0.802	0.804	0.674	0.250	0.462	0.909	0.885	0.897	0.813	0.814	0.814	0.673	0.266	0.470
CLIP query stance	**0.932**	**0.932**	**0.932**	0.836	0.824	0.830	0.662	0.256	0.459	**0.937**	**0.934**	**0.935**	**0.843**	0.830	**0.836**	0.667	0.267	0.467
Aramis Formula	0.920	0.814	0.867	**0.838**	0.742	0.790	**0.690**	0.216	0.453	0.920	0.837	0.878	0.835	0.757	0.796	**0.685**	0.239	0.462
Both-sides baseline	0.926	0.926	0.926	0.832	**0.832**	**0.832**	0.662	0.232	0.447	0.928	0.928	0.928	0.831	**0.831**	0.831	0.658	0.246	0.452
Neural text+image 3class	0.924	0.866	0.895	0.830	0.800	0.815	0.660	0.226	0.443	0.928	0.878	0.903	0.831	0.805	0.818	0.657	0.234	0.446
Random baseline	0.894	0.888	0.891	0.816	0.812	0.814	0.664	0.222	0.443	0.908	0.895	0.901	0.823	0.815	0.819	0.654	0.239	0.447
Aramis Neural	0.694	0.676	0.685	0.668	0.640	0.654	0.588	0.278	0.433	0.733	0.708	0.721	0.703	0.668	0.686	0.602	0.303	0.453
Best of Touché'22 (Boromir)	0.884	0.872	0.878	0.782	0.754	0.768	0.594	0.256	0.425	0.895	0.877	0.886	0.787	0.746	0.767	0.609	0.260	0.435
Crawl query stance	0.830	0.728	0.779	0.744	0.694	0.719	0.610	0.214	0.412	0.842	0.761	0.801	0.761	0.720	0.740	0.612	0.227	0.420
AFINN text sentiment	0.766	0.908	0.837	0.708	0.814	0.761	0.564	0.222	0.393	0.797	0.904	0.851	0.735	0.809	0.772	0.587	0.241	0.414
Neural text+page 3class	0.644	0.616	0.630	0.598	0.560	0.579	0.504	0.154	0.329	0.691	0.675	0.683	0.649	0.611	0.630	0.541	0.176	0.358
Neural text 3class	0.668	0.668	0.668	0.602	0.602	0.602	0.458	0.190	0.324	0.704	0.704	0.704	0.632	0.632	0.632	0.469	0.219	0.344

이미지 입장 탐지에 관한 Challenge

- diagram이 나타내는 입장을 기계는 잘 해석하지 못함
  - 예) can alternative energy effectively replace fossil fuels?
  - 이 문제를 해결하려면 다이어그램을 자연어로 설명할 수 있는 최신 트랜스포머 모델을 통합해야 합

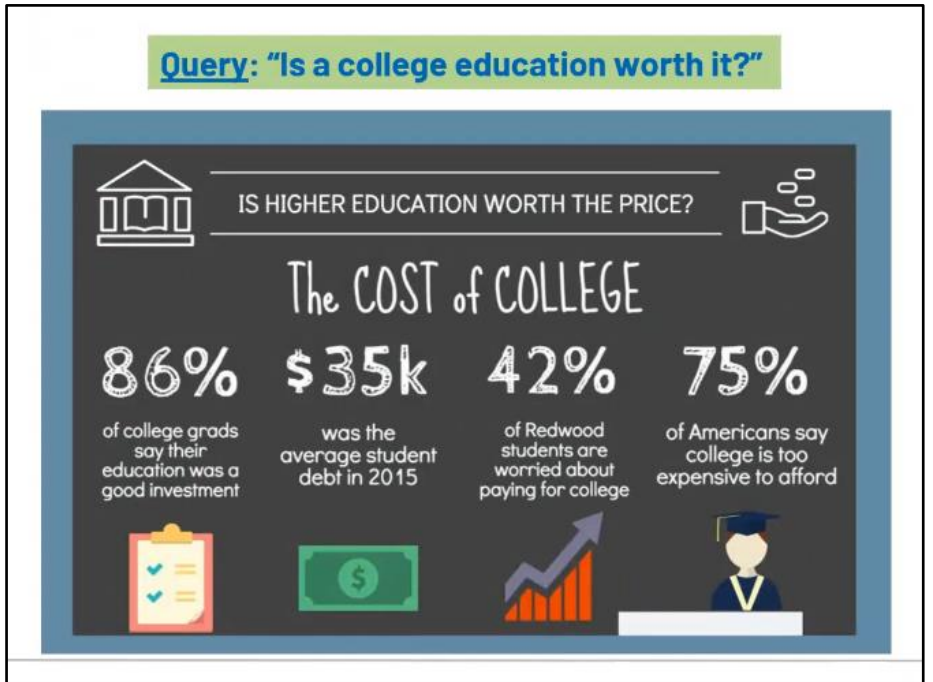


- 주관적일 수 있는 입장의 모호성
  - 예) 낙태를 민주당은 찬성, 공화당은 반대하는 설문 비율이 높은 결과
    개인의 배경이나 의견에 따라 이미지가 찬성으로 보일 수도,
    반대로 보일 수도 있음
  - 이를 해결하려면, 알고리즘이 이러한 문제를 가진 이미지를 식별하거나,
    사용자 프로필을 바탕으로 이미지를 분류하는 방법이 필요

이미지 입장 탐지에 관한 Challenge

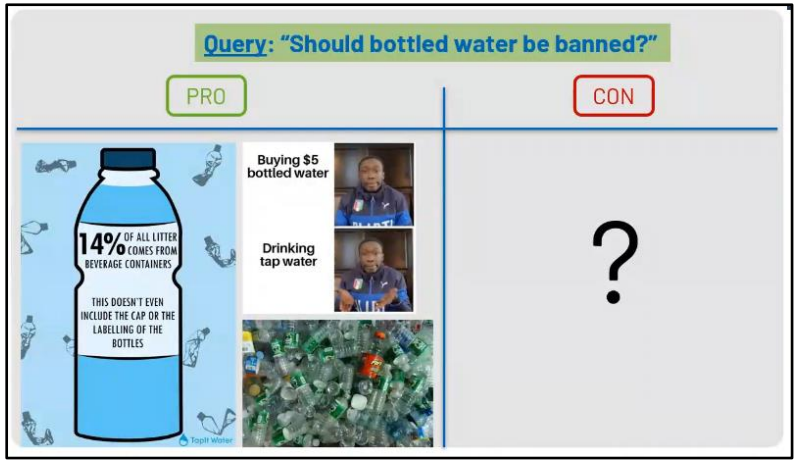
- 이미지 이해를 위해서는 background knowledge가 필요하다
  - 예) 땅을 태우는게 환경에 좋지 않다는 배경지식이 필요
  - 이미지를 사용한 웹 페이지의 맥락을 분석함으로써 관련 배경 지식을 제공할 수 있음

- 지역/문화적 이해가 필요한 이미지
  - 지역별 데이터를 학습시키는 모델을 개발할 수 있으며,
    주석 작업에서는 다양한 지역 출신의 주석자를 포함



이미지 입장 탐지에 관한 Challenge

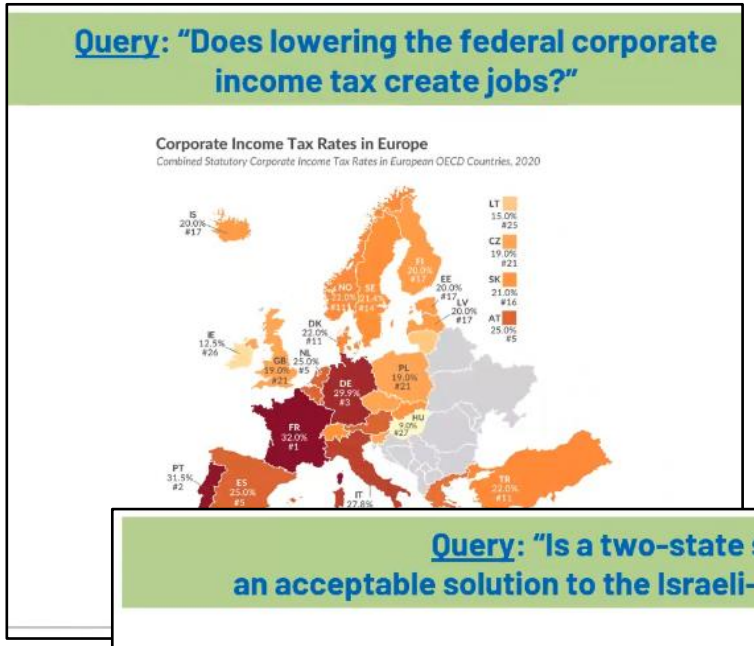
- 데이터셋에서 입장 분포가 불균형한 경우가 존재
  - 학습 데이터셋을 균형 있게 구성하거나 지나치게 불균형한 주제를 제외



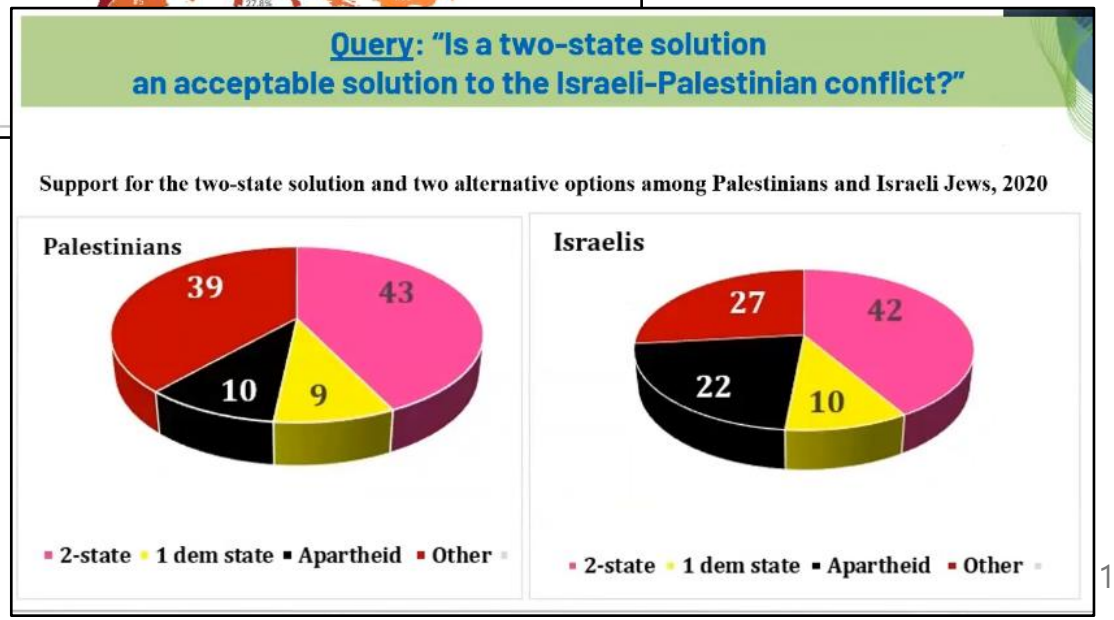
- 한 이미지에 양쪽 입장이 공존하는 경우
  - "양쪽(Both)" 카테고리를 추가

이미지 입장 탐지에 관한 Challenge

- 중립적인 이미지
  - 주제와 관련이 있지만,
    찬성이나 반대 입장을 명확히 나타내지 않음
  - 중립 이미지를 감지할 수 있는 분류기 개발 필요

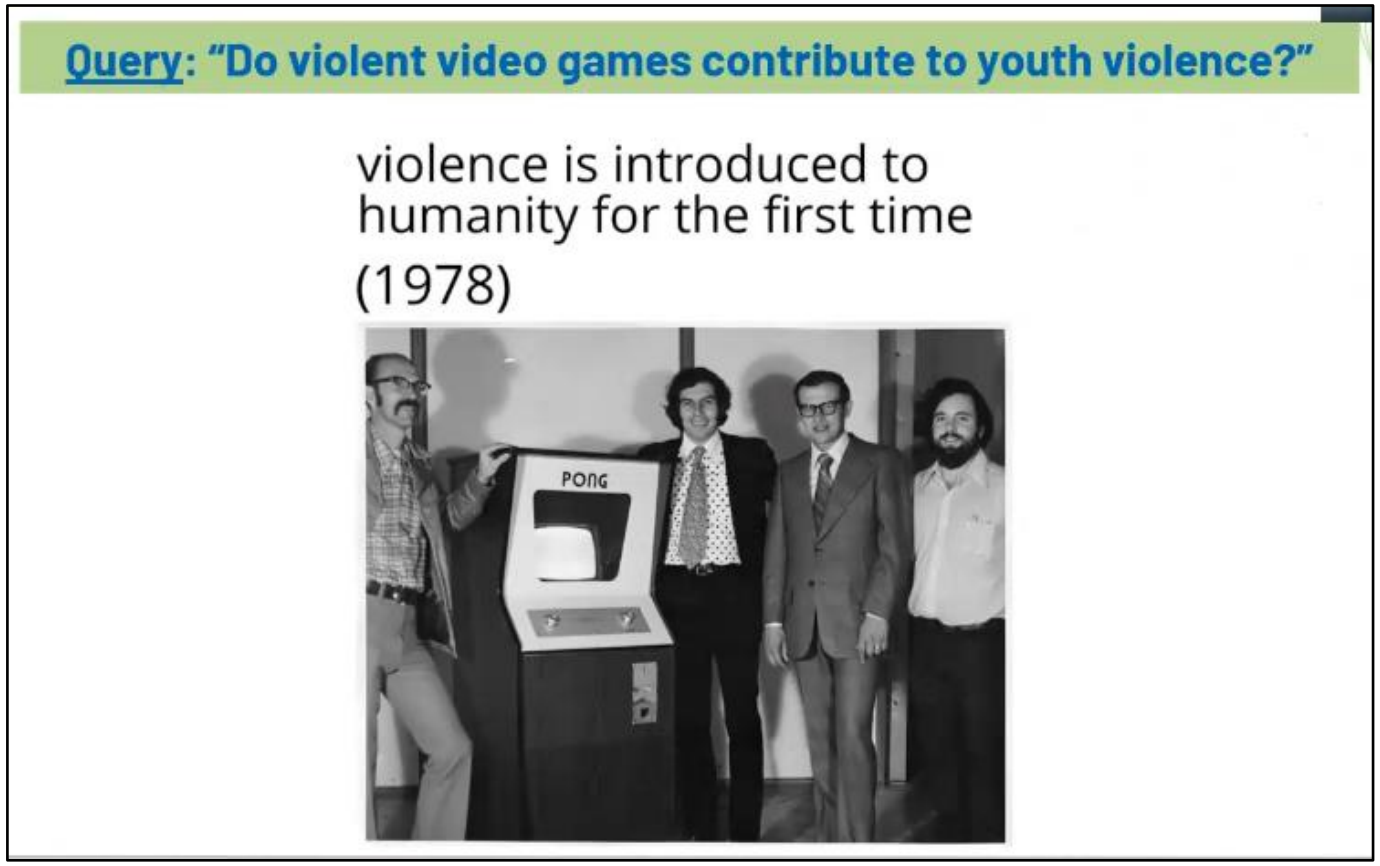


- 찬반으로 나누기 어려운 이미지
  - 찬성 혹은 반대보다 더 복잡한 논쟁
  - 분류하는 대신 클러스터링하는 방법을 사용 가능



이미지 입장 탐지에 관한 Challenge

- 풍자적인 표현이나 농담을 내포한 이미지



결론

- 논쟁을 위한 이미지 검색 작업에서 14가지 접근 방식을 비교
  - 특히 입장 탐지라는 Sub-Task에 중점을 둠
- CLEF의 Touché'22 랩 설정을 재현했으며 분석을 확장
- 모듈형 이미지 검색 시스템을 제안
  - 쿼리와 관련된 이미지를 식별하는 주제 모델
  - 논증에 적합한 이미지를 식별하는 논쟁 모델
  - 이미지를 찬성 및 반대로 분류하는 입장 모델
- 기존 논쟁 모델을 결합한 방식이 해당 작업 부분에서 가장 높은 점수를 달성
  - 논쟁적인 이미지를 찾아내는 정확도가 최대 0.832까지 개선
- 그러나 입장 탐지에 있어서는 낮은 성능
- 입장 탐지에 있어서 발생하는 도전과제를 식별하고 연구 방향을 제공

나의 생각

- 장점
  - 이미지 입장 탐지라는 과제에 있어 발생할 수 있는 Challenge를 잘 정리함
  - 찬성 입장보다 반대 입장이 더 잘 안나오는 이유에 대한 설명이 없음

- 단점
  - 입장 모델 중 LLM API를 사용한 결과가 없음. 사용하면 성능이 더 잘 나올듯 함
  - 논쟁성을 판단하는 방법이 기계적임.
    해당 방법으로 점수를 더 높이기에는 한계가 있을 것 같고
    이미지에 대한 설명을 텍스트로 변환하여 판단하는 등 이미지에 대한 더 깊은 이해가 있어야할 것 같음
    (산불 사진과 같은 예시를 통합하려면)

OPEN QUESTION

- 풍자적인 표현이나 농담을 내포한 이미지에 대한 연구 방향은 논문에서 따로 언급하지 않았는데, 이러한 이미지를 인식하고 입장을 탐지할 수 있으려면 어떻게 해야 할까?