# TEXT MINING ВЪВ ВИРТУАЛНОТО ОБРАЗОВАТЕЛНО ПРОСТРАНСТВО

**Даниела Орозова**
*Бургаски свободен университет*

# TEXT MINING IN VIRTUAL EDUCATIONAL SPACE

**Daniela Orozova**
*Burgas Free University*

***Abstract:*** *The main goal of the report is to show the impact of Data Analytics in the field of educational space. An approach for automatic analysis of documents used by students in the learning process is proposed. By extracting meaningful words, the analyzed documents are defined as related or unrelated to the subject area of study. The algorithm is based on domain ontologies and the creation of classifiers using the Naive Bayes, Random Forest and Logistic regression tools of the Orange Data Mining system. The processing of the obtained data will lead to the study of the basic behavioral patterns and trends that characterize successful learning.*

***Key words:*** *Text mining, Data Analytics, Virtual Education Space, E-learning.*

Поради специфичните характеристики на NET-поколението, обучението с помощта на виртуални образователни платформи е един от предпочитаните начини за получаване на знания. Виртуалните образователни пространства [13] са информационни и социални пространства, които интегрират хетерогенни технологии и различни педагогически подходи. Те са среда за доставка на учебни материали и образователни услуги за различни целеви групи, независимо от времето и пространството. Данните натрупвани от работата на образователното пространство постоянно се увеличават, но случаите на използването им са малко. Въпреки това, много страни и университети по целия свят изграждат инфраструктури за анализиране на тези данни. Siemens & Long [4] определят *Data analytics in Education* като „измерване, събиране, анализ и отчитане на данните за учащите се и техните контексти с цел разбиране и оптимизиране на обучението и околната среда, в която се случва.“ Chatti [5] определя целите на този анализ като: мониторинг, прогнозиране, индивидуализация, намеса в обучаващия процес, оценка и препоръки на обучаемия и обратна връзка с него.

## 1. Науката за данните в обучението

Термините Big Data, Data Analytics и Data Mining описват както самите данни, така и технологиите за събиране, обработка, управление на данните и методите за анализ. *Data Mining* е процеса на търсене на скрити данни и закономерности, предварително неизвестни, нетревиялни и практически полезни, необходими за взимане на решения в различни сфери на човешките дейности. Тук акцентът е не само в извличане на нови факти, но и генериране на хипотези, които могат да бъдат проверявани. Традиционните инструменти на анализа се основават на математическата статистика –

регресия, корелация, клъстеризация, анализ на времеви редове, дървета на решенията и др., а също и техники на изкуствения интелект като: машинно обучение, невронни мрежи, генетични алгоритми, размити логики и др.

*Big Data Analytics* се явява развитие на концепцията Data Mining. Също така е и развитие на решаваните задачи, сфери на приложение, източници на данни, методи и технологии на обработка.

*Data science* съчетава множество подходи и техники, свързани с анализ на данни от областта на статистиката, дейта майнинг и откриване на знания, машинно обучение, изкуствен интелект, програмиране, комуникация др. Науката за данните включва и процесите по изчистване и интеграция на данните, избор и трансформация на данни, извличане на знания, техния анализ, оценяване и представяне. Може да се каже, че *Data science* е „сплав“ от различни дисциплини, технологии и средства за анализ на данните.

Събирането на данни за анализ на обучението се отнася до целия процес и включва всички данни, получени по време на учебните дейности. Това са много и разнообразни типове данни. Международната организация по стандартизация, IMG Global Learning Consortium (IMG Global) класифицира данните, които могат да бъдат събрани и анализирани в областта на образованието в пет вида:

- данни за учебното съдържание;
- данни за учебната дейност;
- оперативни данни;
- данни, свързани с кариерното развитие;
- данни за профила на обучаемия.

Повечето образователни институции предоставят еднопосочно данни, свързани с обучението като: графици, съобщения и данни от ученическия дневник. Но данните показвани в платформите и софтуерът за обучение са в технически формат от различен тип и не са достъпни за лесна обработка. Освен това данните могат да бъдат достъпвани от различни източници, например: данни за оценки, дейности във форуми, посещения на семинари, използване на библиотеката и др. Затова данните трябва да бъдат подходящо съхранени, за да бъдат подлагани на различни анализи с различни цели.

*Predictive Analytics*. Технологията за прогнозиране дава възможности да се предскажат академичните резултати на обучаемите чрез анализ на данни, свързани със специфичния учебен процес. Анализите могат да повлияят върху промяна на метода и стила на обучение, например да се определи дали „студент е в риск“ относно завършване на курса.

*Adaptive Analytics*. Тази технология има за цел да осигури най-подходящото ниво и внимание относно обекта на обучение. Тези анализи могат да насочат обучаемия по време на процеса на учене и да подобрят резултатите му. Например може да се направи определяне на нивото на трудност, като се вземе предвид дейността на обучаемия, особено в критични ситуации. Това се извършва на базата на анализ на данни, свързани с броя часове, през които лицето е извършвало дейности, свързани с конкретния предмет, задавани въпроси, текущи оценки и др.

*Social Network Analytics*. Анализът на социалните мрежи се занимава с извличане на полезна информация за човешките отношения и връзки между хората: косвени отношения или пряко взаимодействие. Анализът на данните за взаимодействието на обучаемите чрез използваната система за управление на съдържанието, електронна поща и дискусии може да доведе до постигане на максимална ефективност на обучението при работа в екипи и съвместно обучение. Провеждането на колаборативно

обучение на студентите насърчава разбирането на учебното съдържание. За даден студент връзката с колеги е в корелация с общите резултати от обучението му [10].

*Discourse Analytics*. Анализ на данните от log-файловете на системата могат да дадат информация кога и колко дълго потребителят е влизал в системата, неговите участия в дискусии и различни други дейности, изпълнявани от него. Анализ на съдържанието на използваните документи, коментарите в дискусиите и създаваните текстове от потребителите, изисква техники за анализ на текстове на естествен език.

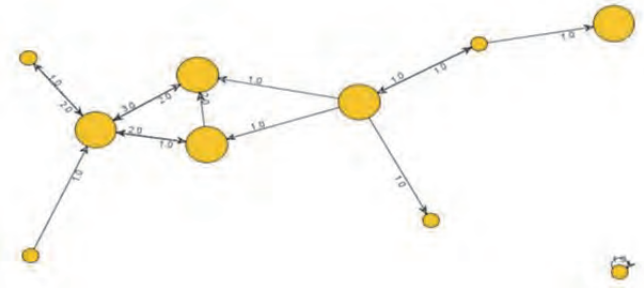


*Fig. 1. Примерен резултат от анализ на дискусии на потребители.*

Техники от областта на Text mining могат да доставят полезна информация от анализ на дискусиите на участниците в онлайн текстов чат, като например: колко и кои са участниците, колко разговора е провел всеки от тях и да се сравняват тези цифри за всяка сесия. Чрез анализ на дискусиите може да се даде оценка на степента на взаимодействие между участниците (фигура 1).

## 2. Анализ на текстови документи, използвани от обучаемите

За да се анализира автоматично документ, трябва да се извърши обобщаване на документа, за да се определи типа на неговото съдържание. Нашата цел е фокусирана към подходи за извличане на подмножество от съществуващи думи, фрази или изречения от оригиналния текст, отразяващи най-важната информация в документа и техния анализ.

Алгоритмите за извеждане на ключови думи и фрази в компютърните програми, използват свойства, описвани чрез примери и дават информацията на обучаващ алгоритъм, който отделя ключови фрази от тези, които не са такива. Свойствата включват честота на различни термини, дължина на примера (брой думи във фразата), относителна позиция на първото появяване (като: първите десет изречения), различни булеви синтактични свойства (като: съдържа само главни букви) и др. Могат да бъдат използвани и евристики, за идентифициране на ключови думи.

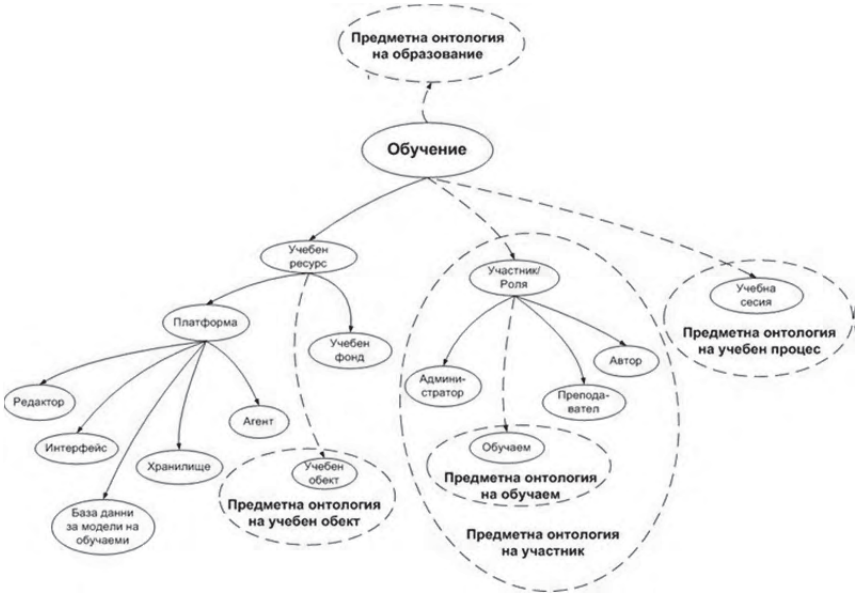
Например алгоритъмът [3] предлага извличане на ключови фрази в един документ като изгражда граф. Използва множеството от текстови единици за върхове, а ребрата се базират на измерване на някакво лексично сходство на върховете. Обикновено ребрата не са насочени и могат да имат тегло, което определя ниво на прилика на текстовите единици във върховете. Друг алгоритъм [8] се основава на намиране на „медицентър” на всички изречения в документа. Останалите изречения се класифицират спрямо приближението им към него. Работата се базира на статистически методи, приложени върху лексикалните единици, като се използва синтактичен филтър. Имплементирайки различни мерки за асоциация между фрази се дава възможност за избор на мярка за асоциация и дължина на анализираните фрази.

Една абстрактна система за извличане на ключови думи обикновено извежда фрази директно от текста. Намирането на важната информация предполага способност да се разбира семантиката на документа и способност да се реорганизира информацията, изразена в различни изречения на оригиналния документ. За прилагането на семантичен анализ в процеса на разпознаване на лексикалните единици, могат да се използват възможностите на лексикално-семантични мрежи като WordNet [7]. Друго важно разширение на анализа е включването на контекста, тъй като в много случаи значението на лексикалната единица се определя от контекста.

В настоящия доклад се представя поход, базиран на използване на онтологии. За реализиране на автоматична обработка е необходима предварителна работа с текста. Той трябва да бъде обработен със система, която да разпознава отделните думи (parser) и морфологичен анализатор, чрез който на всяка дума се приписва съответна граматична информация. Морфологичният анализатор разпознава граматичните характеристики на думата в текста и привежда думата в основната ѝ форма. Прилагайки търсене в предметно-ориентирана онтология може автоматично да се отстрани семантичната многозначност на всяка дума, разглеждайки я в даден контекст. Чрез онтологията могат да се получат нови думи, които не се появяват директно в текста. По този начин се търсят изрази, които са „централни" за текста (изразяват главната идея) и са „разнообразни" (различават се помежду си). Чрез онтология може да се осъществява разширено търсене на синоними на думата или на близки по значение думи. Може да се реализира търсене на синонимни фрази, които се изразяват със съставни думи или търсене по друга семантична връзка.

Онтологиите имат приложение като подход за представяне на знания, който комбинира представянето на данните с връзките между концепциите. Една от най-цитираните дефиниции [1] определя онтологията като експлицитна спецификация на концептуализацията. Концептуализацията предполага описание на множество от обекти и понятия, знания за тях и връзки между тях. Онтологиите могат да бъдат класифицирани в зависимост от различни класификационни признаци. Според зависимостта на онтологията от конкретна предметна област или задача, те се разделят на [6]: Общи онтологии (предметно независими онтологии), Предметни онтологии, Онтологии, ориентирани към задачи и Приложни онтологии (описват концерти, които съответстват на ролите, които играят обектите в предметната област при изпълнение на определена дейност).

Използването на онтологии за представяне на знанията в отделните предметни области осигурява нужната независимост и гъвкавост, а тяхното описание със стандартизирани средства гарантира интелигентност при обработката на семантичната информация и прави създадените ресурси споделяеми. Например: извадка от таксономия на предметна онтология за обучение е представена на фигура 2. Представената диаграма, включва класификация на понятия от информационния модел на процеса на обучение и набор от аксиоми, чрез които се определя семантиката на понятията. За всяка област могат да се създават и използват различни онтологии, според начина на представяне на знанията.




*Фиг. 2. Извадка от таксономия на предметна онтология на „обучение“*

**3. Методология на изследването**

Определени думи в документа са значими за съдържанието му и изреченията, които предават най-важната информация в документа обикновено са тези, които съдържат най-много такива значими думи близо една до друга. За да се определят значимите думи за документа се търси честотата на поява на думите, използвайки софтуера Word Cloud.

Wordclouds.com е безплатен онлайн облак. Поставяйки текст, документ или URL адрес, софтуерът генерира списък на всички думи и изображения в текста и броя на появите им. Могат да бъдат налагани филтри, относно получавания списък с думи и неговия размер. Някой от най-често срещаните думи на практика, не са определящи за неговото съдържание. Често срещани са предлозите и местоименията, но те нямат голямо значение за съдържанието на документа. За тази цел използваме предварително определен списък, състоящ се от думи, които не се взимат под внимание. Извършва се филтриране на думи, които се появяват много често в документа и филтриране на думите, които се появяват твърде рядко. За целта се въвеждат високи и ниски прагове на честота на срещане на думите. Всички думи, които се включват в интервала между праговите стойности са значими и могат да бъдат анализирани за ключови думи по отношение на съдържанието на документа.

На следваща стъпка се извършва търсене на всяка дума от полученото множество, с речника на понятията на свързаната с областта предметна онтология (в случая – морета и водни басейни). Определя се степен на близост за всяка дума от избраното множество, с всички думи от речника на предметната онтология и се взима най-малката получена стойност. За определяне на близостта между думите могат да се използват различни подходи. Например може да се определи чрез прилагане на алгоритъм на Левенщейн, при който, ако са дадени два стринга a и b от една азбука Σ (например множеството на ASCII символите), то Левенщейн разстоянието d(a, b) е равно на минималния брой операции необходими за трансформиране на a в b. Алгоритъмът е представен в [11].

В друго технологично решение степента на близост на ключовите думи от документа с концепциите, заложени в предметната онтология се определя чрез подход, представен в [2], използвайки q-gram метрики. Това е мярка, базирана на символи, която изчислява степен на сходство, на базата на разликата между броя на срещанията на символите в двата сравнявани низа.

Така чрез обтаботка от анализираните документи се създава нов набор от данни с извлечени центални думи и степента им на близост с понятията от областта. 2/3 от тези данни се подават на класификационни алгоритми за обучение, като целевия атрибут е дали документа е свързан с обастта или не. За решаване на класификационния проблем се прилагат техниките на приложението Orange Data Mining System [12]. Създава се работен процес и чрез инструмента „File" се зареждат данните, които предварително са получили чрез експерименти, за ключовите думи, степента им на близост и принадлежността на документа към областта (0/1). Последователно създаваме модели, прилагайки инструментите на системата: Neural Network, Random Forest, Logistic Regression и Naïve Bayes. Работният поток за създаване на моделите е представен на фигура 4.

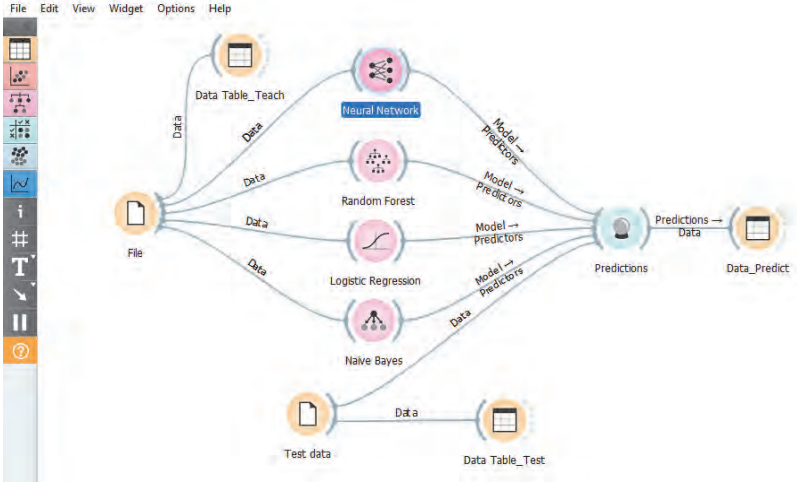



*Fig. 4. Работен поток за създаване на моделите*

Целта е да създадем класификатор, който да се прилага за предвиждане. Чрез него нов документ може да се определя към класовете {несвързан с областта = 0, свързан с областта = 1}. Следващата стъпка е оценяване на точността и прецизността на създадените модели, използвайки останалата 1/3 от данните, с които разполагаме от експеримента. Това се извършва чрез системата Orange с прилагане на инструментите *Test&Score* и *Confusion Matrix* върху създадените модели.

**4. Модел на обучаемия**

През целия период на обучението се натрупват данни от дейностите на обучаемите, когато взаимодействат с учебни материали, изпълняват упражнения, решават тестове и т.н. Значителна част от изследванията днес се фокусират върху обработката на тези данни, с идеята да се научат основните поведенчески модели и тенденции, които характеризират успешното учене. Данните за всеки обучаем се натрупват в модела на обучаемия, които се състои от три основни модула:

- профил на обучаемия (лични данни) – име, пол, възраст, социален статус, образование и др.;
- поведенчески характеристики – данни за стил на учене, степен на концентрация, наклонности, мотивация и т.н.
- ниво на знания – предварителни и текущи знания.

Резултатите от тестовете, решените задачи, курсови проекти и наблюденията върху изследваните от обучаемия документи в пространството по време на обучението, са основа за прогнозиране на темпа и стила на усвояване на знания от обучаемия.

Един от основните недостатъци на стандартните приложения, свързани с обучението е тяхната реактивност – предприемат действия в отговор на вече настъпили събития. Създаването на софтуерни архитектури, които са проактивни, т.е. адаптират се към околната среда, планират и избират начин за постигане на съставения план, дава възможност за премахване на този недостатък. Интелигентните пространства наблюдават какво се случва вътре в тях, моделират поведението си и оперират въз основа на собствените си решения. Те събират, съхраняват и целенасочено анализират данни, оценят различните ситуации и управляват ресурсите.

Това изследване е подкрепено от фонд научни изследвания на Бургаския свободен университет като част от проект Д-9/2020 „Data Science в образователното пространство за синя кариера“.

**Литература**

[1] Gruber, T. R., A translation approach to portable ontology specification. Knowledge Acquisition, 5(2), pp. 199-220, (1993).
[2] Jecheva V., D. Orozova, Ontology-Based Electronic Test Result Evaluation, Third International Conference of Software, Services and Semantic Technologies S3T, Springer, 213-214, (2011).
[3] Erkan G., D. Radev, LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Articial Intelligence Research 22 (2004) 457-479.
[4] Siemens, G, and Long, P, Penetrating the fog-analytics in learning and education. Asynchronous Learning Networks, 2011.
[5] Chatti, M, A, Dychhoff, A, L, Schroeder, U., and Thus, H. A reference model for learning analytics, International journal of Technology Enhanced learning, 2012.
[6] Van Heijst, G., Schreiber, A., & Wielinga, B., Using Explicit Ontologies in KBS Development. International Journal of Human-Computer Studies, 46, pp.183-298, (1997).
[7] WordNet: WordNet: An Electronic Lexical Database. URL: http://wordnet.princeton.edu/ Princeton University.
[8] TermeX (2009): TermeX. URL: http://ktlab.fer.hr/termex/ Faculty of Electrical Engineering è Computing, University of Zagreb. 2009.
[9] NooJ (2002): NooJ. URL: http://www.nooj4nlp.net/ M. Silberztein. 2002.
[10] Bakharia, A., and Dawson, S. SNAPP: A bird’s-eye-view of temporal participant interaction. Proceedings of the First International Conference on Learning Analytics and Knowledge-LAK’11, pp. 168-173, 2011.
[11] The Levenshtein-Algorithm. http://www.levenshtein.net/
[12] Orange system [Online]. https://orange.biolab.si/training/introduction-to-data-mining/
[13] Орозова, Д., С. Стоянов и И. Попчев, „Виртуално образователно пространство“ в Научна конференция с международно участие „Знанието – източник на иновация“, БСУ, Бургас, 2013, pp. 153-159, ISBN 978-954-9370-99-7.