Results of our pilot to predict public procurement risks based on contract notices

Project purpose and general notes

The main aim of the pilot was to predict the riskiness of a procurement call based on the text of the notice. We also aim to give hints to those who are interested in replicating this attempt on similar datasets from other countries using different languages. The Pilot was conducted by Zoltán Varjú, Crow Intelligence.

Our investigations carried out on two different pathways. This first one aimed to detect linguistic clues associated with doctored calls (i.e. call for tenders which had one bidder). The second approach tried to use modern machine learning methods to predict various indicators associated with the calls. Based on our findings, we built a small Proof of Concept (POC) tool, which will help our further investigations.

Summary

Despite our best efforts, we couldn’t find any linguistic clues to identify problematic procurements solely based on their texts. However, modern machine learning methods seem to be promising if there is sufficient amount of high quality data available. Our POC laid the ground for further development and testing by providing an interface to check the quality and usability of the manually set indicators and their configurations.

Data

The project used two very similar datasets, one collected and curated by the redflags.eu portal, and one collected on the tenders.guru platform. Both datasets are based on official Hungarian procurement calls, but they are different in scope and coverage. To cover a longer time period (Hungarian data on tenders.guru are only available from april 2018 on - when the e-procurement platform EKR was launched), we used notices from the redflags.eu portal. Still for a next phase of the pilot, the data on tenders.guru will be valuable, as it contains not only notices but additional documents, such as contracts and evaluation sheets. However these have to be further processed prior to analysis.

Red Flags

As mentioned, our experiments were based on Hungarian procurement data collected by redflags.eu. The data contained 7214 procurements, with the following features:

- Id
- Title
- Short description
- Financial Ability
- Other particular conditions
- Particular Profession
- Personal Situation
- Technical Capacity
- CPV code

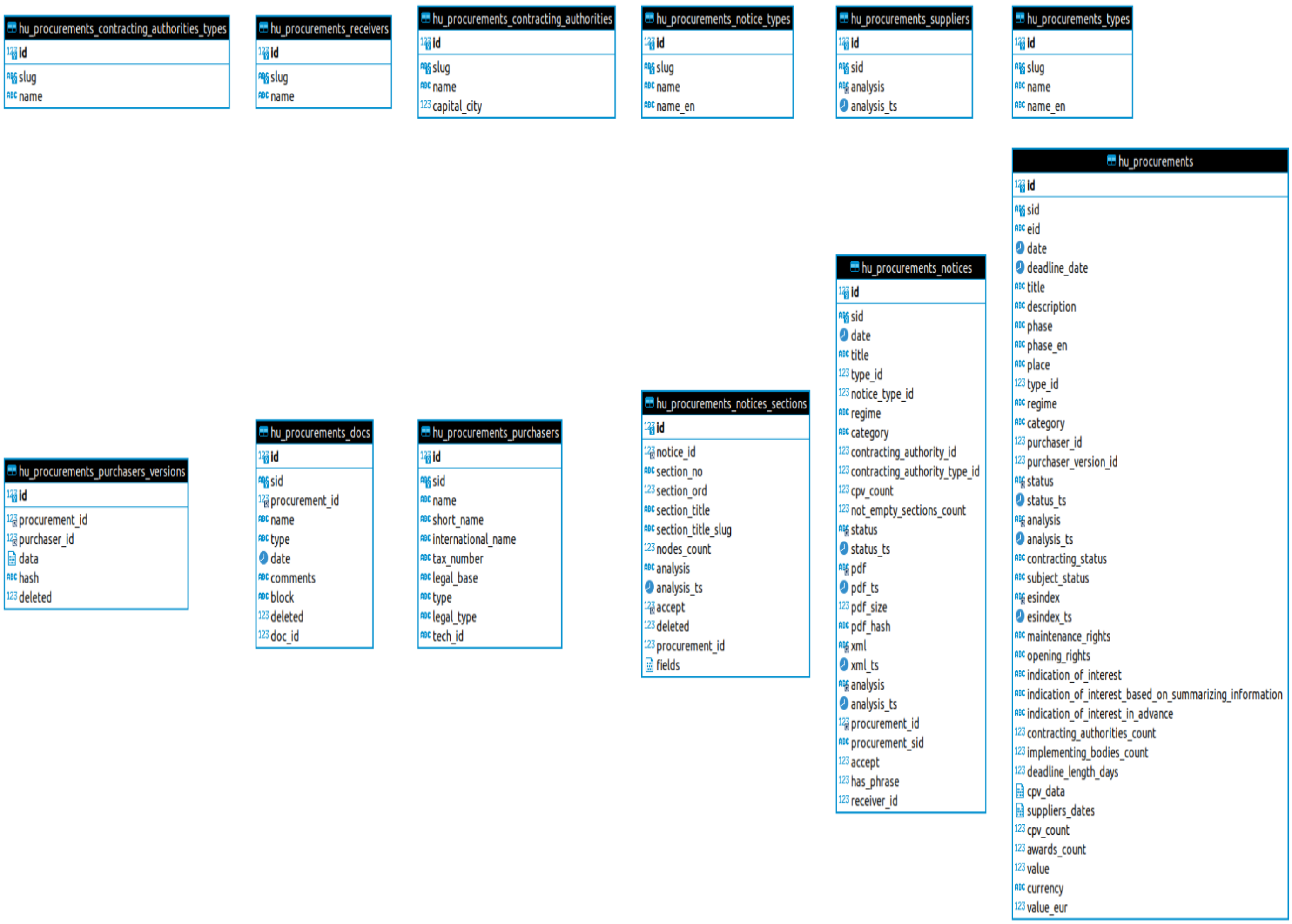
As part of the RedFlag project, the following indicators were identified as possible indicators of risk and added to the features (detailed description of indicators available at docs.redflags.eu):

AwCritLacksIndicator AwCritMethodMissingIndicator
AwCritPaymentDeadlineCondIndicator ContrDescCartellingIndicator
CountOfInvOpsNoCondIndicator DeadlineIsTightIndicator
DurationLongOrIndefiniteIndicator FinAbEquityCondIndicator
FinAbMissingMinCondIndicator FinAbRevenueCondExceedEstimValIndicator
FinAbRevenueCondManyYearsIndicator FwAgFewParticipantsIndicator
FwAgHighEstimatedValueIndicator FwAgLongDurationIndicator FwAgOneParticipantIndicator
HighEstimatedValueIndicator OfferGuaranteeIsHighIndicator
OpeningDateDiffersFromDeadlineIndicator PersSitMissingCondIndicator
ProcTypeAcceleratedIndicator ProcTypeNegotiatedNoJustificationIndicator
RenewalOfContractIndicator TechCapEURefCondIndicator
TechCapExpertsExpCondManyYearsIndicator TechCapGeoCondIndicator
TechCapMissingMinCondIndicator TechCapRefCondExceedEstimValIndicator
TechCapRefCondManyYearsIndicator TechCapSingleContractRefCondIndicator

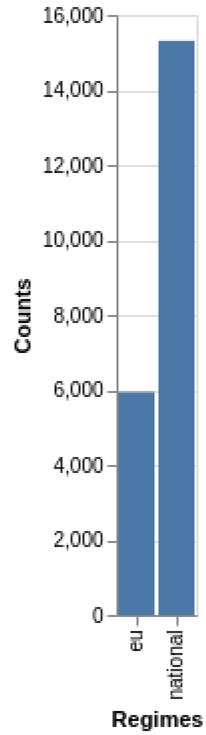
The indicators are applied to concluded procurements. These indicators are rules, which can be checked against the database. The database contains not only the text of a call for tender, but its entire history (e.g. bidders, results, modifications, winners, etc). An indicator has the following values; 0 if it doesn't match, 0.5 if certain parts of it matches and 1 in the case of full matching. We treat a call as risky if the sum of its indicator scores is above 0.5. We label a call "Bad" if it is risky, otherwise "Good".

tenders.guru

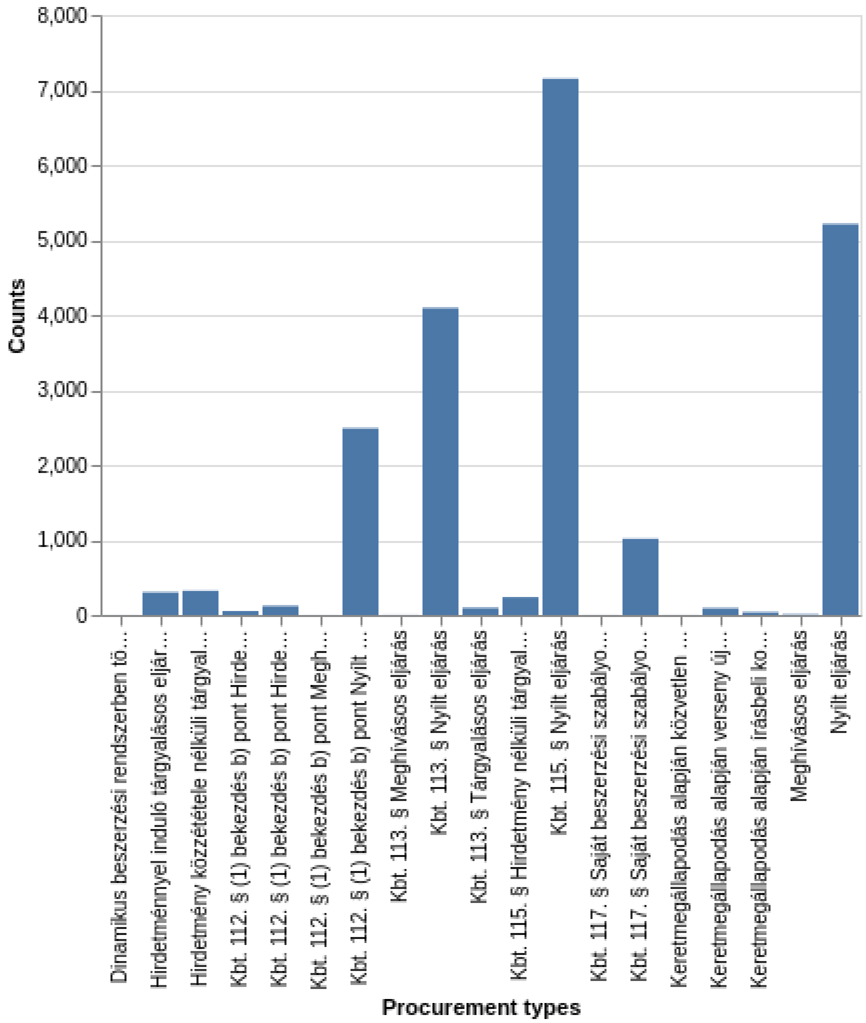
To highlight further possibilities for experiments of this kind, it is worth examining the scheme of the tenders.guru database:



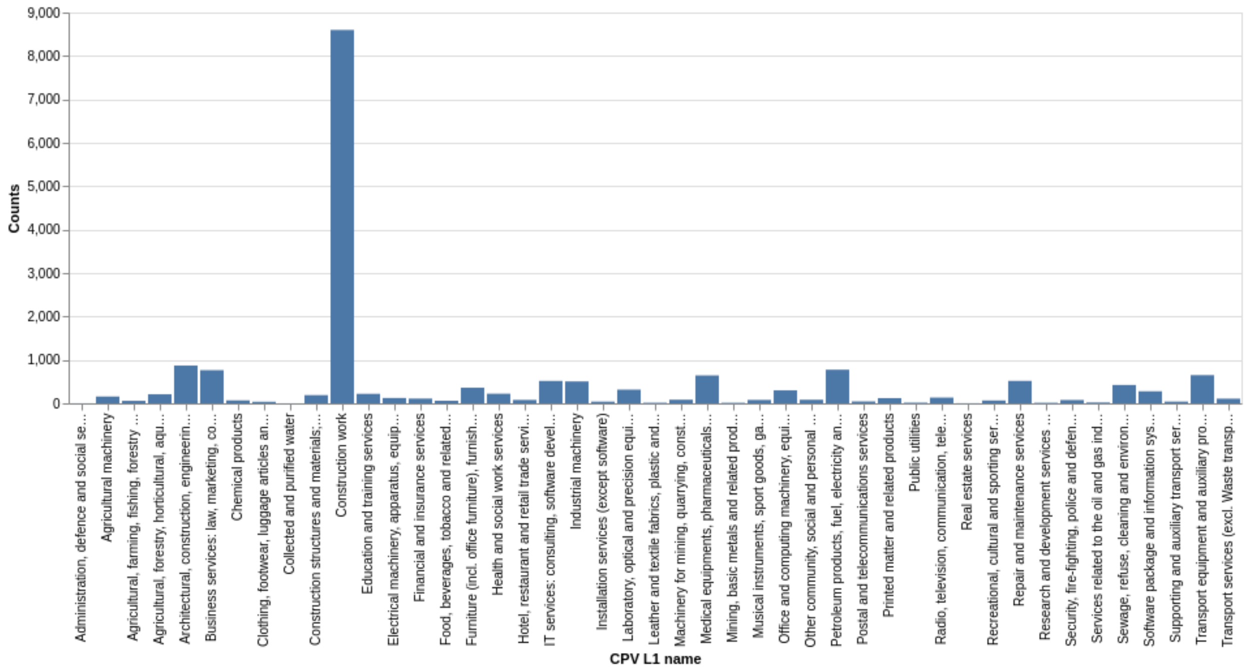
If we have a look at the regimes of the calls, we find the followings:



The number of notices included is inline with the official procurement statistics for this period, around ⅔ of tenders are conducted on the national regime, while ⅓ in the EU regime.



A high share of procurements were conducted in two popular procedure types within the national regime (open tenders according to 113§ and 115§ of the Public Procurement Act, that are only partially open and are based on invitations by contracting authorities that can be joined by further bidders). As these calls may differ significantly in their structure it is much more complicated to analyse them as regular open calls.



Regarding the CPV codes, we find that it is extremely imbalanced, since the overwhelming majority of the calls belong to construction works.

After the preliminary analysis of data from EKR, it became obvious that the data is very imbalanced and it must be processed with caution. For further processing, we created various subcorpora. Each subcorpus consists of at least one thousand entities, enough to divide the dataset into standard test, dev and devtest parts.

Preprocessing

We generated a document embedding for each call based on their textual columns (short description, financial ability, other particular conditions, particular profession, personal situation, technical capacity). Document embedding is a technique which helps to represent texts as a fixed size vector. Usually, these embeddings take how frequently words occur together in various documents, and they use some mechanism to deal with the fact that the length of the documents can highly vary. Columns were concatenated into a single string, then we used the sentence_transformers package and Hugging Face’s Hungarian language model to vectorize each sentence. Document vectors were computed as the mean of the sentence vectors.

Machine Learning Experiments

As mentioned, for the pilot, we finally used the Red Flag dataset based on a simplified, binary classification of the tenders. Our two classes are labeled Good (no indicator or only one indicator matches the call) and Bad (two or more indicators match the call).

Baseline experiments

First, we trained a Gradient Boosting classifier as a baseline model. The following evaluation report shows that it is esp. underperforms in the case of the Bad class.

	precision	recall	f1-score	support
Bad	0.25	0.30	0.27	407
Good	0.74	0.68	0.71	1174
accuracy			0.58	1581
macro avg	0.49	0.49	0.49	1581
weighted avg	0.61	0.58	0.59	1581

If we make subcorpora using the CPV codes, we can get better classification results. the classifier. This shows that there is room for improvement.

CPV 33

	precision	recall	f1-score	support
Bad	0.28	0.28	0.28	81
Good	0.66	0.66	0.66	172
accuracy			0.54	253
macro avg	0.47	0.47	0.47	253
weighted avg	0.54	0.54	0.54	253

CPV 71

	precision	recall	f1-score	support
Bad	0.32	0.14	0.19	50
Good	0.72	0.88	0.79	125
accuracy			0.67	175
macro avg	0.52	0.51	0.49	175
weighted avg	0.60	0.67	0.62	175

CPV 91

	precision	recall	f1-score	support
Bad	0.50	0.09	0.15	22
Good	0.83	0.98	0.90	102
accuracy			0.82	124
macro avg	0.67	0.54	0.53	124
weighted avg	0.77	0.82	0.77	124

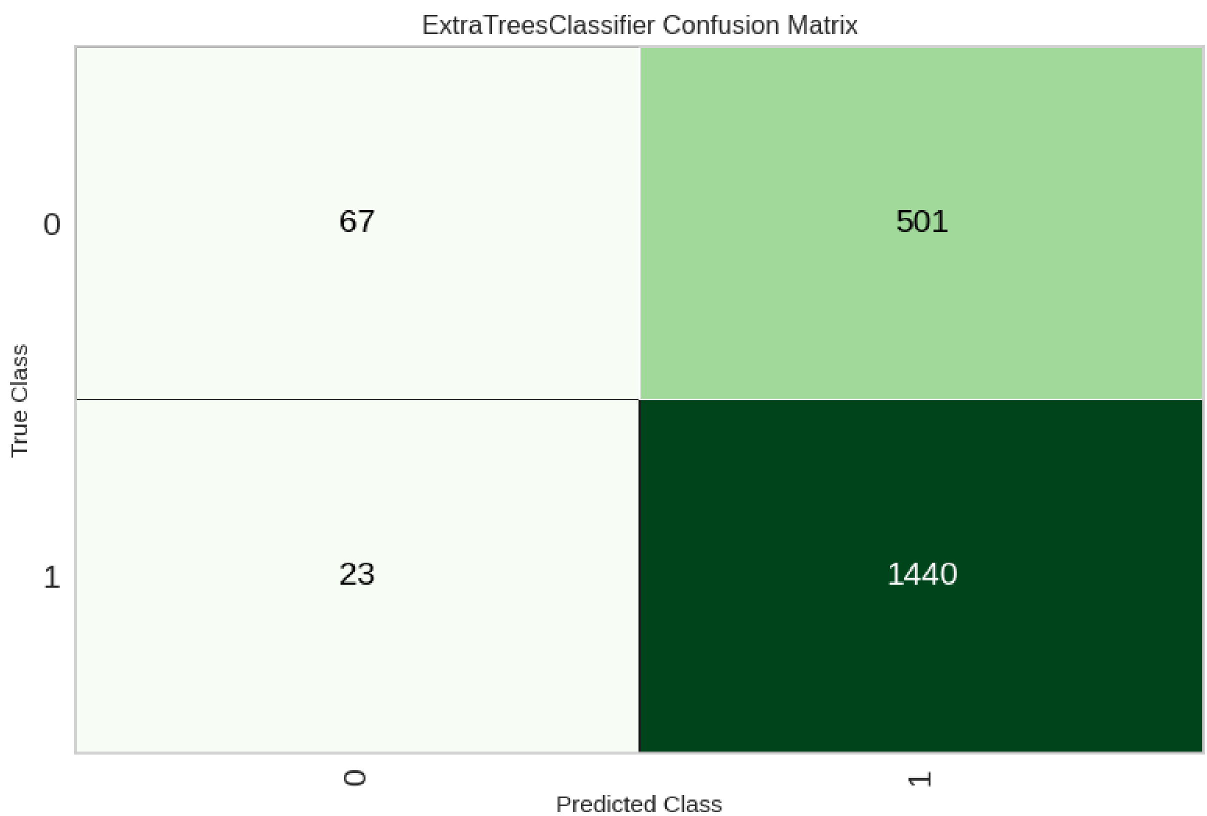
AutoML

AutoML is an umbrella term for machine learning methods which are trying to select and optimize the best algorithm in an automatic way. Usually, they train a handful of machine learning models, optimize their parameters, run standard evaluation techniques and then select the best performing model. This method is especially promising in the case of moderate sized datasets, because it requires only a little development tasks, and runs within a reasonable time on a mid-sized hardware. Since our dataset is ideal for AutoML tasks, we run model selection and parameter tuning using the pycaret package.

Our experiment found that the so-called Extra Trees Classifier is the best fit to the task. The classification report shows that its performance is much better than the baseline Logistic Regression classifier.

	precision	recall	f1-score	support
Bad	0.74	0.12	0.20	568
Good	0.74	0.98	0.85	1463
accuracy			0.74	2031
macro avg	0.74	0.55	0.52	2031
weighted avg	0.74	0.74	0.67	2031

The confusion matrix below shows that the most problematic class is the Bad (or 0) which is often predicted to be Good (1). The diagonal starting from the upper left corner shows the cases in which the classifier classifies the items correctly (i.e. “Bad” as “Bad” and “Good” as “Good”). The other diagonal (starting from the upper right corner) shows that the most common error is identifying texts in the “Bad” category as “Good”. Our investigation shows this type of error has been introduced by using too many indicators for creating the very artificial category of “Bad”. We are quite sure that using more carefully created indicators would dramatically lower this type of error.



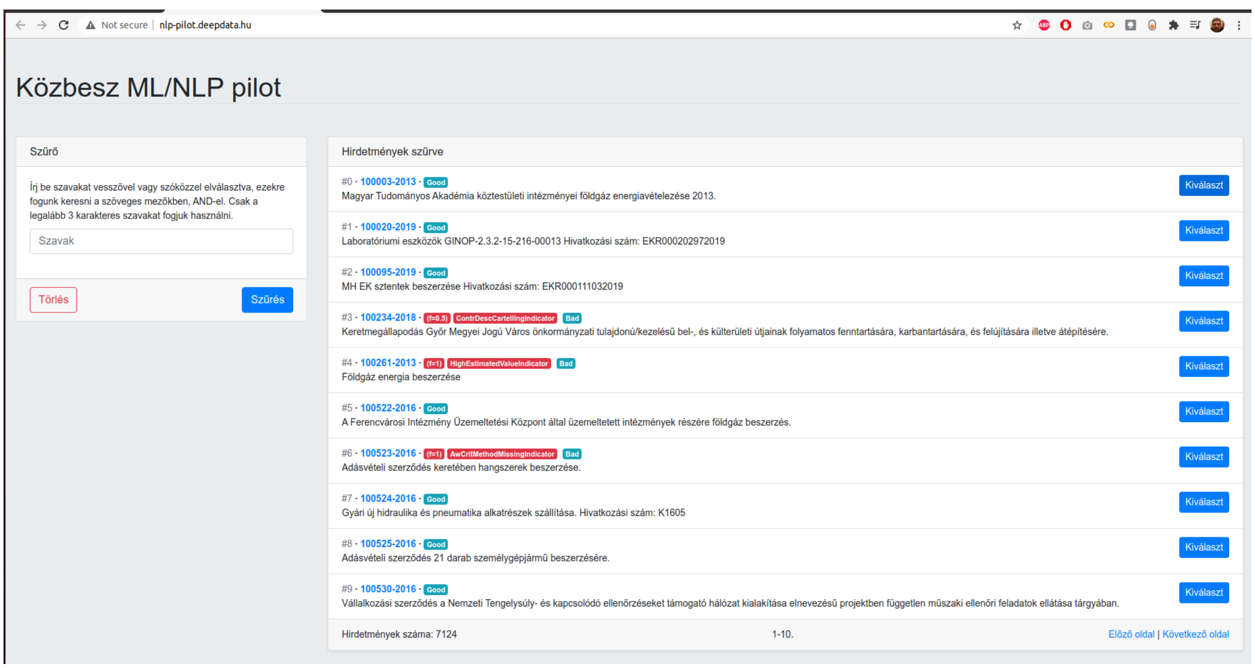
Corpus linguistic investigations

Using the above described model we tried to find linguistic clues to divide suspicious and non-suspicious calls. In a first experiment we created two sub-corpora; one consisting of calls with single bidders, the other containing calls won by firms associated with the ruling elite of

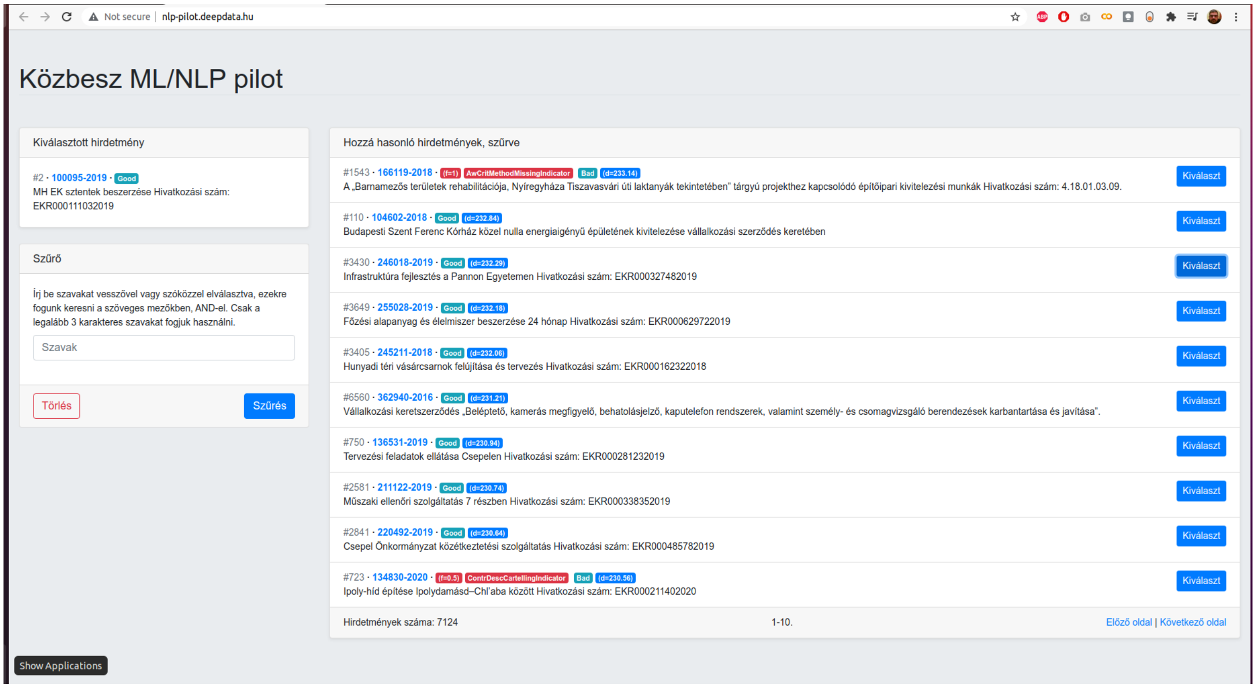
Hungary and its governing party (list in the appendix). We tried out various standard corpus comparison methods, e.g. log-likelihood keyness, TextRank keywords, etc. Despite our best efforts, we haven't found any significant difference between calls that we labeled suspicious and those we did not. This contradicts our hypothesis that there might be significant linguistic differences between such groups. However, methodological changes and a bigger volume of text (eg. involvement of procurement contracts, etc), might lead to different results.

Proof of Concept (POC) tool

Having carefully examined the result of the classifiers, we decided to test the consistency of the categories (aka indicators). If we can predict the indicators based on the text of the calls, linguistically similar calls must have similar indicators, or in our case they must be classified as "Good" or "Bad". Since we have already had the embedding based vectors of the calls, we can easily test our ideas. We can measure the similarity by the so-called Euclidean distance (something everybody learned about in high school, here we apply it to higher dimensions). At first, we designed a simple user interface that can be used as a search tool by typing search terms into the search box on the left side, or one can browse the calls using the panel on the right side.



By selecting a document, the system automatically sorts all documents according to their similarities to the selected one.



Using the interface, one can check if the labels are consistent (e.g. in our case they are mostly consistent, but except for the most similar one). The next step will be adding user management and feedback possibilities, so our procurement experts can easily check and correct the labels of the documents. Also, we can dramatically reduce the time needed to generate high quality and consistent training data for further investigations. (This is the so-called human-in-the-loop, active learning method). This is another path K-Monitor aims to further investigate in the future.

Conclusions

However our experiments were limited in scope, it can be drawn from them that using transformer based models to vectorize texts is a good starting point and AutoML solutions can find a model that dramatically outperforms baseline methods. Given the imbalanced nature of the data available, we are hopeful that by including a broader range of data into the analysis and further refining indicators, we can achieve better results in the next pilot.

Tools used

- Python https://www.python.org/
- Scikit-learn https://scikit-learn.org/stable/
- Transformers https://github.com/huggingface/transformers
- Imbalance-learn https://imbalanced-learn.org/stable/
- PyCaret https://pycaret.org/

- SentenceTransformers https://github.com/UKPLab/sentence-transformers

Indicative literature reviewed for the report

- Hastie et al.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, https://web.stanford.edu/~hastie/ElemStatLearn/ (2021.05.12.)
- Manning et al.: Introduction to Information Retrieval, https://nlp.stanford.edu/IR-book/information-retrieval-book.html (2021.05.12.)
- Robert Monarch, Human-in-the-Loop Machine learning, https://www.manning.com/books/human-in-the-loop-machine-learning (2021.05.12.https://www.manning.com/books/human-in-the-loop-machine-learninghttps://www.manning.com/books/human-in-the-loop-machine-learninghttps://www.manning.com/books/human-in-the-loop-machine-learninghttps://www.manning.com/books/human-in-the-loop-machine-learning)
- T. McEnery - R. Hardie: Corpus Linguistics, Cambridge University Press, 2015

Code repository

- The associated code is available at this link https://github.com/k-monitor/procurement-explorer
- The POC tool is available at this link http://nlp-pilot.deepdata.hu/

Appendix - The list of firms associated with the ruling party (provided by CRCB)

- 4iG Nyrt
- Aranykorona Zrt.
- Duna Aszfalt Kft.
- Elios Innovativ Energetikai Zrt.
- E-OS Energiakereskedo Kereskedelmi es Szolgaltato Zrt.
- ES Holding Zrt
- Euro Publicity Kft.
- Fejér-B.Á.L. Építő és Szolgáltató Zrt.
- Heti Válasz Kiadó Kft.
- Hódút Kft.
- Homlok Építő Zrt.
- Közgép Zrt.
- Közgéphídkorr Kft.
- Lounge Design Kft.
- Magyar Építő Zrt.

- Magyar Vakond Kft.
- MAHIR Cityposter Kft.
- MAHIR Kiállítás és Rendezvény Kft.
- Market Építő Zrt.
- Market Épületszervíz Kft.
- Mészáros és Mészáros Ipari, Kereskedelmi és Szolgáltató Kft.
- MET Magyarország Zrt.
- Mobil Adat Kft.
- NEMZET Lap- és Könyvkiadó Kft.
- Network 360 Reklámügynökség Kft (HG 360 Kft.)
- New Land Média Kft.
- PBE Energiamenedzsment Kft
- PBE Építő Kft.
- Publimont Kft.
- R-Kord Építőipari Kft.
- Sistrade Kft.
- Tief Terra Kft.
- Trinity International Communications kft.
- T-Systems Magyarország Zrt.
- Vakond kft.
- Vakond Via kft.
- Vasútvill Kft.
- V-Híd Zrt.
- Vivienvíz Kft
- West Hungária Bau Kft.
- Young & Partners kft.
- ZÁÉV Zrt.

This report was funded by the European Union’s Internal Security Fund — Police.