JPM444 - Arsenii Rybchenko - Final Exam - Assignment 3

Introduction

The goal of this assignment is to create a machine learning model to perform two natural language processing binary classification tasks. We are given a historical dataset of a Belgian outlet with news articles. The first classifier should distinguish between whether the article is (a_1) about a domestic issue or (b_1) international. The second classifier should distinguish whether the topic of the article is politics (a_2) or something else (b_2). I will be using different models and a few approaches.

Theoretical and conceptual background

In this assignment we are mostly working within a text-as-data framework, where each article is a datapoint that has a label conveying some sort of class that this datapoint belongs to. Most of the documents are represented within an abstract vector space, where the syntax and word order are ignored (Imai, 2018). Linear classifiers 'learn' from the data to find proper boundaries of each class, assigning probabilities of the article to belong to a certain class, which we then validate on out-of-sample data (Imai, 2018). The main metric by which the model validity will be measured is the F1-score, which is a harmonic mean of precision and recall.

Selected methods and their use

For our data and two binary labels, we consider three approaches, which would be the most fitting for the task and are expected to yield high results. After cleaning and examining the dataset, we performed Exploratory Data Analysis to see how the labels differ from year-to-year and potential confounding.

The first approach is to fit some very simple machine learning models, using bag-of-words/TF-IDF and linear classifiers such as logistic regression and SVM. Such models are very computationally cheap and highly interpretable even in NLP tasks, which makes them a good reference point (Wang & Manning, 2012; Lin et al., 2023). Due to cheap computations, we also run hundreds of tests with different hyperparameters, implementing a grid search for the best fit. Simultaneously, some data engineering was considered and some of the runs were done with polynomial degree features engineered.

For the second approach we fine-tune transformer language models with what is called transfer learning, which has shown to improve political text classification, especially when domain cues are subtle (Tereschenko et al., 2020). The XLM-RoBERTa model was chosen due to the ability to process multilingual texts.

Finally, we run prompt-based classification with a large language model API (ChatGPT), providing context and requesting constrained JSON outputs. This follows work on LLMs as coding experts (Brown et al., 2020; Haseltine & Clemm von Hohenberg, 2023). Due to financial constraints, not a lot of tests were conducted, but all of the predictions were logged to ensure reproducible results. The approach is zero-shot, with two prompts of varying detail for each label.

These three methods give us a very broad field for investigation, yet each approach is considered very strong for text classification.

Data description

The dataset consists of 19339 articles, with each article having the following features:

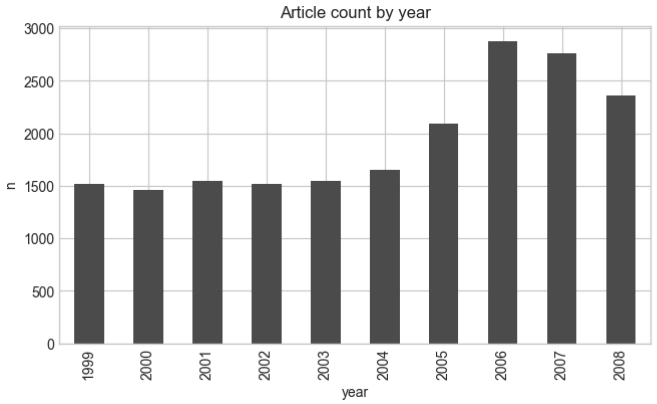
Id feature: identifier of the article. Description: the inner text of the article. Date: the article was published. Headline: the name of the article in the outlet. Domestic: binary classification label, which is set to 1, if the article concerns a domestic issue and not an international one. And finally the political label: classifies the article as of a political topic if set to one.

Feature	Description	NA-count
`id`	id of the article	11
`description`	the text of the article	11
`date`	date of publishing the article	11
`headline`	the headline of the article	243
`domestic`	binary true label, 1 if domestic	11
`political`	binary true label, 1 if political	20

`Table 1: Features in the dataset`

The outlet is fully non-English, and so is the inner text of each article (`description`) and the `headline` . The earliest observation is in 1999 and the latest is in 2008. Plotting the article count by year we can see that from 2005 the article count was rising, almost doubling in 2006.

The ratio of positive labels for domestic issue is more than 60% for all years and positive political labels are fluctuating around 45-50%.



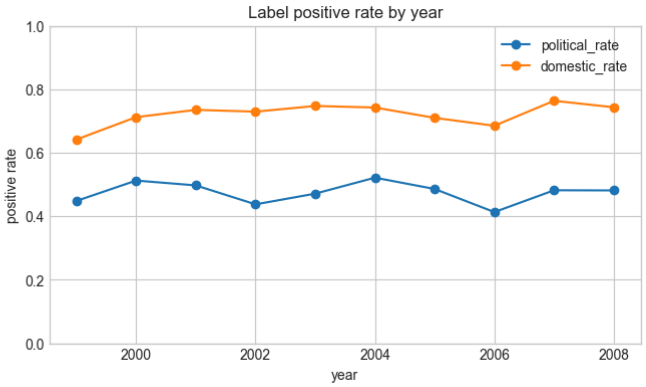


Figure 1: Article count by year.
Figure 2: Count of positive label per label by year.

Models using only description consistently outperformed models using any combination of the year, description and headline features.

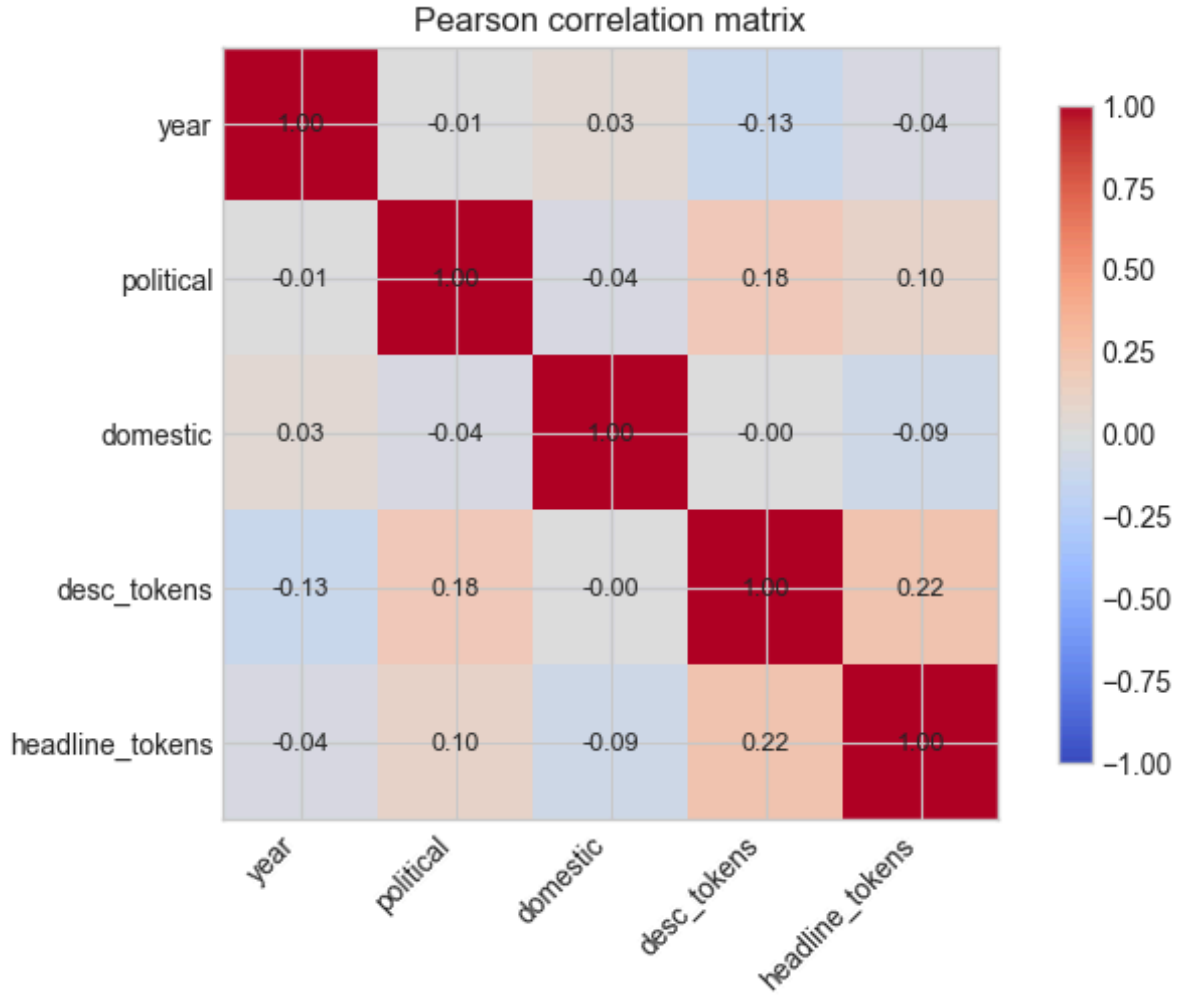


Figure 3 Correlation matrix amongst non-engineered features

Year feature was also not informative for predictions in more complex models. And the model including year, headline and description was worse than the pure description model.

Results and interpretation in the light of the assignment

Many different specifications and hyperparameters were tried, such as feature engineering, grid search for best parameters, different prompts for LLM, using different transformers. The table below shows the best model taken from each approach. We used an 80/20 split. Baselines were tuned with 5-fold cross-validation, then evaluated on a test set. Transformers were trained on the training set using a validation split for early stopping, and evaluated on the held-out test set. LLM prompts were run only on the test set.

Approach	Specification	F1-score	Label	Training/Inference time
Baseline (scikit-learn)	tfidf_log_reg	0.8242	political	~1500s
Baseline (scikit-learn)	tfidf_linear_svm	0.9447	domestic	~1400s
Transformer fine-tune	xlm-roberta-base	0.8578	political	~32400s
Transformer fine-tune	xlm-roberta-base	0.9589	domestic	~32400s
API LLM	gpt-5.1, p2	0.8055	political	~15800s
API LLM	gpt-5-mini, p2	0.8206	political	~15800s
API LLM	gpt-5.1, p2	0.8901	domestic	~15800s
API LLM	gpt-5-mini, p2	0.8795	domestic	~15800s

`Table 2: Performance comparison of different approaches.`

As we have seen from the data part, the domestic label is very dominant in the data, with most years having more than 60% of the articles being about domestic issues. More than that, when compared to political, it might be much easier on an abstract level to tell whether the news are domestic or international vs telling whether the news are political or not. Which is supported by the fact that F1 score drops significantly for the models that tried to classify the political label.

The best model in its basic configuration was the fine-tuned XLM-RoBERTa model that got 0.9589 F1 score on test data. Interestingly enough, a simple linear Support Vector Machine with TF-IDF got very close with an F1 of 0.9447. ChatGPT did pretty well, receiving 0.8901, but was worse than the best approach of other methods.

The ranking is very similar for the political label, with the fine-tuned transformer getting the best score of 0.8578, followed by a logistic regression (as contrary to SVM in the previous task) and finally the GPT model.

Discussion

Overall, fine-tuned transformers performed better than other models, while baseline linear models were extremely competitive and LLM prompting did relatively well, but did not outperform supervised approaches. Across both tasks, performance was consistently higher for the domestic labels.

Data-related issues

First of all, the main validity concern arises from **selection bias**, as we were mostly classifying the articles from only one outlet, the predictive power would be significantly lower for other countries and outlets.

A few concerning details might have affected the end-results and could be considered for a better model. First of all, the `political` label appeared to be very noisy, perhaps articles about economics, or anything relating to public services might have been misclassified as politics. Which is apparently not the case for the `domestic` label.

Thirdly, there is a great class imbalance, especially for the `domestic` label, with a lot of positive labels, which could also affect the F1 score to a certain degree, depending on whether the F1 is macro, micro or weighted.

Strengths and weaknesses

Each of the approaches had its own set of strengths and weaknesses.

TF-IDF + Logistic Regression/Linear SVM

- **Strengths:** Very easy to retrain, low compute time and easy to run hyperparameter grid search. More interpretable with top n-grams, easily reproducible. Very high performance for its cost of training and inference.
- **Weaknesses:** Not generalizable to other domains. Prone to labeling biases. Struggles with subtle political framing.

Fine-tuned XLM-RoBERTa

- **Strengths:** Best overall F1. Better at capturing subtle linguistic cues and context.
- **Weaknesses:** Much higher computational cost, and far longer training. High hyperparameter sensitivity. Reproducibility might be weaker, if seeds are not set.

Prompt-based API classification (ChatGPT)

- **Strengths:** Competitive performance with minimal training. Convenient if the data is poorly labeled. Very flexible with prompt. Low compute time for small n sizes. No training. Adaptable label definitions.
- **Weaknesses:** Very sensitive to prompt wording. Higher variance. Limited transparency.

Bibliography

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). *Language models are few-shot learners*. *Advances in Neural Information Processing Systems*, 33. arXiv:2005.14165.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747
- Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1), 1–10. doi:10.1177/20531680241236239
- Imai, K. (2018). *Quantitative social science: An introduction*. Princeton University Press.
- Lin, Y.-C., Chen, S.-A., Liu, J.-J., & Lin, C.-J. (2023). Linear classifier: An often-forgotten baseline for text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1876–1888). Association for Computational Linguistics. doi:10.18653/v1/2023.acl-short.160
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Terechshenko, Z., Linder, F., Padmakumar, V., Liu, F., Nagler, J., Tucker, J. A., & Bonneau, R. (2020, October 20). *A comparison of methods in political science text classification: Transfer learning language models for politics* (SSRN Working Paper No. 3724644). SSRN. doi:10.2139/ssrn.3724644
- Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 90–94). Association for Computational Linguistics. (Available via ACL Anthology, P12-2018.)