IPTC News categorization

Final Project Presentation

Jan Wojtas

Paulina Szymanek

Łukasz Zalewski

Mikołaj Zalewski

Project goals

This project focuses on automatically categorizing articles according to the IPTC taxonomy.

By analyzing article content, the app assigns appropriate IPTC labels, enhancing content organization and retrieval processes for media organizations and content providers.

Labeling methods are inspired by SOTA solutions.

Contributions

Preparing a labelled test data set based on the STA News Article dataset

Providing results for Ada Embeddings and AnglE Embeddings on English STA dataset

Dockerized system for IPTC categorization and article retrieval

Findings

H1: Is the unsupervised learning method, using only comparison (e.g. cosine similarity) between embeddings of articles and categories, sufficient to effectively solve IPTC news categorisation problem?

Embedding Metric	Ada	Angle
Accuracy	85.67%	64.00%
F1 score	80.40%	51.49%

Ada Embeddings achieved better results than AnglE Embeddings

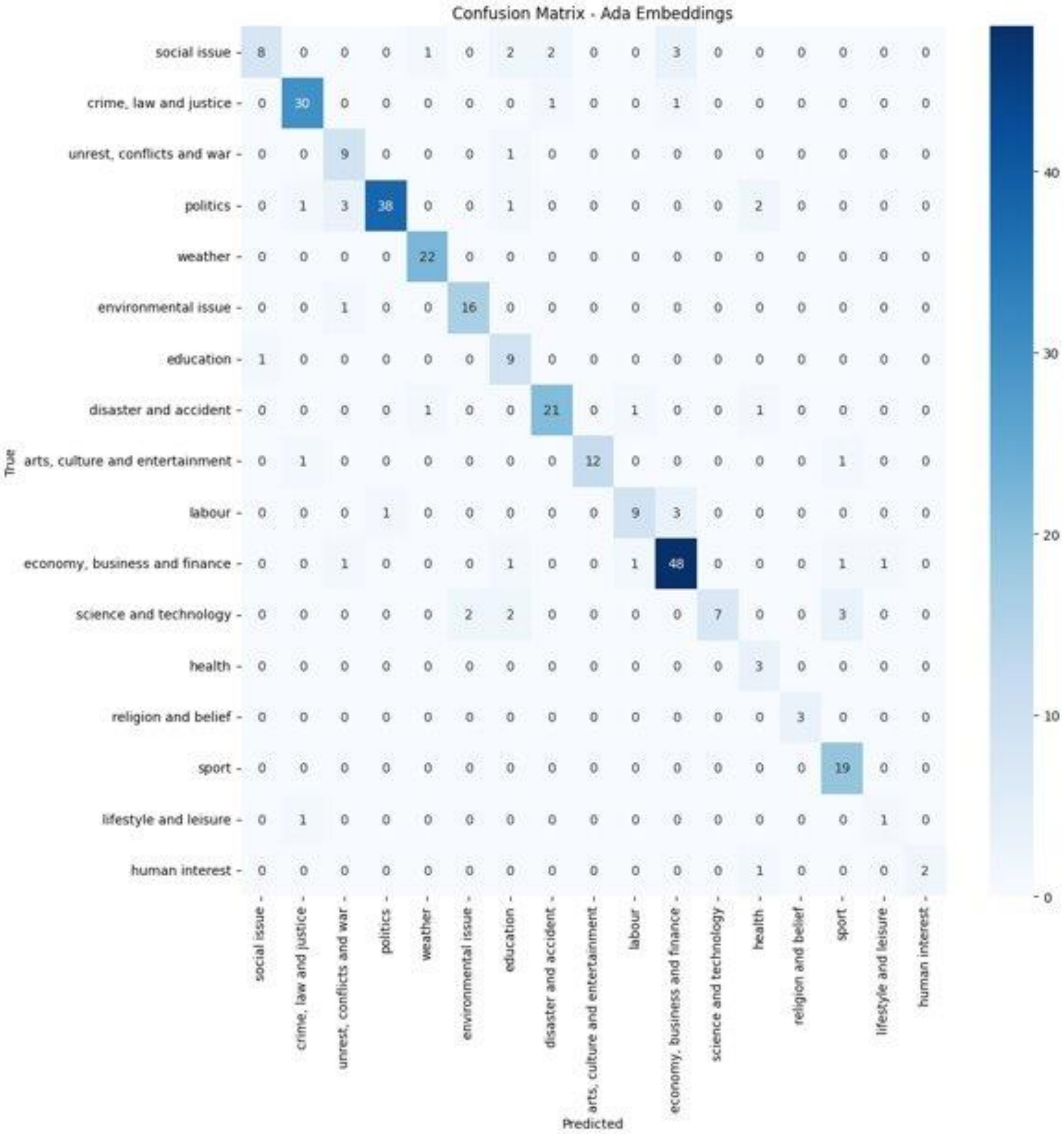
Prediction summary

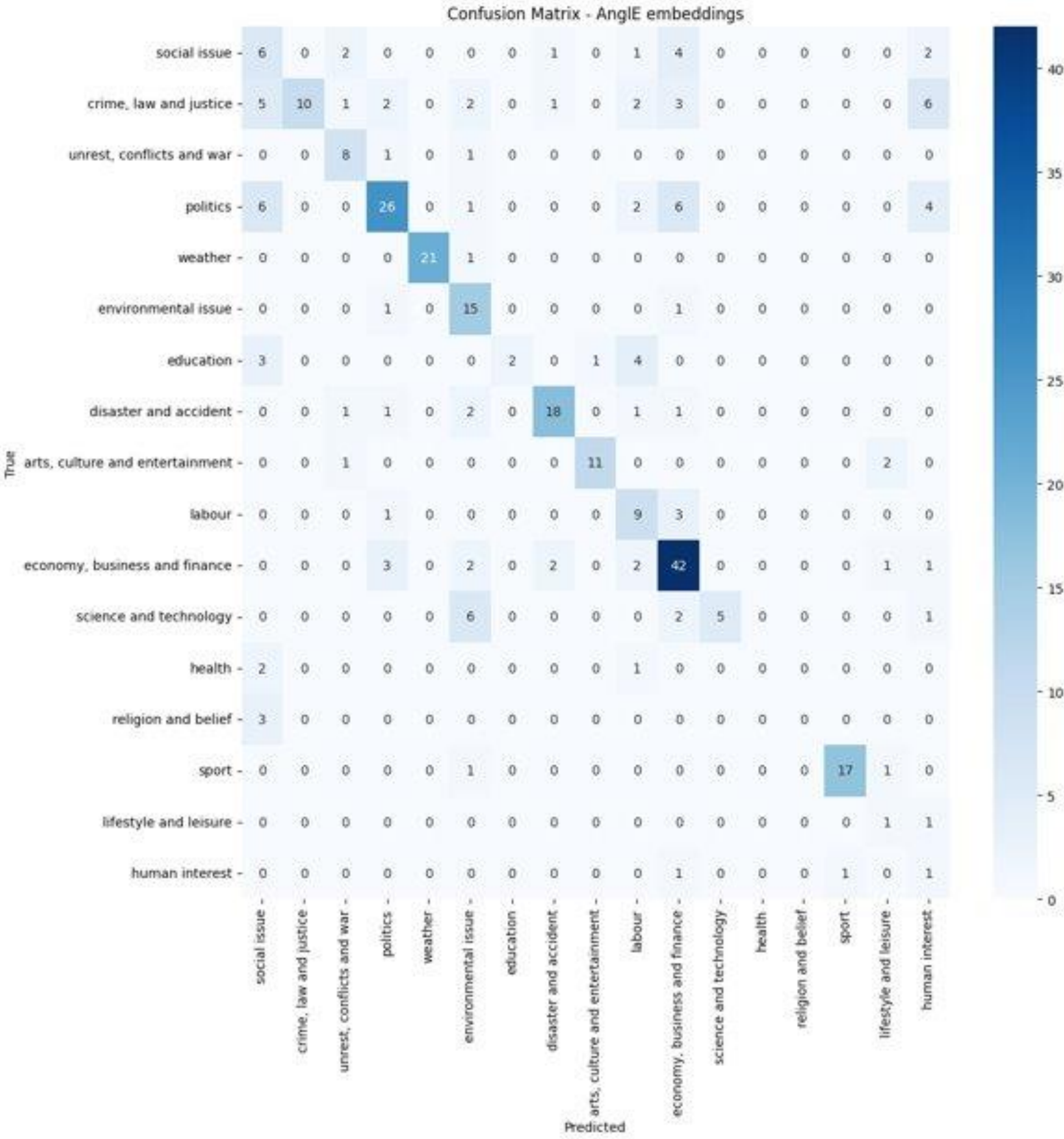
The classes are not balanced

Need for multi-category classification, as many articles do not fit into single categories

Broad categories, like "human interest" or "social issues" are not chosen by the model often, in favour of the more straightforward ones

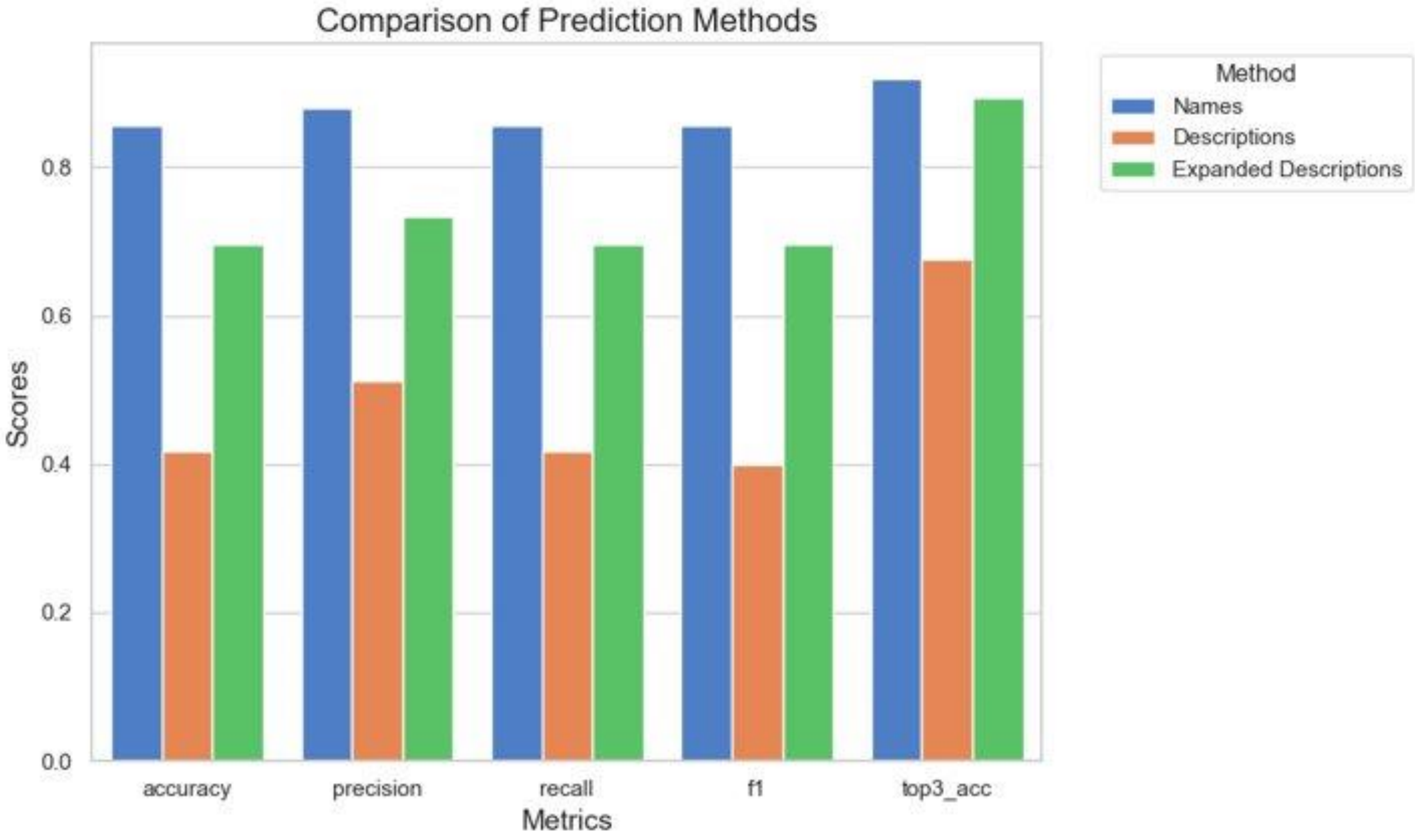
Articles that fit more than one category are often misclassified


Confusion Matrix - Ada Embeddings
True
Predicted
social issue 8 0 0 0 1 0 2 2 0 0 3 0 0 0 0 0 0
crime, law and justice 0 30 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0
unrest, conflicts and war 0 0 9 0 0 0 1 0 0 0 0 0 0 0 0 0 0
politics 0 1 3 38 0 0 1 0 0 0 0 0 2 0 0 0 0
weather 0 0 0 0 22 0 0 0 0 0 0 0 0 0 0 0 0
environmental issue 0 0 1 0 0 16 0 0 0 0 0 0 0 0 0 0 0
education 1 0 0 0 0 0 9 0 0 0 0 0 0 0 0 0 0
disaster and accident 0 0 0 0 1 0 0 21 0 1 0 0 1 0 0 0 0
arts, culture and entertainment 0 1 0 0 0 0 0 0 12 0 0 0 0 0 1 0 0
labour 0 0 0 1 0 0 0 0 0 9 3 0 0 0 0 0 0
economy, business and finance 0 0 1 0 0 0 1 0 0 1 48 0 0 0 1 1 0
science and technology 0 0 0 0 0 2 2 0 0 0 0 7 0 0 3 0 0
health 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0
religion and belief 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0
sport 0 0 0 0 0 0 0 0 0 0 0 0 0 0 19 0 0
lifestyle and leisure 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
human interest 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 2
social issue
crime, law and justice
unrest, conflicts and war
politics
weather
environmental issue
education
disaster and accident
arts, culture and entertainment
labour
economy, business and finance
science and technology
health
religion and belief
sport
lifestyle and leisure
human interest
40
30
20
10
0


Confusion Matrix - AnglE embeddings
True
Predicted
social issue 6 0 2 0 0 0 0 1 0 1 4 0 0 0 0 0 2
crime, law and justice 5 10 1 2 0 2 0 1 0 2 3 0 0 0 0 0 6
unrest, conflicts and war 0 0 8 1 0 1 0 0 0 0 0 0 0 0 0 0 0
politics 6 0 0 26 0 1 0 0 0 2 6 0 0 0 0 0 4
weather 0 0 0 0 21 1 0 0 0 0 0 0 0 0 0 0 0
environmental issue 0 0 0 1 0 15 0 0 0 0 1 0 0 0 0 0 0
education 3 0 0 0 0 0 2 0 1 4 0 0 0 0 0 0 0
disaster and accident 0 0 1 1 0 2 0 18 0 1 1 0 0 0 0 0 0
arts, culture and entertainment 0 0 1 0 0 0 0 0 11 0 0 0 0 0 0 2 0
labour 0 0 0 1 0 0 0 0 0 9 3 0 0 0 0 0 0
economy, business and finance 0 0 0 3 0 2 0 2 0 2 42 0 0 0 0 1 1
science and technology 0 0 0 0 0 6 0 0 0 0 2 5 0 0 0 0 1
health 2 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
religion and belief 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
sport 0 0 0 0 0 1 0 0 0 0 0 0 0 0 17 1 0
lifestyle and leisure 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1
human interest 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1
social issue
crime, law and justice
unrest, conflicts and war
politics
weather
environmental issue
education
disaster and accident
arts, culture and entertainment
labour
economy, business and finance
science and technology
health
religion and belief
sport
lifestyle and leisure
human interest
40
35
30
25
20
15
10
5
0

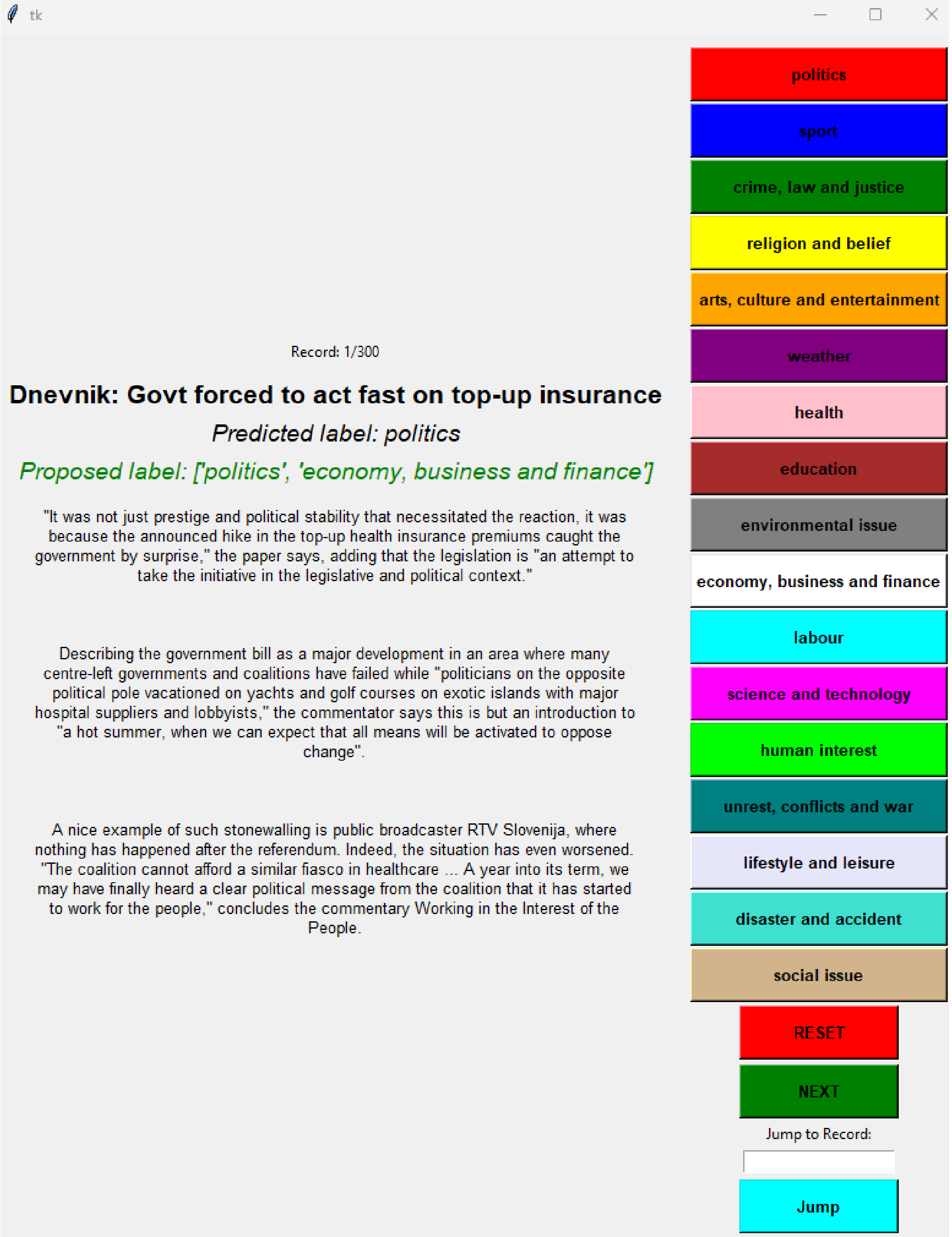
H2: Can additional prompt engineering using GPT-4 for IPTC categories or articles enhance the overall quality of the solution?

Predictions based only on category names achieve the best performance



Improvement: Multi-class labeling

- 300 news with manually assigned labels
- Selected with random sampling
- Labeling done using our own, dedicated application



Record: 1/300

Dnevnik: Govt forced to act fast on top-up insurance

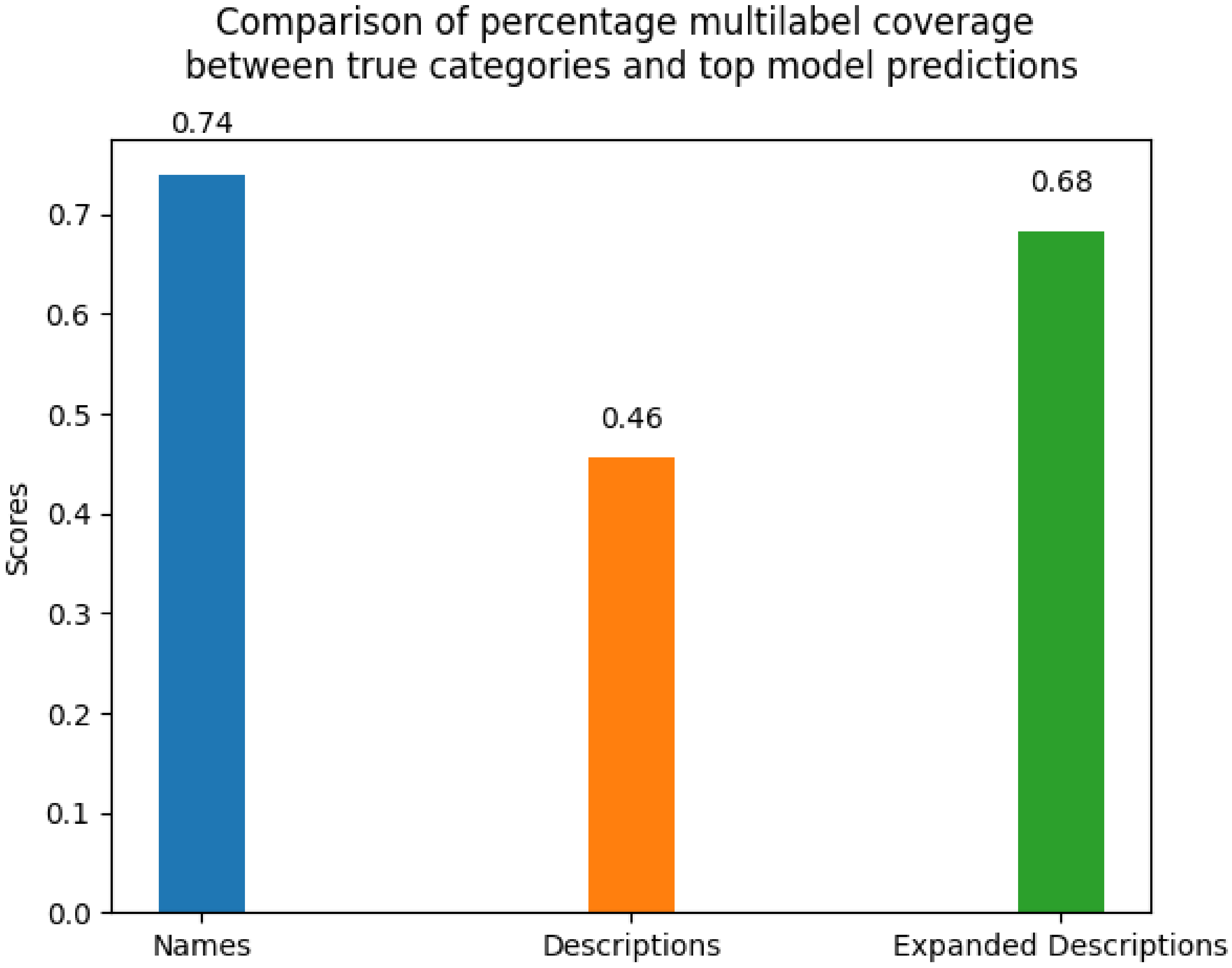
Predicted label: politics

Proposed label: ['politics', 'economy, business and finance']

"It was not just prestige and political stability that necessitated the reaction, it was because the announced hike in the top-up health insurance premiums caught the government by surprise," the paper says, adding that the legislation is "an attempt to take the initiative in the legislative and political context."

Describing the government bill as a major development in an area where many centre-left governments and coalitions have failed while "politicians on the opposite political pole vacationed on yachts and golf courses on exotic islands with major hospital suppliers and lobbyists," the commentator says this is but an introduction to "a hot summer, when we can expect that all means will be activated to oppose change".

A nice example of such stonewalling is public broadcaster RTV Slovenija, where nothing has happened after the referendum. Indeed, the situation has even worsened. "The coalition cannot afford a similar fiasco in healthcare ... A year into its term, we may have finally heard a clear political message from the coalition that it has started to work for the people," concludes the commentary Working in the Interest of the People.


Comparison of percentage multilabel coverage
between true categories and top model predictions
Scores
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.74
0.46
0.68
Names
Descriptions
Expanded Descriptions

Improvements

Quality article preprocessing

- ✓ Remove HTML tags
- ✓ Incorporate headlines
- ✓ Embed per text sections

Vector Database

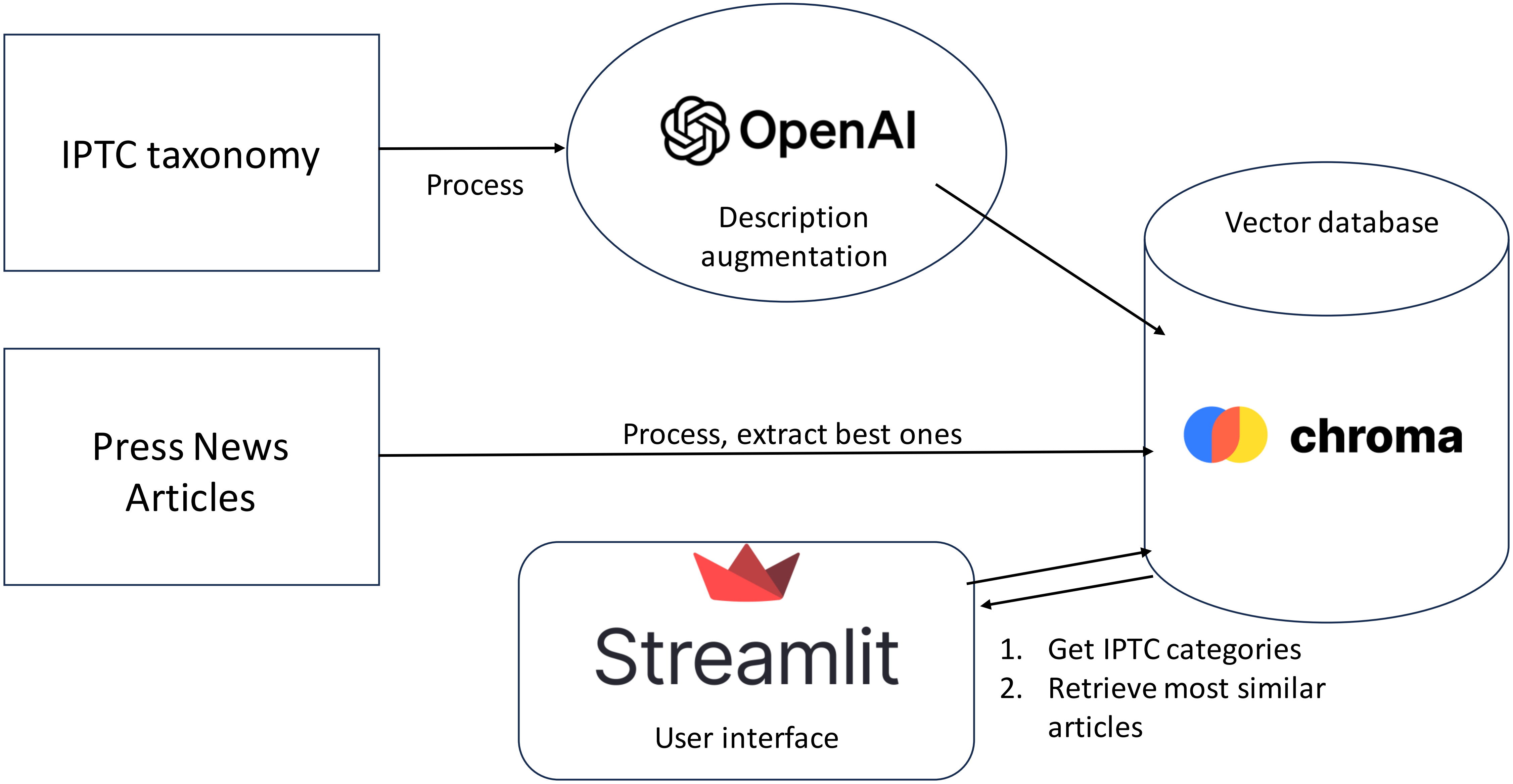
- ✓ Embedding storage
- ✓ Effective querying, inserting
- ✓ Scalable

Retrieval system

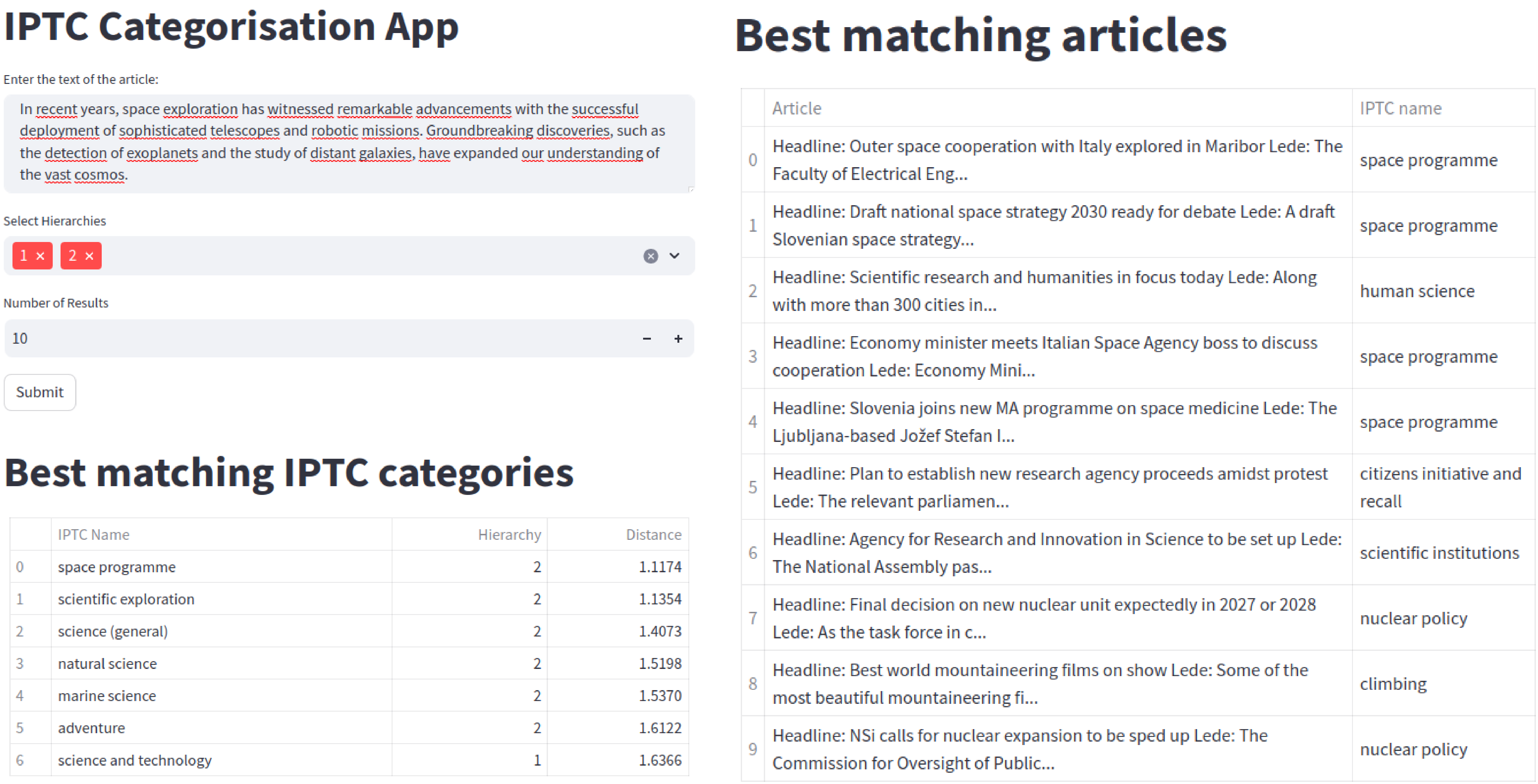
- ✓ Most similar articles
- ✓ Most similar categories

User interface

System design



1. Get IPTC categories
2. Retrieve most similar articles



IPTC Categorisation App

Enter the text of the article:

In recent years, space exploration has witnessed remarkable advancements with the successful deployment of sophisticated telescopes and robotic missions. Groundbreaking discoveries, such as the detection of exoplanets and the study of distant galaxies, have expanded our understanding of the vast cosmos.

Select Hierarchies

1 × 2 ×

Number of Results

10

Best matching IPTC categories

	IPTC Name	Hierarchy	Distance
0	space programme	2	1.1174
1	scientific exploration	2	1.1354
2	science (general)	2	1.4073
3	natural science	2	1.5198
4	marine science	2	1.5370
5	adventure	2	1.6122
6	science and technology	1	1.6366

Best matching articles

	Article	IPTC name
0	Headline: Outer space cooperation with Italy explored in Maribor Lede: The Faculty of Electrical Eng...	space programme
1	Headline: Draft national space strategy 2030 ready for debate Lede: A draft Slovenian space strategy...	space programme
2	Headline: Scientific research and humanities in focus today Lede: Along with more than 300 cities in...	human science
3	Headline: Economy minister meets Italian Space Agency boss to discuss cooperation Lede: Economy Mini...	space programme
4	Headline: Slovenia joins new MA programme on space medicine Lede: The Ljubljana-based Jožef Stefan I...	space programme
5	Headline: Plan to establish new research agency proceeds amidst protest Lede: The relevant parliamen...	citizens initiative and recall
6	Headline: Agency for Research and Innovation in Science to be set up Lede: The National Assembly pas...	scientific institutions
7	Headline: Final decision on new nuclear unit expectedly in 2027 or 2028 Lede: As the task force in c...	nuclear policy
8	Headline: Best world mountaineering films on show Lede: Some of the most beautiful mountaineering fi...	climbing
9	Headline: NSi calls for nuclear expansion to be sped up Lede: The Commission for Oversight of Public...	nuclear policy

Lessons learnt

- Single-class predictions are not the best choice in case of news articles,
- Manual labelling affects the results,
- Sometimes simple methods are also the best methods

QUESTIONS