Data science: opportunities to transform education

Nataliia P. Volkova[1][0000-0003-1258-7251], Nina O. Rizun[2][0000-0002-4343-9713]
and Maryna V. Nehrey[3][0000-0001-9243-1534]

[1] Alfred Nobel University, 18, Sicheslavska Naberezhna Str., Dnipro, 49000, Ukraine
npvolkova@yahoo.com
[2] Gdańsk University of Technology, 11/12, Gabriela Narutowicza, 80-233, Gdańsk, Poland
nina.rizun@pg.edu.pl
[3] National University of Life and Environmental Sciences of Ukraine,
15, Heroiv Oborony Str., Kyiv, 03041, Ukraine
marina.nehrey@gmail.com

Abstract. The article concerns the issue of data science tools implementation, including the text mining and natural language processing algorithms for increasing the value of high education for development modern and technologically flexible society. Data science is the field of study that involves tools, algorithms, and knowledge of math 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. However, data science applications in education have been relatively limited, and many opportunities for advancing the fields still unexplored.

Keywords: data science, high education, clustering, natural language processing, text mining.

1 Introduction

Nowadays the world is changing rapidly: globalization, digitalization, new technologies, etc. These processes are accompanied by the emergence of new types of data and increasing data. To effectively use the benefits of the modern world, you need to be able to use data correctly, model processes and make decisions using modern methods and technologies.

Data science is the field of study that involves tools, algorithms, and knowledge of math 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, etc. However, data science applications in education have been relatively limited, and many opportunities for advancing the fields still unexplored.

Data science should be used in education to solve science problems, for example, in behaviors research in economics, psychology, biology and so on, in predicting different processes, analyzing complex systems, etc.

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 [3], Big data and Data Science methods presented in [4], [8], [12], [23], [25], machine learning used [16], [11], Monte Carlo method presented [17], game theory and genetic algorithms combined [18], Artificial Intelligence presented in [19], [21], [22], etc. 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 decision-making in business [9], [13].

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: the management and processing of data and analytical methods for analysis and modeling. 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.

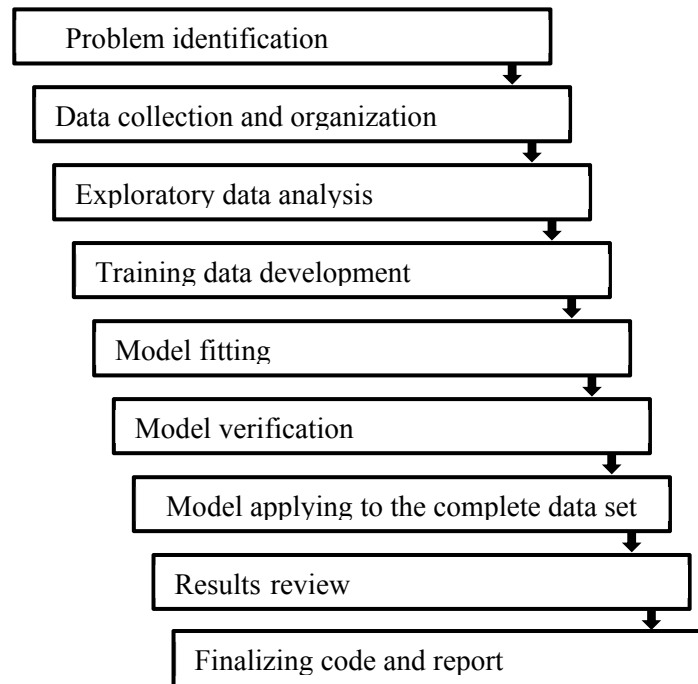Data Science process includes next steps (Figure 1):

```
┌─────────────────────────────────────────┐
│         Problem identification          │
└─────────────────────────────────────────┘
     ┌─────────────────────────────────────────┐
     │     Data collection and organization    │
     └─────────────────────────────────────────┘
          ┌─────────────────────────────────────┐
          │       Exploratory data analysis     │
          └─────────────────────────────────────┘
               ┌────────────────────────────────┐
               │     Training data development   │
               └────────────────────────────────┘
                    ┌───────────────────────────┐
                    │        Model fitting       │
                    └───────────────────────────┘
                         ┌──────────────────────┐
                         │    Model verification │
                         └──────────────────────┘
                              ┌──────────────────────────────┐
                              │ Model applying to the complete data set │
                              └──────────────────────────────┘
                                   ┌──────────────────────┐
                                   │     Results review    │
                                   └──────────────────────┘
                                        ┌──────────────────────┐
                                        │ Finalizing code and report │
                                        └──────────────────────┘
```

Fig. 1. Data Science process

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 algorithm. 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 relationship between variables is expressed by a linear function:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n,$$

x_i – factor i, based on which the forecast is based,

b_i – parameter of the model, the influence of the factor,

y – dependent variable for which the forecast is constructed.

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 business processes, the linear relationship between changes is rare).

As the example, linear regression is trends for time series when time values or index values are taken for an independent variable (for example, from 1 to n, where n is the number of elements in the time series). Trend allows you to predict the value for the next period. For example, research on real climate change has been conducted based on the analysis of the average monthly temperature of soil and air in various areas for 1990-2011 (Figure 2) [9]. It is important to analyze this trend in order to model in the biology, geography, economics.
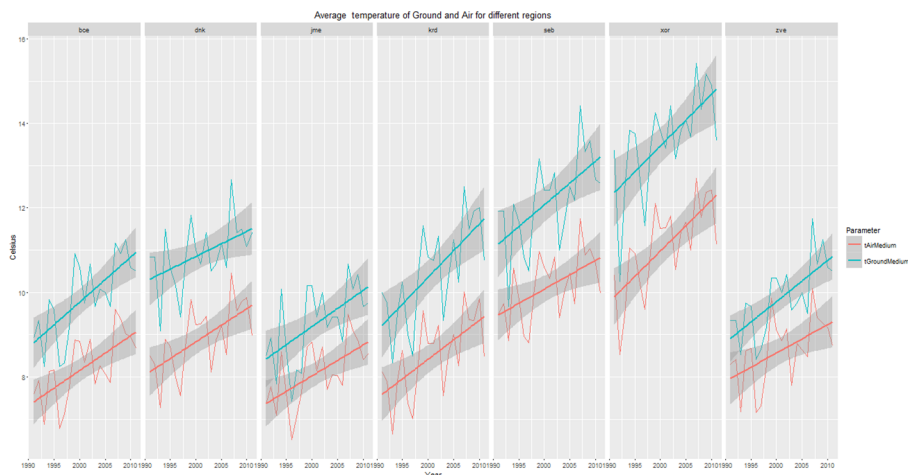
Fig. 2. Average monthly ground and air temperature in Ukraine

Another example of applying linear regression is the optimizing prices. Consider an example of optimizing tickets prices per week in one of the flights. Having historical data on price and demand, this task can be solved in several stages:

1. To forecast the demand for a product by analyzing the time series for the next period, taking into account seasonal characteristics and growth of consumers.
2. Estimate the linear function depending on demand from the price, first calculating the demand level for the next week based on the forecast. In addition, it is possible to add such dependent variables as the promotional product, the presence of ads in booklets/displays, on the Internet, etc. (Figure 3). In complex cases, such dependence can be expressed as a nonlinear function (for example, sigmoid).
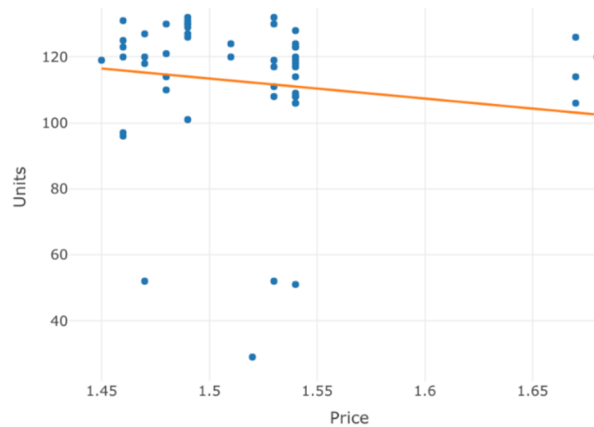
Fig. 3. Price – Demand function of tickets (thousand hryvnyas)

Taking into account the estimated function of demand, optimization of the quadratic function of revenue (Figure 4):

$$D = pu = p(a + bp) = bp^2 + ap,$$

where D – profit, p − price, u – demand.

   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 combinations 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]:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}},$$

$x_i - i$ factor, a base for the forecast,
$b_i - i$ parameter of the model,
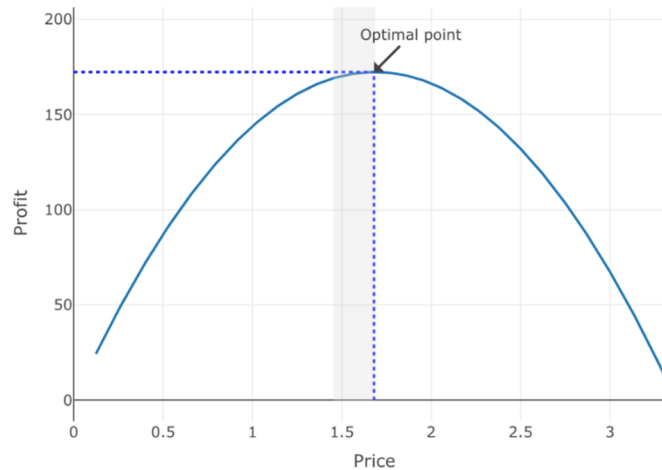p – a probability of belonging to the class "1".

Fig. 4. The function of profit of the tickets

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. In Figure 4 shows a simple model of indicators, which predicts a loan client based on two factors: the age of clients and the term of the loan. This model is based on 1000 copies of the data set "German Credit Risk". As can be seen from the Figure 5, the model assumes higher creditworthiness of clients with a term of lending up to 2 years and at the age of 30–40 years. The accuracy of such a model is ~ 60%, the construction of logistic regression across all 20 attributes, can achieve the accuracy of up to 80%. The black line on the graph reflects the boundary of the model's decision: it has a greater probability of a positive response > 50%.

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 [10].
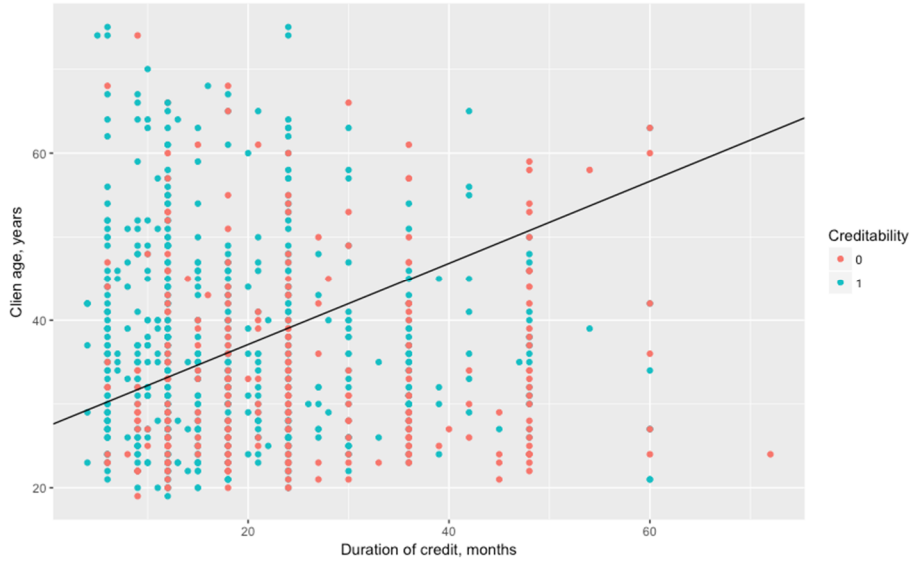
Fig. 5. Scoring model of the creditworthiness of clients

An example of applying a decision tree is the definition of the companies client classification algorithm – the construction of Loyalty Matrix. All clients are divided into 4 groups (TTruly Loyal, Accessible, Trapped, High risk) based on the answer to questions1 to 5 questions. In Figure 6 shows a tree that, based on three questions, allows us to predict the client's class with the accuracy of 98%.
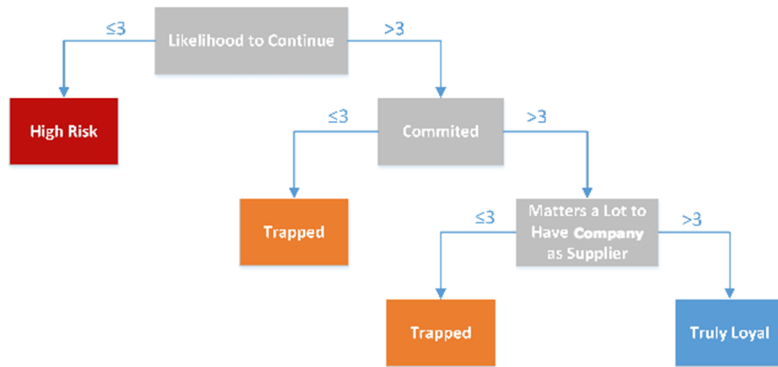

Fig. 6. Decision tree of the classification of company's clients

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).

Figure 7 shows an example of using k-means for clustering users of the Internet service by coordinates [13]. It allows you to split them into groups and form a delivery zone.
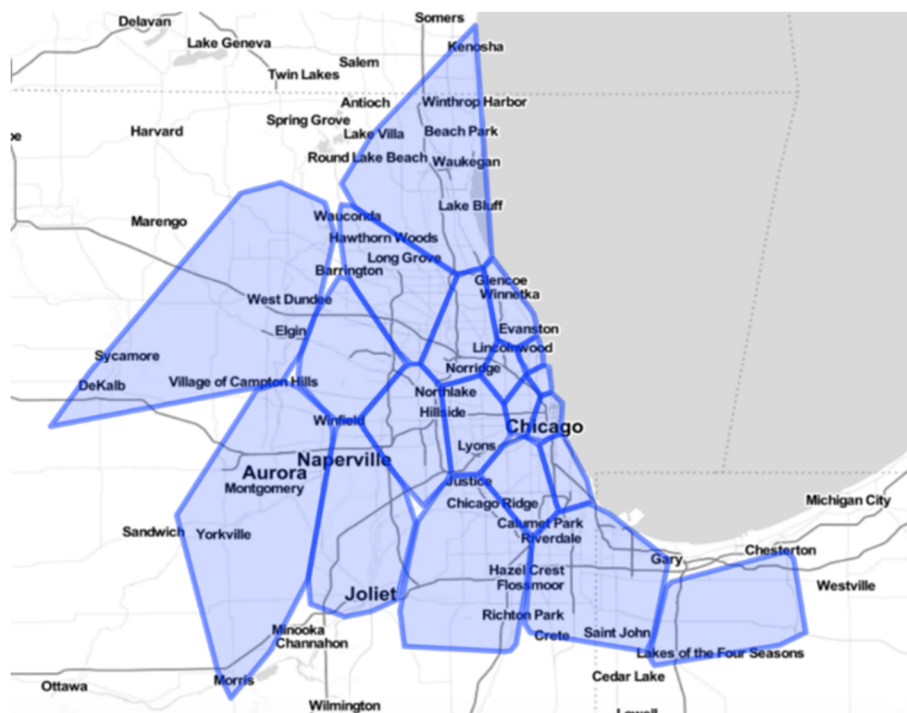

Fig. 7. Clustering of Internet clients based on their coordinates

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.

Figure 8 shows a dendrogram of customer segmentation based on features such as the number of weekends/weekdays transactions, the average number of purchases per week, and so on. Segmentation allows you to select groups of "similar" clients, for example, those who make purchases only on the weekend; those who buy mostly discounted goods, etc. This algorithm allows improving targeted marketing.
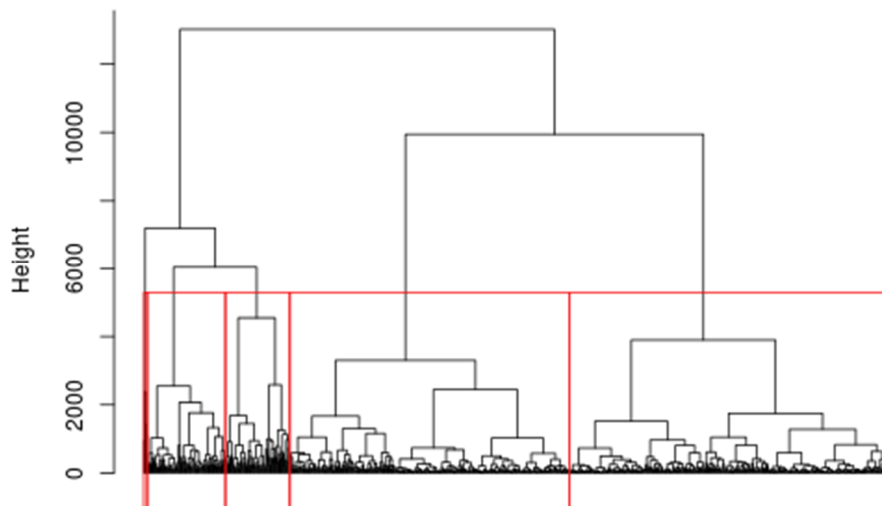

Fig. 8. Dendrogram of customer segmentation

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 [14]. 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 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 a complexity of learning a model in many seasonal conditions.

Figure 9 shows an example of forecasting for 1 week the number of orders in a restaurant [9]. One can clearly see seasonality in one day, which is inherent in the series of this kind.
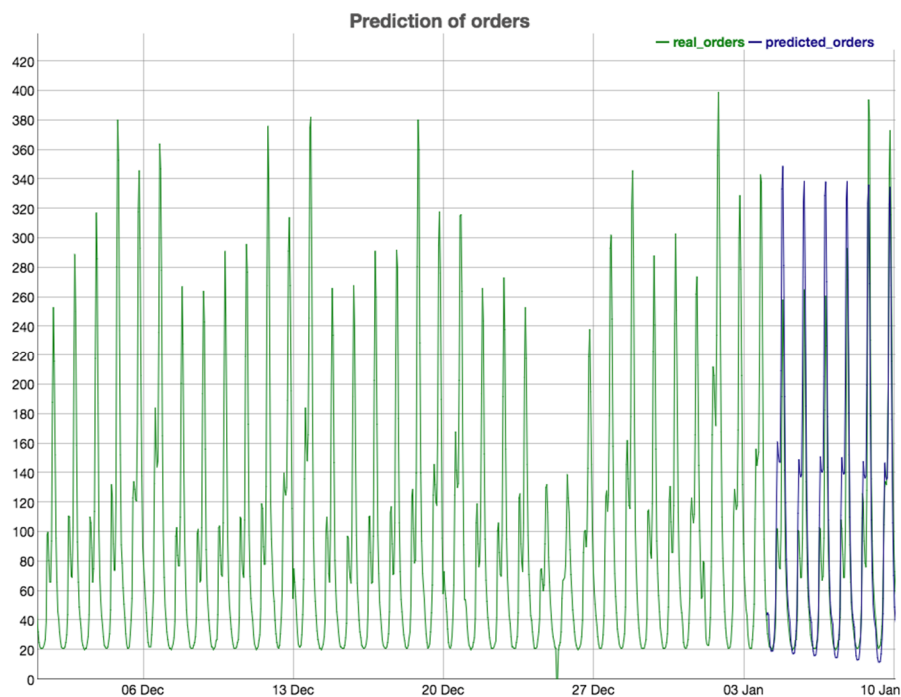

Fig. 9. ARIMA for forecasting tickets demand

Algorithm **Prophet** was developed by Facebook in the beginning of 2017 for forecasting based on time series [14]. 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, 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.

Figures 10, 11 shows an example of prediction with the Prophet. On the first of the charts – forecasting the entire category of goods, on the second – those products that are bought for Christmas. In the second case, only the seasonal components are taken into account and the "holiday" component is not modeled.
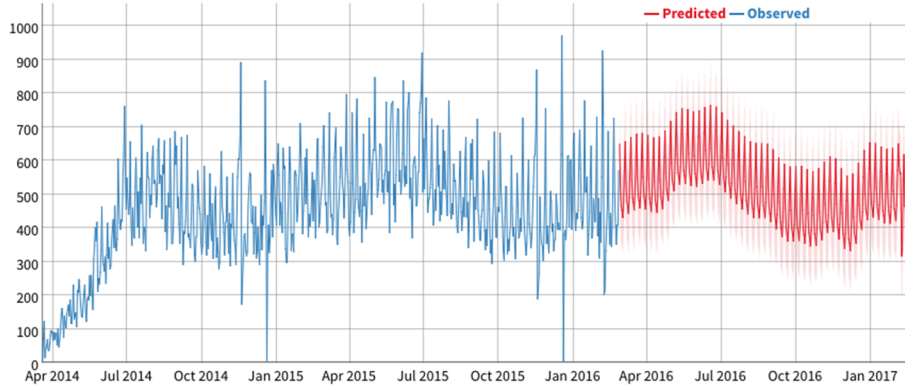
Fig. 10. Forecasting the entire category of goods using the Prophet
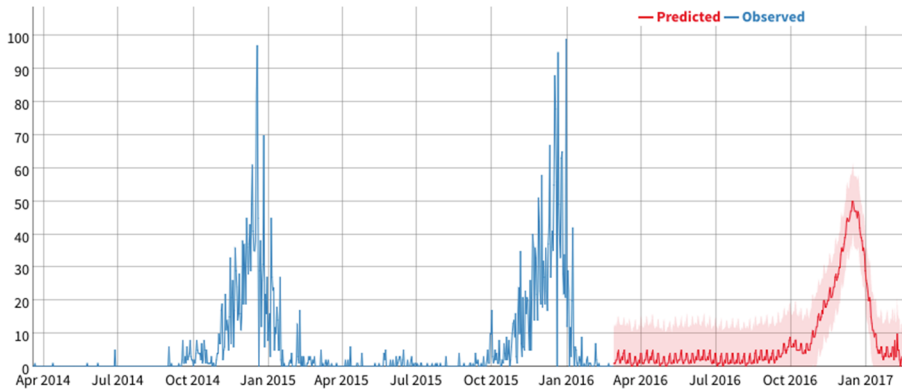

Fig. 11. Time series analysis using the Prophet

3.3 Text mining algorithms

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 their clustering (definition of groups of the similar documents). More and more often statistical topical methods are being applied [24].

The topics are presented as discrete distributions on a number of words, and the documents – as discrete distribution on a number of topics [24]. Topical methods perform a "non-precise" clustering of words and documents, which means that a word or a document can be referred to a few topics with different probabilities simultaneously. The synonyms with higher probability will appear in the same topics since they are frequently used in the same documents. At the same time, the homonyms

(words different in meaning, but similar in writing) will be placed in different topics because they are used in different contexts [7].

Vector Space Model

Topical methods, as a rule, apply the method of a "bad of words", where each document is considered as a set of words not connected to each other. Before the topics are defined, the text is processed – its morphologic analysis is conducted with the objective to define the initial form of words and their meanings in the speech context. The method of processing words in a machine-readable natural language, as a rule, is based on the vector-space method of data description (Vector Space Model) [15], suggested by Gerard A. Salton, Andy Wong, Chung-Shu Yang in [20]. Within the framework of the method each word in a document has its particular weight. Thus, each document is presented as a vector and its dimension is equal to the total number of words in the document.

Similarity of a document and a topic is evaluated as a scalar product of a few information vectors. The weight of separate words (terms) can be calculated both applying the absolute frequency of a word appearing in the text and the relative (normalized) frequency:

$$tf(w,t) = \frac{k(w,t)}{df} \ , \tag{1}$$

where $k(w, L_t)$ is the number of w-word occurrences in the text t; df – total number of words in the text t.

The weight of a word, calculated by the formula (1), in documents is usually put as TF (Term Frequency).

However, this approach does not take into consideration the frequency, with which the word is used in the whole massive of documents – i.e. the so-called discrimination strength of the word. That is why, in case when the statistics for word usage in the whole document is available, it is more efficient to use the other method:

$$TF \times IDF = tf(w,t) \cdot \log_2 \cdot \frac{D}{df}$$

where D – total number of documents in the collection.

TF×IDF method of weighting words shows not the frequency of words appearing in the document, but the measure, inverse to the number of documents in the massive containing this particular word (inverse document frequency).

The Vector Space Model of data presentation provides the systems, which are based on it, with the following functions: creation of professional systems and databases; increase of the level of specialists' competence by means of obtaining an effective possibility of directed search and filtration of text documents; automatic summarization of documents' texts.

Latent Semantic Analysis

In "soft" clustering each word and document refer to a few topics with particular probabilities simultaneously. Semantic description of a word or a document is a

probability distribution on a number of topics. The process of finding these distributions is called "topical modelling".

One of the best methods of "soft" clustering is the Latent Semantic Analysis (LSA), which reflects documents and separate words in the so-called "semantic space", where all the further comparisons are conducted [7].

In this process, the following assumptions are made: documents are a set of words, the order of which is ignored; it is only important how many times a word appears in the text; semantic meaning of a document is defined by the set of words, which, as a rule, go together; each word has a single meaning.

LSA is the method of processing information in natural language, analyzing interconnections between massifs of documents and words, appearing in them, as well as associates topics with documents (words).

The LSA method is based on principles of revealing latent connections of the studied phenomena and objects. In classification/clustering of documents this method is applied to extract context-dependent meanings of lexical units by means of statistical processing of very large text massifs. As the initial information in LSA the matrix "word-document" is used, which describes the set of data, used for system's training.

Elements of this matrix contain weights that consider frequencies of using every word in every document and participation of a word in all documents (*TF×IDF*). The most widely-used variant of LSA is based on using decomposition of a diagonal matrix by singular values (SVD – Singular Value Decomposition).

With the help of SVD any matrix can be decomposed on many orthogonal matrices, the linear combination of which is a rather precise approximation to the initial matrix.

Mathematical basis of the method is as follows:

Formally let A be a $m{\times}n$ words-document matrix of a documents collection. Each column of A corresponds to a document. The values of the matrix elements $A[i, j]$ represent the frequency identifications $tf(w, t)$ of the word occurrence w_i in the document t_j: $A[i, j] = tf(w, t)$. The dimensions of A, m and n correspond to the number of words and documents, respectively, in the collection.

In this case $B = A^T A$ is the document-document matrix. If the documents i and j have b words in common, then $B[i, j] = b$. On the other hand, $C = AA^T$ is the word-word matrix. If the words i and j occur together in c documents, then $C[i, j] = c$. Clearly, both B and C are square and symmetric; B is a $m{\times}m$ matrix, whereas C is an $n{\times}n$ matrix. Now, we perform the Singular Value Decomposition on A using matrices B and C as described in the previous section:

$$A = U\Sigma V^T,$$

where U is the matrix of the eigenvectors of B; V is the matrix of the eigenvectors of C; Σ is the diagonal matrix of singular values obtained as square roots of the eigenvalues of B.

In LSA we ignore these small singular values and replace them by 0. Let us say that we only keep k singular values in Σ. Then Σ will be all zeros except the first k entries along its diagonal. We can reduce the matrix Σ into Σ_k which is a $k{\times}k$ matrix containing only the k singular values that we keep, and also reduce U and V^T, into U_K and V_k^T, to have k columns and rows, respectively. Of course, all these matrix parts that we throw

out would have been zeroed anyway by the zeros in Σ. Matrix A is now approximated by:

$$X_{t \times d} \approx X_{K_{t \times d}} = U_{K_{t \times d}} \Sigma_{K_{t \times d}} \left(V_{K_{t \times d}} \right)^T$$

Intuitively, the k remaining ingredients of the eigenvectors in U and V could be interpreted as a k "hidden concepts" where the words and documents participate. The words and documents now have a new representation in words of these hidden concepts. The results of the LSA method are following:

Document comparison: $Z_k = \Sigma_{K_{t \times d}} \left(V_{K_{t \times d}} \right)^T$ represents docs (cols) in semantic space (scaling with singular values). Documents d_i and d_j can be compared using cosine distance on i and j columns of Z_k.

Word comparison $Y_k = U_{K_{t \times d}} \Sigma_{K_{t \times d}}$ represents words (cols) in semantic space. Words t_i and t_j can be compared as cosine distance on i and j columns of Y_k.

Topic analysis. Left singular vectors $U_{K_{t \times d}}$ map between k words and "semantic dimensions" (topics). Then column k of this vector "describes" topic by giving strength of association with each word. Right singular vectors $V_{K_{t \times d}}$ map between topics and documents could in principle tell us what a document was "about". As with words, one document can be associated with many topics.

Latent Dirichlet Allocation

To get rid of the above-mentioned disadvantages the probability LSA is conducted, based on the multinomial distribution – in particular, on the algorithm of Latent Dirichlet allocation (LDA) (David M. Blei [1; 2], Andrew Y. Ng [2], Michael I. Jordan [2]).

The LDA presupposes that each word in a document is created by a certain latent topic; at the same time distribution of words in each of them is used in a clear form, as well as the prior distribution of words in the document. Topics of all the words in the document are supposed to be independent. In LDA, as well as in LSA, a document can correspond to a few topics. However, LSA sets the algorithm of generation of both words and documents, that is why there appears an additional possibility to evaluate probabilities of documents outside text massive using the algorithm of variation Gibbs sampling.

Unlike LSA, in the LDA the number of parameters does not increase with the growth of number of documents in the studied massive. The applied extensions of the LDA algorithm eliminate some of its limitations and improve productivity for particular tasks. LSA is generating algorithm only for words, but not for documents. The LDA algorithm overcomes this limitation.

The main idea of LDA consists in the fact that the documents are presented as a mix of distributions of latent topics, where each topic is defined by a probability distribution on the set of words. LDA reflects hidden connections between the words by means of topics; it also allows to set probabilities for new documents, which were not included into the training set, applying the algorithm of Variational Bayesian method.

In fact, LDA is a three-stage Bayesian network, which generates a document from a mix of topics. At the first stage for each document d a random vector with the parameter α (usually α is taken as $50/T$) is selected from the Dirichlet distribution. At the second stage a topic z_{di} is selected form the multinomial distribution with the parameter θ_d. Finally, in accordance with the selected topic z a word w_{di} is chosen from the distribution Φ_{zdi}, which is the Dirichlet distribution with the parameter β (usually the parameter β is 0.1; its increase leads to more sparse topics).

4   Example of text mining tools application in education

The printed newspaper becomes a rarity. At the same time, that does not mean that there is less written news, but on the contrary, with the strong online growth of recent decades, more and more people have switched from the classic newspaper to online news. Every written online word can be read and interpreted by a machine. There are numerous application examples in which a machine could improve the news world. Today, everyone wants to read the news, but only the shortest summarized news if possible, as less and less time remains to read long texts. Also, pre-sorting of certain texts in areas has its appeal to news companies because they could then focus entirely on writing messages rather than being biased in the category. On top of that, it would also be great to see a sentiment analysis of each news which will be released, because it would show the reader what kind of news center they are dealing with. It could be that the news center is manipulating its "customers" by choosing certain words, and with a neutral system, which would show the political sorting or sentiment, the user could easily choose what kind of text with the same content he is willing to read. Some of these approaches are already live and some are still future talk. Coming back to the categorization, it is not only a topic for news center but for a variety of businesses or even for the private user, who wants to classify some of his personal texts.

The aim of exemple of text mining tools application in Education could be identification and assigning the news article into categories.

Research Questions

Users gain more power when it comes to advertising a product by reviewing them. A company should therefore carefully watch all the reviews they get. However, if they have too many products and getting too many reviews it is impossible to check every review one by one. So, with the help of text mining techniques it should support the companies to give them a better overview whether a review is important or not. In our opinion companies should not only check for the rating itself but for the written text, because there could be a hidden bug or feature which they could promote better afterwards. Also, we recommend not only looking at the negative reviews, but also at the positive ones, to even strengthen the good quality. So the research question is, whether it is possible to cluster reviews and gain a new insight.

Research Plan

This term paper is based on the programming language R. R is an open source development environment for statistical analysis comparable to other statistical software packages such as MATLAB, the SAS Enterprise Miner or SPSS Statistics. It

is based on its own scripting language, which is optimized for mathematical calculations. R allows you to load records from many data sources, transform them, and then examine them. Insights gained in this way are valuable and can often be developed into predictive models. R also provides a set of domain-specific extensions for special statistical methods or visualizations.

The steps of Research plan are following:

1. Find/Get reliable data for processing our text mining techniques. We gathered data from the BBC is a (online) news center. We decided to take 2225 news, which are categorized into 5 categories. We will go into more detail about the data in the result chapter.
2. Our second step is to preprocess the gathered data in Excel.
3. The next step is to gather more information about the corpus, which we get with the frequency of words. These are defined by the document term matrix (DTM) and the term document matrix (TDM). With this technique we can get an idea of high frequently words and words which are only used rarely. Also, the TF-IDF Transformation is giving us more information about the importance of terms within the corpus.
4. The fourth step is to cluster the news with several algorithms. We will describe the steps more detailed in the results of the experiment chapter and compare each algorithm by their results and mention their advantages and disadvantages.
5. At the end we will give a conclusion.

Results of the experiment

For providing the experiment were took news articles from BBC, which are already labeled into five categories. The following five categories are listed: business; entertainment; politics; sports; tech.

The dataset consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005. Source: http://mlg.ucd.ie/files/datasets/bbc-fulltext.zip.

Firstly, it is possible to split the files into 300 files per class label, which results in a total dataset of 1500 files which we will cluster with all the learned cluster algorithms. After that we renamed all the files to each of the label (to have a better visual comprehension of the clustering plotted images). For example, all the labeled text files have a name like "business (1)", "business (2)", "business (3)" etc. (Figure 12)

Business (18).txt	09.12.2018 11:40	Textdokument	1 KB
Business (19).txt	09.12.2018 11:40	Textdokument	2 KB
Business (20).txt	09.12.2018 11:40	Textdokument	4 KB
entertainment (1).txt	09.12.2018 11:40	Textdokument	2 KB
entertainment (2).txt	09.12.2018 11:40	Textdokument	2 KB
entertainment (3).txt	09.12.2018 11:40	Textdokument	2 KB
entertainment (4).txt	09.12.2018 11:40	Textdokument	2 KB
entertainment (5).txt	09.12.2018 11:40	Textdokument	2 KB

Fig. 12. File names

The corpus has a 100×4346 dimension matrix with a sparsity of 97%, which means that 409476 rows have zeros. This means that the files are not very similar in their content (choice of words), which can be seen in the extract in Figure 12. There you can see an example of the zeros for each row.

Just from that data, we cannot really say something about the corpus, but we are eager to see where this will take us. The tail means, that these words are only used once. We also tried to check the association between some words (Figure 13), we picked some words, which we would take as a representation for the label:

- world – politics;
- music – entertainment;
- sport – sport.

```
> findAssocs(dtmr,"world",0.5)
$`world`
 champion medallist      mark    indoor    silver   finland  helsinki   lewisfr
     0.60      0.60      0.57      0.53      0.53      0.52      0.52      0.52
    relay   fastest     olymp
     0.52      0.51      0.50

> findAssocs(dtmr,"music",0.5)
$`music`
   wonder     spend     stori     frank     opera treatment     stage      turn
     0.68      0.64      0.63      0.60      0.60      0.60      0.55      0.52

> findAssocs(dtmr,"sport",0.5)
$`sport`
  iaaf athlet    dope
  0.60   0.54    0.53
```

Fig. 13. Association of words

The word "world" is most associated with "champion" which comes from "world champion" which would be classified into the sport sector not the politics sector, how we "predicted". However, the word "world" is mostly a universal word, which occur in most of the labels, that is why, it is the second most frequent word in the corpus. To get a better visual, immediate understanding of the used words in the corpus, we plotted a world cloud with three colors. "Year" is the most frequently used word. The same explanation like for the word "world" it is a universal word, which will be used for all of the categories.

A dendrogram shows a hierarchical clustering, it illustrates the arrangement of the clusters produced by the corresponding analyses. Well, as one can see in Figure 14, we clustered the corpus into 5 parts, but we do not get a good result, because in our example there is no hierarchical possibility to cluster the files. It is not like some files are built on others. So, that is why we get a bad result. The advantage of the dendrogram is to see some hierarchical cluster, mostly the cluster algorithm is used for chemical or biological studies to see the relationship between some elements or species. For clustering "similarities" the algorithm is not very effective.

We chose to create a table with the clustered values (Table 1), because just from the Figure 14 it is very difficult to see which files are in which cluster.
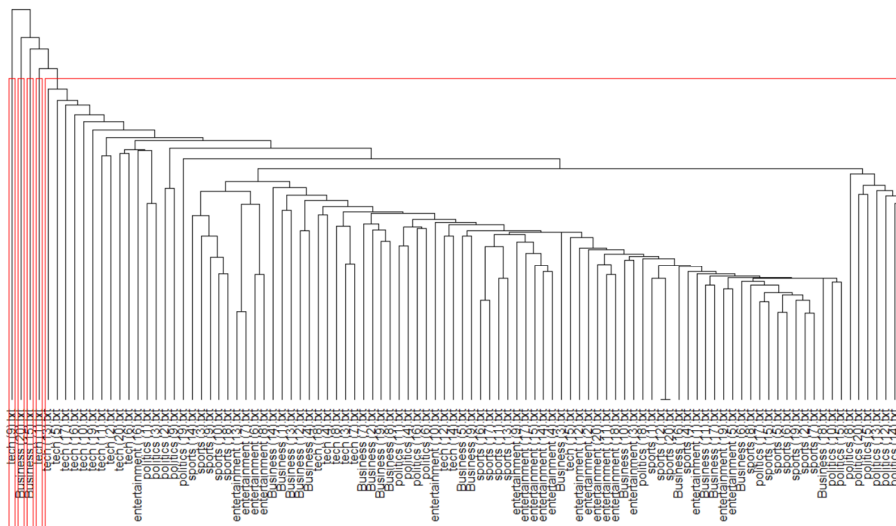
Fig. 14. Cluster dendrogram

Table 1. Clastering results table

	1	2	3	4	5
Business	1	11	0	0	8
entertainment	0	14	0	0	6
politics	0	6	0	7	7
sports	0	16	0	0	4
tech	0	5	4	8	3

The table shows that most of the files are located into the second cluster, followed by cluster 5. All the topics are scrambled into all the cluster (Figure 15). So, the k-means algorithm is also not a good algorithm for our case. The disadvantage for the k-means algorithm is that it is not good with outliers, and if there are some outlier, they will be put into one cluster like business 20, or some of the tech articles in cluster 3. However, the other texts are likely to be "too similar" to distinguish them from the others.

Community Detection

Since we have a lot of files, it is not our goal to see the "biggest" influenceable hub for every sub community but see how the communities are detected. First, we have built the communities itself, like seen in Figure 16. Alone from that one can see a clear cluster like the sport-cluster in the bottom left corner or entertainment in the upper right corner. The community detection is based on the cosine similarity, as we described in the topic Document similarity.

Now we can build the community detection, which is based on the algorithm greedy optimization of modularity. Figure 17 shows clearly that there are more than 5 sub-communities. There is already the first disadvantage, it is very difficult to find the best attributes in order to get exactly 5 cluster. However, in our opinion the community
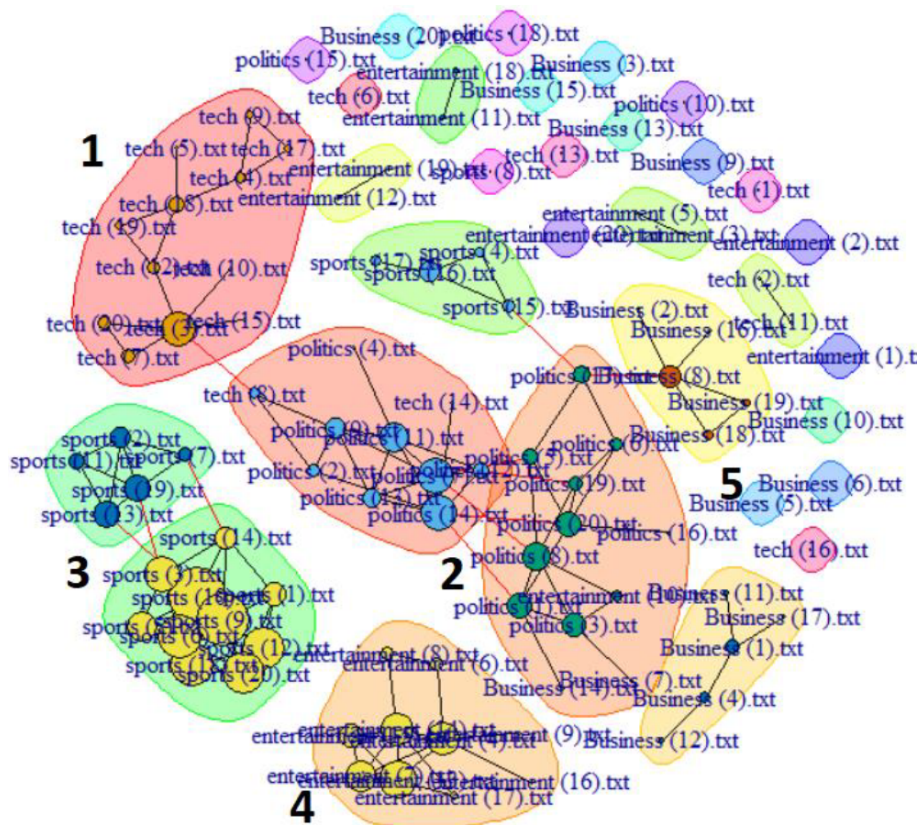
detection did cluster the files very good. We added 5 numbers for every cluster, we will focus on.

Fig. 15. K-Means plot

Fig. 16. Cosine similarity-based graph

Fig. 17. Community detection results

Cluster 1 is mostly having tech files in them. Cluster 2 is politics and a little business and tech; they are definitely in a relationship. Business and politics are areas which are not that different, so it is obvious that they have a connection to each other. Sports on the other hand, cluster 3, is not in a relationship with other news areas, but it is split into 2 and a half sub communities (green). Only two of them are having a strong tie. Entertainment (cluster 4) is similar to sports, they don't have a strong bond to other news categories, but not all of the entertainment news are within this sub-community. The last cluster (5) is business, which only has a low number of files in them. Also, there are several files which have a "single"-degree number, they must be outliers or do not have much in common with other files.

Latent Semantic Analysis (LSA)

As one can already see in the two-dimensional graph in Figures 18 and 19 LSA, the clustering is not very good, even when using the k-means to get a better visual understanding, the results are not getting better.

We only made some file names visible, because otherwise it would get too chaotic. We predefined 5 cluster, as always, but won't get a good result. We processed the data in Excel and got the Table 2. There you can clearly see that most of the files are

clustered into community 5 and 4, there is no real distinguish. Only some outliers are into another category. We assume that it is difficult to predefine 5 clusters, if we test the attributes with a lower cluster size, we assume that the algorithm is getting better. However, the result is useless for our case.
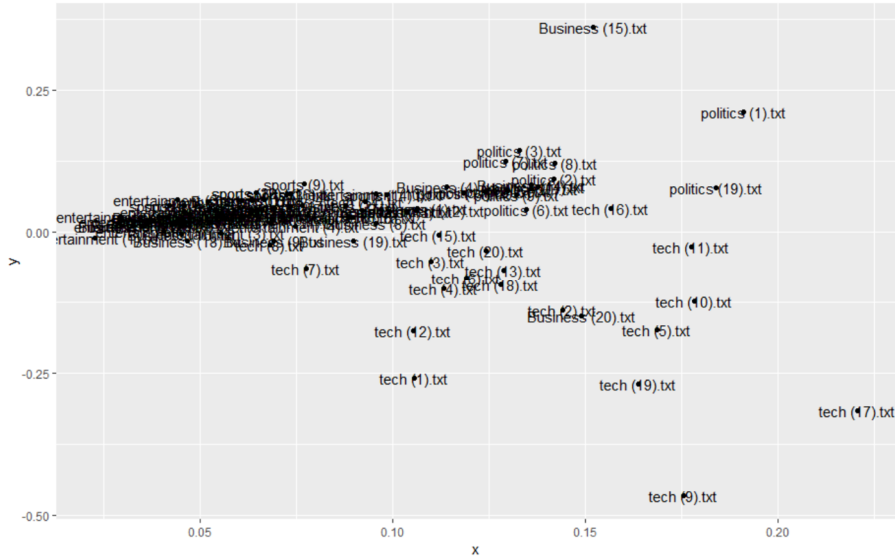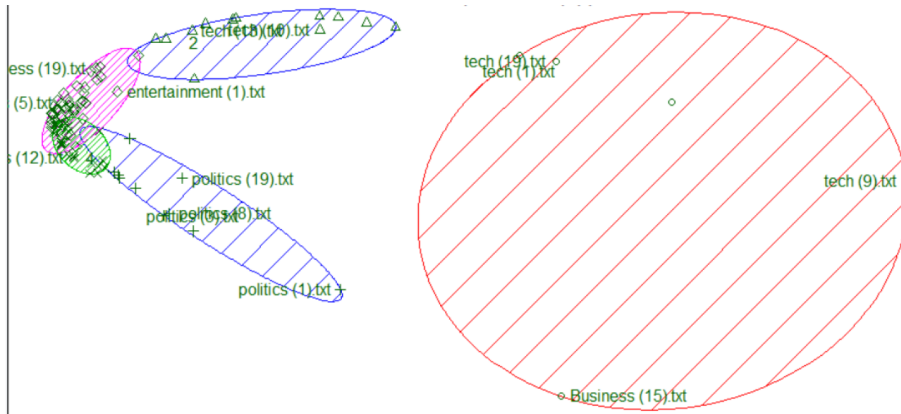

Fig. 18. LSA results

Fig. 19. LSA clustering results

Topic Modeling

Like the other algorithm, we predefined 5 categories, and try to assign the topics, which are seen in Figure 20, based on the most important words for each news category.

Table 2. LSA Result Table

	1	2	3	4	5
Business	1	1	1	6	11
entertainment	0	0	0	2	18
politics	0	0	8	6	6
sports	0	0	0	10	10
tech	4	11	1	1	3

1. Topic 1 is most likely to be the tech category, because we have "phone", "user" "call" those words are most likely to be used when writing a news article in the tech sector.
2. As already seen before the word "world" in topic 2 is most associated with champion which is only relevant for the category sport, also time and last could be very clear terms for sport.
3. Topic 3 is easy, because we have terms like "book", "film", "award" which clearly matches with entertainment.
4. Topic 4 and 5 are not so easy to separate because business and politics are also quite similar when using specific terms. In topic 5 we have words like "elect" and "parti" which is relevant for the election and can only be matched with politics, so topic 4 must be the business category.

```
> ldaOut.terms
     Topic 1  Topic 2    Topic 3   Topic 4    Topic 5
[1,] "call"   "world"    "includ"  "month"    "new"
[2,] "peopl"  "last"     "book"    "govern"   "women"
[3,] "now"    "time"     "film"    "year"     "elect"
[4,] "make"   "take"     "award"   "compani"  "parti"
[5,] "phone"  "european" "year"    "plan"     "blair"
[6,] "user"   "set"      "show"    "expect"   "email"
> |
```

Fig. 20. Topic model terms

After processing the values in excel, we get Table 3. We can clearly see that our prediction is quite similar to the outcome. Even our trouble with topic 4 and topic 5 is seen here, some of the politics files are assigned to topic 4 (in the politics category). Apart from that difficulty the algorithm did a great job clustering all the files in the exact appropriate category.

We also created an error rate table, where one can see that the algorithm is quite accuracy, if we exclude the politics/business similarity (Table 4).

Table 3. Topic model result table

	1	2	3	4	5
Business	1	0	0	18	1
entertainment	0	0	19	0	1
politics	0	1	0	6	13
sports	0	20	0	0	0
tech	16	0	1	0	3

Table 4. Topic model error rate

	Business	entertainment	politics	sports	tech
Errors	2	1	7	0	4
Error Rate	10%	5%	35%	0%	20%

Comparison

Like already described in each algorithm, we saw that there is a great difference in the accuracy for each algorithm. Some resulted in a not so good cluster, but others were quite good. The Table 5 shows a summary of the results with some disadvantages and advantages for our case. For other cases it could be that there are other advantages and disadvantages.

Table 5. Summary of the research results

Algorithm	Result	Advantages	Disadvantages
Dendrogram	Not good	• Easy to understand/ read when having few files	• no existing hierarchical files, unsuitable • when having a lot of files, hard to read
K-Means	Not good	• Easy to understand/ read • Fast algorithm • efficient	• Starting configuration is crucial • not good with outliers • different size of cluster
Community Detection	Quite good	• easy visualization • easy to read • good for specific clustering	• cannot specify the cluster size • difficult to change attributes • no breakdown in R for every cluster
LSA	Not good	• fast algorithm • not sensitive for starting configuration	• Representation is dense • distributional model, not efficient
Topic Modeling	Very good	• NLP-model • tailored for text mining	• only good for large texts • static

Topic Model and Community detection were the best algorithm for our case. For the Topic Modelling it is not surprising because it is a machine learning and natural language processing mode, which cluster the document with the help of statistical model for discovering the abstract "topics" that occur in them. This is the only model which is tailored for text mining usages.

5   Conclusions

Machine Learning techniques are getting more popular in the recent year, even artificial intelligence is getting more attention each year. Governments and businesses investing billions into the area. However, text mining and NLP which are both a part of the area seem to get overlooked. We want to emphasize that text mining is a powerful field, which is far from perfect, and could bear more attention than artificial intelligence. With the growing online news turnover reducing the normal paper, mostly every news will be written online, which means that they are easily analyzed by text mining algorithms. So even news centers are forced to deal with text mining when it comes to learning something about their writing style. Although the results are still taking a lot of human interaction, and the cluster algorithm are still not the best, it is necessary of the companies to invest more in that area, so they get an advantage towards their competition. Our clustering model was partly successful but could be even better. However, we are still in a learning phase and think, with more time and a better model we would be able to get a better accuracy than guessing.

As mentioned in the comparison it is crucial to use a suitable algorithm for the given case. We cannot use a dendrogram to cluster the category, because the purpose is not relevant for that issue. However, if we take the topic model, we can definitely come to a great cluster. The topic model is very good with large texts, but not very good with small texts like reviews or tweets, so you also have to be careful what kind of texts you have. When the texts are quite similar it is always difficult for a machine to differentiate them, at least for now. We think that the NLP model will become better and better, because there are a lot of possible application for text mining. When we look at the impact on social media, before an election or when a company wants to check their reviews by users, there are plenty of application possibilities for text mining, therefore algorithms are getting better and more precise.

6   Future research directions

Described approaches and algorithms are just some basic for business processes modeling, which could be applied to solve the different decision-making problems. There are multiple examples of how all these methods could be used in education. For example, with time series analysis we could predict future demand for tickets, using regression models we could determine the loyalty of the customers and so on.

Nowadays there are much more algorithms, which could be applied in this area. Like complicated non-liners algorithm for regression predictions. As an example, it could be

a random forest, XGBoost, neural networks. With such method, we could build models for maintenance prediction, what is very crucial in education.

Education should correspond to the modern development of the digital economy, digital society, innovation, and creative entrepreneurship. The use of data science in education should be of a multiplatform nature, that is, to be used not only in the study of a subject, but in the training of all subjects, the interaction of students with each other and with teachers, real experts, research, and individual learning.

References

1. Blei, D.M. Probabilistic Topic Models. Communications of the ACM **55**(4), 77–84 (2012). doi:10.1145/2133806.2133826
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**, 993–1022. http://jmlr.org/papers/v3/blei03a.html (2003)
3. Brunner, R.J., Kim, E.J. Teaching data science. Procedia Computer Science **80**, 1947–1956 (2016). doi:10.1016/j.procs.2016.05
4. Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS quarterly: Management Information Systems **36**(4), 1165–1188 (2012)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science **41**(6), 391–407 (1990). doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
6. Deerwester, S.C., Dumais, S.T., Furnas, G.W., Harshman, R.A., Landauer, T.K., Lochbaum, K.E., Streeter, L.A.: Computer information retrieval using latent semantic structure. US Patent 4,839,853, 13 June 1989
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1), 1–38 (1977)
8. George, G., Osinga, E.C., Lavie, D., Scott, B.A.: Big data and data science methods for management research: From the Editors. Academy of Management Journal **59**(5), 1493–1507 (2016). doi:10.5465/amj.2016.4005
9. Hnot, T.V., Nehrey, M.V.: Alhorytmy Data Science u modeliuvanni biznes-protsesiv (Data Science Algorithms in Modeling Business Processes). Ekonomika i suspilstvo 12, 743–751 (2017)
10. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Application in R. Springer, New York (2013)
11. Kiv, A., Semerikov, S., Soloviev, V., Kibalnyk, L., Danylchuk, H., Matviychuk, A.: Experimental Economics and Machine Learning for Prediction of Emergent Economy Dynamics. In: Kiv, A., Semerikov, S., Soloviev, V., Kibalnyk, L., Danylchuk, H., Matviychuk, A. (eds.) Experimental Economics and Machine Learning for Prediction of Emergent Economy Dynamics, Proceedings of the Selected Papers of the 8th International Conference on Monitoring, Modeling & Management of Emergent Economy (M3E2 2019), Odessa, Ukraine, May 22-24, 2019. CEUR Workshop Proceedings **2422**, 1–4. http://ceur-ws.org/Vol-2422/paper00.pdf (2019). Accessed 1 Aug 2019
12. Kucherov, D.P.: The Synthesis of Adaptive Terminal Control Algorithm for Inertial Secondary Order System with Bounded Noises. Journal of Automation and Information Sciences **39**(9), 16–25 (2007). doi:10.1615/JAutomatInfScien.v39.i9.20
13. Nehrey, M., Hnot, T.: Data Science Tools Application for Business Processes Modelling in Aviation. In: Shmelova, T., Sikirda, Yu., Rizun, N., Kucherov, D. (eds.) Cases on Modern

Computer Systems in Aviation, 176-190. IGI Global, Hershey (2019). doi:10.4018/978-1-5225-7588-7.ch006

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

15. Nokel, M.A., Loukachevitch, N.V.: Tematicheskie modeli: dobavlenie bigramm i uchet skhodstva mezhdu unigrammami i bigrammami (Topic models: adding bigrams and taking account of the similarity between unigrams and bigrams). Computational methods and programming **16**(2), 215–234 (2015). doi:10.26089/NumMet.v16r222

16. Parish, E.J., Duraisamy, K.: A paradigm for data-driven predictive modeling using field inversion and machine learning. Journal of Computational Physics **305**, 758–774 (2016). doi:10.1016/j.jcp.2015.11.012

17. Patriarca, R., Di Gravio, G., Costantino, F.: A Monte Carlo evolution of the Functional Resonance Analysis Method (FRAM) to assess performance variability in complex systems. Safety science **91**, 49-60 (2017). doi:10.1016/j.ssci.2016.07.016

18. Périaux, J., Chen, H.Q., Mantel, B., Sefrioui, M., Sui, H.T.: Combining game theory and genetic algorithms with application to DDM-nozzle optimization problems. Finite elements in analysis and design **37**(5), 417–429 (2001). doi:10.1016/S0168-874X(00)00055-X

19. Rizun, N., Shmelova, T.: Decision-Making Models of the Human-Operator as an Element of the Socio-Technical Systems. In: Batko, R., Szopa, A. (eds.) Strategic Imperatives and Core Competencies in the Era of Robotics and Artificial Intelligence, pp. 167–204. IGI Global, Hershey (2017). doi:10.4018/978-1-5225-1656-9.ch009

20. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM **18**(11), 613–620 (1975)

21. Semerikov, S.O., Teplytskyi, I.O., Yechkalo, Yu.V., Kiv, A.E.: Computer Simulation of Neural Networks Using Spreadsheets: The Dawn of the Age of Camelot. In: Kiv, A.E., Soloviev, V.N. (eds.) Proceedings of the 1st International Workshop on Augmented Reality in Education (AREdu 2018), Kryvyi Rih, Ukraine, October 2, 2018. CEUR Workshop Proceedings **2257**, 122–147. http://ceur-ws.org/Vol-2257/paper14.pdf (2018). Accessed 30 Nov 2018

22. Semerikov, S.O., Teplytskyi, I.O., Yechkalo, Yu.V., Markova, O.M., Soloviev, V.N., Kiv, A.E.: Computer Simulation of Neural Networks Using Spreadsheets: Dr. Anderson, Welcome Back. In: Ermolayev, V., Mallet, F., Yakovyna, V., Kharchenko, V., Kobets, V., Korniłowicz, A., Kravtsov, H., Nikitchenko, M., Semerikov, S., Spivakovsky, A. (eds.) Proceedings of the 15th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer (ICTERI, 2019), Kherson, Ukraine, June 12-15 2019, vol. II: Workshops. CEUR Workshop Proceedings **2393**, 833–848. http://ceur-ws.org/Vol-2393/paper_348.pdf (2019). Accessed 30 Jun 2019

23. Shoro, A.G., Soomro, T.R.: Big Data Analysis: Apache Spark Perspective. Global Journal of Computer Science and Technology **15**(1-C). https://computerresearch.org/index.php/computer/article/view/1137 (2015)

24. Vorontsov, K.V., Potapenko, A.A.: Modifikatcii EM-algoritma dlia veroiatnostnogo tematicheskogo modelirovaniia (EM-like algorithms for probabilistic topic modeling). Machine Learning and Data Analysis **1**(6), 657–686 (2013)

25. Xiong, J., Yu, G., Zhang, X.: Research on Governance Structure of Big Data of Civil Aviation. Journal of Computer and Communications **5**(5), 112–118 (2017). doi:10.4236/jcc.2017.55009