IPTC News Categorisation
Project Proposal 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 proposal serves as an introduction to our research on the automation of news categorization in line with the IPTC taxonomy. Our initial investigations have highlighted the potential of NLP techniques to transform the efficiency and accuracy of news classification. 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. We present a conceptual framework for our approach, which we anticipate refining through iterative development and testing. This document encapsulates our preliminary findings and the foundational ideas that will guide the subsequent stages of our project.

1 Scientific Goal of the Project

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 International Press Telecommunications Council (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.

1.1 Description of the Problem

Manual categorization of news content 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 IPTC taxonomy, while comprehensive, requires an understanding of complex news narratives to apply accurately, a task that is well-suited to the capabilities of advanced NLP methods.

1.2 Research Questions

The project is guided by the following research questions:

- Can state-of-the-art NLP techniques, including deep learning and transformer-based models, effectively automate the categorization of news articles in line with the IPTC taxonomy?
- What are the best practices for training and evaluating NLP models on the task of news categorization to achieve results that are both accurate and scalable?
- 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?

1.3 Hypotheses

The project is predicated on two core hypotheses:

1. The application of advanced NLP models will significantly outperform traditional keyword-based and rule-based methods in the task of categorizing news articles according to the IPTC taxonomy.
2. Automating the news categorization process will lead to a substantial reduction in manual labor and time expenditure for news organizations, thereby increasing the overall throughput of news content processing.

2 Significance of the Project

The significance of this project lies in its potential to revolutionize the field of news categorization by incorporating the IPTC taxonomy with the latest NLP techniques. This integration addresses a pressing need within the digital news landscape for more efficient content management systems. By advancing the state of the art in automated news categorization, the project stands to make substantial contributions to both the practice of journalism and the field of NLP.

2.1 Justification for the Scientific Problem

The proliferation of digital news sources has made it difficult for news organizations to quickly and accurately categorize content. As the volume of information grows, so does the need for improved classification systems. This project seeks to fill that gap, providing a scientifically sound approach to automating news categorization that can keep pace with the rate of information production.

2.2 Impact of Project Results

The expected impact of the project is multifaceted:

- It will increase the accuracy and consistency of categorization, improving news discoverability.
- The findings might contribute to academic research in NLP, offering insights into the application of machine learning in real-world text classification tasks.
- If successful, the methodology could be adapted to other domains that require text categorization, broadening the project's relevance.

3 Concept and Work Plan

The concept of this project is centered on the development of categorization system that is efficient, robust and accurately assigns IPTC taxonomy categories. This system will be based on state-of-the-art NLP models, embeddings and machine learning algorithms.

3.1 General Work Plan

We divided our project into 4 main parts, that are presented in the table 1.

Date	Stage Name	Description
8.11.2023	Project Proposal	Literature review, solution concept, and proposal
15.11.2023	EDA	Exploratory data analysis performed on STA data
22.11.2023	Proof of Concept	Preliminary ML models and solutions
13.12.2022	Final Project	Full solution and prepared product

Table 1: Project Timeline

3.2 Specific Research Goals

The project's specific research goals include:

1. To establish a baseline for IPTC news article categorization using traditional machine learning models.
2. To assess the influence of embeddings for the overall evaluation.
3. To investigate and implement advanced deep learning techniques for improved classification performance.
4. To evaluate and compare the effectiveness of different NLP models in the context of IPTC taxonomy.

4 Approach & research methodology

4.1 State-of-the-art

4.1.1 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 un-

like 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 (Z. Dai et al. 2019) for capturing long-term dependencies.

4.1.2 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 pre-trained BERT (Devlin et al. 2018) 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 following requirements:
  - Low distance to category centers.
  - 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) and yielded highest scores in terms of accuracy for text classification benchmark datasets.

4.1.3 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 Yu 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 Seed-TopicMine 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.

4.1.4 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" by Bogery et al. 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.

4.1.5 State-of-the-art language model embeddings

The "Massive Text Embedding Benchmark (MTEB)" by Muennighoff et al. 2022 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.

4.2 Solution concept & methodology

1. At first, exploratory data analysis with data preprocessing and visualization will be conducted. Our main goal is to get a good

grip on understanding of the STA dataset and the underlying schemas for the articles. The scope of this analysis would include creating wordclouds for the articles, analyzing the frequency and type of words appearing in the news, sentiment analysis in different article groups (which might be beneficial for the categorisation itself and might bring some valuable insights), outlier detection and more.

2. The next step would include building test set, which would serve as an indicator of our model quality. We decided that the test set would consist of 300 samples coming from STA data and manual data labeling will be performed. We believe that, at least for top-hierarchy of the IPTC taxonomy, the size of the test set is appropriate and might be extended, if we decide to explore lower levels of the taxonomy.

3. Proof of concept: we decided to apply SOTA OpenAI Ada embeddings for category prediction (which are still considered state of the art across variety of NLP tasks according to results from Muennighoff et al. 2022), together with article-category cosine similarity and taxonomy dataset. Exploratory analysis of the results will be performed, as well as model preliminary evaluation (including relevant scoring metrics and confusion matrix). For further work, we plan to:

    - Try generating embeddings for category descriptions instead of category names. This has a chance to increase the precision of the categorization.
    - Use language models to generate a set of diversified descriptions for each category, based on the original description. Then we can generate their embeddings and approach the whole problem through as a sort of voting process. The are many ways to approach this so the details of such solutions are yet to be defined.

    Other embedding method will also be explored in comparison to Ada embeddings.

4. Possible use of other methods for comparison:

    - topic discovery methods (e.g. Seed-Guided methods, BERTopic)
    - Few-Shot learning with labeling small portion of data (e.g. Mask-guided BERT)
    - Label bigger data corpus and try supervised learning approaches along with transfer-learning approaches (e.g. classical machine learning methods, XL-Net)
    - Large Language Models for topic discovery

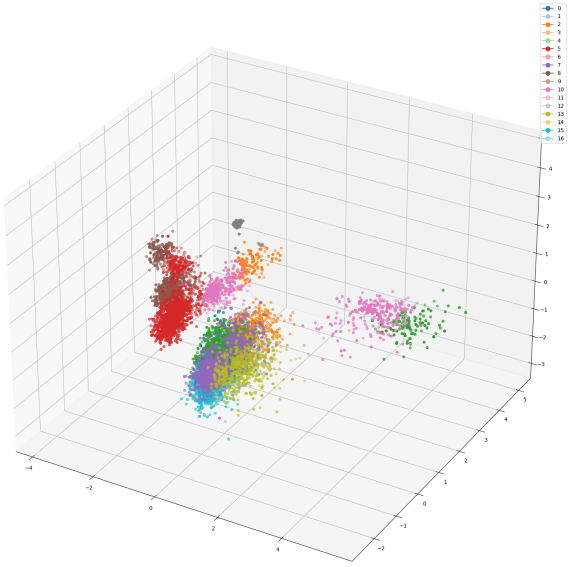

Figure 1: Clustering of STA article embeddings

5 Datasets

5.1 STA News Dataset

Consists of Slovenian and English news articles from STA channel, with both text and metadata. Contains such information as headline, keywords, categories and lede.

5.2 Taxonomy

A scheme used for classification. Contains names of IPTC categories with their description, their mapping to STA data and topic.

5.3 Text classification datasets

1. **AG News** - Consists of news articles categorized into four classes: World, Sports, Business, and Sci/Tech. It is often used for news topic classification and text categorization tasks, (Yang et al. 2019, Liao et al. 2023)

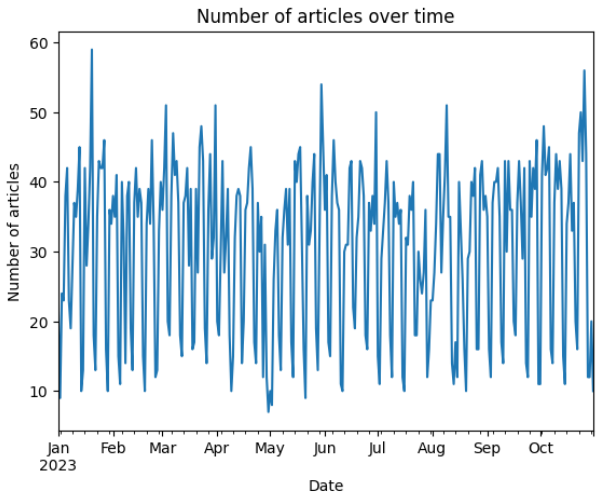

Figure 2: Distribution of STA English articles over time.

2. **DBpedia14** - ontology classification dataset, containing 14 different classes (used in Liao et al. 2023)

6 Rebuttals

In this section, we address several concerns and comments raised during the review of our report. Our responses are structured to clarify these points and provide a comprehensive understanding of our approach and the limitations we faced.

6.1 Dataset Description

Due to the Non-Disclosure Agreement (NDA) we had to sign, the description of the dataset is brief. The NDA restricts the amount of detail we can provide publicly about the dataset. This limitation is not a reflection of inadequate reporting but a legal obligation we are bound to respect.

6.2 Lack of Risk Analysis in the Report

While it is noted that a dedicated risk analysis section is not present in the main body of our report, it is important to clarify that such an analysis was indeed conducted and included in the accompanying presentation. After thorough discussions within our team, we collectively decided that at this stage of the project, the conclusions drawn from the risk analysis were too general and lacked practical depth. We believed it was not beneficial to forcibly include these preliminary and broad findings in the report. Our decision was guided by a commitment to maintaining the quality and relevance of the information presented in the report.

6.3 No Reference to First Two Datasets

Regarding the first two datasets used in our study, it is important to note that these datasets do not have explicit references available. However, in our report, we have clearly indicated the source of the data used, specifying that they were provided by STA.

Regarding the other two datasets which might be used as additional training resources, we admit that the references are not given explicitly, although we mention the papers that utilize them. The AGNews dataset might be found here `http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html` and the DBPedia14 here `https://huggingface.co/datasets/dbpedia_14/tree/main`.

6.4 Comprehensive Analysis of Related Works

We disagree with the comment that our report lacks a comprehensive analysis of related works. Our literature review is extensive and closely related to the topic at hand. We believe this criticism may stem from a misunderstanding or a cursory reading of our report. We have cross-referenced numerous sources and believe our literature review adequately covers the scope of the research.

6.5 No Exploratory Data Analysis

The absence of an exploratory data analysis in our report is directly due to the restrictions imposed by the NDA. The NDA limits the extent to which we can analyze and publicly share findings derived from the dataset. This constraint significantly impacted our ability to conduct a traditional exploratory data analysis.

7 Team Member Contribution

- Jan Wojtas - Research questions, approach and research methodology (XLNet, Mask-guided BERT), Solution concept, Datasets (AGNews, DBPedia), Exploratory Data Analysis & PoC results with Ada Embeddings
- Paulina Szymanek - solution concept and methodology, Seed-Guided methods, datasets, rebuttals
- Mikołaj Zalewski - STA API code, data labeling, overview of classical ML methods,

IPTC taxonomy research, research of alternative IPTC news APIs

- Łukasz Zalewski - solution concept and methodology, overview of SotA language model embeddings

References

Bogery, Raghad et al. (2022). "Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study". In: *Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University*.

Dai, Zihang et al. (2019). "Transformer-xl: Attentive language models beyond a fixed-length context". In: *arXiv preprint arXiv:1901.02860*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Liao, Wenxiong et al. (2023). "Mask-guided bert for few shot text classification". In: *arXiv preprint arXiv:2302.10447*.

Muennighoff, Niklas et al. (2022). "MTEB: Massive text embedding benchmark". In: *arXiv preprint arXiv:2210.07316*.

Yang, Zhilin et al. (2019). "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32.

Zhang, Yu et al. (2023). *Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts*.