194.093 Natural Language Processing and Information Extraction

Topic 3: Extraction of Narratives from Online News (‘propaganda detection’)

Subtask 2: Narrative Classification

Ahmed Šabanović (12330648)
Ivan Babiy (12142160)
Théo Hauray (12404716)
Tibor Čuš (12325298)

Management Summary

Overview

The goal of this task is to identify and characterize narratives in online news articles to help detect potential propaganda manipulation attempts. This task involves analyzing articles from two domains: the war in Ukraine and climate change. The dataset includes multilingual articles in five languages: Bulgarian, English, Hindi, Portuguese, and Russian.

For this task, we have chosen the subtask of **Narrative Classification**. Given a news article, the task's main purpose is to assign appropriate narrative label(s) based on a two-level taxonomy of narrative categories, which (most of the time) include sub-narratives. This classification will allow for the identification of the underlying narratives and potential biases in the articles, a key step in detecting and preventing propaganda.

The task is a multi-label multi-class document classification problem, in other words, each article can be assigned multiple (sub-)narrative labels.

Challenges

Several challenges were encountered during the project:

- **Multilingual Data:** Handling articles in five different languages (Bulgarian, English, Hindi, Portuguese, and Russian) added complexity, especially with differences in syntactic structures and cultural nuances.
- **Complex Narratives:** Each domain contains over 10 narratives and almost 40 sub-narratives, with ideological and nuanced statements, requiring deeper contextual understanding.
- **Model Complexity:** Selecting and fine-tuning the right NLP models for multi-label classification, especially given the amount of sub-narratives, required considerable experimentation.
- **Data Imbalance:** Limited training samples for certain labels (some sub-narratives are labeled in very few articles) severely impacted performance, especially for deep learning models.
- **Ambiguity in Labels:** Overlapping or vague categories (e.g.: "Other") reduced classification accuracy.

External Resources

Several tools and libraries were used to implement, improve, and evaluate the final solution, including:

- **Python 3.12**: the latest version of Python was used.
- **Jupyter Notebook**: for interactive development, analysis, and visualization of results in a clear and modular manner.
- **Poetry**: enabled dependency management and virtual environment creation, ensuring reproducibility.
- **Numpy, Pandas, & Matplotlib**: for processing, visualization, and analysis of performance metrics and other results.
- **Scikit-learn**: for implementing traditional machine learning models (i.e.: Multinomial Naive Bayes, Random Forest), as well as for metrics calculation.
- **Transformers Library by Hugging Face**: for implementing the deep learning models i.e.: BERT, RoBERTa), including fine-tuning and evaluation.

Solution Implemented

To classify the (sub-)narratives in news articles, we implemented both traditional machine learning and deep learning approaches. The traditional methods, including Multinomial Naive Bayes and Random Forest, relied on simple word count features (i.e.: Bag of Words) for text representation. These methods were quick to train and evaluate, the Multinomial Naive Bayes models performed consistently better than the Random Forest models due to its compatibility with sparse datasets.

On the other hand, we fine-tuned state-of-the-art deep learning models like BERT and RoBERTa for multi-label classification. To address the challenges of class imbalance and improve model learning, we incorporated a weighted random sampler during data loading and a custom Binary Cross-Entropy (BCE) loss function with dynamic class weighting during training. Additionally, we applied threshold optimization to fine-tune the classification decision boundary and maximize macro F1 scores. To further enhance performance, we employed a hierarchical modeling approach, where one model was trained to classify narratives at a higher level, while a second model specialized in identifying sub-narratives. Despite these enhancements, the deep learning models struggled with the dataset's small size and uneven class distribution, leading to poor overall performance compared to the simpler (traditional) methods.

Among all approaches, MultinomialNB emerged as the most effective, achieving the highest macro F1 scores. This outcome emphasizes that, for this complex task, straightforward methods can be more reliable than sophisticated models when working with limited data.

Solution Limitations

The BERT-based solution struggles to model the hierarchical relationship between narratives and sub-narratives, treating them equivalently rather than as structured dependencies. Severe class imbalance, particularly the dominance of the "Other" class and the under-representation of most categories, leads to poor generalization, with some classes exhibiting complete failure. The model suffers from over-prediction, with high recall but extremely low precision, generating many false positives.

While combining narrative and sub-narrative classification simplifies the architecture, it prevents specialization, limiting performance. Despite employing threshold optimization and dynamic class weighting, these techniques fail to fully mitigate the imbalance and improve learning for minority classes. Additionally, the independent treatment of each label in the multi-label setting neglects potential correlations between related classes, further impacting classification accuracy.

Conducting a qualitative analysis corroborated the suspicions raised during the quantitative metric analysis. More often that not, the model predicts labels as *False* when the ground truth is *True*. *Other* is a recurring label, being frequently misclassified, suggesting the inability of the model to learn the underlying distribution of the sub-narratives (probably due to the relatively small amount of training data). Unsurprisingly, articles mentioning specific persons or policies are always misclassified, indicating the difficulty for the model to "interpret" subtle language and multi-layered arguments.

Possible Next Steps

Further possible steps, in order to mitigate the limitations faced, are:

1. **Incorporate New Data**
   - Collect additional labeled data for underrepresented classes, particularly subnarratives, to address class imbalance.
   - Expand the dataset with more diverse articles from each language, ensuring broader coverage of narrative categories.
   - Include additional Russian-language data to balance the multilingual representation.
2. **Refine the Models**
   - Improve hierarchical modeling by explicitly modeling relationships between narratives and subnarratives.
   - Explore transformer-based multilingual models (e.g., XLM-RoBERTa) to better handle multilingual data.
   - Experiment with label correlation techniques, such as conditional dependencies, to enhance multi-label classification.
   - Regularize overprediction using techniques like focal loss or confidence-based prediction thresholds.
3. **Quantitative and Qualitative Analysis of the Final Model**
   - Compare the performance of the refined model against both baselines (e.g., MultinomialNB) and earlier iterations.
   - Conduct a more comprehensive evaluation across all languages and labels, to find recurring patterns that may serve as clues on what the models are missing and how to improve the results.
   - Perform error analysis to identify recurring misclassifications and fine-tune the model accordingly.