Data Science in Economics Education: Examples and Opportunities

Nina O. Rizun[1][a], Maryna V. Nehrey[2][b] and Nataliia P. Volkova[3][c]
[1]*Gdańsk University of Technology, 11/12 Gabriela Narutowicza, 80-233 Gdańsk, Poland*
[2]*National University of Life and Environmental Sciences of Ukraine, 15 Heroyiv Oborony Str., Kyiv, 03041, Ukraine*
[3]*Alfred Nobel University, 18 Naberezhna Sicheslavska Str., Dnipro, 49000, Ukraine*

Keywords: Data Science, High Education, Economics, Machine Learning, Topic Modelling, Natural Language Processing.

Abstract: Data science is the field of study that involves tools, algorithms, and knowledge of maths and statistics to discover knowledge from the raw data. Data science is developing fast and penetrating all spheres of life. More people understand the importance of the science of data and the need for implementation in everyday life. Data science is used in business for business analytics and production, in sales for offerings and, for sales forecasting, in marketing for customizing customers, and recommendations on purchasing, digital marketing, in banking and insurance for risk assessment, fraud detection, scoring, and in medicine for disease forecasting, process automation, and patient health monitoring, in tourism in the field of price analysis, flight safety, opinion mining, etc. This article concerns the issue of data science tools implementation, including the Text Mining and Natural Language Processing algorithms for increasing the value of Economics Education for the development of modern and technologically flexible society. The article deeply discusses the opportunities of using Text Analytics and Topic modeling for conducting scientific studies and applying them in the educational process. Presented examples demonstrate the nature of tasks and approaches which could develop students' research skills in the public perception analysis. Such approaches also allow students to gain practical experience in the study and interpretation of the influence of additional metadata, characterizing the comments authors, on differences in their opinions about events, companies, goods, and services. Finally, the Data science study programs for economics at top-20 universities are selected and discovered.

1 INTRODUCTION

2020 was a turning point for the whole world. COVID-19 and the resulting pandemic have identified weaknesses in society and opened up opportunities for development in many areas. The education sector has also felt the significant impact of the pandemic: the digitalization of the educational process, the transition to online learning and the abolition of educational activities – all this forces to seek effective solutions and adapt to new conditions. The field of economics has also undergone significant changes, accompanied by the digitalization of processes, the transition to remote work and changes in service and communication with customers (Soloviev et al., 2020b). The fast-growing world has become even more digital. Therefore, the skill is becoming increasingly popular use data correctly, model processes and make decisions using modern methods and technologies.

Data science is a field of study that includes tools, algorithms, and knowledge of mathematics and statistics to identify knowledge from raw data. Data science is evolving rapidly and is penetrating all walks of life. More people understand the importance of data science and the need to implement it in everyday life. Data science is used in business for business intelligence and manufacturing, for sales offerings and for sales forecasting, marketing for customer customization and procurement recommendations, digital marketing, banking and insurance for risk assessment, fraud detection, valuation and in medicine for forecasting diseases, process automation and monitoring of patients' health, in tourism in the field of price analysis, flight safety, etc. However, the application of data science in education has been relatively limited, and many opportunities for advancing industries have not yet been explored.

[a] https://orcid.org/0000-0002-4343-9713
[b] https://orcid.org/0000-0001-9243-1534
[c] https://orcid.org/0000-0003-1258-7251

Rizun, N., Nehrey, M. and Volkova, N.
Data Science in Economics Education: Examples and Opportunities.
DOI: 10.5220/0010926100003364
In *Proceedings of the 1st Symposium on Advances in Educational Technology (AET 2020) - Volume 1*, pages 550-564
ISBN: 978-989-758-558-6

Data science should be used in education to solve scientific problems, for example, in the study of behaviour in economics, in macro- and microeconomics, marketing, finance, agriculture, environmental and ecological economics and so on.

2 LITERATURE REVIEW

Data Science has a big list of tools: Linear Regression, Logistic Regression, Density Estimation, Confidence Interval, Test of Hypotheses, Pattern Recognition, Clustering, Supervised Learning, Time Series, Decision Trees, Monte-Carlo Simulation, Naive Bayes, Principal Component Analysis, Neural Networks, k-means, Recommendation Engine, Collaborative Filtering, Association Rules, Scoring Engine, Segmentation, Predictive Modeling, Graphs, Deep Learning, Game Theory, Arbitrage, Cross-Validation, Model Fitting, etc. Some of these tools were used in the next researches.

Teaching data science, for example, were introduced in (Brunner and Kim, 2016), Big data and Data Science methods presented in (Chen et al., 2012; George et al., 2016; Shoro and Soomro, 2015; Xiong et al., 2017; Cao, 2017; Ignatyuk et al., 2020), machine learning used in (Parish and Duraisamy, 2016; Derbentsev et al., 2020; Guryanova et al., 2020b; Babenko et al., 2021; Nosratabadi et al., 2020; Zelinska, 2020), Monte Carlo method presented in (Balabay and Chernonog, 2007; Patriarca et al., 2017), Artificial Intelligence presented in (Rizun and Shmelova, 2017). Data Science is fast developing. A large volume of information that grows with each passing year makes it possible to build high-precision models that simplify and partially automate the decision-making process. Models are being developed that implement the key data science algorithms for different areas of economics: financial Data Science (Bielinskyi et al., 2021; Brooks et al., 2019; De Prado, 2018; Danylchuk et al., 2019; Soloviev and Belinskiy, 2019; Soloviev et al., 2020a; Guryanova et al., 2020a; Kuzmenko et al., 2020; Klymenko et al., 2019), for institutional economics – (Prüfer and Prüfer, 2018; Hrabovskyi et al., 2020; Ilchuk et al., 2019; Oliskevych et al., 2018; Shi et al., 2020; Matviychuk et al., 2019), for agriculture – (Kaminskyi et al., 2020; Nehrey et al., 2019; Voronenko et al., 2020), for taxation – (Ausloos et al., 2017), and labor market – (Oliskevych and Lukianenko, 2019).

Data Science developing for education discussed in (National Academies of Sciences, Engineering, and Medicine et al., 2018; Volkova et al., 2019; Perevozova et al., 2020; Dimitrov et al., 2019).

3 DATA SCIENCE: PRINCIPLES AND TOOLS

Data Science in education is a multidisciplinary approach to technologies, processes, and systems for extract knowledge, understanding of data, and supports decision-making under uncertainty. Data science deals with mathematics, statistics, statistical modeling, signal processing, computer science & programming, database technologies, data modeling, machine learning, natural language processing, predictive analytics, visualization, etc. Data Science in education has two aspects of the application: (i) the management and processing of data and (ii) analytical methods for analysis and modeling, and includes nine main steps (figure 1). The first aspect includes data systems and their preparation, including databases facilities, data cleansing, engineering, visualization, monitoring, and reporting. The second aspect includes data analytics data mining, machine learning, text analytics, probability theory, optimization, and visualization. The basis of the learning process is the availability of relevant data that is of sufficient quality, appropriately organized for the task. Primary data often requires pre-processing. First of all, it is necessary to investigate the availability of the necessary data and how they can be obtained. The data search ends with the creation of a data set in which data coexistence is to be provided. Data science has a wide range of tools for data evaluation and preparation, in particular for data mining, data manipulation (value conversion, data aggregation and reordering, table aggregation, breakdown or merge of values, etc.) and validation of data (checking format, ranges of test values and search in legal values tables). The problem of missing values is solved by using different analytical methods: simulation, inserting default values, statistical simulation. Data science provides broad opportunities for text analytics. In addition, the use of data science tools facilitates work with big data. The main approaches in Data Science are Supervised learning models and Unsupervised learning models.

3.1 Supervised Learning Models

Supervised learning is one of the methods of machine learning, in which the model learns on the basis of labeled data. Using Supervised learning is possible to decide on two types of tasks: regression and classification. The main difference between them is the type of variance that is predicted by the corresponding al-

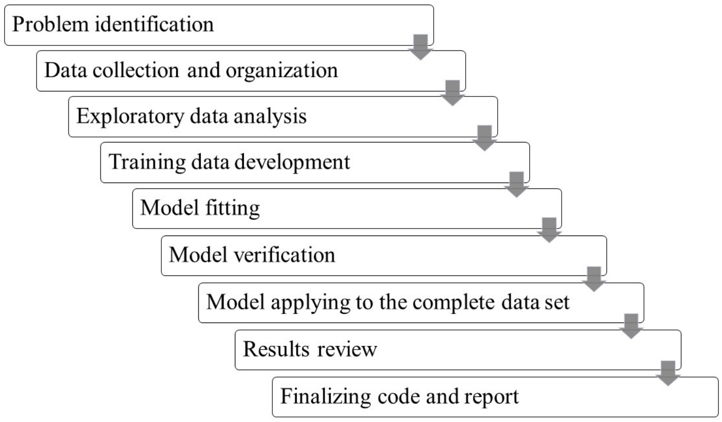

Figure 1: Data Science process.

gorithm. In regression training, it is a continuous variable, in the classification, it is a categorical variable. To solve these problems, many algorithms have been developed. One of the most common is a linear and logistic regression, a decision tree.

Linear regression. Regression analysis can be considered as the basis of statistical research. This approach involves a wide range of algorithms for forecasting a dependent variable using one or more factors (independent variables). The advantage of applying such an approach to modeling is the simplicity and clarity of the results, the speed of learning, and the release of the forecast. The disadvantage is not always sufficiently high precision (since in economics and finances, the linear relationship between changes is rare).

Logistic regression is used when it is necessary to predict the release of a binary variable using a dataset of continuous or categorical variables. Situations, where the parent variable has more than 2 possible values, can be simulated by a one-vs-all approach when constructing a logistic classifier for a possible output, or one-vs-one when constructing logistic classifiers for each possible combination of categories of the original variable. The dependence between the independent and the logarithmic variable in logistic regression is linear, the only difference with linear regression is sigmoidal functions, which converts a linear result in the probability of belonging to a class within [0; 1]. The advantages and disadvantages of logistic regression are due to the advantages and disadvantages of linear regression. This is the speed of the algorithm and the possible interpretation of the results, on the one hand, and a little accuracy – on the other. Logistic regression is often used to construct vote-counting models. An important factor in this is the interpretation of its results. The influence of each factor is clearly expressed by the magnitude of the coefficient b, which allows it to be clearly defined which of them positively and to what extent influence the decision.

A **decision tree** is an approach to both regression and classification. It is widely used in intelligent data analysis. The decision tree consists of “nodes” and “branches”. The tree nodes have attributes that are used to make decisions. In order to make a decision, it is needed to go down to the bottom of the decision tree. The sequence of attributes in a tree, as well as the values that divide the leaves into branches, depends on such parameters as the amount of information or entropy that the attribute adds to the prediction variable. The advantages of decision trees are the simplicity of interpretation, greater accuracy in decision-making simulation compared with regression models, the simplicity of visualization, natural modeling of categorical variables (in regression models it is needed to be coded by artificial variables). However, the decision trees have one significant drawback – low predictive accuracy (James et al., 2013).

3.2 Unsupervised Learning

Unsupervised learning describes a more complex situation in which, for each observation $i = 1, ..., n$, observation of the measurement vector x_i, but without any variables in the output y_i. In such data, the construction of linear or logistic regression models is impossible, since there are no predictive variables. In such a situation, a so-called “blind” analysis is conducted. Such a task belongs to the class of tasks of unsupervised learning, due to the absence of an output variable that guided the analysis. Unsupervised learning algorithms can be divided into algorithms for space reduction and clustering algorithms. The main task of clustering is to find patterns in the data that allow you to divide the data into groups and then in a certain way analyze them and give them an interpretation.

K-means is one of the most popular clustering algorithms, whose main task is to divide *n* observations into *k* clusters. The minimum sum of squares is the distance of each observation to the center of the corresponding cluster. This algorithm is iterative, at each step the cluster centers are re-indexed and redistributed observation between them until a stable result is achieved. The benefits of such an algorithm of clustering are the simplicity, speed, and the ability to process large amounts of data. But the user must specify the number of clusters he wants to use for clustering before computing; the instability of the result (it depends on the initial separation of points between the clusters).

Hierarchical clustering is an alternative approach to clustering, which does not require a preliminary determination of the number of clusters. Moreover, the hierarchical clustering ensures the stability of the result and gives the output an attractive visualization based on the tree-like structure of observations/clusters – dendrogram. This clustering algorithm uses different distance metrics and cluster agglomeration cluster criteria, which makes it very flexible to the data on which clustering is performed. However, the disadvantage of hierarchical clustering is the need to calculate the matrices of the distance between observations before agglomeration, which complicates the application of this algorithm for large data and data with many dimensions.

Time series analysis. A time series is built by observations that have been collected with a fixed interval. It could be daily demand, or monthly profit growth rates, number of flights, etc. The time series analysis takes an important part in the analysis of data that covers the region, from the analysis of exchange rates to sales forecasting (Nehrey and Hnot, 2017; Voronenko et al., 2021). One of the tasks of time series analysis is the allocation of trend and seasonal components and the construction of the forecast. There are many algorithms that have been developed, and we consider models such as ARIMA and Prophet.

The **ARIMA** algorithm is one of the most common algorithms for forecasting time series. The basic idea is to use the previous time series values to predict the future. This can use any number of lags, which makes such an approach difficult in setting because it is necessary to select the parameter so as to minimize the error and not override the model. ARIMA is often used for short-term forecasting. A disadvantage is the complexity of learning a model in many seasonal conditions.

Algorithm Prophet was developed by Facebook at the beginning of 2017 for forecasting based on time series (Nehrey and Hnot, 2017). It is based on an additive model in which nonlinear trends are of annual and weekly seasonality. This approach also allows to model holidays and weekends, thereby allowing to predict residuals in a time series. Also, the Prophet is insensitive to missed values, the bias in the trend, and significant residuals, which is an important advantage over ARIMA. Another advantage is the rather high speed of training, as well as the ability to use large-scale time series.

4 TOPIC MODELING IN DATA SCIENCE

Under the notion of texts mining in natural language we understand the application of methods of texts computer analysis and presentation in order to achieve the quality, which corresponds to the "manual" processing for further usage in various tasks and applications. One of the actual tasks of automatic texts mining is topic modelling.

4.1 Latent Dirichlet Allocation

Topic modelling is a statistical approach to extract the hidden semantics that occurs in a collection of documents or reviews. *Latent Dirichlet Allocation (LDA)* model proposed by (Blei et al., 2003) is one of the most notable approach for unsupervised topic modeling, which assumes documents and the words within them are derived from a "generative probabilistic model". Within the class of unsupervised statistical topic models, themes are defined as distributions over a vocabulary of words that represent semantically interpretable "topic" (Roberts et al., 2014). 'Meaning' of those topics (usually, in the form of topic Label and topic Description) is an emergent quality of the relationship between words (Robinson, 2019; DiMaggio et al., 2013). The task of topic meaning recognizing is often fraught with difficulty and requires the application of a triangular approach to its implementation, namely: (i) a literature review of existing topics found in the analyzed problem domain; (ii) independent work of experts on assigning labels to topics; (iii) conducting joint expert discussions in order to compare and revise the obtained labelling results.

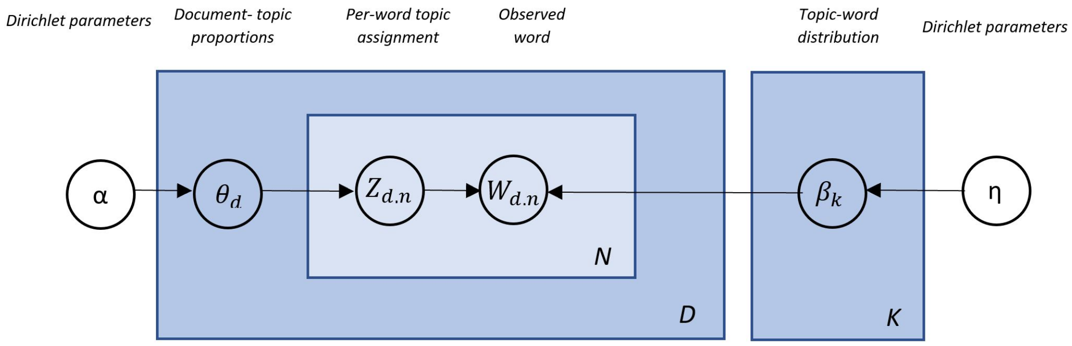

Figure 2: Latent Dirichlet allocation model (Blei, 2012).

As for main assumption of LDA method, there are the following (Roberts et al., 2016): (i) document is represented as a mixture of topics; (ii) each topic are present in many documents; (iii) each word within a given document belonging to exactly one topic; (iv) each document can be represented as a vector of proportions that denote what fraction of the words belong to each topic.

The basic LDA model is shown in figure 2.

Figure 2 serves as a visual explanation of the model and could be described as follows: (i) we have *D* documents and *K* topics; (ii) each topic presented by β_k words distribution over the vocabulary within the topic *k*; (iii) each document is presented by θ_d topic proportions within the document, where $\theta_{d,k}$ is the topic proportion for topic *k* in document *d*. Finally, we have (iv) for each n^{th} word in the document *d* – topic assignments $z_{d,n}$ (depends on the per-document topic proportions θ_d) and (v) for each d^{th} document – observed words $w_{d,n}$ which is an element from the fixed vocabulary (depends on the topic assignment $z_{d,n}$ and all of the topics $\beta_{1:k}$) (Blei, 2012).

In is obviously that data scientist in cooperation with other science domains increasingly seek ways to apply NLP and especially LDA topic modelling techniques to extract, organize, recognize, label and classify customers opinions and experiences (Kobayashi et al., 2018a). Next examples demonstrate the possibilities to sol the apply LDA topic modelling for solving: (i) human resources management, (ii) service quality assessment, (iii) research & development policy coordination tasks and (iv) strategic planning in universities.

Kobayashi et al. (Kobayashi et al., 2018b) used topic modelling to summarize the worker attributes and find worker attribute constructs and use these to cluster jobs. 140 main topics were identified, and such skills, as, for example, interpersonal communication (vocabulary of words: communication, written, oral, verbal, interpersonal, presentation, effective, listening); analytical and problem-solving (vocabulary of words: problem, solving, analytical, solver, troubleshooting, approach, abilities, capabilities); data analytical skills (vocabulary of words: data, Analysis, quantitative, research, statistics, economics, statistical, modeling); willingness to travel and the ability to operate on a flexible work schedule (vocabulary of words: travel, willingness, willing, work, time, needed, internationally, international) and other. As authors mentioned, topic modelling showed that it is not only possible to classify job information from vacancies but that we can also derive behavioral characteristics that are valued or required by employers from potential or existing job holders. Moreover, as a further analysis of this research was planned the analysing trends of worker attributes required by organizations (i) over time, (ii) occupations, companies, and (iii) geographical regions, and also (iv) possibility to build a network of work activities to examine relationship among tasks.

Wallace et al. (Wallace et al., 2014), Sharma et al. (Sharma et al., 2016) captured the main positive and negative words within latent aspects (topics), which characterise interpersonal manner, technical competence, and systems issues (López et al., 2012) from online physician reviews. Similar with previous work, James et al. (James et al., 2017) based on López et al. (López et al., 2012) categorization, examined unstructured textual feedback of physicians in order to determine: (i) how the extracted sentiment and topics compared to traditional identified dimensions of service quality in healthcare and (ii) what tone and topic elements were driving patients' service quality ratings. As a main finding were the following list of topics and their tone: (1) Negative system quality: Staff and Timeliness (vocabulary of words: office, staff, time, doctor, wait, appointment); (2) Positive interpersonal quality: Physician Compassion (vocabulary of words: doctor, caring, great, knowledgeable, excellent, recommend); (3) Negative system quality: Experience (vocabulary of words: told, don't, doctor, ask, bad, money, call); (4) Positive Technical quality: Family (vocabulary of words: doctor, questions, staff, practice, children, son, pregnancy); (5) Positive Technical quality: Surgery (vocabulary of words: surgery, pain, procedure, staff, hospital, knee, cancer, age); (6) Negative Technical quality: Diagnosis (vocabulary of words: years, treatment, medical, patient, conditions, test, diagnosis, time, treated). The obtained results allowed the authors to establish the dependence on the degree of influence of the identified aspects (topics) on the general perception of the physician's quality, as well as the behavioural characteristics of patients when choosing a doctor online, depending on the content of comments and overall rating.

4.2 Structural Topic Modelling

When conducting research on the basis of textual documents or customers comments, researchers often have a more of information "about the text" than "about the content of the text". From the perspective of topic modelling as a statistical approach, the existence of such information "about the text" (metadata) allows and initiates the inclusion in the model of additional covariates that could influence the following components of the topic model: (1) Proportion of the document devoted to the topic ("prevalence of the topic"). For example, we can know that "clients who buy products online are more likely to talk about delivery problems than clients who buy offline". (2) Word rates used in the discussing of the topic ("topical content"). For example, we can clarify that "when clients talking about delivery problems, clients who buy products online are more likely discuss the problems about products returning, but patients clients who buy offline are more likely discuss staff rudeness issues" (Roberts et al., 2019). Such possibilities are proposed by *Structural topic modelling (STM)* as an extension of the LDA framework (Robinson, 2019; Roberts et al., 2019, 2013) .

Drawing analogies with LDA: (i) each document in STM arises as a mixture over K topics; (ii) topic proportions (θ_d) can be correlated (LDA limitation 1); (iii) topics prevalence θ_d can be influenced by set of covariates X through a standard regression model with covariates; (ii) for each w_n word in the document d (iii) a topic $Z_{d,n}$ is drawn from the document-specific distribution, and (iv) conditional on that topic, a word is chosen from a multinomial distribution over words parameterized by $\beta_{d,k,v}$, where $k = Z_{d,n}$. This distribution can include a second set of covariates Y (Roberts et al., 2019). Thus, the main differences between the LSA and STM models (figure 3) are that the prevalence (content) parameters determined in the LDA by the general a priori Dirichlet parameters $\alpha(\eta)$ in the STM model are replaced with prior structures specified in the form of generalized linear models parameterized by document specific covariates $X(Y)$ (Hu et al., 2019) These covariates inform either the topic prevalence (covariates X) or the topical content (covariates Y) latent variables with information "about the text" (metadata).

5 EXAMPLE OF STRUCTURAL MODELLING ALGORITHMS APPLICATION IN EDUCATION

In order to study customer perception of the quality of services, assess their satisfaction with goods or services received, as well as identify factors that influence customer acceptance of new offers on the market, students were asked to use STM tools. As a data source 610 textual comments about hospitals from the site http://www.ratemyhospital.ie/ (over the past two years – 2018–2019) were used. STM package allows to use all additional variables to demonstrate the power of meta-data for topic modelling. With this aim, textual comments data was extended by information about (1) hospital ownership (private, public), (2) sentiment (positive or negative) (table 1) (Ojo and Rizun, 2020). After that, all steps of text pre-processing were performed.

First, the STM model's setup were performed. To determine the optimal number of topics, STM models from 10 till 30 topics were built were analyzed. Semantic coherence is maximized when the most probable words in a given topic frequently co-occur together, and it is a metric that correlates well with a human judgment of topic quality. Having high semantic coherence is relatively easy, though, if we only have a few topics dominated by very common words, so we wanted to look at both semantic coherence and exclusivity of words to topics. So, the most valuable number of topics should be very coherent and also very exclusive. Looking at figure 4, we draw the conclusion that the 15 topics suit the most to these criteria. Most of the topics, in this case, are above the average of exclusivity and have high coherence, especially compared to the other number of topics which are often spread out on both axes. 15-topic STM model was selected based on subjectively optimal combination of the average semantic coherence and exclusivity outcomes.

As a result, for 15-topic model, we received the (i) topic-words distribution β; (ii) document-topic pro-

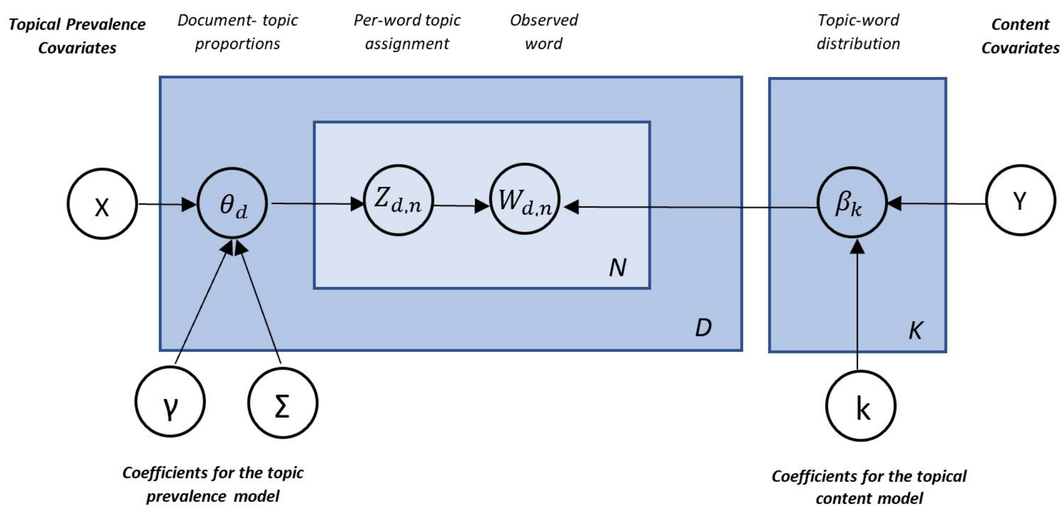

Figure 3: A graphical illustration of the structural topic model (Roberts et al., 2016).

Table 1: Comments before pre-processing.

Comments	Hospital Ownership	Sentiment
A lovely friendly patient-focussed hospital	Public	Positive
Consultant I found seriously lacking compassion for my mother the patient. Sniggered while informing us that while my mother's condition is uncomfortable, it is not life threatening. To be frank, consultant spoke down to us.	Public	Negative
Tullamore is a very clean hospital and looks very well. All staff I had the pleasure of meeting were lovely and very professional at all times. The staff in all capacities do not receive enough thanks for the jobs they do	Private	Positive

portions θ; (iii) list of Highest probability-, FREX-, Lift- and Score-keywords (*Highest Prob*: are the words within each topic with the highest probability; *FREX*: are the words that are both frequent and exclusive, identifying words that distinguish topics; *Lift*: give more weight to words that appear less frequently in other topics by dividing their frequency into other topics; *Score*: score words are weighted by dividing the log frequency of the word in the topic by the log frequency in other topics (Roberts et al., 2013; Chang, 2015; Griffiths and Steyvers, 2004)); (iv) set of documents, mostly associated with this topic. The figure 5 allows us to get information on the share of the different topics at the overall corpus.

Second, students needed to realize the *Topics labelling* step. For that: (1) two students independently labelled the topics to produce the first version of labels based on top weighted keywords; (2) two students discussed the labels and resolved discrepancies in labelling; (3) two students independently refined topic labels based on the computationally guided deep reading 20 of the most representative tweets of the topics; (5) two students agreed on final 15 topic labels and jointly developed the topics descriptions (short summarization of the topic content) (Ojo and Rizun, 2020). The result of topic labelling is presented in the table 2.

Third, the STM covariate analysis could be performed. In this stage, we aimed the evaluating the *Sentiment* effect on the formation of more positively and more negatively oriented aspects of hospitals service quality (HSQ). Thus, we use Sentiment metadata as Covariate in the STM model. Formally, we can identify an aspect as negative if, according to the results of effect estimation, the proportion of this aspect in negative comments (Sentiment = Negative) is significantly higher than in comments in positive comments (Sentiment = Positive). According to the results of our experiment, 5 topics (33.33%) are positive (right side of figure 6), and 10 topics (66.66%) are negative (left side of figure 6).

The dots in the figure 6 indicated the mean values of the estimated proportion differences (power of influence, PI) with 95% confidence intervals, allows us

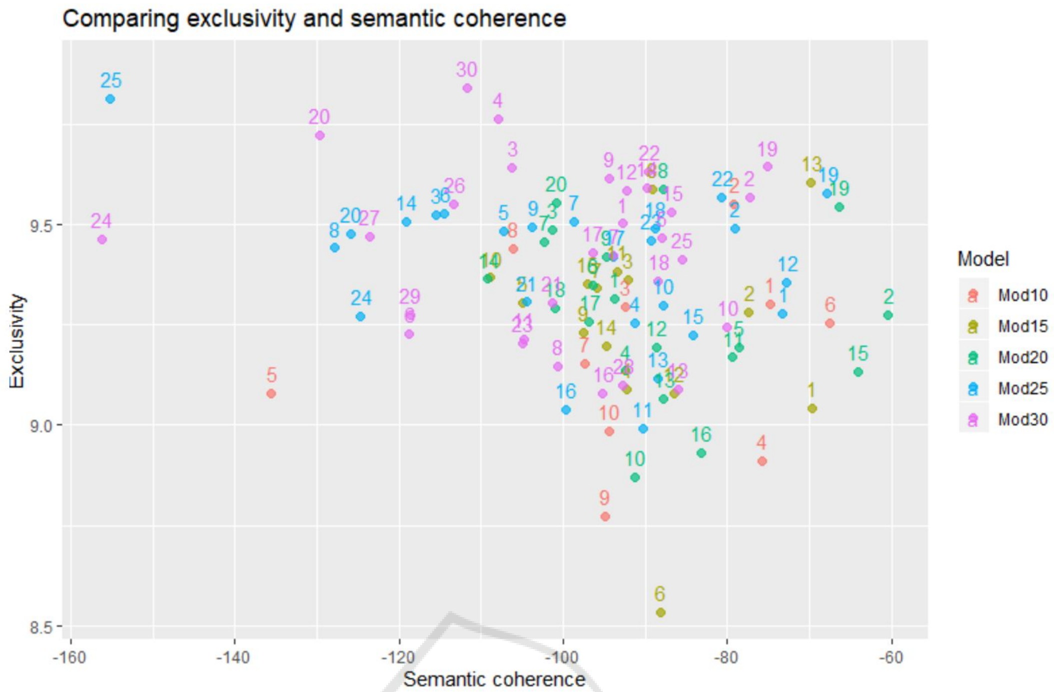

Figure 4: Semantic coherence and exclusivity of STM models.

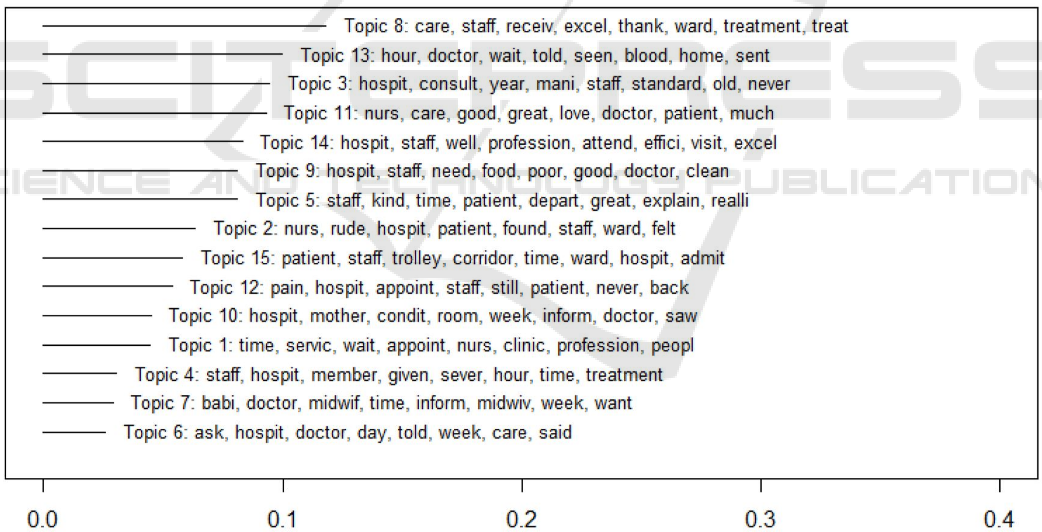

Figure 5: Expected topic proportions over corpus.

to evaluate the relative degree of influence of sentiment on of hospitals service quality aspects. For example, the five most negative Topic of are (1) *Information Exchange with Patient/Family* (Topic 13) with highest power of negative influence; (2) *Communication Skills* (Topic 2); (3) *A&E/Admission* (Topic 12), (4) *Waiting Time* (Topic 4) and (5) *Patient-Focusing Service* (Topic 6). In turn two most positive topics are (1) *Service Rapidness* (Topic 14); (2) *Personnel Reliability/Treatment* (Topic 8). Knowledge about Topics with a positive and negative impact of comments Sentiment allow to indicate the strength of patient satisfaction/dissatisfaction with the hospitals service quality.

Fourth, the power of *Time* influence on positive and negative Topics dynamics (from 2018 to 2019) using the STM model (with Year and Sentiment as a Covariates) should be performed. In terms of the

Table 2: Topics labels.

No	Topics label	Topic keywords	Topic proportion, %
1	Appointment Time Reliability	time, service, wait, appoint, nurses, clinic, profession	4.47
2	Communication Skills	nurses, rude, hospital, patient, found, staff, ward	6.34
3	Service Standards	hospital, consult, year, many, staff, standard, old	9.45
4	Waiting Time	staff, hospital, member, given, sever, hour, time	3.03
5	Staff Feedback/Explanation	staff, kind, time, patient, depart, great, explain	8.09
6	Patient-Focusing Service	ask, hospital, doctor, day, told, week, care	2.56
7	Maternity Unit/Care	baby, doctor, midwife, time, inform, midwife, week	2.89
8	Personnel Reliability / Treatment	scare, staff, receive, excel, thank, ward, treatment	11.81
9	Food Service	hospital, staff, need, food, poor, good, doctor	8.10
10	Hospital Environment	hospital, mother, conditions, room, week, inform, doctor	4.48
11	Care and Recovery	nursed, care, good, great, love, doctor, patient	9.29
12	A&E/Admission	pain, hospital, appoint, staff, still, patient, never	5.37
13	Information Exchange with Patient/Family	hour, doctor, wait, told, seen, blood, home	9.99
14	Service Rapidness	hospital, staff, well, profession, attend, efficiency, visit	8.31
15	Ward/Hospital's Facilities	patient, staff, trolley, corridor, time, ward, hospital	5.82

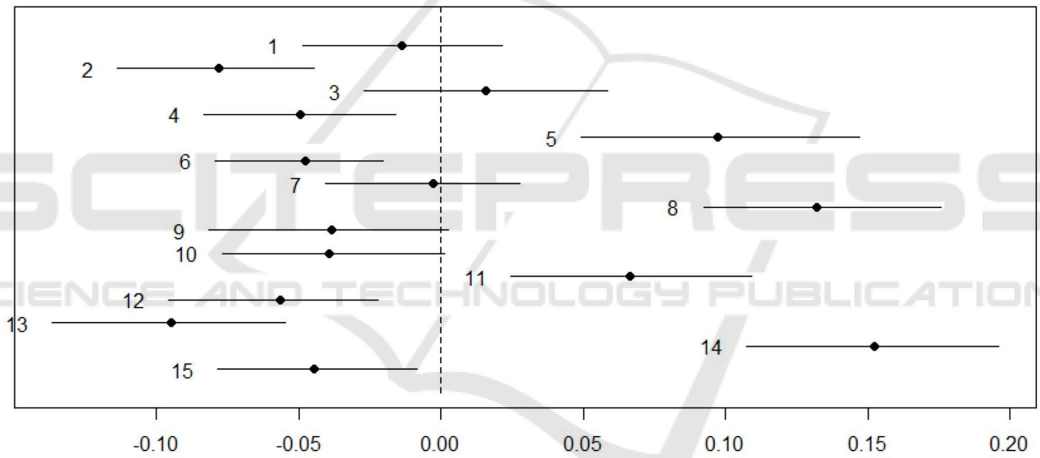

Figure 6: Difference in the power of Sentiment influence on topic proportion.

Influence of the Time Factor on the Service Quality, the following four groups of HSQ Topics can be distinguished: (1) Topics causing the growth of patient satisfaction with the Service Quality over the time: positive topics with a positive dynamic over the time; (2) Topics causing a recession in patient satisfaction with the hospitals service quality (HSQ) over the time: positive topics with a negative dynamic over the time; (3) Topics causing the growth of patient dissatisfaction with the HSQ over the time: negative topics with a positive dynamic over the time (4) Topics causing a recession in patient dissatisfaction with the HSQ over the time: negative topics with a negative dynamic over the time.

As an indicator that allows us to identify the direction and growth rate (GR) of change in the level of positive or negative comments describing the Topic, the slope of the regression (dependence between the proportion of Positive/Negative Aspects and Time) will be used. The presented four charts (figure 7 a, b, c, d) show examples of four possible types of Influence of the Time Factor on the Service Quality:

1. Positive impact on Service Quality over the time: Service Rapidness topic characterized by growth rate (GR=1.100763) of patient satisfaction with the HSQ over the time (figure 7, b);
2. Worsening of Service Quality over the time: Personnel Reliability/Treatment topic characterized by and recession (GR=0.821713) in patient satisfaction with the HSQ over the time (figure 7, a);
3. Negative impact on Service Quality over the time:

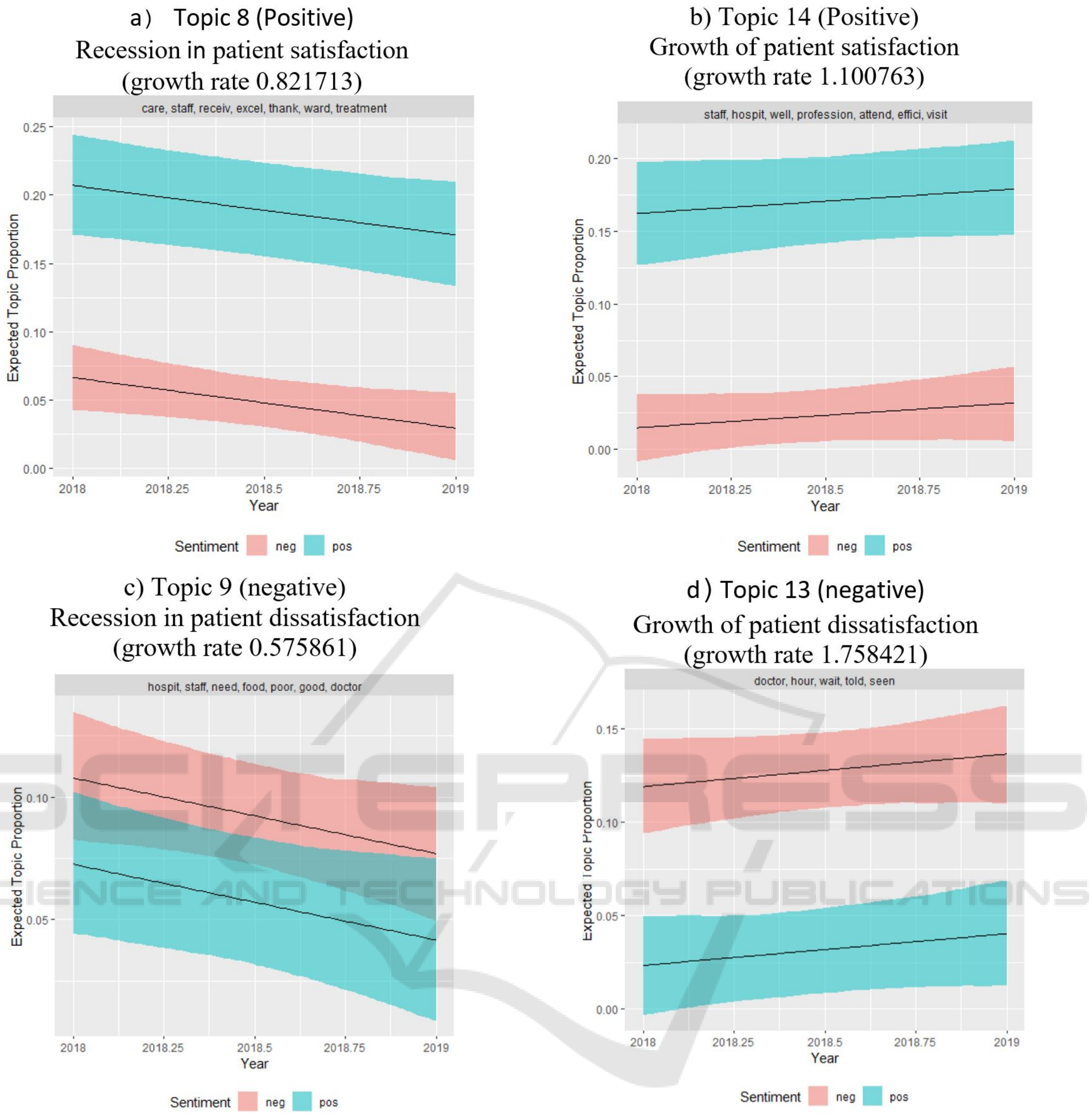

Figure 7: Examples of identification the influence of the Years Metadata.

Information Exchange with Patient/Family topic characterized by growth (GR= 1.758421) of patient dissatisfaction with the HSQ over the time (figure 7, d);

4. Improvement of Service Quality over the time: Food Service topic causing a recession in customer dissatisfaction (GR= 0.575861) with the HSQ over the time (figure 7, c).

As a result, student could see that the largest number of aspects (37.5%) has a negative impact on the HSQ. The highest degree of growth in patient dissatisfaction is characterized by *A$E/Waiting Time* topic. Moreover, this growth rate is not only the largest in the category of Negative impact, but in all analyzed topics. The most rapid (within the whole set of topics) decrease in the number of positive comments is characterized by the aspect of *Maternity Unit/Care*. The group of topics on which improvement in their quality is noted is 25.1%. At the same time, the Hospital Environment is characterized by the highest rate of improvement. 16.7% of topics have a positive effect on the HSQ, among which *Service Rapidness* and *Maternity Unit/Treatment* have the largest increase in the number of positive comments.

Fifth, students may identify the influencing the *Hospital Ownership* on more positively and more

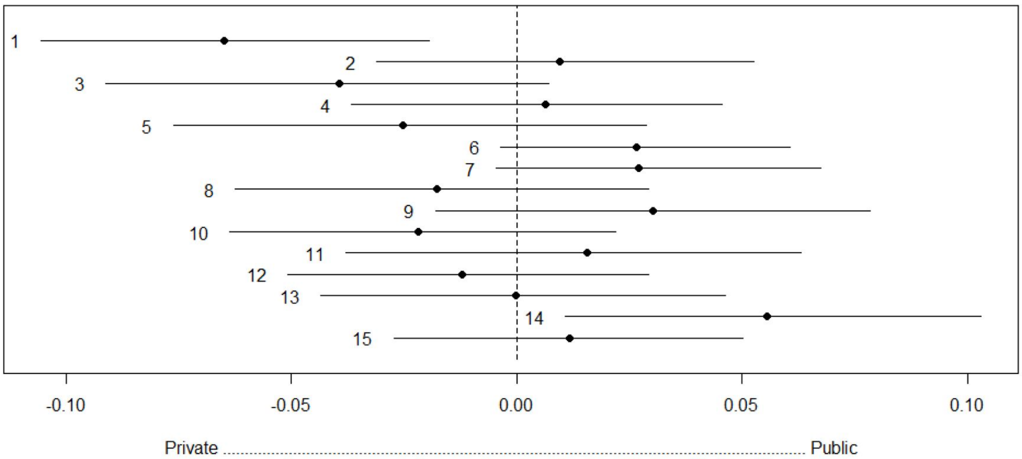

Figure 8: Difference in the power of Hospital Ownership influence on Topic Proportion.

negatively oriented HSQ aspects structure (using the Sentiment and Hospital Ownership factors as in the Covariates STM model). For this purpose, the following interpretation of the results could be proposed: (1) the Topics, more related to Public Hospital Ownership according to the results of effect estimation, in which the proportion of this Topics in comments about Public hospitals (Hospital Ownership = Public) is significantly higher than in comments about Private hospitals and vice versa; (2) the direction (positive or negative) of Hospital Ownership influencing on HSQ. For reaching the first purpose, the Hospital Ownership effect estimation was performed for revealing the aspects in which the proportion of the comments about Public hospitals (Hospital Ownership = Public) is significantly higher than comments about Private hospitals and vice versa.

For formalization the rules for second purpose reaching, in terms of discovering the Influence of the Hospital Ownership on the Service Quality, the following groups of aspects proposed to be distinguished: (1) Topics causing the growth the level of patients satisfaction with Service Quality in Public hospitals: positive topics with a positive dynamic from Private to Public; (2) Topics causing the growth in the level of patients satisfaction with Service Quality in Private hospitals: positive topics with a positive dynamic from Public to Private; (3) Topics causing the growth the level of patients dissatisfaction with Service Quality in Public hospitals: negative topics with a positive dynamic from Private to Public; (4) Topics causing the growth in the level of patients dissatisfaction with Service Quality in Private hospitals: negative topics with a positive dynamic from Public to Private.

According to the results of our experiment, 8 Topics are more associated with Public Hospitals (right side of figure 8), and 6 Topics are more associated with Private Hospitals (left side of figure 8), and one topic (Topic 13) is for both types of hospitals. Based on received results, we can conclude that the four topics (one positive and 3 negative), which more characterize the Public Hospital Ownership are (1) *Service Rapidness* (positive); (2) *Food Service* (negative) (3) *Maternity Unit/Care* (negative) and (4) *Patient-Focusing Service* (negative). In turn five Aspects, which more characterize the Private Hospital Ownership (two positive and two negative) are (1) *Appointment Time Reliability* (negative); (2)*Service Standards* (positive); (3) *Staff Feedback/Explanation* (positive) and (4) *Hospital Environment* (negative).

Thus, this example of the use of STM modeling in teaching students shows how versatile and in-depth research can be carried out using data science. Presented examples demonstrate the nature of tasks and approaches which could develop students' technical and research skills in the public perception analysis. Such approaches also allow students to gain *practical experience* in the study and interpretation the influence of additional metadata, characterizing the comments authors, on differences in their opinions about events, companies, goods, and services.

6 DATA SCIENCE STUDY PROGRAMS IN ECONOMICS FIELD

Classical methods of statistical analysis, modeling methods, and data mining are used in economics. The analysis of data in these areas is aimed at the study of

Table 3: Data Science courses and programs for economics at top-20 universities.

University	Location	Programs, courses
Massachusetts Institute of Technology (MIT)	United States	MicroMasters Program in Data, Economics, and Development; Policy Computer Science, Economics and Data Science – course
Stanford University	United States	M.S. in Statistics: Data Science; Tackling Big Questions Using Social Data Science – course
Harvard University	United States	Data Science for Business – course; Using Big Data Solve Economic and Social Problems – course
California Institute of Technology	United States	Business Analytics – course
University of Oxford	United Kingdom	MSc in Social Data Science
ETH Zurich - Swiss Federal Institute of Technology	Switzerland	Data Science in Techno-Socio-Economic Systems – course
University of Cambridge	United Kingdom	Economics: Data Science and Policy – course
Imperial College London	United Kingdom	MSc Business Analytics
University of Chicago	United States	Economic Policy Analysis – course
UCL	United Kingdom	Economics and Statistics BSc; Social Sciences with Data Science BSc
National University of Singapore	Singapore	Master of Science in Business Analytics
Princeton University	United States	Statistics and Machine Learning – course
Nanyang Technological University	Singapore	Master of Science in Analytics
EPFL	Switzerland	Master's program in Data science
Tsinghua University	China (Mainland)	Master's Program in Data Science
University of Pennsylvania	United States	Master of Information Systems Management, Business Intelligence and Data Analytics; MS in Information Technology, Business Intelligence and Data Analytics; Online Master of Science in Business Analytics
Yale University	United States	Applied Econometrics: Politics, Sports, Microeconomics; Applied Econometrics: Macroeconomic and Finance Forecasting
Cornell University	United States	Introduction to Data Science – course
Columbia University	United States	Data Science for Social Good – summer program
The University of Edinburgh	United Kingdom	Statistics with Data Science MSc

causation. In economics, current issues include policy development, determining the impact of a decision, long-term and short-term planning and forecasting, choosing the best solution from many possible, and many others. Drawing conclusions is also important in economics. In addition, the modern economy and finance are characterized using big data, so it is not always possible to use classical methods. Therefore, the methods of data science are precisely those methods that should be used in economics, which gives positive results and effect. Data Science methods were first used in economic research and gradually penetrated into practice. Today, economics need specialists who have knowledge in these areas and are able to apply Data Science methods. In response to this market need, universities have begun to implement Data Science courses and programs for students of economics. The table 3 presents the courses and programs of the top 20 universities in the world.

A study programs in economic field in Ukrainian universities has shown that Data Science courses and programs are still being introduced in Ukraine. Currently, there are separate programs for studying Data Science, mainly for computer science. Therefore, we believe that the prospects that Data Science opens for modern economists necessitate the introduction of courses and programs in Data Science.

7 CONCLUSIONS

Data science annually extends to more and more areas is used in various areas of research, in society, and in business processes. Businesses and governments are making huge investments in this area. Education, accordingly, must keep pace with the times and teach students new modern technologies. However, as research has shown, the process of studying Data science for economists is at an early stage. Leading universities are gradually introducing new courses and programs to study Data science in economics, but this phenomenon has not yet become widespread and needs to be developed.

As an example of the implementation of Data science methods, we have shown the use of STM-modeling in teaching students. The application of such approaches promotes the development of technology and research skills of students, demonstrates work with big data, and allows to gain experience in studying and interpreting the influence of additional metadata characterizing the authors' comments on differences in their opinions about events, companies, goods, and services.

The described methods and algorithms are just some of the basics of modeling and analysis of economic processes. There are many examples of how all these methods can be used in education. For example, using time series analysis, we could predict the future value of a cryptocurrency, using regression models, we could determine customer loyalty or the likelihood of customer insolvency, and so on. Today, there are many more algorithms that can be applied in economics.

Education must meet the modern development of the digital economy, digital society, innovation, and creative entrepreneurship. The use of Data science in education should be multi-platform, ie used not only in the study of the subject but in the teaching of all subjects, interaction of students with each other and with teachers, real experts, research, and individual learning.

REFERENCES

Ausloos, M., Cerqueti, R., and Mir, T. A. (2017). Data science for assessing possible tax income manipulation: The case of Italy. *Chaos, Solitons & Fractals*, 104:238–256.

Babenko, V., Panchyshyn, A., Zomchak, L., Nehrey, M., Artym-Drohomyretska, Z., and Lahotskyi, T. (2021). Classical machine learning methods in economics research: Macro and micro level examples. *WSEAS Transactions on Business and Economics*, (18):209–217.

Balabay, R. and Chernonog, E. (2007). Alteration on the surface of the pore walls of the porous silicon under high temperature ageing: Computer simulation. *Radiation Measurements*, 42(4-5):739–741.

Bielinskyi, A., Khvostina, I., Mamanazarov, A., Matviychuk, A., Semerikov, S., Serdyuk, O., Solovieva, V., and Soloviev, V. (2021). Predictors of oil shocks. Econophysical approach in environmental science. *IOP Conference Series: Earth and Environmental Science*, 628(1):012019.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Brooks, C., Hoepner, A. G. F., McMillan, D., Vivian, A., and Simen, C. W. (2019). Financial data science: the birth of a new financial research paradigm complementing econometrics? *The European Journal of Finance*, 25(17):1627–1636.

Brunner, R. J. and Kim, E. J. (2016). Teaching data science. *Procedia Computer Science*, 80:1947–1956.

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys*, 50(3):1–42.

Chang, J. (2015). lda: Collapsed Gibbs sampling methods for topic models. https://rdrr.io/cran/lda/.

Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4):1165–1188. http://www.jstor.org/stable/41703503.

Danylchuk, H., Chebanova, N., Reznik, N., and Vitkovskyi, Y. (2019). Modeling of investment attractiveness of countries using entropy analysis of regional stock markets. *Global Journal of Environmental Science and Management*, 5(Special Issue):227–235. https://www.gjesm.net/article_35558.html.

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

Derbentsev, V., Matviychuk, A., Datsenko, N., Bezkorovainyi, V., and Azaryan, A. (2020). Machine learning approaches for financial time series forecasting. *CEUR Workshop Proceedings*, 2713:434–450.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41(6):570–606.

Dimitrov, I., Davydenko, N., Lotko, A., and Dimitrova, A. (2019). Comparative study of main determinants of entrepreneurship intentions of business students. In *2019 International Conference on Creative Business for Smart and Sustainable Growth (CREBUS)*, pages 1–4. IEEE.

George, G., Osinga, E. C., Lavie, D., and Scott, B. A. (2016). Big data and data science methods for management research. *The Academy of Management Journal*, 59(5):1493–1507.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Guryanova, L., Bogachkova, L., Zyma, O., Novosel, M., Poluektova, N., and Gvozdytskyi, V. (2020a). Models of estimation and analysis of a systemic risk in the banking sector. In *2020 IEEE 2nd International Conference on System Analysis Intelligent Computing (SAIC)*, pages 1–6.

Guryanova, L., Yatsenko, R., Dubrovina, N., and Babenko, V. (2020b). Machine learning methods and models, predictive analytics and applications. *CEUR Workshop Proceedings*, 2649:1–5.

Hrabovskyi, Y., Babenko, V., Al'Boschiy, O., and Gerasimenko, V. (2020). Development of a Technology for Automation of Work with Sources of Information on the Internet. *WSEAS Transactions on Business and Economics*, 17(25):231–240.

Hu, N., Zhang, T., Gao, B., and Bose, I. (2019). What do hotel customers complain about? text analysis using structural topic model. *Tourism Management*, 72:417–426.

Ignatyuk, A., Liubkina, O., Murovana, T., and Magomedova, A. (2020). FinTech as an innovation challenge: From big data to sustainable development. *E3S Web of Conferences*, 166:13027.

Ilchuk, M., Davydenko, N., and Nehoda, Y. (2019). Scenario modeling of financial resources at the enterprise. *Intellectual Economics*, 13(2).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning: with Applications in R*, volume 112 of *Springer Texts in Statistics*. Springer, New York.

James, T. L., Calderon, E. D. V., and Cook, D. F. (2017). Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Systems with Applications*, 71:479–492.

Kaminskyi, A., Nehrey, M., and Rizun, N. (2020). The impact of COVID-induced shock on the risk-return correspondence of agricultural ETFs. *CEUR Workshop Proceedings*, 2713:204–218.

Klymenko, N., Nosovets, O., Sokolenko, L., Hryshchenko, O., and Pisochenko, T. (2019). Off-balance accounting in the modern information system of an enterprise. *Academy of Accounting and Financial Studies Journal*, 23(2). https://www.abacademies.org/articles/offbalance-accounting-in-the-modern-information-system-of-an-enterprise-8403.html.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., and Den Hartog, D. N. (2018a). Text classification for organizational researchers: A tutorial. *Organizational research methods*, 21(3):766–799.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., and Den Hartog, D. N. (2018b). Text mining in organizational research. *Organizational research methods*, 21(3):733–765.

Kuzmenko, O. V., Lieonov, S. V., and Boiko, A. O. (2020). Data mining and bifurcation analysis of the risk of money laundering with the involvement of financial institutions. *Journal of International Studies*, 13(3). https://www.jois.eu/files/22_871_Kuzmenko%20et%20al.pdf.

López, A., Detz, A., Ratanawongsa, N., and Sarkar, U. (2012). What patients say about their doctors online: a qualitative content analysis. *Journal of general internal medicine*, 27(6):685–692.

Matviychuk, A., Strelchenko, I., Vashchaiev, S., and Velykoivanenko, H. (2019). Simulation of the crisis contagion process between countries with different levels of socio-economic development. *CEUR Workshop Proceedings*, 2393:485–496.

National Academies of Sciences, Engineering, and Medicine, Division on Engineering and Physical Sciences, Division of Behavioral and Social Sciences and Education, Computer Science and Telecommunications Board, Board on Mathematical Sciences and Analytics, Committee on Applied and Theoretical Statistics, Board on Science Education, and Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective (2018). *Data science for undergraduates: Opportunities and options*. The National Academies Press, Washington, DC.

Nehrey, M. and Hnot, T. (2017). Using recommendation approaches for ratings matrixes in online marketing. *Studia Ekonomiczne*, (342):115–130.

Nehrey, M., Kaminskyi, A., and Komar, M. (2019). Agro-economic models: a review and directions for research. *Periodicals of Engineering and Natural Sciences*, 7(2):702–711. http://pen.ius.edu.ba/index.php/pen/article/view/579.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., Reuter, U., Gama, J., and Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. *Mathematics*, 8(10):1799.

Ojo, A. and Rizun, N. (2020). Structural and temporal topic models of feedbacks on service quality–a path to theory development? In *Americas Conference on Information Systems (AMCIS 2020). Healthcare Informatics & Health Information Tech (SIGHealth)*, volume 15.

Oliskevych, M., Beregova, G., and Tokarchuk, V. (2018). Fuel consumption in Ukraine: Evidence from vector error correction model. *International Journal of Energy Economics and Policy*, 8(5). https://www.econjournals.com/index.php/ijeep/article/view/6825/3925.

Oliskevych, M. and Lukianenko, I. (2019). Labor force participation in Eastern European countries: nonlinear modeling. *Journal of Economic Studies*, 46(6):1258–1279.

Parish, E. J. and Duraisamy, K. (2016). A paradigm for data-driven predictive modeling using field inversion and machine learning. *Journal of Computational Physics*, 305:758–774.

Patriarca, R., Di Gravio, G., and Costantino, F. (2017). A Monte Carlo evolution of the Functional Resonance Analysis Method (FRAM) to assess performance variability in complex systems. *Safety science*, 91:49–60.

Perevozova, I., Babenko, V., Krykhovetska, Z., and Popadynets, I. (2020). Holistic approach based assessment of social efficiency of research conducted by higher educational establishments. *E3S Web of Conferences*, 166:13022.

Prüfer, J. and Prüfer, P. (2018). Data science for institutional and organizational economics. Technical report.

Rizun, N. and Shmelova, T. (2017). Decision-making models of the human-operator as an element of the socio-technical systems. In *Strategic Imperatives and Core Competencies in the Era of Robotics and Artificial Intelligence*, pages 167–204. IGI Global.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91(1):1–40.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. (2013). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Albertson, B., Gadarian, S., and Rand, D. (2014). Topic models for open ended survey responses with applications to experiments. *American Journal of Political Science*, 58:1064–82.

Robinson, S. D. (2019). Temporal topic modeling applied to aviation safety reports: A subject matter expert review. *Safety science*, 116:275–286.

Sharma, R. D., Tripathi, S., Sahu, S. K., Mittal, S., and Anand, A. (2016). Predicting online doctor ratings from user reviews using convolutional neural networks. *International Journal of Machine Learning and Computing*, 6(2):149.

Shi, Y., Zhu, J., and Charles, V. (2020). Data science and productivity: A bibliometric review of data science applications and approaches in productivity evaluations. *Journal of the Operational Research Society*, 72(5):975–988.

Shoro, A. G. and Soomro, T. R. (2015). Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology: C Software & Data Engineering*, 15(1):7–14.

Soloviev, V. and Belinskiy, A. (2019). Complex systems theory and crashes of cryptocurrency market. *Communications in Computer and Information Science*, 1007:276–297.

Soloviev, V., Serdiuk, O., Semerikov, S., and Kiv, A. (2020a). Recurrence plot-based analysis of financial-economic crashes. *CEUR Workshop Proceedings*, 2713:21–40.

Soloviev, V. N., Bielinskyi, A. O., and Kharadzjan, N. A. (2020b). Coverage of the coronavirus pandemic through entropy measures. *CEUR Workshop Proceedings*, 2832:24–42.

Volkova, N., Rizun, N., and Nehrey, M. (2019). Data science: Opportunities to transform education. *CEUR Workshop Proceedings*, 2433:48–73.

Voronenko, I., Nehrey, M., Kostenko, S., Lashchyk, I., and Niziaieva, V. (2021). Advertising strategy management in Internet marketing. *Journal of Information Technology Management*, 13(Special Issue):35–47.

Voronenko, I., Skrypnyk, A., Klymenko, N., Zherlitsyn, D., and Starychenko, Y. (2020). Food security risk in Ukraine: assessment and forecast. *Agricultural and Resource Economics: International Scientific E-Journal*, 6(4):63–75.

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., and Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, 21(6):1098–1103.

Xiong, J., Yu, G., and Zhang, X. (2017). Research on governance structure of big data of civil aviation. *Journal of Computer and Communications*, 5(5):112–118.

Zelinska, S. (2020). Machine learning: Technologies and potential application at mining companies. *E3S Web of Conferences*, 166:03007.