Krukovets, Dmytro

Article

Data science opportunities at central banks: overview

Visnyk Nacional'noho Banku Ukraïny

Provided in Cooperation with:
ZBW OAS

Reference: Krukovets, Dmytro (2020). Data science opportunities at central banks: overview. In: Visnyk Nacional'noho Banku Ukraïny (249), S. 13 - 24. https://journal.bank.gov.ua/download/article/2020/249/02/en. doi:10.26531/vnbu2020.249.02.

DATA SCIENCE OPPORTUNITIES AT CENTRAL BANKS: OVERVIEW

DMYTRO KRUKOVETS[a]

[a]National Bank of Ukraine
E-mail: dmytro.krukovets@bank.gov.ua

Abstract This paper reviews the main streams of Data Science algorithm usage at central banks and shows their rising popularity over time. It contains an overview of use cases for macroeconomic and financial forecasting, text analysis (newspapers, social networks, and various types of reports), and other techniques based on or connected to large amounts of data. The author also pays attention to the recent achievements of the National Bank of Ukraine in this area. This study contributes to the building of the vector for research the role of Data Science for central banking.

JEL Codes C45, C53, C82, E27, E37

Keywords Data Science, Machine Learning, Natural-Language Processing, macroeconomics, forecasting

1. INTRODUCTION

Data Science techniques are an innovative way to solve traditional central bank problems. It is a broad term that unites Machine Learning and Data Processing. The first one is a collection of tools that learn with the given data, and understand patterns and interactions between series and values. They can observe relations when people are unable to do so (because of a large amount of data, and connections complexity). The second expresses the possible set of actions with the data itself: collection, manipulation, preparation, and visualization. Data Science brings more ability to work with a non-linear relationship in the system, contrary to econometrics that concentrates on solving non-linearity bias problems in the linear form. Another difference is that econometrics concentrate on method robustness, while Machine Learning algorithms become popularized with their outstanding performance.[1] Among major drawback of advanced Data Science techniques is a lack of interpretability. Thus, it is not always possible to use these techniques because central bankers are often required to explain the results, being specialists in other fields (Kuhn and Johnson, 2013).

The new rise of Data Science started at the beginning of 2010 when high-quality models for image recognition were created, computational power grew enough, and people in many areas realized the full potential of such an approach. Thus, the paper will pay special attention to instruments that were rarely used in economics and the financial sector before.

New features and tools start infiltrating into research activity and routine processes at central banks. Advanced predictive models, based on high-frequency data, improve the forecasting performance of the current toolbox and could be a decent compliment or even a substitute for existing models. Forecasting is not the only possible application field for Data Science algorithms. Natural Language Processing (NLP) techniques are a part of Data Science for text analysis. They can support analysis of the text (news, social networks) to evaluate public reaction to central bank policy and actions, or in their explanatory research work.

The paper aims to elaborate on the use of Data Science techniques in central banks and to show possible use cases. The focus will be on the techniques that are used in research activity, modelling, and forecasting. This paper will not dig deeply into popular Big Data technologies: web scraping, as well as Supervisory and Regulatory Technologies (SupTech and RegTech). The reason is that they focus on technical implementation and computer science, rather than statistics (econometrics) and mathematics.

The article is structured as follows. First, we'll go through the motivation for Data Science tools and algorithm usage at central banks. Then it will be an overview of forecasting models and approaches within the Data Science framework, which is continued by papers about text analysis and a wide range of use cases for this technology by central banks. We need an overview of other supplementary techniques to complete the picture. Finally, there will be a list of achievements by NBU researchers in the area and a short

[1] "A very revealing example is the XGBoost algorithm, which owes its success to its domination over several Machine Learning competitions rather than its mathematical demonstration". Citation from the article, which is available here: https://towardsdatascience.com/from-econometrics-to-machine-learning-ee182f3a45d7

Available at https://doi.org/10.26531/vnbu2020.249.02

summary of the abovementioned findings. All Data Science techniques, which are used in the referred papers, have a short description in Appendix A.

2. OVERVIEW OF DATA SCIENCE IN CENTRAL BANKS

Central banks are interested in Data Science for several reasons. The first argument is its novelty and potential to give more precise results. Second, such models are good for micro-level, rich, and granular data. Moreover, Data Science gives a breath to innovative sources of information that have not been used much (such as texts) for a better approximation of people sentiments and their expectations from the economy, central bank actions, and related matters. Finally, more data contains not any less information; the question is how to squeeze it out.

There are already a few pioneers of Machine Learning implementation in central banking, which have given a nice start for researchers in the area. For example, the central banks of England, Canada, Poland, and Indonesia. However, an overwhelming number of banks exhibit no signs of using these techniques, which supports the claim that Data Science in central banking is in the embryo stage.

The majority of materials about Data Science in central banking are in an overview format. A solid example is the presentation of a Bank of England representative, Paul Robinson, in 2018. It is a discussion about policymaking issues such as bad measurements, too complex models and the imperfect theories behind them, and internal frictions. Broadly speaking, Big Data approaches and corresponding methods could assist in solving all these problems.

The presentation shows the possibilities that Machine Learning could provide as a complement to traditional models. As an example, the labour market was approximated and nowcasted with job ads and job search density data that was loaded via web-scraping tools. Despite these solid results, the statistics take into account only those users who are Google consumers, rather than those of Bing, Yahoo, or those who do not use the Internet at all. But even with this issue, this sub-sample offered adequate representation and could be adjusted using the basic aggregated statistics of Google users.

Unfortunately, with new approaches come new risks. First, abundant data does not necessarily mean lots of fresh information. Some events, such as extremely high inflation, liquidity traps, financial instability, and bank failures, are quite rare. Part of the reason is that central banks act in a manner to avoid them. No wonder information about these events is scarce, no matter how much data there is.

Second, Data Science models are mostly "black boxes" without much interior interpretability, which could be unacceptable for central banks, contrary to IT companies, the regular beneficiaries of such techniques. Finally, the more granular dataset increases the probability of confidential information popping-up, which imposes additional requirements at the security level.

A working paper from the Bank of England by Chakraborty and Joseph (2017) offers technical details and real use cases. The 90-page text provides an overview, in simple words and formulas, of data transformation, evaluation tools, and modern Machine Learning techniques: Naive Bayes, k-Nearest-Neighbours (k-NN), Neural Network, Support Vector Machine (SVM), K-means and others. The paper also discusses policy implementation and use cases.

The first example is a prediction of banking supervision results, a classical anomaly detection exercise. In this exercise, the model is trained to detect abnormal behaviour and capture outliers (anomalies). To build a robust model, the authors removed part of the data, which was used to create a target variable. Under these conditions, Random Forest turns out to be the best model after all evaluations (accuracy, precision, recall, F1-score).

The second case is a UK CPI forecasting, one of the most classical exercises for the majority of central banks. Data Science algorithms mostly outperform traditional econometric models and the best one is the combination of NN and SVM. However, the authors emphasize that these models are computationally expensive and require tons of data, which is not common enough in macroeconomics.

The last case in the article is about “unicorns” in financial technology. It tells a story about the top technology-driven firms that changed the rules for the whole sector. Examples are Uber (taxi driving), AirBnB (hotel industry), and Glovo (delivery business). Their activity changed the sector and the whole economy to some extent. Authors built a clustering model based on the CrunchBase[2] database and obtained a cluster, which consists mostly of "unicorns". However, even in this particular cluster, there are a lot of non-unicorns. Thus, the model is positive in understanding the necessary conditions for a firm to be successful, but not sufficient.

According to Per Nymand-Andersen (2017), European Central Bank (ECB) advisor, the "data service evolution" provides a great new field of possibilities. Central bankers should not miss it. Nowadays, the data stream could give almost a real-time snapshot of the economy, which might be used to develop short-run strategies, and adjust them for the long run. Financial institutions are practically forced to use more micro-level data to be competitive. The financial regulator might use this data as well to better understand the behaviour of financial agents and effectively perform supervisory tasks.

To summarize, Data Science is a viable complement to existing techniques. It gives new opportunities in forecasting, analysis, and processing of data. In the next chapters, we will dive into details about these algorithms and broaden our pool of use cases.

3. MACROECONOMIC FORECASTING AND MODELLING

Macroeconomic researchers have built many various econometric models throughout the decades. Models have grown into more sophisticated ones, able to deal with different biases. However, recent Machine Learning algorithms have become involved and taken the place of promising additions or alternatives. These new algorithms are more demanding in terms of data, but they have become less problematic in the current trends of world progress.

[2] Platforms with information about business and private companies collected a large and comprehensive dataset. https://data.crunchbase.com/docs/getting-started

We will start with inflation-forecasting, one of the most popular forecasting exercises in central banking, especially in these days of inflation-targeting as one of the main functions of the monetary policy regime.[3] Nakamura (2005) wrote the oldest paper in this review. In those times, neural networks were not a popular instrument, thus the number of papers was scarce. The author uses quarterly data from 1960 to 2003. His neural network is simple, with only two pairs of univariate equations, connected one-by-one. The method to find the best coefficients is significantly different from a modern one, taking a hundred random initial values and choosing between them, instead of a backpropagation optimization solution. Even with such a primitive approach, the neural network showed a better performance than the AutoRegressive (AR) benchmark on the forecasting horizon of one to four quarters ahead. This comes from the ability of neural networks to capture nonlinearity. The main conclusion is that neural networks might be a solid addition to the pool of forecasting models already in use.

The complexity of networks and methods builds up with the popularity of the field. Choudhary and Haider (2012) shows neural network performance in different countries' datasets and compares it to the AR(1). This paper has more sophisticated architectures than the previous one: two networks are called hybrid-network and dynamic-network (the latter is closer to the Recurrent Neural Network (RNN), but simpler), followed by two of their combinations. As a result, using the database with monthly inflation from 07.1991 to 06.2008 for 28 OECD countries, the neural network mostly outperforms the AR(1) model in short-term forecasting. The author claims that the continuous comparison of econometric and other models is a preferable strategy because of the results' instability. Thus, the building of a toolbox with a wide range of instruments fits the strategy well.

Nevertheless, model complexity is not the only difference. The world is on its way to the "Big Data Land", where the quality and amount of data increased. It influences the forecasting process and predictive models. Medeiros et al. (2018) used a very rich monthly-based dataset called FRED-MD, which contains hundreds of features to forecast U.S. inflation. The paper goes through some models, from benchmarks and traditional econometric to Data Science models. First, algorithms are barely able to capture non-linearities that ML models can, for example, the relationship between inflation and employment. Thus, Random Forest has worked best on most of the horizons, while Ridge/Lasso regressions have performed decently too. Most models produced a sub-product: a list of features that were selected as the most important for variance explanation for each horizon. The results among the models were quite different. Lasso regression produced output and prices as significant variables for inflation explanation. In its turn, Random Forest and Ridge regression have employment, prices, and interest rate. There is a lot of space for analysis and comparison of different model results. That is why it is always beneficial to widen the range of models in use, even when they employ the same variable set, in order to investigate the matter from different perspectives.

Jung et al. (2018) offers a comprehensive example of using a few Machine Learning techniques – namely Elastic Net, Super Learner and RNN – to forecast GDP growth in several representative countries. The main interest is to compare these models' performance with official WEO predictions that are based on more traditional models. Elastic Net and Super Learner have rapidly better accuracy rates (35 to 80 percent higher than the benchmark) for one quarter ahead. But on an annual basis, there was much less certainty (RNN was better for the U.S., U.K., and Germany, while WEO for Spain, Mexico, and Vietnam). These algorithms work well for short-run forecasting and might be useful in long-run cases.

Nowcasting is a technique to forecast the value of variables "in a moment", if they are published with a large time gap, such as GDP. This exercise is widely described in the literature, for example by Richardson et al. in 2019 (based on their paper in 2018, but more advanced and with more data used). The available toolkit is quite rich: Ridge/Lasso regression, Bayesian VAR, neural network, boosting algorithms, SVM, and k-NN. The dataset is rich too: domestic and international statistics, surveys, and financial data. Data frequency varies from daily to quarterly. This paper represents how much work can be done in the area of nowcasting of GDP. The methodology can be translated to other macroeconomic variables too.

One more example of GDP nowcasting is given by Bolhuis and Rayner, 2020. The dataset consists of hundreds of variables about the Turkish economy, some of which are represented several times with different transformations. There are classic unemployment or current account variables that are paired with different confidence indexes and survey-based variables. The method is a combination of several standard Machine Learning techniques: SVM, GBM, and RF. Authors claim that these models complement each other and a combination of their nowcasts leads to reduced error. There were different types of combinations: with equal weights and based on the relative RMSE of the single models. The result is the complete outperformance of ensembles to single methods and to the benchmark (traditional for the exercise of GDP nowcasting – dynamic factor model).

A few steps aside from such a classical application as inflation or macroeconomic forecasting leads to the various ways of Machine Learning usage. Gogas et al. (2014) explain the output and inflationary gap via different yield curves. The SVM model forecasts future deviations, which allows swift and appropriate policy responses. In the same spirit, research by Gogas et al. (2019) estimates an SVM model using a monthly dataset based on the Eurocoin Index and money aggregates (M1, M2, M3). It is made by Gogas et al. (2019). The authors show that money supply data can forecast euro-area economic activity. It means a significant dependency between economic activity and money supply. The hypothesis that monetary policy is effective cannot be rejected for the euro-area.

There is not only macroeconomics but also the financial side of central bank activity. Petropoulos et al. (2018) discussed tougher supervisory actions by regulators. There is a classical trade-off between less restrictive rules and financial safety. Thus, algorithms for analysis and prevention for different kinds of risk are in high demand at the moment. The authors produced a rich, combined semi-annual dataset with loan data from over 10 years and 354 time series of financial ratios and macroeconomic variables. A ratio between the number of features and observations leads to the "curse of dimensionality", so as to the poor

[3] According to the IMF Annual Report on Exchange Arrangements and Exchange Restrictions, 2018.

generalization. The solution is to reduce the number of features with the Boruta algorithm, which is based on the Random Forest, to extract the most important 65 variables. Then the authors used eXtreme Gradient Boosting (XGBoost) and Deep Neural Network (DNN), and compared them with Latent Dirichlet Analysis (LDA) and logit models. XGBoost showed the best performance (measured with AUROC) and gave a variable importance table. The most important were returns on equity, availability of working capital, and interest expense coverage. These results might be used in both risk-scoring and further analysis of the area.

Scoring in the financial sector is one of the most promising vectors for Data Science usage. There are several reasons for that, among them are a lot of high-frequency data, a complex structure of the data, alternative sources and types of data that might not be easily met in classical econometric models. Bazarbash (2019) presented a deep discussion about the pros and cons of ML algorithms in credit scoring, especially for emerging economies with weak financial institutions, that was accompanied by a general methodology overview. Authors highlighted several strengths of Data Science in credit scoring: small costs to analyze a small borrower with decent accuracy when it's unprofitable to hire a financial analyst; the ability to harden soft information, making it more quantitative; and capturing non-linearities and reducing the sharpness of information asymmetry.[4] Among the weaknesses, there are privacy and ethical issues, and classical data-driven models issues such as bad responses for structural breaks.

Debatable papers, to be useful, should be reinforced with some technical ones that produce a measurable result. Munkhdalai et al. (2019) have tested a set of Data Science models to build a credit-scoring system and compare it with judgmental credit scoring (FICO). In the process of system building, authors used an automated grid search to find hyper-parameters for DS models (brute force over several options for all algorithms), and distinct feature engineering algorithms (to reduce their amount and to avoid overfitting problems). The result of the system is, expectedly, better than for benchmarks.

Another approach to perform Data Science in the financial market is to support decision-making in the credit market. Arora et al. (2019) write about horizontal or vertical slicing during portfolio building. These strategies support correct liquidity management, which is important because it holds the risks of large redemptions that can destabilize the financial sector. One of the tools is the ability to quantify the impact of sales on the market state. The Random Forest method works well for this purpose, predicting the response relatively better than the more mainstream model.

The last paper for this chapter is a quite unusual forecasting exercise that was studied by Hatko in 2017. The paper deals with nonresponses in a survey for business firms. Such observations are usually ignored or accounted for with a dummy. Another approach is a prediction for the variable of interest that is based on all responses from other companies with similar characteristics and responses for other questions by this firm. The author divided the global problems into a few sub-exercises: unit non-response (when there is no response at all) and item non-response (when there are some unanswered questions). The first one was solved by generating a response probability with a combination of logistic regression and k-means clustering. The second part was about using GBM and XGBoost on the datasets with missed data and its imputation. The quality was evaluated using several mean-based methods (cross-entropy) and out-of-sample forecasting exercises.

Special attention is required by fully developed packages that use several tools described above. They have a simple user interface and give comprehensive solutions to particular problems that a central bank faces. The Mindbridge.ai project has an award as the best Machine Learning solution at the annual Central Banking FinTech RegTech Global Awards.[5] Their product analyzes information about agents and builds their risk score, which is a proxy for the probability of a particular one to be fraudulent. Moreover, they have developed a search engine that allows the regulator to find and compare with the rest of the industry, and to visualize different features of an economic agent. Hereinafter, they have become a part of the Bank of England Fintech Accelerator project and won a Central Banking Award in 2018 as the best innovation, which emphasizes once again the openness of the Bank of England to new technologies such as Machine Learning algorithms. This project supports the further digitalization of supervising processes to increase the quality and speed of the audit by pairing human and AI performance.

A majority of use cases might be done using traditional econometric models. However, Data Science approaches can provide additional accuracy. Most papers claim that new models are solid complements to the current toolbox. The full potential of Data Science will be revealed in the next chapter, in a field where there are no decent alternatives.

To finish this chapter, it's worth reviewing the discussed use cases of Data Science algorithms once more: 1) forecasting of important macroeconomic variables such as inflation, GDP, unemployment and others, possibly using additional datasets; 2) analyzing whether some variables could be predicted with another, meaning the influence of the second variable on the first one; 3) building different indices for decision-making; 4) finding alternatives to expert judgement in areas where classical econometrics models seem to be unable to capture and use all the information provided; and 5) filling in gaps and unobservables in the data.

4. TEXT ANALYSIS

With the rise of Data Science, central banks have become able to use alternative sources of information. As an example, the Monetary Policy Authority of Singapore screens the news to detect events that require further attention (alerts). David R. Hardoon describes this issue in his presentation at IFC 2018. Analysts can do the job of looking through loads of text and discarding irrelevant news. However, it is a rather routine job. Alternatively, machine algorithms can take over, do this exercise faster, and save analysts time.

Unpredictable human sentiments are one of the biggest sources of error in current macroeconomic models. A general way to address this issue is to use aggregated data and make the assumption that, on average, agents act

[4] More about hard and soft information in finances might be found in the paper by Liberti and Petersen, 2019.
[5] Generally described in the article on a Medium: https://medium.com/reciprocal-ventures/mindbridge-analytics-why-we-invested-9cdb2099ba

rationally and behavioural effects can be neglected. These effects might be also approximated with the data from news, social networks and other text-based sources. Such techniques that help to solve the problem will be the main focus of this chapter.

A field of Data Science called Natural Language Processing (NLP) studies algorithms that can "understand" the text instead of simply collecting statistics about it. It is a broad field with several stages and tools for data preparation and processing, which differ upon the issue that this model should solve. Bholat et al. (2015) overview fundamental techniques in this area, as well as the comprehensive motivation for their usage and why are they underestimated. In most of the papers that are presented in this chapter, the used techniques are the same as described in this article: dictionary building, LDA and other.

Text mining is not a new area of study, but previous attempts by central banks to automate working with a text have been not particularly successful. In the IFC 2018 presentation by Hansen, a history of these attempts is shown. In 2007-2011, tools were mostly about finding some particular words or combination of words to determine the emotional exteriors of articles. The critique of such an approach is as follows: an article with the phrase "Central bank has a poor performance," and another one with the phrase, "Many articles say that central bank has a poor performance, but, in fact, it's the opposite," have contrary meanings, but might be judged as similar. Modern literature offers better methodology due to the better computational efficiency and higher interest: it starts from advanced dictionary methods that use psychological insights and goes to the LDA and RNN techniques that can capture the context throughout the text. Real use cases include the impact of a BoE Inflation Report release text on bond prices and the relationship between Fed statements and Romer and Romer shocks.[6] There are decent results (76% accuracy), which allow for applying the model to get complementary information about the economy and its dynamics.

The news is a solid proxy of social reaction. Thus, some methods are called to construct time-series indicators that help explain important macroeconomic variables. In the paper by Nicholas Apergis and Ioannis Pragidis, 2019, the authors built an index based on news sentiments and used it for prediction of stock returns over a database with articles and word-based statistics. The classical for financial purposes EGARCH-X model shows increased performance with news-based indices.

Another case described the model that explains bond spreads with these news indices (on a local and global level) via panel data. It is written by Fulop and Kocsis in 2018. The main output is a model with news that shows a huge increase in the R-squared, compared to a simple macro-based model. It is one of the best examples when NLP techniques and news data give the main contribution, rather than being a simple compliment for an additional few percentage points of explained variance.

Indices for models are not the only purpose of news mining. Rybinski (2019) showed a model that analyzes 20 years of articles about the Narodowy Bank Polski in Rzeczpospolita, the main newspaper in Poland. Text mining helps estimate the link between hot topics in the economy (proxied by news scores) and the Monetary Policy Committee (called RPP in Polish) talks score. NBP is a key decision-maker in several areas (e.g. inflation or interest rates). But in other areas, it is not (e.g. public finance, fiscal sector). The model suggests that the former topics are highlighted by media strongly in the periods of MPC activity, while the latter are not.

"Words are the new numbers: A newsy coincident index of the business cycle" is the speaks-for-itself name for a paper written by Thorsrud in 2016. The core idea is an approximation for the U.S. business cycle using the LDA model. The author investigated business newspaper topics, their dynamics (a measure of the topic heat in the particular day), and quarterly GDP growth in the time-varying dynamic factor model.

On many occasions, central banks need to estimate agents' expectations about the economy to build a strategy. The usual way is based on surveys. However, they are quite expensive, especially when high quality is needed (robustness to many factors, size, and homogeneity of the sample, etc). The ability to estimate expectations via news is a good addition to the central bank toolbox. Zulen and Wibisono (2018) implement such a model, which predicts society expectations about policy rate changes based on the news (four categories: no information, no change, no hike or no cut), and compared the results with the Bloomberg Survey Index. The results were quite satisfactory (up to 84% accuracy for XGBoost in this classification exercise), which is great for such small costs and an advantage in speed. The benefits are crucial for nowcasting and a better understanding of the current economic state.

Monitoring news helps predict shocks, such as armed conflicts. Originally, despite the impressive predicting power of modern forecasting tools, there is a room for shocks that could not be detected earlier as long as they might have had an unusual or rare nature, no seasonal pattern, etc. Mueller and Rauh (2017) wrote a paper about the prediction of political violence, such as armed conflicts or civil wars. The authors took some country news, extracted topics and evaluated whether there would be a conflict inside the country via dynamics exploration, different Bayesian techniques, the Gibbs sampler, and many other tools. The results were remarkable: a model can predict with 70% probability – and only 20% false positives – that there is an armed conflict coming. These kinds of models might strongly improve the quality of structural economic model output and might be used for better scenario design.

Social network data is powerful, but a strongly underused source of the real-time micro-level information that central banks can use. Thus, it is an innovative field of study that might give even better results than news-based models. Currently, there are some unfinished projects in the field, for example, Angelico et al. (2018) criticized surveys for being available on a monthly or lower frequency level, which does not allow analyzing an immediate response for particular events. Thus, they propose to use filtered and prepared data from Twitter to build inflation expectations. The result is highly correlated with survey-based and market-based indexes, however, with real-time feature and low costs. It might complement traditional methods for beliefs estimation.

[6] Romer and Romer shocks (RR shocks) described in the corresponding paper in 2004.

Corea (2016) describes another use case for Twitter data to approximate investor sentiments on the stock market. For example, as of 2016, there were 88,000 tweets about Apple stocks. Their analysis might approximate public expectations, thus forecast behaviour. Unfortunately, the results were are quite mixed, which speaks to the necessity to use complex models and devote more attention to data preparation.

Social-network data is continuous. This feature offers the ability to capture sentiments that are not obtained with surveys regularly. Thus, it might be used not only for forecasting but for research too. Stiefel and Vives (2019) found a significant relationship between the index of intervention by ECB perception to bond spreads. Such data allow for working with sentiment dynamics and with rumours, which is nearly impossible with other instruments. To conclude, using social network data might be used both as a support for forecasting models and as an independent technique for distinct research.

At the end of the section, there are few papers about communication between different agents in the economy. There are government authorities, people who are living in the country and other institutions (international). The first paper, written by Fayad et al., (2020), gives a small insight into the analysis of communication between government authorities and the IMF during Article IV Consultations.[7] About 2,600 staff reports construct the dataset from 2000 to 2018, which allow for analyzing the dynamics too. In the beginning, several "relevant" paragraphs were extracted from each report and topics were assigned to them using a "dictionary method", which offers 89% accuracy (compared to the smaller train set, made by hands). Then, the state-of-the-art NLP solution BERT is applied to extract the sentiments from these paragraphs with an 81% accuracy and did relatively well when authorities agreed or disagreed with IMF advisers, though failing in the case when the answer was mixed. These IMF advisers might find these results helpful in order to improve their program and find the most efficient vectors to work with.

The analysis of internal communications (e.g. MPC discussions) at a central bank could contribute to many areas, including improvements in transparency, which is one of the features that central banks care a lot about. Recently, studies about the transparency effect were limited by using periodical dummies. Nowadays, it is possible to track the effect directly through the topic dynamics in corresponding periods (before, during, and after changes to the more transparent behaviour). A paper about this issue, written by Hansen et al. (2018), shows the proof of hypotheses about the positive discipline effect and negative conformity effect due to the transparency increase, parallel to the hypothesis about its structural changes. Based on the transcripts from the FOMC, the authors found that communication between members increased significantly (discussion of the same topics), as long as there was discipline (preparation before meetings, which increases informativeness) and conformity (avoiding the expression of the true views, which decrease informativeness). The main conclusion, despite the necessity of transparency, is that NLP techniques open a window for research in areas that had been mostly unavailable.

The last paper for the chapter is made by Cedervall and Jansson in 2018. It is strongly connected to the previous paper and serves as an example of topic dynamics analysis. The authors paid special attention to the "business value" of this exercise, such as creating a quick overview of the report via machine techniques. For businesses in the world of available data, those who win understand the data and sign first, rather than obtain completed reports with a delay. Thus, they will be happy to have this raw and pressed information. The conclusion is as follows: transparency depends not only on the quality of one communication stream but also on the diversity of these streams, while some might perceive the information in one or another way better. This instrument helps to build diversity with relatively low costs.

To conclude the chapter, text mining is a broad field of study, even when we are talking about news analysis, but not only about that. For the moment, central banks haven't taken advantage of much of the analysis of social networks to predict behaviour, making it a good area to pioneer research. Many use cases were described throughout the chapter: 1) digging for news to build an index that helps to predict different macroeconomic and financial series, such as bond spreads and stock returns; 2) designing an index of credibility, transparency and other social interactions, based on the news; 3) predicting the probability of shocks, that couldn't be predicted, but could be guessed by experts previously; 4) examining proxy survey results about expectations for different series, for example, inflation expectations; 5) researching about different effects built on communication; 6) easing the routine job process in some cases; and 7) exploring one of the best places to understand individual behaviour and moods – social networks.

5. OTHER ALGORITHMS

Data Science is not bounded by the above-mentioned algorithms. There are a lot of techniques at the junction with data analysis, statistics, and IT. Among the most widely used techniques for researchers is web-scraping. The tool takes real-time data directly from the websites. The well-known macroeconomic project that uses web-scraping is called "Billion Prices Projects", which scrapes prices from retail websites and uses this data instead (or as a complement to) the official price level data. There were many papers about this project, among them by Cavallo (2013). He used the project-produced data to challenge the credibility of the official inflation in Argentina. He aggregated series into components, similar to the official basket to represent some part of the goods basket inflation. The results between the scrapped and official prices might be different in this case due to the inconsistencies in methodology and other issues. But in this paper's case, the scraped inflation was twice as high, which is taken as evidence of manipulations with official data.

Google Trends is another web-based technique, which is close to web-scraping but uses the search data produced by users. It was launched in 2006, but the number of papers in the classical economics environment about its usage started to grow 5-10 years later. Per Nymand-Andersen in IFC 2018 showed a model for predicting car registration data, which comes with a lag in the official stream, with the data from Google Trends (or several searches related to the car buying). The hypothesis was proven that if people start to search for opportunities of buying a car, then many of them will do it soon.

[7] This is a series of consultations with an individual country representative, according to the Article IV in the IMF Article of Agreement. A short description might be found here: https://www.imf.org/external/about/econsurv.htm

Other techniques that are rather about “Big Data” are actively discussed by the central banking community. Building and maintaining a database is a crucial part of the implementation of Data Science algorithms, which corresponds to the Erwin Rijanto key opening speech at the IFC 2018 conference. Renaud Lacroix further discusses this question in his presentation about the project to build a multidisciplinary granular data platform, which took place at the same conference. Soramaki (2018) gives a good example of a widely popular Regtech and Suptech solution. He described the FNA Ltd. product for analyzing, monitoring, and visualizing transactions between companies and institutions, their interaction in financial terms. Several graphical approaches and statistics support the understanding of the network (financial system). A few use cases were about prices for housing right before, during, and after the Global Financial Crisis, fraud detection in money transfers and others. However, it is less about research, thus out of the paper’s scope.

The last example of the Big Data usage is less about central bank activity, but it is a very good example of the complex structure in economics that might be explained or approximated with Data Science techniques. Fan Liang et al. in 2018 described a project of Chinese authorities for scoring the trustworthiness of Chinese citizens and organizations. The model includes numerous factors: what people buy in shops, where they spend their time, whether they paid their loans in time, who are their friends and many others. This data supports building a score, which represents the probability of a credit loan return. There is a system with rewards and punishment for those who have high and low social credit scores, respectively. It is possible to obtain a lot of data about a single person and its interconnections to build this score, which represents a behaviour in a particular situation. It is the ultimate technology that might be used by central banks to understand public attitudes and expectations with overwhelming accuracy. However, questions about costs and ethics arise. The conclusion from this paper and this product as follows: most things, including such unstable ones like personal behaviour, might be approximated and forecasted with great accuracy. Everything depends upon the data.

In conclusion, there are lots of methods and tools that are associated with this topic and should be mentioned, but do not directly represent the paper’s objective.

6. NATIONAL BANK OF UKRAINE PROJECTS

Many DS-based projects are at different stages of completion at the National Bank of Ukraine. This section’s purpose is to highlight them.

A project for official monthly disaggregated inflation data aims to find distances between series and cluster them into several groups. This will challenge the rationality of the current division of headline inflation into four main categories, reveal the series with the highest exchange rate pass-through, and investigate relative prices between tradable and non-tradable goods.

A few small supplementary models for inflation forecasting are also in use. They are based on the Random Forest and GBM approaches (XGBoost in particular), fitted to forecast core inflation components. Further research on forecasting capabilities is still under development.

The attention of the NBU as an oversight authority, concerned about financial stability, is geared towards risks in the credit sector. Researchers investigate them in different contexts of credit risk, ranging from individual agent data to the aggregated indexes over the whole set of clients in a particular bank. The logistic regression serves as a main documented benchmark. One of the reasons is an explanatory power of this approach when there is a need to prohibit credit issuance and offer the reasoning. As for supplementary tools, we can recall a paper by Pokidin, 2015, that investigates whether an SVM is superior to the main benchmark and a few other models in terms of company-level credit risk. The answer is “slightly”. Another case is an XGBoost-based project for SME fraud detection.

A paper by Rashkovan and Pokidin in 2016 reviews banks clustering into a few types, according to their crediting business model. The authors investigate their density during the major changes in the banking sector during 2014-2016 and search to determine which of them were the riskiest. To do so, the authors used an algorithm from the expansions to Neural Networks. In the end, a risk indicator for the banking sector was constructed and it has been able to locate 92% of defaulted banks. Thus, it's an informative and useful product for regulatory authorities.

In the NBU, the question about estimating public sentiment is an open one. Despite numerous different surveys, there is a search for the more effective ways to approximate it from both the cost point of view and accuracy. So, a news analysis algorithm is on its way to completion. The project will consist of a model (and a few research papers based on it) to study the field of the media environment in Ukraine from different angles.

A web-scraping of prices for goods from online retail stores is a hot project with much research and models based on the data. One was already published in 2018 by Faryna, Talavera and Yukhymenko, where they built a model that aggregates these series into components of an inflation basket. This dataset covers 46% of the total official basket of goods determined by Ukrstat. The purpose of the paper was to investigate how well online prices correspond to official statistics and what are the drivers for inflation to be represented well by its scraped counterpart. This project is developed further to treat new problems such as series pairing, like in the Abe, Shinozaki, 2018 project. A similar technique is present in a few more projects about the labour market and densities inside it.

NBU researchers are doing well and are active in most important areas, with projects in different stages of completion. However, there is much more to be done.

7. CONCLUSION

The number of papers completed, even with consistent growth during recent years, leaves much to be desired. A quote from the review to one of the papers presented in the previous chapter is as follows: “The important point to note here being that the poverty of publications in this area is not because of some intrinsic limitations of ML's applicability to the domain, it's just that its early days". This phrase was written on January 11, 2018, and things have gotten better since those times. But there is still room for improvement.

As expected, the most effective and well-described Data Science models in central banking belong to the forecasting

area. As long as the quality and diversity of the data grow, there is an opportunity to improve the predicting toolbox with models that connect different disaggregated series in a complex, "black box-like" manner. The best bet for modelling units is to invest time in this area.

On the other hand, text analysis finds its use too. It is a rather difficult task, especially in countries where English is not the main language. This comes from the necessity to have a robust vocabulary. For developed economies, this area is interesting because it is a new way to further improvement of existing economic understanding, which better accounts for people's behaviour. Nevertheless, for emerging economies, the cost-to-benefit ratio is low and it might not be worth investing too much time here.

New sources of data (web-scraping, Google Trends) are viable and low-cost supplements to the existing toolbox, which might give more opportunities to use advanced Data Science techniques. A similar reason is related to improved communication with other central banks. Joint projects might improve data quality and diversity even further. Conditions for this are favourable, there are big conferences and frequent meet-ups in the area.

Hopefully, this paper will guide and encourage further research activity on the frontier between central banking and Data Science, offering a flavour of what are the most promising vectors.

REFERENCES

Abe, N., Shinozaki, K. (2018). Compilation of experimental price indexes using Big Data and Machine Learning: A comparative analysis and validity verification. Bank of Japan Working Paper Series, No. 18-E-13. Bank of Japan. Retrieved from https://www.boj.or.jp/en/research/wps_rev/wps_2018/data/wp18e13.pdf

Angelico, C., Marcucci, J., Miccoli, M., Quarta, F. (2018). Can we measure inflation expectations using Twitter? Harnessing Big Data & Machine Learning Technologies for Central Banks (Rome, March 26). Retrieved from https://www.bancaditalia.it/pubblicazioni/altri-atti-convegni/2018-bigdata/Miccoli_Presentazione_Twitter_Workshop.pdf

Apergis, N., Pragidis, I. (2019). Stock price reactions to wire news from the European Central Bank: evidence from changes in the sentiment tone and international market indexes. International Advances in Economic Research, 25, 91–112. https://doi.org/10.1007/s11294-019-09721-y

Arora, R., Fan, C., Leblanc, G. (2019). Liquidity management of Canadian corporate bond mutual funds: A Machine Learning approach. Staff Analytical Note, 2019-7. Bank of Canada. Retrieved from https://www.bankofcanada.ca/2019/02/staff-analytical-note-2019-7/

Bazarbash, M. (2019). FinTech in financial inclusion Machine Learning applications in assessing credit risk. IMF Working Papers, WP/19/109. International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2019/05/17/FinTech-in-Financial-Inclusion-Machine-Learning-Applications-in-Assessing-Credit-Risk-46883

Bholat, D., Hansen, S., Santos, P., Schonhardt-Bailey, C. (2015). Text mining for central banks. Centre for Central Banking Studies Publication. Bank of England. Retrieved from https://www.bankofengland.co.uk/-/media/boe/files/ccbs/resources/text-mining-for-central-banks.pdf

Bolhuis, M., Rayner, B. (2020). Deus ex machina? A framework for macro forecasting with Machine Learning. IMF Working Papers, WP/20/45. International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2020/02/28/Deus-ex-Machina-A-Framework-for-Macro-Forecasting-with-Machine-Learning-49094

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics, 60(2), 153-165. http://dx.doi.org/10.1016/j.jmoneco.2012.10.002

Cedervall, A., Jansson, D. (2018). Topic classification of Monetary Policy Minutes from the Swedish Central Bank. Examensarbete Inom Technology, Grundnivå, 15 Hp. Stockholm, Sverige. Retrieved from http://www.diva-portal.org/smash/get/diva2:1272108/FULLTEXT01.pdf

Chakraborty, C., Joseph, A. (2017). Machine Learning at central banks. Staff Working Paper, 674. Bank of England. Retrieved from https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2017/machine-learning-at-central-banks.pdf

Choudhary, A., Haider, A. (2012). Neural network models for inflation forecasting: an appraisal. Applied Economics, 44(20), 2631-2635. https://doi.org/10.1080/00036846.2011.566190

Corea, F. (2016). Can Twitter proxy the investors' sentiment? The case for the technology sector. Big Data Research, 4(C), 70–74. https://dl.acm.org/doi/10.5555/2991306.2991336

Faryna, O., Talavera, O., Yukhymenko, T. (2018). What drives the difference between online and official price indexes? Visnyk of the National Bank of Ukraine, 243, 21-32. https://doi.org/10.26531/vnbu2018.243.021

Fayad, G., Huang, C., Shibuya, Y., Zhao, P. (2020). How do member countries receive IMF policy advice: Results from a state-of-the-art sentiment index. IMF Working Papers, WP/20/7. International Monetary Fund. Retrieved from: https://www.imf.org/en/Publications/WP/Issues/2020/01/17/How-Do-Member-Countries-Receive-IMF-Policy-Advice-Results-from-a-State-of-the-art-Sentiment-48937

Fulop, A., Kocsis, Z. (2018). News-based indices on country fundamentals: do they help explain sovereign credit spread fluctuations. MNB Working Papers, 1. Magyar Nemzeti Bank. Retrieved from https://www.mnb.hu/letoltes/mnb-wp-2018-1-final-1.pdf

Gogas, P., Papadimitriou, T., Matthaiou, M., Chrysanthidou, E. (2014). Yield curve and recession forecasting in a Machine Learning framework. Computational Economics, 45, 635–645. https://doi.org/10.1007/s10614-014-9432-0

Gogas, P., Papadimitriou, T., Sofianos, E. (2019). Money neutrality, monetary aggregates, and machine learning. Algorithms, 12(7), 137. https://doi.org/10.3390/a12070137

Hansen, S. (2018). Measuring market and consumer sentiment and confidence. Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_21.pdf

Hansen, S., McMahon, M., Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. The Quarterly Journal of Economics, 133(2), 801–870. https://doi.org/10.1093/qje/qjx045

Hardoon, D. (2018). Exploring big data to sharpen financial sector risk assessment. Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_28.pdf

Hatko, S. (2017). The Bank of Canada 2015 retailer survey on the cost of payment methods: nonresponse. Technical Report, No. 107. Bank of Canada. Retrieved from: https://www.bankofcanada.ca/wp-content/uploads/2017/03/tr107.pdf

Jung, J., Patnam, M., Ter-Martirosyan, A. (2018). An algorithmic crystal ball: forecasts-based on Machine Learning. IMF Working Papers, WP/20/7. International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2018/11/01/An-Algorithmic-Crystal-Ball-Forecasts-based-on-Machine-Learning-46288

Kuhn, M., Johnson, K. (2013). Applied Predictive Modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lacroix, R. (2018). The Bank of France datalake. Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_26.pdf

Liang, F., Das, V., Kostyuk, N., Hussain, M. (2018). Constructing a data-driven society: China's Social Credit System as a state surveillance infrastructure. Policy & Internet, 10(4), 415-453. https://doi.org/10.1002/poi3.183

Liberti, J., Petersen, M. (2019). Information: hard and soft. The Review of Corporate Finance Studies, 8(1), 1–41. https://doi.org/10.1093/rcfs/cfy009

Medeiros, M., Vasconcelos, G., Veiga, A., Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of Machine Learning methods. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.2019.1637745

Mueller, H., Rauh, C. (2017). Reading between the lines: prediction of political violence using newspaper text. American Political Science Review, 112(2), 358-375. https://doi.org/10.1017/S0003055417000570

Munkhdalai, L., Munkhdalai, T., Namsrai, O., Yun Lee, J., Ho Ryu, K. (2019). An empirical comparison of Machine Learning methods on bank client credit assessments. Sustainability, 11(3), 699. https://doi.org/10.3390/su11030699

Nakamura, E. (2005). Inflation forecasting using a neural network. Economics Letters, 86(3), 373-378. https://doi.org/10.1016/j.econlet.2004.09.003

Nymand-Andersen, P. (2017). Big data in central banks: Central Banking focus report. Central Banking Publications. Retrieved from https://www.centralbanking.com/media/download/24906/download

Nymand-Andersen, P. (2018). Google econometrics: nowcasting euro area car sales and big data quality requirements. Statistics Paper Series, No. 30. European Central Bank. Retrieved from https://www.ecb.europa.eu/pub/pdf/scpsps/ecb.sps30.en.pdf

Petropoulos A., Siakoulis V., Stavroulakis E., Klamargias A. (2018). A robust Machine Learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting. Ninth IFC Conference on "Are post-crisis statistical initiatives completed?" (Basel, 30-31 August 2018). Retrieved from: https://www.bis.org/ifc/publ/ifcb49_49.pdf

Pokidin, D. (2015). National Bank of Ukraine econometric model for the assessment of banks' credit risk and support vector machine alternative. Visnyk of the National Bank of Ukraine, 234, 52-72. https://doi.org/10.26531/vnbu2015.234.052

Rashkovan, V., Pokidin, D. (2016). Ukrainian banks' business models clustering: application of Kohonen neural networks. Visnyk of the National Bank of Ukraine, 238, 13-38. https://doi.org/10.26531/vnbu2016.238.013

Richardson, A., Mulder, T., Vehbi, T. (2019). Nowcasting GDP using Machine Learning algorithms: A real-time assessment. Discussion Paper, 2019-03. Reserve Bank of New Zeland. Retrieved from https://www.rbnz.govt.nz/research-and-publications/discussion-papers/2019/dp2019-03

Rijanto, E. (2018). Opening remarks. International Seminar on Big Data "Building Pathways for Policy-Making with Big Data" (Bali, 26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_02.pdf

Robinson, P. (2018). Big data: new insights for economic policy – The Bank of England Experience. IFC – Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_12.pdf

Romer, C., Romer, D. (2004). A new measure of monetary shocks: Derivation and implications. NBER Working Paper Series, 9866. National Bureau Of Economic Research. Retrieved from https://www.nber.org/papers/w9866.pdf

Rybinski, K. (2019). A Machine Learning framework for automated analysis of central bank communication and media discourse. The case of Narodowy Bank Polski. Bank i Kredyt, 50(1), 1-20. Retrieved from http://bankikredyt.nbp.pl/content/2019/01/BIK_01_2019_01.pdf

Soramaki, K. (2018). Introduction to network science & visualization. IFC – Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_10.pdf

Stiefel, M., Vives, R. (2019). 'Whatever it Takes' to change belief: evidence from Twitter. AMSE Working Papers, Nr. 07. Aix-Marseille School of Economics. Retrieved from https://www.amse-aixmarseille.fr/sites/default/files/working_papers/wp_2019_-_nr_07_0.pdf

Thorsrud, L. (2016). Words are the new numbers: A newsy coincident index of business cycles. Journal of Business & Economic Statistics, 38(2), 393-409. https://doi.org/10.1080/07350015.2018.1506344

Zulen, A., Wibisono, O. (2018). Measuring stakeholders' expectations for the central bank's policy rate. Ninth IFC Conference on "Are post-crisis statistical initiatives completed?" (Basel, 30-31 August 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb49_50.pdf

APPENDIX

Short description of Data Science tools that are used in the papers

Elastic Net – method of regression regularization. It adds into the original OLS minimization problem a product of lambda with coefficient beta and another lambda with beta squared. It helps to penalize too high values of beta because high beta increases loss function value according to the magnitude of lambda.

$$\underbrace{\hat{\beta} \equiv \underset{\beta}{argmin}(\|y - X\beta\|^2)}_{Original\ min.problem} \Rightarrow \underbrace{\hat{\beta} \equiv \underset{\beta}{argmin}(\|y - X\beta\|^2 + \lambda_1\|\beta\| + \lambda_2\|\beta\|^2)}_{Elastic\ Net\ min.problem}$$

Lasso regression – a particular case of the Elastic Net where $\lambda_1 = \lambda$, $\lambda_2 = 0$.

Ridge regression – a particular case of the Elastic Net where $\lambda_1 = 0$, $\lambda_2 = \lambda$.

Bagging – an ensemble technique that divides the total dataset on subsamples uniformly and trains models on each of the subsamples. Then the resulting coefficients should be combined by some kind of averaging (different for the particular method and task to which bagging is applied).

Decision Trees – an algorithm that builds a tree (graph), in which every node is a "question" to the observation features. Answers to these questions leads to the leaf, which represents some value or class.

Random Forest (RF) – a combination of Bagging and Decision Trees, where trees are building for different subsets of features and then combined together.

Gradient Boosting Method (GBM) – an ensemble algorithm, which claims that the aggregated result from a few weak learners might present a solid solution. It trains weak models and adds them to the strong combination iteratively. During each iteration, data is reweighted and gives more weight to those that are badly predicted.

EXtreme Gradient Boosting (XGBoost) – an open-source library for a gradient boosting framework. It has become very popular and has several advantages over other libraries (LightGBM, CatBoost). However, it also has some drawbacks, leaving open the discussion of which is library is better.

Super Learner – algorithm, based on stacking, the third main ensemble technique. It has two stages: train many weak learners (not necessarily homogeneous, i.e. could be different techniques); train a meta-learner that uses outputs from these models to make a real prediction.

Clustering – a set of tools for grouping objects by their similarity.

K-Means – one of the most popular clustering algorithms. It randomly puts k points as cluster centres, then iteratively move them to the centroid of it and the nearest point. This approach minimizes in-cluster variance.

Support Vector Machine (SVM) – a model for a hyperplane construction to separate observations into several groups. It maximizes the distance from a hyperplane to the nearest observations from both sides (set margins).

k-Nearest Neighbours (k-NN) – a classification model. It assigns a class to the new point, according to how many points of this class are in k nearest points overall.

Naïve Bayes – it is a Bayes Theorem, used for the data with an assumption whose features contribute to the probability independently (that is why it is called naive).

Dimensionality reduction – a set of tools to reduce the number of dimensions (feature) without much loss of information.

Neural Network (NN) – this algorithm is composed of a few parts. There are nodes, layers, connections and activation functions. Nodes consist of some values and they form ordered layers. All nodes in the first layer are connected to all nodes in the second, second to the third, etc. The first layer has input data and the last gives an output. A node value is equal to the weighted sum of values in all nodes from the previous layer (connected nodes), transformed with the activation function (which produces, traditionally, a value between 0 and 1, like in the logistics regression). The goal of the algorithm is to calibrate weights in a way to minimize deviations between fitted and real output.

Deep Neural Network (DNN) – it is a NN, but with a much higher number of layers. In some cases, it improves the precision significantly, but the approaches of how to work with an algorithm are slightly different too.

Recurrent Neural Network (RNN) – it is a NN, but some nodes can be recurrent. It means that it takes not only values from the previous layer but its own value too (which corresponds to the definition of recurrence).

Latent Dirichlet Allocation (LDA) – an algorithm, used mostly in NLP and topic modelling. It allocates a set of topics to a set of documents and a set of words to reduce the number of direct connections, according to the Dirichlet distribution.

Bidirectional Encoder Representation from Transformers (BERT) – a state-of-the-art technique for different NLP tasks (at the moment of publishing). It uses transformers that reads the entire sequence of words, embeds them into vectors, replaces some words with a "mask" token (to improve contextual learning) and gives them to the neural network.