information

*Article*

# School in Digital Age: How Big Data Help to Transform the Curriculum

**Svetlana Nikolaevna Vachkova [1], Elena Yurevna Petryaeva [1], Roman B. Kupriyanov [2,*] and Ruslan S. Suleymanov [2]**

[1] Institute of Systems Projects, Moscow City University, 4 Vtoroy Selskohoziajstvenny proezd, 129226 Moscow, Russia; svachkova@mgpu.ru (S.N.V.); PetryaevaEYU@mgpu.ru (E.Y.P.)
[2] Information Technology Department, Moscow City University, 4 Vtoroy Selskohoziajstvenny proezd, 129226 Moscow, Russia; sulejmanovRS@mgpu.ru
* Correspondence: kupriyanovrb@mgpu.ru

**Abstract:** The transition to digital society is characterised by the development of new methods and tools for big data processing. New technologies have a substantial impact on the education sector. The article represents the results of applying big data to analyse and transform the learning content of Moscow's schools. The analysis of the school curriculum comprised the following: (a) identifying one-topic lesson scripts, (b) analysing cross-disciplinary connections between subjects, (c) verifying the compliance of the lesson script digital content to the Federal Educational Standards. The analysed material included 36,644 lesson scripts. The analysis has been conducted using specifically designed digital tools featuring data mining algorithms. The article considers the issue of applying data mining algorithms to analyse school curriculum for the improvement of its quality.

**Keywords:** Moscow electronic school; artificial intelligence; repository; text mining; educational data mining; learning analytics



**Citation:** Vachkova, S.N.; Petryaeva, E.Y.; Kupriyanov, R.B.; Suleymanov, R.S. School in Digital Age: How Big Data Help to Transform the Curriculum. *Information* **2021**, *12*, 33. https://doi.org/10.3390/info12010033

Received: 22 December 2020
Accepted: 12 January 2021
Published: 15 January 2021



## 1. Introduction

The transition to digital society is characterized by the development of new methods and tools for big data processing. New technologies have a substantial impact on the education sector. The current transformation of education is characterized by the application of data mining algorithms and big data technologies to the digital learning materials stored in the learning repositories and learning management systems. For the education system of Moscow, the Moscow Electronic School project (MES) was designed and implemented—a digital management system for the city's school education. By the spring of 2020, the MES offered access to 1,638,275 students, 72,492 educators, and 1,648,362 parents of students of different regions of Russia, including Moscow.

The MES system consists of the electronic diary and the electronic library. The electronic diary is a digital tool that allows teachers, parents and students to obtain such information about the organization of a student's life at school as the curriculum of each academic subject, lesson schedule, extracurricular activities, schedule of additional classes, the progress and results of educational activities (assessments in subjects, topic, teacher comments), analysis of the dynamics of academic performance in the subject and etc. The electronic library is a digital repository of learning objects.

The MES is a unique repository of learning objects open to every teacher, student, and parent. By May 2020, the MES contained 45,321 electronic study guides, 154,311 applications, 167,227 tests, 1,413,646 lesson scripts, as well as several millions of atomic contents.

The process of creating new digital content by Moscow's teacher community is continuous and the volume of the uploaded data is increased daily. This is one of the outcomes of the digital transformation of education. Although a contemporary school education is

not very different from the school education of the analogue age, it may be assumed that for the first time in the Russian education science, the researchers have obtained access to big datasets of learning content.

During recent years, Moscow City University have been engaged in research using the MES data. The issues of differentiation and segmentation of electronic learning resources collections have been described in the article "The Frontiers of the Moscow Electronic School" (Remorenko and Grinshkun) in 2017 [1]. In 2018–2019, research has been conducted on the causes, explaining high demand for lesson scripts and their quality [2,3], network engagement of teachers with E-learning objects stored in the MES [4], search of data in the MES [5,6], development and design of materials for different subjects in the MES, including materials for special needs students [7–13], training of pre-service teachers [14–18], and selection and updating of learning content [19,20]. Big data and its impact on the transformation of business processes and structure in the education sphere have been explained in the article "Building a learning culture for the digital world: lessons from Moscow" by A. Schleicher [21]. While Schleicher describes the use of big data in education in general, in our work, we go further by applying text mining methods to analyse the curriculum of school education and show how the analysis of a large sets of data helps to transform the curriculum of school education through solving practical problems that previously could not be solved before new technologies appeared.

Text mining methods are often used to solve information retrieval problems and, in particular, to cluster scientific articles [22]. There are also examples of the use of text mining methods to the curriculum selection development and validation [23,24]. However, these studies are characterized by the use of data sets limited to a small number of subjects or represented by data from only one educational organization or subject area.

Today, many methods of semantic analysis of texts are known. According to [24–26], the best results are shown by methods based on neural networks.

The MES data is one of the elements of the post-industrial education as a cultural space that has been defined by I. Levin as "Data Intensive Science", along with social media and personality online [27–30]. In the age of ubiquitous access to all kinds of data, it is crucial to conduct research experiments with data and apply new technologies to inform science with larger amounts of data. By using the data "mirror", we can look into the actual embodiment of subject content that is used for teaching and learning at school, analyse its compliance to the Federal State Educational Standards (FSES), discover the existence of cross-disciplinary connections in the learning content of various school subjects, etc.

The purpose of this research is to study the results and prospects of using big data technologies, in particular text mining methods and approaches to transform the curriculum of school education by the example of solving the practical problem of comparing the curriculum of school education in Moscow with a thematic classifier—the topical framework. The novelty of the research is the use of well-established text mining methods for the study of the curriculum of school education on the city scale. Moreover, this is the first example of applying text mining methods to MES educational data.

## 2. Materials and Methods

The research focused on the most popular type of digital content in the MES—lesson scripts. A lesson script describes the content and the course of a lesson of any subject in the electronic format. A lesson script may include interactive tasks, schemes, maps, videoclips, tests, etc. Lesson scripts are produced by teachers and used to conduct in-person and online lessons and classes. When designing a lesson script, the author must specify its title, subject, grade level, education level (primary general, basic general, secondary general education), give a short description, create and upload content for all lesson stages on three screens: teacher's, students' and common screen. Each lesson script has its own ID.

Since June 2020, every new lesson script is assigned to a topic—a didactic item in the topical framework. The topical framework is a unified structured classifier of learning programmes that is based on the FSES and includes topics and didactic items (elements

of learning content). The topical framework was introduced into the MES in May 2020 to solve the issues of structuring learning content and eliminating low-quality learning materials. However, there is still the problem of mapping previously created lesson scripts to the thematic framework. Moreover, as the thematic framework changes dynamically, the problem of regularly comparing new didactic items of the thematic framework to the existing lesson scripts in the MES database remains actual.

A teacher, author of a lesson script, can submit a lesson script for moderation. Moderation is a process of confirming the compliance of a lesson script with the FSES, the current scientific vision, the federal educational legislation, as well as checking for grammar, stylistic or factual mistakes. The lesson scripts that have undergone moderation are uploaded in the MES public space. These lesson scripts are open to all registered users and can be copied, added to favourites, launched or rated.

At the time of the analysis, there were 1,413,646 lesson scripts in the MES e-library, 44,527 of which have undergone moderation. In order to solve the problem of comparing didactic items and titles of lesson scripts the lesson scripts with unique titles have been selected. In total, 36,644 lesson scripts have been selected for the analysis.

The analysis was conducted on the raw data. Tables 1 and 2 show the examples of raw data used in the research. Table 1 shows the data describing the lesson scripts. The data include the unique number of the lesson script in the MES, the subject that the lesson script belongs to, the level of education that the lesson script belongs to, and the name of the lesson script. Table 2 shows semantic data related to the thematic framework: the level of education, the subject, the topic and the didactic item.

**Table 1.** Example of the data assigned to a lesson script in the analysis.

| Script_ID | Subject | Education Level | Title |
|---|---|---|---|
| 34105 | Environment | Primary general education | Geographic map and plan |
| 34147 | Mathematics | Primary general education | Numbers from 10 to 20. Grade 1. Lesson 1. |
| 34214 | Environment | Primary general education | Land navigation. Compass |
| 286028 | Russian language | Primary general education | Consolidating knowledge on noun cases. Grade 3 |
| 60449 | Reading of literature | Primary general education | 4 gr._V.F.Odoyevsky “A town in a tobacco box”. 4 grade |

**Table 2.** Example of the data interconnected in the topical framework in the analysis.

| Education Level | Subject | Topic | Didactic Item |
|---|---|---|---|
| Primary general education | Environment | Human and society | I am primary school student |
| Primary general education | Environment | Human and society | I and my family |
| Primary general education | Environment | Human and society | Rules of safe conduct |
| Primary general education | Russian language | Spelling | Hyphenation |
| Primary general education | Russian language | Spelling | Capital (uppercase) letter at the beginning of a sentence |

To compare the lesson scripts and the thematic framework, we had to solve the problem of mapping each lesson script to a relevant didactic item.

Based on the available data, we developed a text mining model and an algorithm that were both aimed at analysing the data and extracting one-topic lesson scripts and didactic items, as well as providing visualization of the results in two-dimensional space.

The text mining process comprised three steps in working with the data:

1. Pre-processing of data;
2. Defining the proximity of meanings of didactic items and lesson script titles with regard to each other;
3. Visualizing the results of comparing didactic items and lesson script titles.

Step 1. Pre-processing of data

The step of data pre-processing included eliminating irrelevant data and making consistent changes in the data display, such as making the letter cases uniform, removing conjunctions and prepositions, removing punctuation marks such as commas, full stops, dashes, etc.

Step 2. Defining the proximity of meanings of didactic items and lesson script titles with regard to each other

At the second step, to define the proximity of meanings of didactic items and lesson script titles with regard to each other, we used the word2vec method [31,32]. At this step, the artificial neural network was trained on a large text corpus consisting of lesson script titles and didactic item titles.

Step 3. Visualizing the results of comparing didactic items and lesson script titles

The third step included visualizing the results of comparing didactic items and lesson script titles in a convenient mode for further evaluation by experts. We decided to visualize the data in the form of two-dimensional graphs, where the distance between data points reflected the semantic proximity of lesson scripts and didactic items with regard to lesson scripts. However, at the second step, we obtained word embeddings within the 50-dimensional space. Therefore, for adequate visualization of the data, the number of embedding dimensions had to be reduced while maintaining the distance between the embeddings. To do this, we used the algorithm t-distributed stochastic neighbour embedding (t-SNE) [33].

After the data had been processed by the text mining algorithms, we suggested two hypotheses.

The first hypothesis stated that one-topic lesson scripts must have similar titles, and the algorithm can group them into one-topic semantic clusters. This would allow defining topical correlations among school subjects and topics that generate a large amount of lesson scripts.

The second hypothesis stated that the algorithm can generate semantic clusters of lesson script titles according to their proximity to the titles of didactic items within the topical framework. Thus, we would verify the compliance of lesson scripts with the learning content defined by the FSES.

To test the hypotheses, we conducted the text mining of lesson script titles, and the comparative analysis of lesson titles and didactic items for the subjects of the primary general and basic general education.

The research has its limitations. On the one hand, the research was conducted using only the titles of lesson scripts and didactic items. Other types of semantic data included in the lesson scripts were not analysed. Therefore, we assume that the expansion of semantic data i.e., the inclusion of all content of lesson scripts into the analysed material at the following stages of the research, might alter the obtained results. On the other hand, this research was the first attempt to apply text mining algorithms to solve the task of analysing learning content uploaded in the MES. The important aspect at this stage was to evaluate the performance quality of the algorithm.

The evaluation of the algorithm performance was conducted by expert evaluations. The expert community consisted of the faculty members of Moscow City University with expertise in the subjects of primary general and basic general school levels, including Russian, Literature, English, French, German, Mathematics, Algebra, Geometry, Computer Science, History, Social Studies, Geography, Chemistry, Biology, Physics, Physical Education, Music, Fine Arts, Environment. The experts were asked to assess the titles of lesson scripts and their correlation with the didactic items within the semantic clusters that were generated by the algorithm, as well as to indicate the typical mistakes in the algorithm's patterns for further improvement of its performance.

## 3. Results and Discussion

As a result of applying the algorithm, we obtained a vector representation of each lesson script and each didactic item. We also mapped each lesson script to the closest didactic item defined by the cosine similarity:

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

where $A$—vector representation of lesson script, $B$—vector representation of didactic item, $n$—set of real numbers.

Figures 1 and 2 show visualization for 36,644 lesson scripts in the MES, based on the vector representation of each lesson script. Every school subject is shown by a different colour. The stars on the graph show the lesson scripts that have been awarded with grants. The Figures show that most of the lesson scripts are clustered by the school subjects that they are connected with. However, a number of lesson scripts were not included in subject clusters which, we assumed, were cross-subject (cross-disciplinary) lesson scripts.

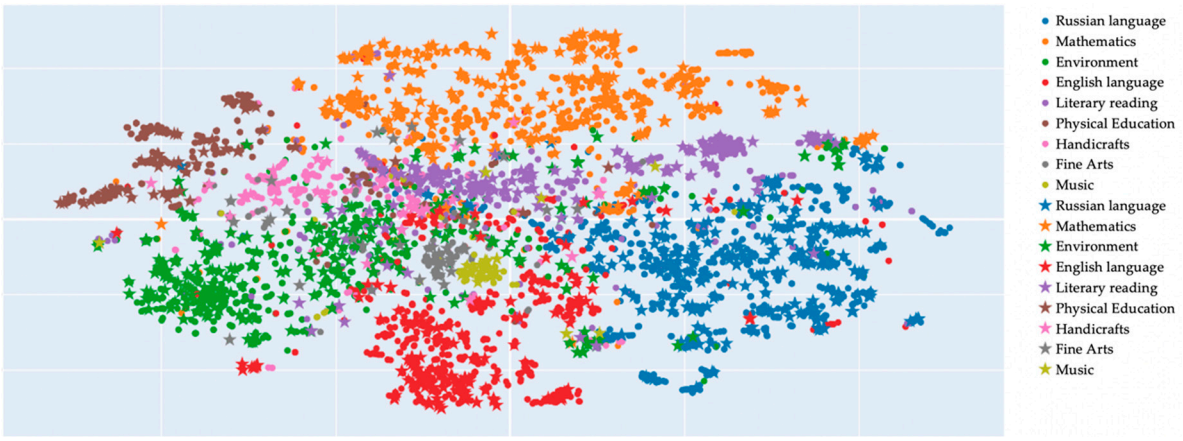


**Figure 1.** Visualization of lesson scripts in the MES by subject topic proximity. Primary general education.

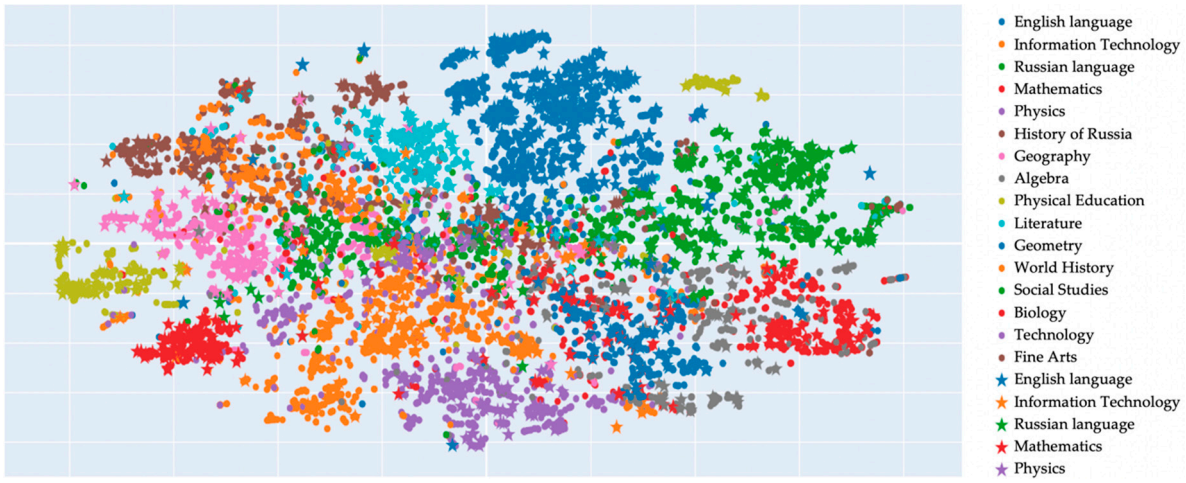


**Figure 2.** Visualization of lesson scripts in the MES by subject topic proximity. Basic general education.

The analysis of the data revealed the following cross-disciplinary topics at the level of primary general education: "The Great Patriotic War", "The surrounding environment", "Letters and sounds", "Regional studies: Moscow". The basic general level comprised such cross-disciplinary topics as "Plant Classification", "Poems about Moscow. Marina

Tsvetayeva", "Neurological technologies", "Healthy lifestyle. Diseases and illnesses", "People in big cities and their lifestyle", "Safety basics for everyday activities", "War". In the semantic cluster "Safety basics for everyday activities", there were titles of lesson scripts in the subjects Physical Education, Handicrafts, Computer Science. The semantic cluster "War" included lesson scripts in such subjects as, on the one hand, the world history and history of Russia and, on the other hand, literature and music. These lesson scripts included, for example, "War consequences: revolution and collapse of the empire", "Russia in the 1st World War", "The birth of song during war".

The comparative text mining of the lesson script titles and didactic item titles allowed obtaining the following visualizations (see examples in Figures 3–6). The visualizations are based on the vector representation of each lesson script and each didactic item. The blue dots on the graphs denote the lesson script titles, the orange dots denote didactic items in the topical framework by subject.

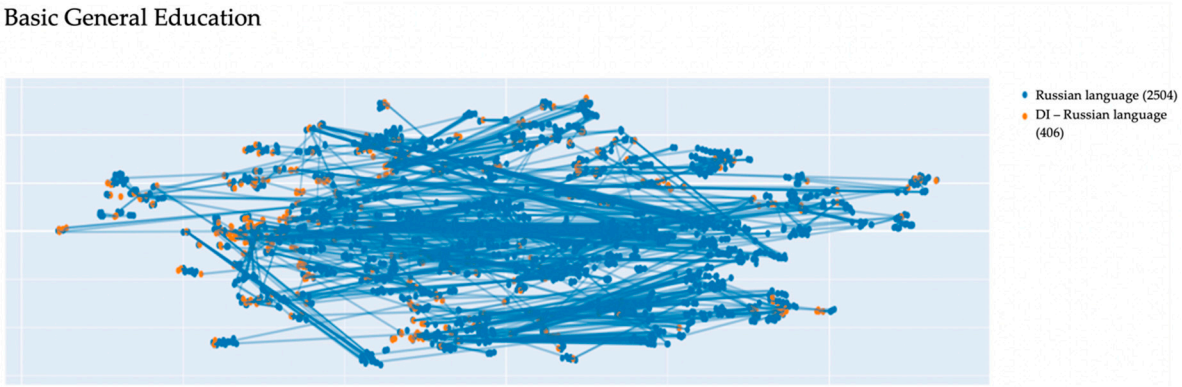


**Figure 3.** Semantic correlations between the didactic item titles in the topical framework and the lesson script titles in the subject Russian language.

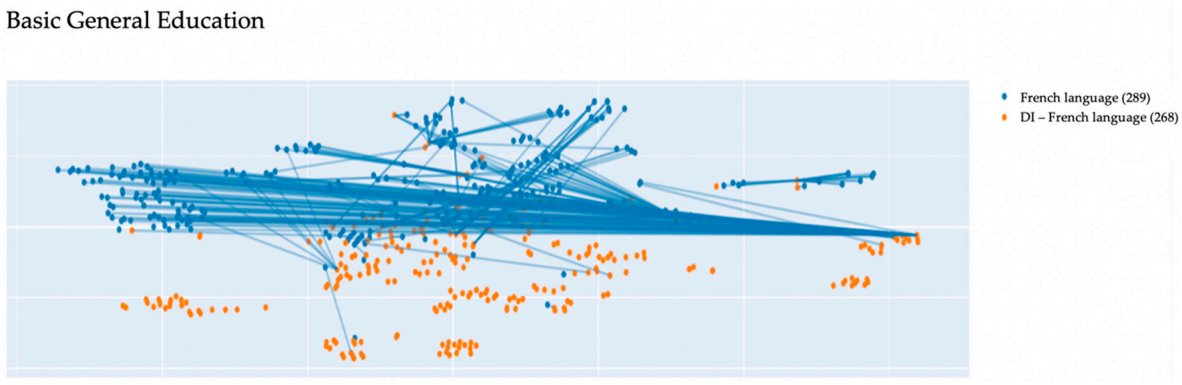


**Figure 4.** Semantic correlations between the didactic item titles in the topical framework and the lesson script titles in the subject French language.

The visualized data represent the semantic proximity of the lesson script titles and the didactic items in the topical framework by subject and the overall compliance of the lesson script dataset with the FSES, since the topical framework and its didactic items have been developed based on the current educational standards by subject.

Having obtained these results, we proceeded to evaluating the quality of the developed model of semantic data mining. As mentioned above, the quality assurance of the model was conducted by expert evaluation using the dataset of lesson script titles and didactic items in the school subjects of basic general education.

The experts were asked to evaluate the clusters generated by the algorithm using the scale 0/1 for every entry "didactic item—lesson script title". The table entries contained

the didactic item titles and the lesson script titles that had the closest proximity to each other (see example in Table 3).

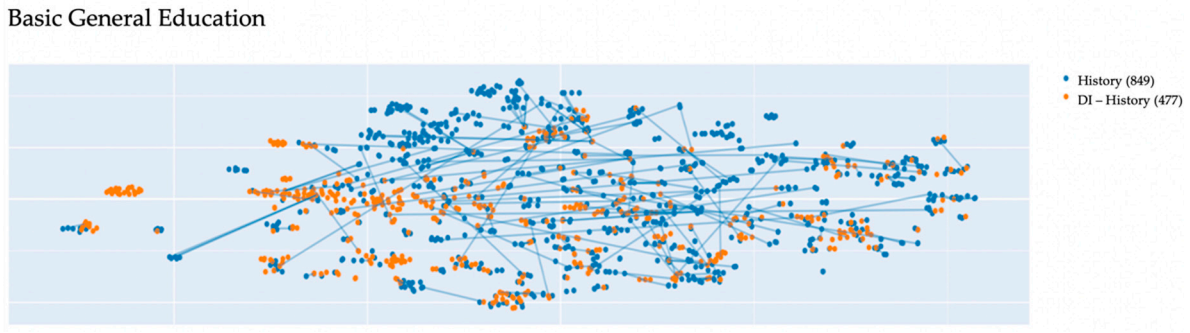


**Figure 5.** Semantic correlations between the didactic item titles in the Topical Framework and the lesson script titles in the subject History.

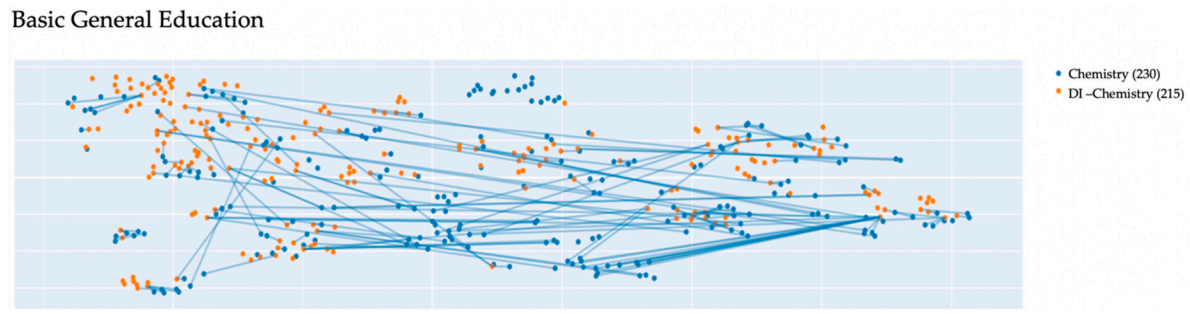


**Figure 6.** Semantic correlations between the didactic item titles in the topical framework and the lesson script titles in the subject Chemistry.

**Table 3.** Example of data mapping representation for expert evaluation.

| Didactic Item Title | Lesson Script Title | Expert Evaluation (0/1) |
|---|---|---|
| Nuclear energetics | 9-grade Physics Atomic nucleus structure | |
| Nuclear energetics | Alternative energetics | |
| Nuclear energetics | Atomic nucleus structure | |
| Barometers | Spotlight 5. Module 6. Extensive Reading. Across the Curriculum: Science | |
| Barometers | "Interconnected vessels" | |
| Brownian motion | Mechanical motion | |
| Brownian motion | Mechanical motion | |
| Brownian motion | Motion | |

We agreed that "0" would mean that the mapping is incorrect, "1" that the mapping is possible. The experts were also asked to describe the reasons of incorrect mapping by the algorithm.

As in the [25,26] studies, accuracy metric is used to measure the performance of lesson scripts to didactic item mapping. It is important to note that, according to the study [25], the accuracy of data mining methods varies from 8% to 77%, depending on the field of knowledge and the algorithm used.

The example statistics of the expert evaluation reflecting the quality of mapping the didactic item titles and the lesson script titles in the basic general education are shown in Table 4. The ratio of correct mapping by the algorithm of the didactic item titles and the lesson script titles varies from 6.01% (History) to 69.25% (Physical Education).

These results are consistent with the text mining algorithms quality reported by the other researchers [25,26].

**Table 4.** Results of expert evaluation of the semantic algorithm performance.

| N° | Subject | Total N° of Entries | Mapping Evaluation—1 | Mapping Evaluation—0 | % of Correct Mapping (Accuracy) |
|---|---|---|---|---|---|
| 1 | English language | 3041 | 584 | 2457 | 19.20 |
| 2 | Biology | 712 | 194 | 518 | 27.24 |
| 3 | Geography | 1418 | 680 | 738 | 47.95 |
| 4 | Fine Arts | 443 | 85 | 358 | 19.19 |
| 5 | Computer Science | 3075 | 264 | 2811 | 8.58 |
| 6 | History | 849 | 51 | 798 | 6.01 |
| 7 | Literature | 1759 | 142 | 1617 | 8.07 |
| 8 | Mathematics (incl. Algebra and Geometry) | 1573 | 822 | 691 | 56.07 |
| 9 | Music | 263 | 80 | 183 | 30.42 |
| 10 | German language | 519 | 335 | 184 | 64.55 |
| 11 | Social studies | 783 | 297 | 486 | 37.93 |
| 12 | Russian language | 2504 | 935 | 1569 | 37.34 |
| 13 | Physics | 1759 | 928 | 684 | 52.76 |
| 14 | Physical Education | 1262 | 874 | 388 | 69.25 |
| 15 | French language | 289 | 116 | 173 | 40.13 |
| 16 | Chemistry | 230 | 141 | 89 | 61.30 |
| **-** | **Total** | **20,479** | **6528** | **13,744** | **36.62** |

The problem of mapping lesson scripts to thematic items is actually a classification problem, where the didactic item is a class. As researchers [34], we can compare the results obtained with the results of classification, if the lesson scripts were mapped to didactic items based on random selection. This comparison is presented in Table 5.

Table 5 shows that the average accuracy of 36.62 % is significantly (7.5 times) higher than the random selection-based approach. In comparison with the random selection approach, the proposed text mining algorithm shows the best accuracy in Chemistry (57 times higher). Moreover, extremely high accuracy is shown in French language and German language (37 and 31 times higher, respectively). The high efficiency of application of the used method is achieved in disciplines, such as Mathematics, Biology, Social Studies, Physics and Music (20, 12, 11, 10 and 9 times higher, respectively). A good quality of assignment of lesson scripts to didactic items is shown in Geography and Russian language (seven and six times higher, respectively). Acceptable results can also be considered in Fine Arts, History and Physical Education (more than three times higher, then random selection). Low results were achieved in subjects such as English Language and Literature (marginally better than random selection). The worst accuracy of mapping lesson scripts to didactic items is shown in Computer Science. As it is worse than the random selection-based approach, a separate study in this discipline is required.

**Table 5.** The comparison of the proposed text mining method with random selection.

| N° | Subject | Total N° of Didactic Items (Classes) | Total N° of Lesson Scripts | Random Selection Accuracy, % | Text Mining Mapping Accuracy, % |
|---|---|---|---|---|---|
| 1 | English language | 180 | 3041 | 16.89 | 19.20 |
| 2 | Biology | 314 | 712 | 2.27 | 27.24 |
| 3 | Geography | 210 | 1418 | 6.75 | 47.95 |
| 4 | Fine Arts | 88 | 443 | 5.03 | 19.19 |
| 5 | Computer Science | 154 | 3075 | 19.96 | 8.58 |
| 6 | History | 477 | 849 | 1.78 | 6.01 |
| 7 | Literature | 328 | 1759 | 5.36 | 8.07 |
| 8 | Mathematics (incl. Algebra and Geometry) | 568 | 1573 | 2.77 | 56.07 |
| 9 | Music | 82 | 263 | 3.2 | 30.42 |
| 10 | German language | 254 | 519 | 2.04 | 64.55 |
| 11 | Social studies | 240 | 783 | 3.26 | 37.93 |
| 12 | Russian language | 406 | 2504 | 6.17 | 37.34 |
| 13 | Physics | 357 | 1759 | 4.93 | 52.76 |
| 14 | Physical Education | 58 | 1262 | 21.76 | 69.25 |
| 15 | French language | 268 | 289 | 1.08 | 40.13 |
| 16 | Chemistry | 215 | 230 | 1.07 | 61.30 |
| **-** | **Total** | **4199** | **20,479** | **4.88** | **36.62** |

According to the expert evaluation results, the most typical mistakes of the algorithm included the clustering of lesson script titles into semantic clusters based on:

- similar word, one-rooted words, interference with prepositions;
- similar lemmas;
- close meaning of words;
- proper names;
- generalized definitions;
- prepositions;
- word abbreviations;
- arithmetic operation symbols;
- Roman numbers;
- overlapping of letter sets.

The algorithm also makes mistakes in clustering lesson script titles when they incorporate several meanings or ambivalent wordings. An example of this may be the lesson script, with the title of a famous Russian fable "The Dragonfly and the Ant".

The large number of mistakes made by the algorithm shows that the algorithm lacks sufficient "knowledge" to cluster incomplete titles, and that lexical topics are mapped to the grammar code, which is not applicable.

Besides, it was impossible to map the lesson script titles in foreign languages to the didactic item titles in Russian. This is why the visualizations of French language and German language subjects show that the didactic items are located far from the lesson scripts (see Figure 4).

Within the course of the research, we discovered that the distance between the objects (titles of didactic items and lesson scripts) inside the semantic clusters depends on the structure of the lesson script titles.

Here is an example of the semantic cluster in the subject English language, generated by the algorithm with regard to the didactic item "Healthy Lifestyle" (Figures 7 and 8).

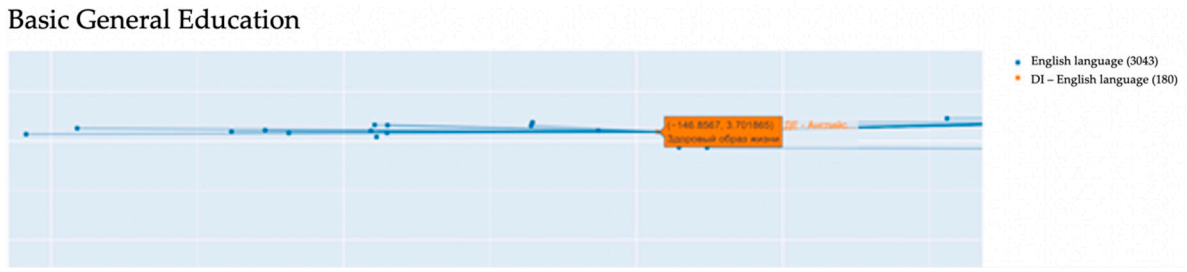


**Figure 7.** Semantic cluster "Healthy Lifestyle".

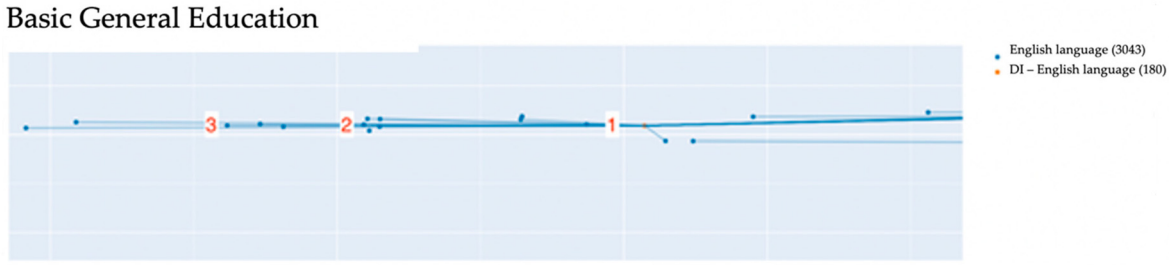


**Figure 8.** Semantic cluster "Healthy Lifestyle" Distance between objects.

Area 1 of this semantic cluster includes lesson scripts with titles "Health is more important than wealth". "Healthy lifestyle. Grade 6", "Grade 7. Healthy lifestyle. Control lesson (two lesson scripts)". Area 2 includes lesson scripts with titles "Healthy lifestyle. Food", "Healthy lifestyle. Healthy nutrition", "Grade 7. Healthy lifestyle. Conclusions", "Grade 5. Healthy lifestyle. Conclusions". Area 3 includes lesson scripts with titles "Healthy lifestyle. Food-5", "Healthy lifestyle. Food-6", "Healthy lifestyle. Food-3", "Healthy lifestyle. Sports exercises". The distance between the objects depends on the semantic emphasis in the topic "Healthy Lifestyle". In this example, we see that all the names of lesson scripts containing the words "Healthy lifestyle" are combined by the algorithm into a separate cluster and the distance of the names of lesson scripts from didactic items depends on the composition of words in the names of lesson scripts, which demonstrates the correct operation of the developed algorithm.

Another example shows the semantic cluster of Physical Education. Figure 9 represents the semantic cluster "Sports games components for physical games".

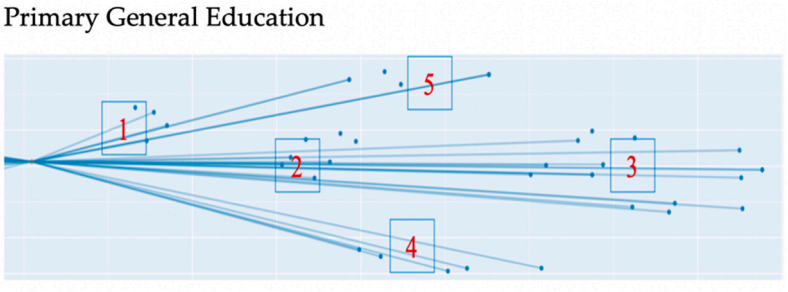


**Figure 9.** Semantic cluster "Sports games components for physical games".

Area 1 of this semantic cluster comprises such lesson scripts titles as "Components of Basketball. Active sports game. Five strikes", "Basketball components for physical games", "Football components for physical games". Area 2 includes lesson scripts with titles "Basketball components for physical games. Grade 2 (3 lessons)", "Basketball components for physical games. Grade 3", "Basketball components for physical games". Area 3 comprises lesson scripts with titles that include the phrase "Active games with volleyball elements". Area 4 comprises lesson scripts with titles that include the phrases "Pitching components

for physical games", "Track-and-field athletics components for physical games". Area 4 comprises lesson scripts with titles that include the phrase "Active games with components of pitching". Area 5 comprises such lesson scripts titles as "Outdoor games with elements of football". It should be noted that since the model has been developed by means of training the neural network, the quality of algorithm performance in terms of clustering one-topic lesson scripts can be further improved by ensuring that it is trained on larger text corpora.

We have obtained these results for all subjects. However, there have also been some "peculiar" results of mapping didactic item titles to lesson script titles. For example, in the subject History, the semantic clusters were formed not around one didactic item, but around semantic clusters containing didactic items and lesson script titles.

Here is an example of a semantic cluster that has been generated by the algorithm (see Figure 10) with regard to the didactic items of the semantic cluster "Rebellions".

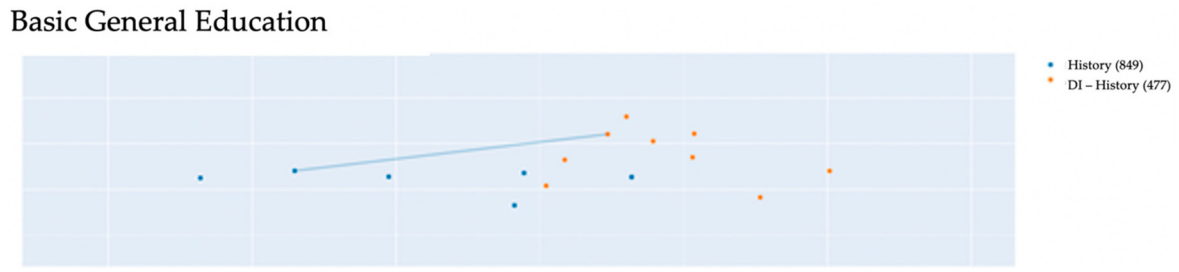


**Figure 10.** Semantic cluster "Rebellions".

The didactic items and lesson scripts in this cluster include the word "rebellion" in their titles. The didactic items comprise "Pugachev's Rebellion", "Rebellion on Don lead by K. Bulavin, Jacquerie", "Rebellion of W. Tyler", "Astrakhan Rebellion", "Bolotnikov's Rebellion", "Decembrist rebellion", "Rebellion in Bashkiria", "Rebellion of Stepan Razin", "Spartak's Rebellion". We can see here that the algorithm includes one-topic semantic cluster didactic items with identical words. This semantic cluster also includes (without line connections) lesson scripts with the titles "Decembrist rebellion. Grade 8", "Decembrist rebellion on December 14, 1825", "Decembrist rebellion (consolidating learned material and conclusions)", "Spartak's rebellion (three lesson scripts)".

Figure 11 shows the semantic cluster generated by the algorithm with didactic items that incorrectly include different historic persons. For example, M. Loris-Melikov, Zh. Garibaldi, Voltaire, Miltiades, V. Kornilov, P. Milyukov, and M. Skobelev have been grouped into a one-topic cluster.

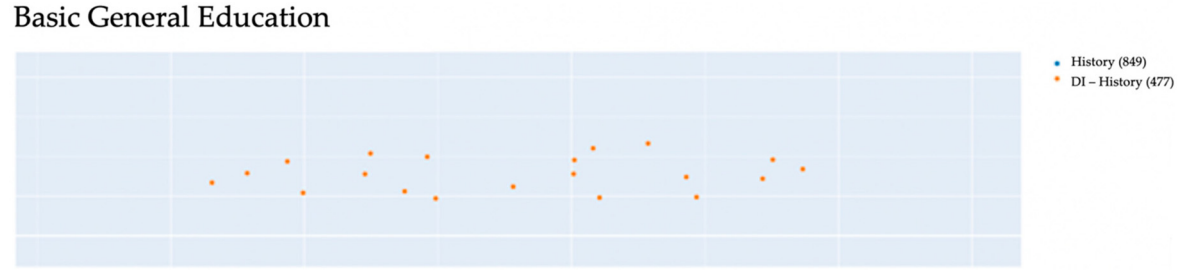


**Figure 11.** Semantic cluster including different historic persons.

The semantic analysis has also revealed that titles of certain didactic items are so broad that neither algorithms nor experts are able to correctly assign lesson script titles to didactic items. This is the case when didactic item titles do not relate to concepts but are worded as, for example, "Relationship of music and art".

## 4. Conclusions

The developed algorithm for semantic data mining and visual representation of the data allowed analysing the structure of lesson script topics in the MES learning repository. Large semantic clusters unite topics that are popular among teachers in terms of creating lesson scripts, for example: lesson scripts in " Addition and subtraction" (Mathematics, PGE); lesson scripts on topics related to tenses of the verb (English language, PGE); lesson scripts on general topics such as "Spotlight", "Starlight", "City" (English, BGE); lesson scripts on the topic "Track-and-filed athletics" (Physical education, BGE); lesson scripts on the topic "Punctuation marks" (Russian language, BGE).

In the course of the research, we have discovered popular cross-disciplinary topics such as "War", "Healthy lifestyle", "Rules of safe conduct", "City", etc. However, the structure of lesson script titles in the common semantic field shows that school subjects tend to be semantically isolated from each other.

The semantic analysis of the titles of lesson scripts of the basic general education subjects revealed the topical closeness of lesson script titles, as well as the clustering of similar lesson scripts with titles that begin with words that do not always have semantic significancy for understanding the meaning or topic of a lesson script. For example, the gene- rated semantic clusters included one-topic lesson scripts, as well as lesson scripts with titles that begin similarly or have common words. We also noted the clustering of lesson script topics by common meanings or similar topics. Nevertheless, there are semantic clusters that have been generated due to technical peculiarities of data display.

The comparison of lesson script titles and didactic item titles visualized the compliance of lesson scripts with the FSES (in the semantic aspect). At the current stage of the research, the accuracy of the results produced by the text mining algorithm is quite low, however, in case of further training and development of the algorithm, we will be able to receive instant visualization, showing compliance of learning content to the FSES.

Therefore, it is essential to train the algorithm: create a list of stop-words, keywords, as well as other methods that can boost the algorithm's performance.

The analysis of the semantic mapping results shows that it is necessary to improve the topical framework. The topical framework must be structured based on the ontology of concepts, not just items of learning content.

Besides, the application of text mining methods to the learning content contained in the MES e-library will allow more advanced development (including requirement specification) of the topical framework for the whole school curriculum, and transformation of approaches to structuring learning content using big ideas. The results of the research can be applied to structure the learning content of different digital learning systems in school education. This will streamline integration of learning content from other systems into the MES or the exchange of learning content between different digital systems in school education which, in the long run, will improve the quality of learning content, facilitate the indication of various types of content and accelerate resources with high teaching potential, as well as raise the overall quality of learning materials.

The analysis of the MES data clearly shows that the application of new methods and approaches to big data processing, as well as using artificial intelligence technologies, can change our views on education, even today. With the help of these technologies, we can discover new unconventional knowledge about the course and content of teaching and learning processes, which leads to the changes in the educational system and reflects the transition to digital society.

Further research is planned to improve the quality of comparison of lesson scripts to didactic items by:

- expanding the semantic data set by including annotations and full texts of lesson scripts in the analysis;
- translating topics and didactic units of the thematic framework into foreign languages (Chinese, English, German, French) for semantic analysis of the names of lesson scripts in foreign languages.

In the future, it is assumed that the developed algorithm will be able to automatically map educational content uploaded to the MES with the thematic framework, determine the "best" version of the lesson script among similar ones (most accurately revealing the topic), and discard low-quality content.

**Author Contributions:** S.N.V. development of the research design, interpretation of the research results, writing part of the article; E.Y.P. literature review, interpretation of the research results, writing part of the article; R.B.K. literature review, research data analysis, development of the model and algorithms for data analysis, writing part of the article; R.S.S. interpretation of the research results, writing part of the article. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Department of Education and Science of the City of Moscow.

**Data Availability Statement:** Data sharing not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Grinshkun, V.V.; Remorenko, I.M. Frontiers of Moscow Electronic School. *Inform. Educ.* **2017**, *7*, 3–8.
2. Vachkova, S.N.; Obydenkova, V.K.; Zaslavskiy, A.A.; Kats, S.V. About causes for the "Moscow E-School" lessons scripts relevance. *Vestn. Mosc. City Univ. Pedagog. Psychol. Ser.* **2020**, *1*, 8–24.
3. Vachkova, S.N.; Patarakin, E.D.; Petryaeva, E.Y. Content Quality of Lesson Scenarios in Moscow E-School. Theory and Practice of Project Management in Education: Horizons and Risks. In Proceedings of the International Scientific and Practical Conference, Moscow, Russia, 17 April 2020; p. 1017.
4. Patarakin, E.D.; Vachkova, S.N. Network analysis of collective operations on the digital education units. *Vestn. Mosc. City Univ. Pedagog. Psychol. Ser.* **2019**, *4*, 101–112.
5. Zaslavskaya, O.Y.; Makhotin, D.A.; Kats, S.V. Recommendations on searching interactive lesson scripts in the Moscow Electronic School (on the example of the subject Handicrafts). *Interact. Educ.* **2019**, *6*, 36–42.
6. Levchenko, I.V.; Sadykova, A.R. Approaches to solving the problem of search of scenarios of lessons on informatics for basic school in the Moscow E-School library. *Rudn J. Informatiz. Educ.* **2019**, *16*, 231–242. [CrossRef]
7. Azevich, A.I. Interactive Lesson in the «Moscow Electronic School»: From Concept to Implementation. *Vestn. Mosc. City Univ. Ser. Inf. Technol. Informatiz. Educ.* **2018**, *3*, 64–73.
8. Afanasieva, Y.A. Development of a modern lesson by means of the Moscow Electronic School (MES) for students with special needs. *Curr. Sci. Issues* **2018**, *46*, 103–107.
9. Grushina, T.P. Designing a Lesson with the Use of Digital Educational Resources. *Vestn. Mosc. City Univ. Ser. Nat. Sci.* **2018**, *4*, 93–101.
10. Zaslavskaya, O.Y.; Makhotin, D.A.; Kats, S.V. Approaches to the Description of the Model for Designing Scenarios of Technology Lessons on the Moscow E-School Portal. *Vestn. Mosc. City Univ. Ser. Inf. Technol. Informatiz. Educ.* **2019**, *4*, 64–72.
11. Zaslavskij, A.A. Extension of Content Types for the Project «Moscow Electronic School». *Vestn. Mosc. City Univ. Ser. Inf. Technol. Informatiz. Educ.* **2020**, *2*, 49–52.
12. Kanunnikova, I.A. Modeling Scenarios of Literature Lessons in the Moscow E-School System: Methodological Aspect. Russian word in the multi-cultural world. In Proceedings of the 14th Congress of the International Association of Teachers of Russian Language and Literature, Nur-Sultan, Kazakhstan, 29 April–3 May 2019; pp. 2015–2019.
13. Kohanova, V.A.; Kanunnikova, I.A. Designing Lesson Plans for Literature Classes Within the Information Space «Moscow E-school». *Vestn. Mosc. City Univ. Ser. Philol. Theory Lang. Lang. Educ.* **2019**, *3*, 92–99.
14. Bazhenova, S.A. Approaches to improving the training of teachers working under the International Baccalaureate programs in the field of education informatization. *Rudn J. Informatiz. Educ.* **2020**, *17*, 123–133. [CrossRef]
15. Zaslavskaya, O.Y. Organization of interaction between the teacher and students in the preparation for the creation and use of electronic educational materials. *Rudn J. Informatiz. Educ.* **2018**, *4*, 351–362. [CrossRef]
16. Zinovyeva, T.I.; Afanasyeva, Z.V.; Bogdanova, A.V. "Moscow E-School" as a Resource of the Future Teacher's Training to the Innovative Activities. *Nizhny Novgorod Educ.* **2018**, *3*, 121–127.
17. Smirnova, M.S.; Dobrotin, D.Y. The results of the design and approbation of the new content and forms of organizing teaching practices of future teachers at primary school with the use of the MES. *J. Inst. Pedagog. Psychol. Educ.* **2018**, *2*, 19–27.
18. Tsaplina, O.V. Training of Teachers to Evaluating the Quality of Educational Content and Resource «Moscow Electronic School». *J. Inst. Pedagog. Psychol. Educ.* **2017**, *3*, 21–25.
19. Mashinyan, A.A.; Kochergina, N.V. The Concept of Updating the Content of Physical and Mathematical Education by Dynamic Means of Graphic Visualization in the Conditions of E-School. Innovative Scientific Research: Theory, Methodology, Practice. In Proceedings of the 17th International Research and Practice Conference, Penza, Russia, 27 May 2019; Volume 2, pp. 187–190.

20. Vodolazov, D.M. Opportunities of the "Moscow Electronic School" at the Selection of Contents to the Individual Project on the Topic "The Moscow Battle". The Current Issues of the Humanities: Theory, Methodology, Practice. In Proceedings of the Collection of Articles for the 20th Anniversary of the Department of Teaching History, Social Sciences and Law, Moscow, Russia, 3 April 2019; pp. 193–200.
21. Schleicher, A. Building a Learning Culture for the Digital World: Lessons from Moscow. Available online: https://oecdedutoday.com/learning-digital-world-technology-education-moscow/ (accessed on 16 December 2020).
22. Parhomenko, P.A.; Grigorev, A.A.; Astrakhantsev, N.A. A survey and an experimental comparison of methods for text clustering: Application to scientific articles. *Proc. ISP RAS* **2017**, *29*, 161–200. [CrossRef]
23. Fortino, A.; Lowrance, R.; Zhong, Q.; Huang, W. RightJob: Application of Text Data Mining to Curriculum Selection and Development. *Acad. Manag. Proc.* **2019**, *1*, 10848. [CrossRef]
24. West, J. Validating curriculum development using text mining. *Curric. J.* **2016**, *28*, 389–402. [CrossRef]
25. Xu, J.; Xu, B.; Wang, P.; Zheng, S.; Tian, G.; Zhao, J.; Xu, B. Self-taught convolutional neural networks for short text clustering. *Neural Netw.* **2017**, *88*, 22–31. [CrossRef]
26. Lapshin, S.V.; Lebedev, I.S.; Spivak, A.I. Text clustering powered by semantico-syntactic features. *Sci. Tech. J. Inf. Technol. Mech. Opt.* **2019**, *19*, 1058–1063. [CrossRef]
27. Levin, I. Main Trends in the Development of Teaching and Learning Processea at School of the Post-Industrial Society. Available online: https://www.tau.ac.il/~{}ilia1/publications/education-trends.pdf (accessed on 16 December 2020).
28. Levin, I. Academic Education in Era of Digital Culture. Available online: https://www.tau.ac.il/~{}ilia1/publications/academic-trends-4.pdf (accessed on 16 December 2020).
29. Levin, I. Cultural trends in a digital society. In Proceedings of the TMCE 2014, Budapest, Hungary, 19–23 May 2014; Available online: https://www.tau.ac.il/~{}ilia1/publications/100-levin-2014.pdf (accessed on 16 December 2020).
30. Levin, I. Cyber-physical Systems as a Cultural Phenomenon. *Int. J. Des. Sci. Technol.* **2016**, *22*. Available online: https://www.tau.ac.il/~{}ilia1/levin_i_cyber-physical_syst.pdf (accessed on 16 December 2020).
31. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop at ICLR, Scottsdale, AZ, USA, 2−4 May 2013.
32. Mikolov, T.; Yih, W.-T.; Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the NAACL HLT, Atlanta, GA, USA, 10–12 June 2013; pp. 746–751.
33. Van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
34. Kassarnig, V.; Mones, E.; Bjerre-Nielsen, A.; Sapiezynski, P.; Lassen, D.; Lehmann, S. Academic Performance and Behavioral Patterns. *EPJ Data Sci.* **2018**, *7*. [CrossRef]