Extraction of Narratives from Online News

Management Summary

Leon Baumgärtner - Sergej Bensack - Niklas Kessler - Jonas Kruse

1 Task and Challenges

A reliable method for narrative classification, to which we aim to contribute with this task, has numerous benefits, including fighting misinformation and improving news transparency on online platforms.

The task can be framed as follows: Given an article, the prediction goal is the set of narratives contained. For this, we are given a fixed two-level taxonomy of narrative and sub-narrative labels, such as "Downplaying climate change" - "Climate cycles are natural". The articles cover the topics Ukraine-Russia war and climate change. Articles of different languages are contained in the dataset, including English, Russian, Portuguese, Hindi and Bulgarian. [2]

The main challenges include the sparsity of data for training classifiers in relation to the task complexity. The task complexity is driven by two key aspects: A large number of possible narratives to choose from when classifying an article and the imbalanced occurrence of articles that belong to specific narratives. There are also semantic challenges, namely the question of what is classified as a narrative, how is an article classified that is not considered a narrative and the subjectivity involved in assigning a narrative to an article.

2 Implemented Solution

2.1 Methods

The applied methods are Support Vector Machines (SVM) as a traditional technique and BERT and LLAMA as deep-learning techniques. Within these methods, different approaches of solving the specific problem were applied. These include training separately for the topics Ukraine war and Climate change as well as using the complete data for training. The rationale behind this is lowering the task complexity by splitting it into easier to handle sub-tasks.

For the SVM method, different input representations of the text data in the articles were used to identify the most effective one, including Bag-of-Words, TF-IDF and multilingual embeddings produced by BERT [1]. The input representation has a direct impact on the performance that can be hoped to be achieved.

For the LLaMa training and predictions we used a model from OpenLM-Research "openlm-research/open_llama_3b". It utilizes a two-phase training strategy. The approach includes an initial classification head training phase followed by a complete model fine-tuning phase, which helps to adapt the pre-trained model to the specific task.

We use the metrics precision, recall and F1-Score, where a high recall is the desired outcome for this task, as we are interested in catching actual narratives rather than being confident that each and every predicted narrative actually applies (e.g. flagging online articles with a disclaimer).

2.2 Results

The comparison of final model performances is shown in Table 1. Regarding the vectorization methods used with SVM, we note that TF-IDF yields the best results across almost all datasets and metrics, outperforming multilingual embeddings despite multiple languages being present in the dataset.

All methods yield overall much better results across metrics on the Ukraine-War dataset compared to the Climate-Change dataset, with the results of the full dataset somewhere in between. Therefore, the hypothesis that splitting the dataset to reduce task complexity improves performance can not be generally confirmed. However, in case of the UA dataset, the results were considerably improved compared to the full dataset.

The BERT model generally lends towards higher recall, while the SVM favors precision. LLama demonstrates strong performance on the Climate Change (CC) dataset with the highest F1 score of 0.398 and superior recall of 0.528. However, its performance varies significantly across datasets, showing notably weaker results on the full dataset where it achieves the lowest F1 score (0.098). This variance suggests that LLama's effectiveness is highly dependent on the specific dataset characteristics.

Word-level analysis of correctly and incorrectly classified articles revealed some insights about how the model makes decisions. For correctly classified articles of the class "Hidden plots by secret schemes of powerful groups" -

Dataset	Method	Vectorization Method	F1	Precision	Recall
UA	SVM	Bag-Of-Words	0.169	0.256	0.137
	SVM	TF-IDF	0.209	0.436	0.146
	SVM	Multilingual Embeddings	0.207	0.365	0.157
	BERT	-	0.113	0.077	0.213
	LLama	-	0.221	0.194	0.306
CC	SVM	Bag-Of-Words	0.248	0.275	0.249
	SVM	TF-IDF	0.264	0.381	0.231
	SVM	Multilingual Embeddings	0.285	0.335	0.267
	BERT	-	0.323	0.261	0.437
	LLama	-	0.398	0.321	0.528
Full	SVM	Bag-Of-Words	0.141	0.232	0.121
	SVM	TF-IDF	0.225	0.438	0.171
	SVM	Multilingual Embeddings	0.206	0.351	0.166
	BERT	-	0.146	0.115	0.254
	LLama	-	0.098	0.064	0.212

Table 1: Combined performance metrics (SVM, BERT, LLama) across datasets and methods.

"Blaming global elites", for instance, words clearly tying an article to the class include "globalist" and "extremist", with little off-topic content. Incorrectly classified articles were often diluted by words from other topics not directly related to the class.

2.3 Limitations

All methods show a performance insufficient to use in automated decision making based on the predictions. Our findings can therefore rather be seen as prototype towards improving ML-based narrative classification and experimentation with topic-independent classification.

The imbalance of the dataset has a significant influence on the measured results. Table 2 shows the significant decrease in performance when examining macro-scores, highlighting how poorly the models handle the most underrepresented classes.

Score Calculation	F1	Precision	Recall
Weighted Avg	0.225	0.438	0.171
Macro Avg	0.087	0.192	0.064

Table 2: SVM with TF-IDF on full data set

Lastly, it is important to keep in mind, that narrative classification based on a fixed taxonomy is not synonymous with propaganda detection. Correct classification does not necessarily imply propaganda, yet it is an important step that could support effective propaganda detection, which furthermore requires the definition of narratives that indeed imply propaganda.

3 Conclusion and Recommendations

Although it could not be definitely confirmed that splitting the task by topic increases performance, we still recommend trying the approach, as we showed that it can lead to improvements in some cases. Another approach that could potentially lead to further improvements is applying a two step modeling approach, where a first model predicts the narrative and a second the subnarrative, incorporating the prediction of the first. This is another way of splitting the task into easier subtasks.

Our findings show that a final model recommendation is highly dependent on the specific application. We recommend applying LLama or BERT for cases where recall is of interest, and SVMs with TF-IDF where precision might be more relevant.

References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[2] Jakub Piskorski et al. *SemEval 2025 Task 10: Multilingual Characterization and Extraction of Narratives from Online News*. `https://semeval2025-task10.org/`. 2025. URL: `https://semeval2025-task10.org/`.