IPTC News Categorisation
Final Report for NLP Course, Winter 2023

Jan Wojtas
WUT
01151523@pw.edu.pl

Paulina Szymanek
WUT
01186057@pw.edu.pl

Mikołaj Zalewski
WUT
01151710@pw.edu.pl

Łukasz Zalewski
WUT
01186057@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This project aims to develop an efficient and automated process for categorizing news articles in line with the IPTC taxonomy. Our initial investigation has highlighted the potential of NLP techniques to create a new quality standard for news classification. We highlight methods based on language model embeddings, which can be used in an unsupervised manner. The research, forming a part of an NLP course project, delves into the intricacies of text analysis and the application of machine learning algorithms for the identification and assignment of IPTC categories. A conceptual framework for our approach is presented - along with data preparation, initial results and discussion.

1 Introduction

In today's digital age, staying informed is of utmost importance. With increasing number of articles available online, efficient categorization and article filtering has become an essential feature for both readers and content creators. To address this need, the International Press Telecommunications Council (IPTC) taxonomy was developed, providing a systematic and standardized approach for categorizing news content. However, manual categorization of news articles is labor-intensive, time-consuming, and subject to human error. The exponential growth in digital news output exacerbates these issues, creating a pressing need for an automated solution.

The primary scientific goal of this project is to leverage Natural Language Processing (NLP) techniques to automate the categorization of news articles according to the IPTC taxonomy. This venture aims to address the dual challenges of efficiency and accuracy in the classification process, which are currently hampered by manual methods. By designing and implementing a machine learning-driven approach, the project seeks to enhance the speed of categorization without compromising the granularity and precision that the IPTC taxonomy demands.

Our study prioritizes the exploration of language model embeddings due to their universal applicability - particularly in scenarios involving modifications of the IPTC taxonomy. Moreover, no labeled data is needed which makes our method data-independent. The investigation is guided by the following research questions:

- 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?
- What are the key differences of examined embedding methods and how do these distinctions impact their performances in the categorization task?"
- Can additional prompt engineering for IPTC categories or articles enhance the overall quality of the solution?

2 Related works

2.1 Classical machine learning methods

Classical machine learning approaches have been extensively employed in the categorization of news articles, with methodologies that typically involve the transformation of text data into vector space representations. The essence of these methods lies in the conversion of textual information into a set of features, often using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings before applying various machine learning algorithms.

A pivotal study in this area is the paper titled "Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study" (Bogery et al., 2022). The authors of this paper provide a comprehensive examination of ensemble machine learning techniques to automate the semantic categorization of news headlines. Their research underscores the efficacy of combining multiple machine learning models to improve classification accuracy.

The paper presents several key findings:

- Ensemble methods, which combine the predictions of several base estimators, often outperform single-model approaches, especially in tasks where the decision boundary is not linear or clear-cut.
- Machine learning models such as Support Vector Machines (SVM), Random Forests, and Naïve Bayes classifiers have been successfully applied to text classification problems, with varying degrees of success contingent upon the nature of the dataset and the complexity of the classification task.
- The preprocessing of text data, including the removal of stop words, stemming, and lemmatization, plays a crucial role in the performance of these models.

The study by Bogery et al. is instrumental in demonstrating that classical machine learning methods, despite the emergence of deep learning, remain relevant and powerful tools for news categorization. Their work supports the notion that with appropriate data preprocessing and model selection, classical algorithms can serve as a strong baseline or component in a more complex news categorization system.

That suggest that a carefully tuned Multinomial Naive Bayes classifier may be particularly effective for text classification tasks, achieving high accuracy and recall. The success of these classifiers hinges on the transformation of text into meaningful vector representations, utilizing methodologies like Word2Vec combined with TF-IDF vectorization to preserve semantic information.

2.2 XLNet

XLNet is a pretraining approach for language understanding models introduced in (Yang et al., 2019). Its main idea is to overcome the limitations of Autoregressive (AR) and BERT-based models.

The most important aspect of XLNet is the permutation-based training technique. Instead of forward or backward text factorization used in AR models, XLNet predicts words based on random permutation of the input. Therefore, the model is capable of capturing bidirectional context and the factorization allows model to consider all words in a sentence and their relationships, providing a better representation of the entire context.

Secondly, XLNet does not use masks for predicted words and does not introduce the mismatch in pretraining-fine-tuning discrepancy unlike BERT-based models, which utilize Masked-Language Modelling (MLM).

The model uses 2-stream self-attention mechanism, utilizing content and query representations and incorporates recurrence mechanisms from Transformer-XL ((Dai et al., 2019)) for capturing long-term dependencies.

2.3 Mask-guided BERT

Mask-guided BERT is a framework for Few-Shot-Learning (FSL) proposed in (Liao et al., 2023). It is a suitable method for dealing with the situation, when there are few labeled observations at our disposal. The Mask-BERT pipeline can be divided into 4 key parts:

- **Fine-tuning with base dataset** - the pretrained BERT (Devlin et al., 2019) architecture or its variant is used for initial fine-tuning on the base dataset. The base dataset is a large corpus of data and plays a supporting role for the FSL task on "novel" dataset.
- **Select anchor observations** - the fine-tuned BERT model is used for feature extraction of base and novel samples. Then, a set of "anchor" samples is selected out of the base dataset. They should meet the requirements of having relatively low distance to category centers, along with low distance to the few-shot novel observations.
- **Mask anchor samples** - XAI method (e.g. Integrated Gradients) is used for creating token masks for anchor samples. The main purpose of this step is to keep only relevant fragments of anchor samples (with regard to the novel dataset).
- **Final tuning** The final tuning of the model is used on the union of anchor and novel

dataset. The model is trained by minimizing loss function, being combination of cross-entropy and contrastive loss.

The Mask-guided BERT method was compared with BERT-based approaches (e.g. BERT, CNN-BERT) and other NLP techniques (CPFT). It yielded highest scores in terms of accuracy for text classification benchmark datasets, such as AGNews (Zhang et al., 2015) or DBPedia14 (Lehmann et al., 2015).

2.4 Seed-Guided methods

Seed-Guided methods utilize concept of a seed - a unigram or a phrase under which a set of therms that form a coherent topic may be found. Mentioned terms can be a unigram or a phrase as well. SeedTopicMine is a framework proposed is (Zhang et al., 2023) that uses the concept of seed in an iterative manner for topic discovery. There are three key modules in the framework:

- Initial Term Ranking that uses seed-guided text embeddings and PLM-based representations to find terms relevant to each category. For each category, top-r terms are found.
- Topic-Indicative Sentence Retrieval uses topic-indicative terms from previous module to find set of topic-indicative sentences.
- Ensemble of Multiple Types of Context calculates measure for semantic proximity between term and category based on topic-indicative sentences from previous module. Terms are ranked in descending order, additionally having embedding and PLM-based rank positions. Based on those three positions, rank ensemble is performed by calculating mean reciprocal rank (MRR). The updated term set contains only terms whose MRR score exceeds a certain threshold.

In the mentioned paper, SeedTopicMine is compared to other seed-guided modeling methods:

- SeededLDA - LDA method that biases each topic and document to generate more seeds and select topics relevant to the seeds respectively,
- Anchored CorEx - method that instead of relying on generative assumptions, leverages seeds by balancing between compressing the input corpus and preserving seed-related information,
- KeyETM - embedding-based model that modifies objective of ETM to utilize seeds in form of topic-level priors over vocabulary,
- CatE - embedding method for discriminative topic mining that jointly learn term embedding and specificity from input corpus. After that terms are selected based on embedding similarity with the seeds and specifity.

The study concludes that proposed SeedTopicMine framework outperforms existing seed-guided topic discovery methods in both term accuracy and topic coherence. It is suggested that the results may benefit keyword-based text classification via expanding the seed word semantics and prompt-based methods via enriching their verbalizers. It can also be extended to model input seeds organized in a hierarchical manner by injecting regularization or discovering topics beyond the provided seeds by incorporating latent topic learning in the corpus modeling process.

2.5 LLM prompting

Prompt engineering has gained significance in the context of recent advancements in Large Language Models. It focuses on the formulation of precise instructions or queries to elicit desired responses from models like GPT. Such approach was explored for IPTC categorization problem (Fatemi et al., 2023), where authors utilized GPT-3.5 Turbo architecture to model the first and second IPTC hierarchy.

2.6 State-of-the-art language model embeddings

The ”Massive Text Embedding Benchmark (MTEB)” by (Muennighoff et al., 2023) is a comprehensive evaluation of text embedding methods that spans eight embedding tasks, covering a total of 58 datasets and 112 languages.

The paper has benchmarked 33 models and finds that no single text embedding method is superior across all tasks, suggesting that the field has yet to converge on a universal text embedding method that can provide state-of-the-art results on all embedding tasks.

Classification within the MTEB is executed by training a logistic regression classifier on top of embeddings extracted from language model. The main metric for classification tasks is accuracy, with average precision and F1 score also provided as additional metrics.

The benchmark found ST5 models dominate the multilingual classification task across most datasets. ST5-XXL has the highest average performance, 3% ahead of the best non-ST5 model - OpenAI's Ada.

3 IPTC text classification through LLM embeddings

3.1 Solution framework

We propose a method for classifying news articles with the usage of embeddings generated with Large Language Model. Our framework consists of 3 steps (fig 12):

1. Generating embedding for each category name or description.

2. Generating embeddings for article using generally pretrained Large Language Model.

3. Calculating cosine similarity between each article and each category name. The article is assigned the closest category in the embedding space.

Language model embeddings capture the semantic meaning of the whole text fragment and can be used in multilingual setting depending on the model. For the sake of fast prototyping, we decided to apply OpenAI ADA Embeddings, which are still considered state of the art across variety of NLP tasks according to results from 2023 paper "MTEB: Massive text embedding benchmark" (Fatemi et al., 2023).

3.2 Solution advantages

The main advantages of solution are:

1. **Speed/Scalibility:** Classification requires only one forward-pass through transformer model.

2. **Expendability:** Can be easily extended to new categories since it doesn't require model fine-tuning.

3. **Multilinguality:** Since solution can use any generally pretrained LLM, it can also work for various other languages depending on the capability of base LLM.

4 IPTC taxonomy and datasets

4.1 IPTC taxonomy

The IPTC (International Press Telecommunications Council) Taxonomy is a standardized system used for categorizing and tagging news content and media assets in journalism and media industries. It provides a structured set of controlled vocabulary terms for consistent content management, efficient search, and cross-platform publishing.

Categories

The IPTC Taxonomy covers various categories, including topics, locations, media types, rights, people, organizations, events, and dates, ensuring efficient and consistent news content handling.

4.2 STA dataset

4.2.1 Slovenian Press Agency (STA)

The Slovenian Press Agency (STA) is a leading news agency headquartered in Slovenia, operating under the abbreviation STA. It serves as a primary source of news and information about Slovenia and the surrounding region.

4.2.2 Avaiable STA's datasets

STA is a reputable news agency engaged in various activities, including:

- **News Reporting:** STA provides comprehensive news coverage, reporting on topics such as politics, economics, culture, and sports, with a focus on Slovenia and the region.

- **Media Services:** STA offers media services, including text, photo, and video content for various platforms.

- **Public Information:** It serves as a vital source of public information, ensuring transparency and accessibility to news and events in Slovenia.

- **International Reach:** STA covers international news, making it a valuable resource for global audiences interested in the region.

- **API Access:** We have access to STA's API, allowing us to retrieve news articles enriched

with manually assigned categories, titles, and additional information.

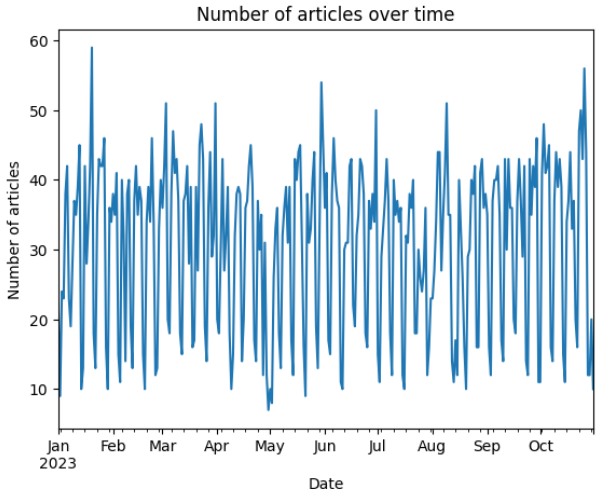

Figure 1: Distribution of STA English articles over time.

STA's commitment to accurate and timely reporting has established it as a trusted news source for both domestic and international audiences.

The STA News Dataset used in this project contains 8778 English articles from 2023 alone. Among all fields available in the data, the most relevant for this research were ones containing headline, keywords and text of the article. The most common words in those articles, as shown in figure 2, were LJUBLJANA, Slovenia, said and year. As for the keywords, most prominent were Slovenia, Press and Events. All most frequent keywords are visalized in figure 3.

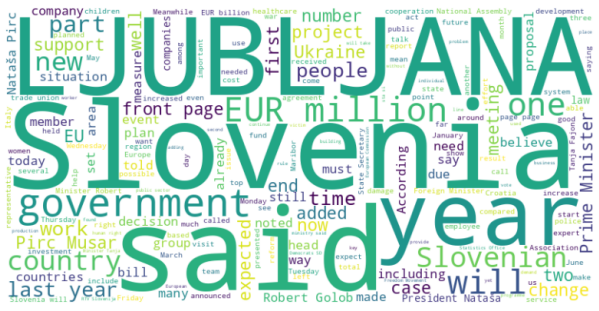

Figure 2: Wordcloud of most common words across all articles.

4.3 Related Datasets

There are multiple open-source datasets related to the news categorization task. Such datasets might

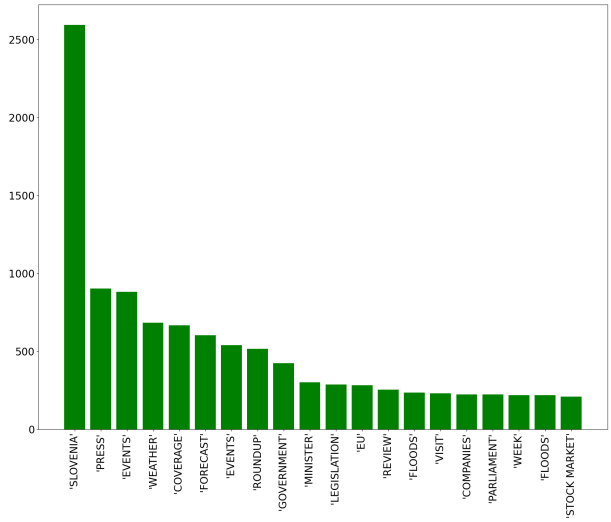

Figure 3: Histogram of most frequent keywords provided in the dataset across all articles.

also provide valuable resources for IPTC classification and are as follows:

4.3.1 AGNews Dataset

The AGNews dataset is a widely used collection of news articles categorized into four distinct classes: World, Sports, Business, and Science/Technology. This dataset is commonly employed in text classification and natural language processing (NLP) tasks. It can be accessed at `http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html`.

4.3.2 DBPedia14 Dataset

The DBPedia14 dataset is derived from DBpedia, a large-scale multilingual knowledge base extracted from Wikipedia. It consists of articles that cover a wide range of topics, each labeled with a specific category. DBPedia14 is a comprehensive resource for text classification and knowledge-based NLP tasks. The DBPedia14 dataset is available through the Hugging Face Datasets library and can be accessed via `https://huggingface.co/datasets/dbpedia_14/tree/main`.

4.4 Test set preparation

The test set consists of 300 samples from the STA data. Considering the lack of ground truth labels, observations would be labeled manually. The samples were chosen from the English articles, excluding articles that do not contain text or ones that would be difficult to assign to just one cate-

gory. An example of such article would be article containing schedule for a specific day. It contains information related to multiple events, of which all of them may be assigned to other category.

There are three methods used to sample the data:

- **Balanced sampling:** the categories are sampled in a way, that each of them has the same number of samples.
- **Stratified sampling:** the number of samples from each category has same proportion of samples as the original dataset.
- **Random sampling:** the samples are chosen randomly.

The first two approaches are done in reference to predicted classes, not the ground truth labels. It may add bias to the resulting test sets.

Data labeling is done using an app developed during this research. The app shows the article's text along with it's headline, one article at a time. Below it, predicted label is shown. In case the article was previously annotated, the chosen label is shown as well. In this version, one of the eighteen categories of the highest hierarchy can be chosen as a label. The visual interface of the created application is visualized in figure 4.

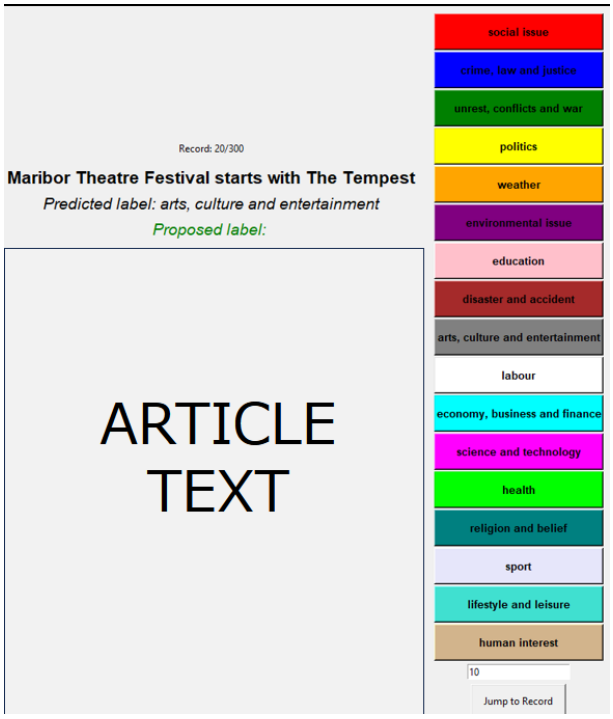

Figure 4: Labeling application interface.

The final test set contains the article's headline, text, the predicted high label and chosen label.

5 Experiments

5.1 Experimental settings

We conduct two experiments:

- Comparison of performance of closed-source OpenAI Ada embeddings with open-source AnglE embeddings.
- Comparison of performance between category embeddings generated with various strategies: name, name + description, name + description + keywords.

5.2 Taxonomy preparation

The taxonomy data is sourced from the original IPTC taxonomy file and processed to construct a structured file. Few rows from the original file are shown in figure 5. First, based on the available data, the hierarchy of each IPTC category is extracted. It is done to provide a multi-level representation of taxonomic relationships. As shown in table 1, the higher the hierarchy, the less categories it contains. This stage of the project focused on the highest possible category due to its lowest count of unique categories. It allowed for better data labeling and more observations in all of the categories.

	code	hierarchy	name	description
0	subj:01000000	1	arts, culture and entertainment	Matters pertaining to the advancement and refi...
1	subj:01001000	2	archaeology	Probing the past through ruins and artefacts
2	subj:01002000	2	architecture	Designing of buildings, monuments and the spac...
3	subj:01003000	2	bullfighting	Classical contest pitting man against the bull
4	subj:01004000	2	festive event (including carnival)	Parades, parties, celebrations and the like no...

Figure 5: Overview of taxonomy file.

Hierarchy	Count
1	17
2	385
3	535

Table 1: Hierarchy count.

Next, embeddings are created, to enhance the representation of the data. It is done using OpenAI'a Ada Embeddings and AnglE Embeddings,

both described above. Embeddings are made separately for both name of the category, as well as for its description.

Resulting dataset consists of code, hierarchy, name of the category, its description and created embeddings.

5.3 OpenAI Ada vs AnglE

Predictions are done using embeddings created with OpenAI's Ada and AnglE Embeddings. The embeddings are done for column containing article's text. Using them, as well as embeddings created during taxonomy dataset preprocessing, cosine similarity is calculated. The chosen label is category with highest cosine similarity value.

Achieved performance is shown in table 2. Ada Embeddings accomplished both higher accuracy and F1 score. On top of that, taking a look at confusion matrixes in figures 6 and 7, the misclassification often happened in categories related to each other in some way. It may have happened due to the fact that not all of the articles could have been assigned to just one category, but more of them may have been correct category for it.

Metric\Embedding	Ada	AnglE
Accuracy	85.67%	64.00%
F1 score	80.40%	51.49%

Table 2: Embeddings performance.

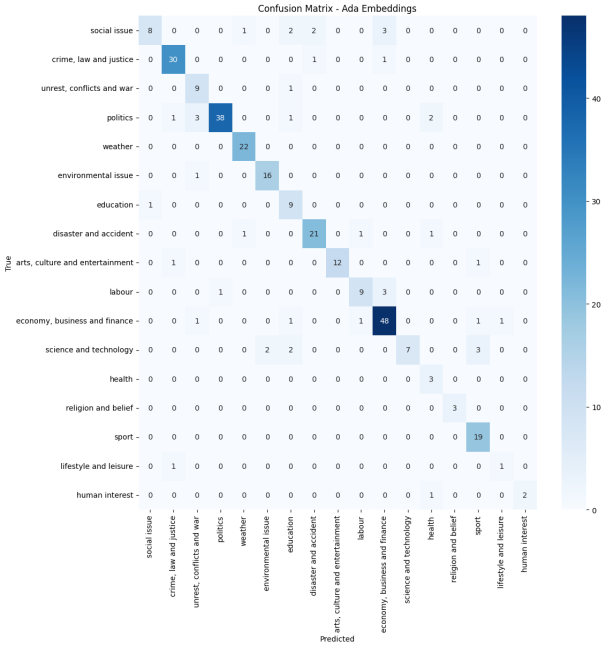

Figure 6: Ada Embeddings confusion matrix.

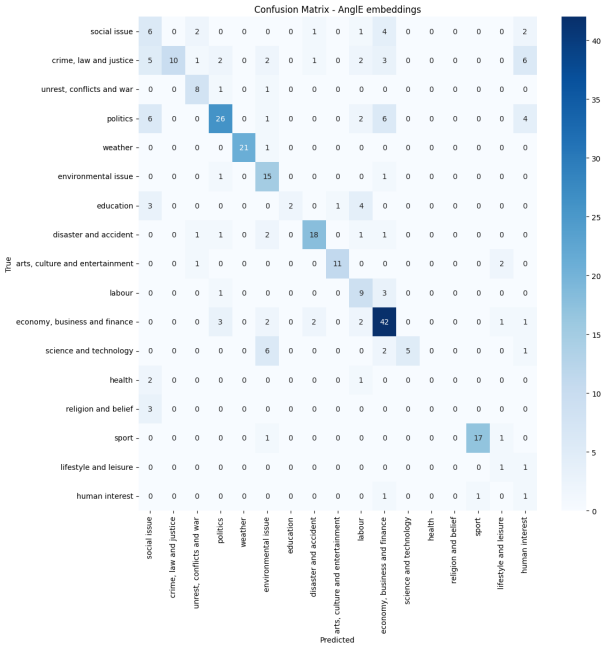

Figure 7: AnglE Embeddings confusion matrix.

```
Generate as many semantically diversified keywords as you can,
related to the following category of news articles:
{category} - {description}

Format:
```
{{
   "keywords": [...]
}}
```
```

Figure 8: Prompt for generating keywords based on category description.

The classes themselves are not balanced, as shown in figure 10. The most common are related to economy, politics and weather. Classes related to religion, lifestyle and human interest were not predicted often, which may support the thesis that the misclassification happened where categories were too broad for model to understand them. The articles themselves were not much different in length. The only category that has significantly less words on average than the rest in weather, as seen in figure 9.

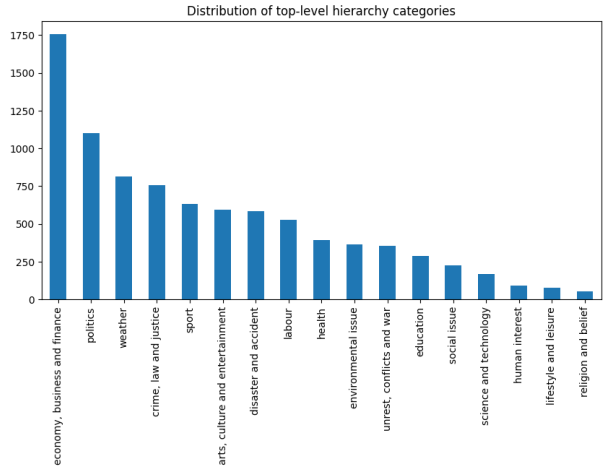

Figure 9: Average Word Count per Category.

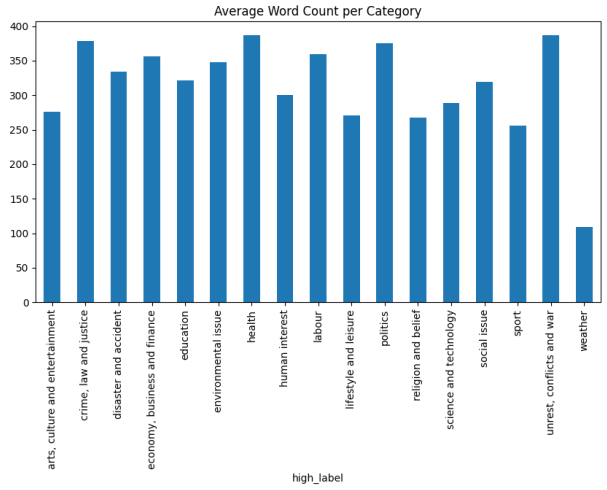

Figure 10: Distribution of top-level hierarchy categories.

5.4 Name vs description embeddings

Categories were predicted in three ways. First, only the embedding of category name was considered. Second, the original IPTC category description was used. Lastly, the category description augmented with GPT-4 generated keywords was utilized. The prompt for generating keywords is depicted on figure 8. As shown in figure 11, using only the name of the category yielded the best results in most categories. Only for top three accuracy, expanded descriptions achieved better score. Descriptions performed the worst in all of used metrics.

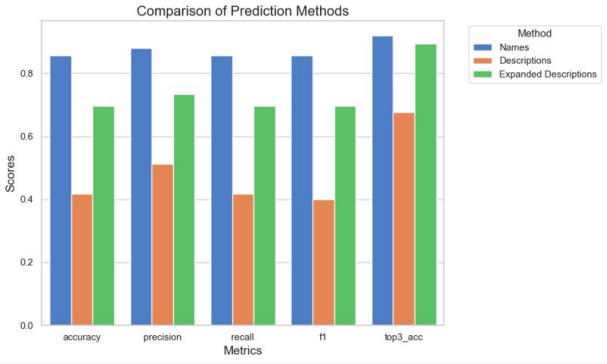

Figure 11: Prediction methods.

6 Future Works

As we continue to develop and refine our project, several areas of improvement and expansion have been identified. We focus on accomplishing the following tasks:

6.1 Vector Databases

The implementation of vector database is a key step in enhancing the performance and efficiency of our model. This involves the establishment of optimized database to store vector representations of IPTC categories for quicker retrieval and processing. Moreover the database could be extended with news articles and their headlines, which are classified on high confidence level and could enhance the quality of queries.

For now we decided to go with ChromaDB, an open-source solution which can be easily integrated with Python applications. The database is deployed locally, however the ultimate goal is to make it available for public usage.

6.2 Quality Article Pre-processing

In our pursuit of delivering high-quality categorized news articles, we plan to enhance article preprocessing with the following steps:

- **Remove HTML Tags:** Improve text cleanliness by removing HTML tags from the article content.

- **Split Texts into Sections:** Segment articles into meaningful sections and calculate the average embeddings for each section to capture the nuances of different parts of an article effectively.

- **Incorporate headlines:** The headlines of articles are not considered in the embedding process. It leads to information loss, as they represent the most relevant and essential part of text. We might consider two additional strategies: categorisation based on the headlines only and on the merging product of both the headlines and article bodies.

6.3 Self-improvement system

Our system would include a self-improving mechanism through continuous integration of new articles into the vector database. Articles surpassing a predefined threshold of similarity with IPTC categoriy embeddings present in the database would be incorporated. Such process holds the potential to enhance the efficiency and effectiveness of the database querying system, as new incoming articles could be matched not only with category embeddings, but also with high quality article embeddings.

6.4 Final Architecture Diagram

To provide a comprehensive view of our system, we created a final architecture diagram. This diagram illustrates the components, data flow, and interactions within the final solution, making it easier to understand the overall structure and functionality.

These future developments aim to optimize our project, improve the accuracy of category assignment, and enhance the user experience when classifying and accessing categorized news articles.

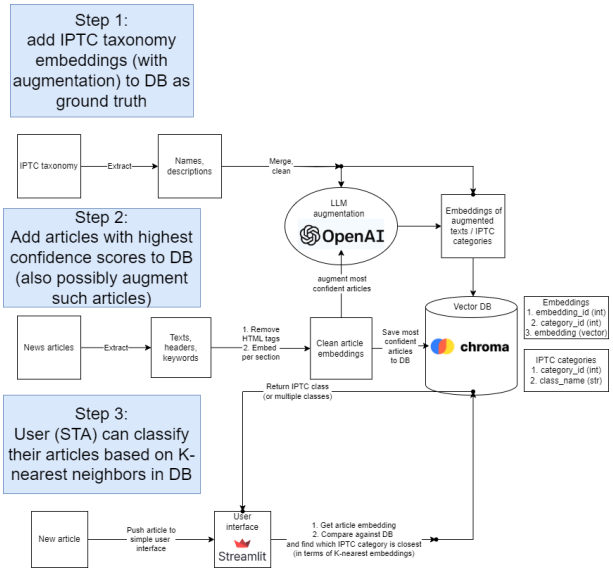

Figure 12: Final Architecture Diagram

References

Raghad Bogery, Nora Al Babtain, Nida Aslam, Nada Alkabour, Yara Al Hashim, and Irfan Ullah Khan. 2022. Automatic semantic categorization of news headlines using ensemble machine learning: A comparative study. *Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bahareh Fatemi, Fazle Rabbi, and Andreas L Opdahl. 2023. Evaluating the effectiveness of gpt large language model for news classification in the iptc news ontology. *IEEE Access*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Yuzhong Chen, Xi Jiang, Dajiang Zhu, Tianming Liu, et al. 2023. Mask-guided bert for few shot text classification. *arXiv preprint arXiv:2302.10447*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association*

for Computational Linguistics, pages 2014–2037, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective seed-guided topic discovery by integrating multiple types of contexts.