NJË VËSHTRIM I PËRGJITHSHËM I TEXT MINING, TEKNIKAVE DHE ZBATIMEVE TË SAJ

ROLAND VASILI.[1], ENDRI XHINA.[2], ILIA NINKA.[3]

[1]Universiteti "Eqrem Çabej", Fakulteti i Shkencave të Natyrës, Departamenti i Matematikës, Informatikës & Fizikës

[2,3]Universiteti i Tiranës, Fakulteti i Shkencave të Natyrës, Departamenti i Informatikës

email: rvasili@uogj.edu.al

Përmbledhje

Text Mining është një teknologji që përdoret për të ekstraktuar informacion të dobishëm ose dije nga tekste të pastrukturuar ose gjysëm të strukturuar. Sasia e të dhënave të strukturuara është jo më shumë se 20% e të dhënave të prodhura sot në botë. Duke patur parasysh që të dhënat rriten me shpejtësi të jashtëzakonshme, lind nevoja e shfrytëzimit të tyre. Për plotësimin e kësaj nevoje, dhe konkretisht për të ekstraktuar informacion kuptimplotë nga këto sasi vigane të dhënash, përdoren teknika dhe metoda nga fusha e Text Mining. Ky punim fokuson në paraqitjen e procesit të Text Mining, teknikave dhe metodave që përdoren në këtë teknologji si dhe zbatimeve të tyre praktike.

Fjalëkyçe: Text Mining, Teknika Text Mining, Aplikime Text Mining.

Abstract

Text Mining is a technology used for the extraction of useful information or knowledge from ustructured or semistructured texts. No more than 20 percent of the present information is made out of structured data. Bearing in mind that data grows at extraordinary speeds, the need for their exploitation arises. To supplement this need, and specifically to extract meaningful information from these massive amounts of data, the techniques and methods used in the Text Mining field are employed. This paper focuses on presenting the Text Mining process, techniques and methods used in this technology as well as some of their practical applications.

Key words: Text Mining, Text Mining Techniques, Text Mining Applications.

Hyrje

Jetojmë në «Epokën e informacionit», që mbështetet në grumbullimin, menaxhimin dhe përpunimin e informacioneve dhe përbën faktor vendimtar për suksesin e kërkimit shkencor, të veprimtarisë së biznesit dhe në përgjithësi të zhvillimit shoqëror. Në këtë epokë, që bindja themelore është se informacioni ofron fuqi dhe sukses, është bërë karakteristikë koleksionimi intensiv i të dhënave dhe informacioneve vigane.

Përditë, në shërbime publike apo private prodhohen dhe ruhen, në baza të dhënash voluminoze, miliona të dhëna tekstuale. Këto të dhëna mund të jenë shkresa të formateve heterogjene, dhe të jenë shkruar në më shumë se një gjuhë. Kështu, përditë shpenzohet kohë dhe përpjekje në kërkime informacioni në këto baza të dhënash. Kërkime të cilat, për arsye të formës

së pastrukturuar dhe të ndryshme të të dhënave, bëhen më të vështira dhe kohëngrenëse. Njëkohësisht, të dhënat e disponueshme vazhdojnë të rriten.

Që nga fillimet e dekades së viteve ’90 ekzistojnë referenca ku përmendet se volumi i informacionit të grumbulluar dyfishohet afërsisht çdo 20 muaj (*Frawley, Piatetsky-Shapiro & Matheus 1992*). Vitet e fundit është venë re një rritje akoma më e madhe e volumit të të dhenave që regjistrohen në gjithë botën. Sipas një raporti nga *IBM Marketing Cloud, "10 Key Marketing Trends For 2017,"* 90% e të dhënave të 2017 në botë janë krijuar vetëm dy vitet e fundit, me një ritëm $2.5 \cdot 10^{30}$ bytes të dhënash në ditë!

Prandaj, nevoja për ekstraktim të automatizuar të informacionit të dobishëm nga baza kolosale të dhënash, që përmbajnë kryesisht text është mëse e dukshme. Zbulimi i dijes në tekst (Knowledge Discovery in Text, shkurt KDT) dhe Text Mining (shkurt TM) (*Karanikas & Mavroudakis, 2005*) janë teknikat më të automatizuara që synojnë zbulimin e informacioneve të nivelit të lartë nga baza të mëdha të të dhënave të ruajtur në formë tekstuale.

KDT ose TM është një fushë e re kërkimi, që përdor teknika Data Mining (shkurt DM), të të mësuarit e makinës (Machine Learning), të përpunimit të gjuhës natyrale (Natural Language Processing), të tërheqjes së informacionit (Information Retrieval), të ekstraktimit të informacionit (Information Extraction) dhe të menaxhimit të dijes (Knowledge Management). Është një fushë me sfida mjaft të mëdha dhe fusha të pashtershme zbatimi.

Materiali dhe metodat

Ky punim ka për synim shqyrtimin e literaturës aktuale për teknologjinë TM, duke u fokusuar në përqasje të reja të saj, si dhe zbatimet e teknikave të saj në fusha të ndryshme, duke kontribuar kështu në përcaktimin se cilat prej tyre mund të përdoren potencialisht për studime në këto fusha. Fillimisht diskutohen përkufizimet e teknologjisë TM dhe teknikat kryesore të saj. Më pas shqyrtohen zbatimet e metodave dhe teknikave të TM në fusha të ndryshme.

Përkufizimi i Text Mining

Kohët e fundit Text Mining është bërë mjaft i njohur. Njihet gjithashtu edhe si Text Data Mining *(Hearst, 1999)* si edhe zbulimi i dijes në tekst (KDT). Duke u përpjekur për të gjetur një përkufizim për TM apo DM, studiuesit kanë konkluduar në shumë teori. Aktualisht, literatura ofron një larmi përkufizimesh në lidhje me TM, të tilla si ai nga *Hearst (1999)* që e përkufizon atë si "zbulimin nga kompjuteri të një informacioni të ri, të panjohur më parë, nëpërmjet ekstraktimit automatik të tij nga burime të ndryshme të shkruara". Megjithatë, ky dhe shumë përkufizime të tjerë në lidhje me TM kanë rrënjët e tyre në përkufizimin e *Feldman & Dagan (1995)* që e përkufizojnë atë si "proces të zbulimit të dijes nga baza të dhënash teksti" (*Gupta & Lehal, 2009; Tan, 1999*). Ky përkufizim në një farë mënyre paraqet rrënjët e TM, meqë ai rrjedh nga përkufizimi i zbulimit të dijes (Knowledge Discovery, shkurt KD) nga *Frawley, Piatetsky-Shapiro*

& Matheus (1992), të cilët e përcaktojnë atë si "ekstraktim jo të parëndësishëm të informacionit të nënkuptuar, të panjohur më parë dhe potencialisht të dobishëm prej të dhënave aktuale".

Karanikas & Mavroudakis (2005) e përkufizojnë Text Mining si: "Një hap në procesin KDT që përbëhet nga algoritma të veçantë të Data Mining dhe të përpunimit të gjuhës natyrale, që nën disa kufizime të pranueshme llogaritëse të performancës, prodhojnë një numër të veçantë motivesh (modelesh) prej një bashkësie të dhënash të pastrukturuar teksti".

Si përkufizim të TM mund të japim të vijuarin: "Text Mining është procesi i ekstraktimi të informacionit të ri nëpërmjet të cilit përdoruesi ndërvepron me një koleksion tekstesh duke përdorur një grup mjetesh analizimi".

Synimi i TM është të zbulohen informacione të panjohura dhe gjer tani, të fshehur mirë në tekste që janë të ruajtur në baza kolosale të dhënash dhe qendrojnë "të heshtur".

Kritere për Text Mining

Sharp (2001) në studimin e tij pohon se një model real Text Mining duhet të plotësojë disa kondita si:

(1) Të jetë funksional në koleksione shumë të mëdha tekstesh, të shkruar në gjuhë natyrale. **(2)** Të bazohet në përdorim algoritmash. **(3)** Të ekstraktojë njësi informacioni si p.sh. motive (patterns). **(4)** Më e rëndësishmja prej të gjithave, të zbulojë dije të re.

Hapat e procesit Text Mining

Procesi TM, në thelb, mund të përmblidhet në tre (3) hapat e mëposhtëm (*Karanikas & Mavroudakis, 2005*):

- **Koleksionimi i dokumenteve që kanë të bëjnë me problemin nën studim (Document Collection):** Fillimisht do të duhet të përcaktohet burimi nga i cili do të tërhiqen dokumentet. Në vijim, bëhet përzgjedhja përfundimtare e dokumenteve dhe zotërimi (tërheqja) i tyre.
- **Para-përpunimi i dokumenteve (Pre-processing):** Gjatë kësaj faze ekzekutohen procedura të ndryshme transformimi, duke synuar që dokumentet që u morren të marrin formën e deshiruar për përpunimin e tyre. Pastaj, dokumentët e përftuar bëhen objekt përpunimi që të ofrojnë informacione të dobishme për përdoruesin.
- **Zbulimi i Dijes (Text Mining Operations):** Informacionet e ekstraktuar, nga ana tjetër bëhen të dhëna të reja (meta-të-dhëna). Midis tyre zbulohen marrëdhënie dhe ngjashmëri që na çojnë në konkluzionet përfundimtare dhe në zbulimin e dijes së re.

Figura 1: Tre hapat e zbulimit të dijes nga teksti (TM)

Ndërsa hapat që ndjek procesi text mining persa i përket prodhimit të rezultateve paraqitet në figurën 2:

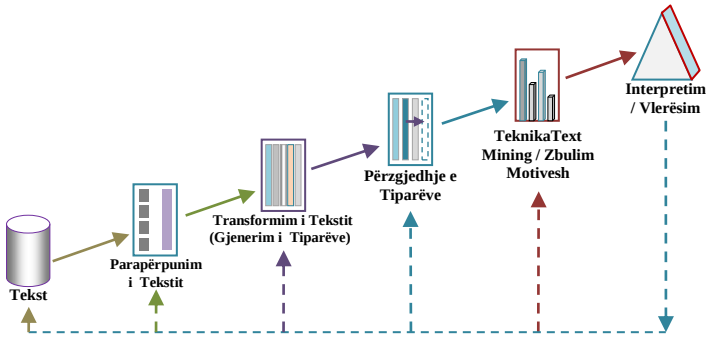

Figura 2. Hapat e Text Mining (Sipas procesit të zbulimit të dijes)

Një sistem TM merr si hyrje një koleksion dokumentesh, pastaj para-përpunon çdo dokument duke kontrolluar formatin dhe setin e karakterëve të tij. Në vijim, këto dokumente të para-përpunuar kalojnë në fazën e analizimit të tekstit, disa herë duke përsëritur teknikat, gjersa të ekstraktohet informacioni i kërkuar. Në figurën 3 duken tre teknikat e analizimit të tekstit,

Figura 3. Hapat e Text Mining

por mund të përdoren edhe kombinime të tjera teknikash, në vartësi të synimeve dhe të korporatës. Informacioni që rrjedh nga ekstraktimi mund të bëhet hyrje për një sistem menaxhimi informacioni, duke prodhuar një sasi të pasur dijesh për përdoruesin e këtij sistemi. Figura 4 eksploron detajet e hapave të procesimit që ndjek një sistem tipik TM.

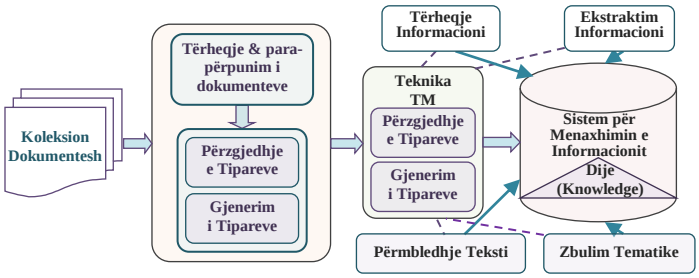

Figura 4. Sistem Text Mining

Koleksionimi i Dokumenteve-Teksteve

Komponenti kryesor i TM janë koleksionet e dokumenteve, ku secili prej tyre përbëhet nga një numër çfarëdo dokumentesh teksti. Numri i teksteve në koleksione të tilla mund të variojë nga disa mijëra gjer në disa milionë.

Koleksionet e tekstit mund të jenë *statike*, ose *dinamike*. Koleksionet jashtëzakonisht të mëdha dhe koleksionet me ritëm të lartë të ndryshimit të teksteve, konsiderohen sfida dhe përbëjnë objekt kryesor të sistemeve TM.
Shembull karakteristik i një koleksioni të madh dinamik tekstesh, të cilën e përdorin miliona përdorues nga e gjithë bota, është *PubMed (2018)*. Përbën një burim rrjetor, i cili përfshin mbi 25.000.000 raporte kërkimorë në fushën biomjekësore në të cilat shtohen, me përafërsi, 35.000 me 40.000 artikuj të rinj çdo muaj.
Prandaj, për të filluar procesi TM, përdoruesi do të thirret të zgjedhë koleksionin e deshiruar të teksteve mbi të cilin do të mbështetet procedura në fjalë si dhe larminë e teksteve që do të përbëjnë burimin e të dhënave.
Procesi në vijim i takon sistemit TM, i cili ka mundësinë, me ndihmën e algoritmeve të zbulimit të dijes, të identifikojë shpejt dhe me efikasitet motive midis një numri të madh tekstesh të gjuhës natyrale.

Por, realizimi i kësaj kërkon ekzistencën e koleksioneve të përpunuar të tekstit. Për këtë arsye procedura më e rendësishme e TM është faza e para-përpunimit të teksteve nën ekzaminim dhe në vazhdim zbatimi i suksesshëm i algoritmeve të zbulimit të dijes.

Para-përpunimi i Teksteve

Procedura e para-përpunimit të teksteve përbën procesin më të rëndësishëm të një sistemi Text Mining. Synimi i kësaj procedure është optimizimi dhe rritja e efikasitetit të ekstraktimit të informacionit përmes një bashkësie të dhënash tekstuale, nëpërmjet zvogëlimit të fjalorit dhe si rrjedhim të madhësisë së indeksit të teksteve.

Si rezultat të para-përpunimit të teksteve kemi ekstraktimin e termave karakteristikë të çdo teksti, të cilat janë të përshtatshëm për përfaqësimin e përmbajtjes së çdo teksti. Para-përpunimi i teksteve përbëhet nga fazat vijuese (Figura 5):

(a) Heqja e strukturës së teksteve, **(b)** Lematizimi (Tokenization),
(c) Heqje e Stopwords, **(d)** Heqja e termave në bazë të gjatësisë së tyre,
(e) Heqja e termave në bazë të frekuencës së tyre, **(f) A**naliza Leksikore (POS Tagging)**, (g)** Stemming, për shembull algoritmi Porter.

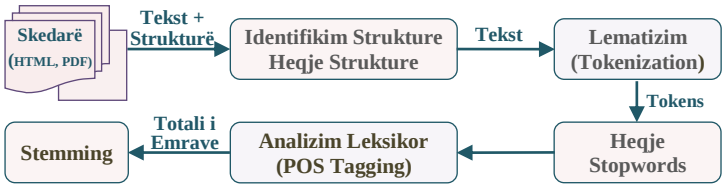

Figura 5: Hapat e përpunimit të teksteve

- **N- grams.**
- **Ekuilibrimi (term weighting).**

Paraqitja e Teksteve

Që të zbatohen teknikat TM do të duhet që tekstet të paraqiten në një formë të përpunueshme. Do të mund të thoshim se metoda më e njohur e paraqitjes së teksteve është paraqitja vektoriale. Në këtë paraqitje krijohet një hapësirë vektoriale, ku çdo tekst paraqitet si një vektor. Kjo hapësirë vektoriale përbëhet nga aq përmasa sa janë dhe termat unikë të teksteve. Prandaj, në paraqitjen vektoriale, vektori i çdo teksti paraqitet me një sasi termash. Kështu, në mënyrë që të lokalizohen termat unikë që do të karakterizojnë semantikën e tekstit dhe në vazhdim do të përbëjnë përmasat e hapësirës vektoriale, para-përpunimi i teksteve i paraprin paraqitjes.

Ka dy mënyra kryesore që përdoren për paraqitjen vektoriale të teksteve:

(a) Modeli Bulean (***Boolean Model***),

(b) Modeli Term i Peshuar (***Term-Weight Model***).

Metoda Teknike të Text Mining

Rritja e vrullshme në koleksionin e të dhënave të pastrukturuar nxiti shkencëtarët të shqyrtojnë dhe të zhvillojnë teknika që do t'u mundësojnë atyre të shfrytëzojnë këto lloje të të dhënave. Sipas *Xu et al. (2003)*, këto teknika janë thelbësore për të mundësuar organizimin efikas, navigimin, tërheqjen dhe përmbledhjen e një korpusi me dokumente të mëdha. Përfitimet nga TM kanë qënë gjithnjë të njohura, por vetëm gjatë dy dekadave të fundit përdoruesit arritën ta shfrytëzojnë këtë teknologji. Në punimin *Fan et al. (2006)* thuhet se qëllimi i TM është të krijojë teknologji që do të kombinojnë shpejtësinë dhe saktësinë e kompjuterave me aftësitë linguistike të njeriut.

Synimi kryesor i TM është të ndihmojë përdoruesit të ekstraktojnë informacione nga burime të mëdha tekstuale. Metodat kryesore të TM që rekomandohen në këtë objektiv janë:

- Ekstraktim Informacioni **(*Information Extraction*)**
- Klasifikim **(*Categorization*)**
- Grupim **(*Clustering*)**
- Përmbledhje **(*Summarization*)**
- Vizualizim Informacioni **(*Information Visualization*)**
- Ndërlidhje Koncepti **(*Concept Linkage*)**
- Modelim Tematike **(*Topic Modelling*)**
- Pyetje/Përgjigje **(*Question Answering*)**
- Ekstraktim Ontologjie **(*Ontology Extraction*)**
- Identifikim Gjuhe dhe Përcaktim i Autorit të Tekstit **(*Language & Author Identification*)**

- Asociacione **(*Associations*)**

10. Zbatime të Text Mining

E konsideruara si vala e ardhshme e Zbulimit të Dijes , teknologjia TM ka një vlerë komerciale shumë të lartë. Është një teknologji në zhvillim që shërbën për analizimin e koleksioneve të mëdha të dokumenteve të pastrukturuar me qëllim ekstraktimin e motiveve ose dijeve interesante, jo të parëndësishme. Mjetet që përdorin zbatimet e teknologjisë Text Mining mund të organizohen gjerësisht në dy grupe:
Mjete për ekplorimin e dokumentit - Ato organizojnë dokumentet duke u bazuar në përmbajtjen e tekstit dhe ofrojnë për përdoruesin një mjedis për navigim dhe shfletim në një hapësirë dokumenti apo koncepti. Një përqasje e përhapur është kryerja e grupimit (*clustering*) mbi bazë dukumentesh mbi ngjashmëritë e tyre në përmbajtje dhe pasqyrimi i grupeve ose grumbujve të përftuar të dokumenteve në një paraqitje grafike të caktuar.

Mjete për analizimin e dokumentit - Ato analizojnë përmbajtjen e tekstit të dokumenteve dhe zbulojnë marrëdhëniet midis koncepteve ose entiteteve që përshkruhen në dokumente. Mbështeten kryesisht në teknika të procesimit të gjuhës natyrale, duke përfshirë analizimin e tekstit (text analysis), kategorizimin e tekstit, extraktimin e informacionit (information extraction), dhe përmbledhjen (summarization) e tekstit.

Nga fushat më aktive të zbatimit të Text Mining janë fusha biomjekësore dhe bio-shkencat (*Gonzalez et al., 2016*).
Një tjetër fushë zbatimi të TM është ajo e sigurisë. Shumë paketa software TM u drejtohen aplikimeve të sigurisë, veçanërisht monitorimit dhe analizimit të burimeve të tekstit online (lajme në Internet, Blogs, etj) për arsye të sigurisë kombëtare (*Zanasi, 2009*). Ndërsa vitet e fundit TM gjen zbatime edhe në Marketing (*Amado et al., 2018*).
Ka shumë fusha të mundshme aplikative ku mund të zbatohet **teknologjia Text Mining**. Ne shkurtimisht do të përmendim disa prej tyre:

Analizimi i profilit të Klientëve, p.sh., përdorimi i TM prej firmave për të parë ngjarje dhe instanca të një termi kyç në blloqe të mëdha teksti që mund të vijnë nga blogs, artikuj, faqe web, forume apo në mesazhet e reagimit apo të ankesave që vijnë nga posta elektronike e klientëve (*Chen, 2009*).

Analizimi i Patentave, p.sh., analizimi i databazës së patentave për faktorë të rendësishëm teknologjie, tendenca, dhe shanse.

Shpërndarja e Informacionit (*Information dissemination*), p.sh., organizimi dhe përmbledhja e lajmeve tregtare dhe raportet për shërbime të përsonalizuar informimi.

Planifikimi i burimeve të ndërmarrjes (*Company resource planning*), p.sh., ekzaminimin e raporteve dhe korrespondencës së ndërmarrjes për aktivitetet, gjendjen, dhe problemet e raportuar.

Çeshtje Sigurie, p.sh., analizimi i burimeve të thjeshta tekstuale si lajmet e internetit, blog-et etj (*Zanasi, 2009*).

Parandalimi i krimit kybernetik. Natyra anonime e internetit dhe shumë karakteristika komunikimi të operuara prej tij kontribuojnë në rritjen e riskut të krimeve të bazuara në internet (*Zanasi, 2009*).

Përgjigjet e anketimeve të hapura (*Open-ended survey responses*), p.sh., duke analizuar një set konkret fjalësh apo termash që përdoren zakonisht në anketime për të përshkruar avantazhet apo disavantazhet e një produkti ose shërbimi (nën shqyrtim), ose duke e përdorur këtë informacion në marketing (*Grimes, 2005; Amado et al., 2018*).

Klasifikimi i Tekstit (*Text classification*), p.sh., filtrimi automatik i postës elektronike ("junk email") duke u bazuar në terma konkret ose fjalë që nuk janë të përshtatshme për t'u pasqyruar në mesazhe zyrtare.

Inteligjenca Konkuruese (*Competitive Intelligence*), p.sh., aftëson kompanitë të organizojnë ose transformojnë strategjitë e kompanive sipas nevojave aktuale të tregut (*Grimes, 2005*).

Në mjedise software (*Software Environment*). Teknika dhe software TM studiohen dhe zhvillohen edhe nga fima të mëdha si IBM dhe Microsoft, për të shqyrtuar dhe automatizuar proceset. Ndërsa në sektorin publik i është dhënë prioritet krijimit të software për gjurmimin dhe monitorimin e aktiviteteve terroriste.

Menaxhimi i Marrëdhënieve me Konsumatorët (*CRM*), p.sh., riadresimi automatik i kërkesave specifike tek shërbimi i duhur ose dhënia e një përgjigje të menjëhershme pyetjeve më të shpeshta.

Aplikime multigjuhësore të përpunimit të gjuhës natyrale (*Multilingual Applications of Natural Language Processing*), p.sh., identifikimi dhe analizimi i faqeve të internetit të publikuara në gjuhë të ndryshme.

Monitorimi i teknologjisë (*Technology watch*), p.sh., identifikimi i literaturës së lidhur me teknologjinë dhe shkencën, dhe ekstraktimin e informacionit të kërkuar prej kësaj literature në mënyrë efikase.

Përmbledhje teksti (*Text summarization*), p.sh., krijimi i një versioni të shkurtuar të një dokumenti ose të një koleksioni me dokumente (*multi-document summarization*) që duhet të përmbajë temat (çeshtjet) më të rëndësishme *(Yao et al., 2018).*

Njohje bio-entiteti (*Bio-entity recognition*), p.sh., identifikimi dhe klasifikimi i termave teknikë në fushën e biologjisë molekulare që i përkasin shembujve të koncepteve që janë me interes për biologët. Shembuj të entiteteve të tillë përfshijnë emrat e proteinave, gjenet dhe hapësira e aktivitetit të tyre si emrat e qelizave apo organizmave *(Giorgi et al., 2018).*

Organizimi i magazinave të dokumenteve të lidhur me meta-informata, p.sh., metodat e klasifikimi automatik të tekstit përdoren për të krijuar meta-

të-dhëna të strukturuar që shërbëjnë për kërkimin dhe tërheqjen e dokumenteve përkatës të bazuar në një pyetje.

Fitimi i dijeve rreth tendencave, marrëdhënieve midis njerëzve / vendeve / organizatave, p.sh., grumbullimin dhe krahasimin e informacionit të ekstraktuar automatikisht prej dokumentëve të një tipi konkret si posta elektronike hyrëse, letrat e konsumatorëve, raportet e lajmeve etj.

Akoma, zbulimi i dijes në tekst ndeshet edhe në zbatime akademike (*Ojo & Adeyemo, 2017*), në ekonomi, në drejtësi dhe shumë fusha të tjera të jetës së përditshme, nga të cilat mund të theksonim ato që e kanë adoptuar gjerësish, si: (a) Industria e automakinave (menaxhimi i garancisë), (b) Industria e kujdesit shëndetësor, (c) Industria bankare (menaxhimi i kartave të kreditit) dhe mjaft të tjera, për të cilat nuk kemi komoditetin të zgjerohemi.

Konkluzione

Në epokën tonë, ka një rritje të vrullshme të informacionit të ardhur nga burime të ndryshme dhe që disponohet në mënyra të ndryshme. Informacioni dixhital është kolosal, dhe duke shtuar dhe arkivat tradicionale të dixhitalizuara përftohet një sasi shumë e madhe informacioni. Nevojat për një mënyrë për të lexuar, organizuar dhe analizuar këtë informacion ka çuar në inovacione teknologjike që bënë jetën e kërkuesve shkencorë, institucioneve publike apo private, bizneseve të çdo lloji si dhe individëve shumë më të lehtë. Text Mining është një prej inovacioneve të tilla që ekstraktojnë informacion prej të dhënave tekstuale. Fraza "Text Mining" përdoret për deklarimin e çfarëdo sistemi që analizon sasi të mëdha nga burime të ndryshme tekstuale të gjuhës natyrale, gjurmon motive lektike (lexical) apo linguistike përdorimi në një përpjekje për të ekstraktuar motive të vlefshme (edhe pse vetëm me probabilitet saktësie) dhe asociacione.
Text Mining u ka dhënë njerëzve një shans, për të zbuluar lidhje të reja midis informacionit dhe të dhënave, të cilat kërkuesit ishte e pamundur t'i zbulonin vet. Siç kemi theksuar ka mjaft përkufizime të ndryshme për TM, disa prej të cilave mbulojnë funksionet teknologjike të TM, ndërsa të tjerat janë më shumë të lidhur me procesin e përdorur për TM.
Nga ana praktike, përdorimi i TM për identifikimin dhe interpolimin e informacionit mund të reduktojë probabilitetin e mashtrimit, të rrisë prosperitetin e biznesit, të ndihmojë në parandalimin e terrorizmit dhe është e sigurt që risku do të menaxhohet në mënyrë të mirëfilltë dhe praktike.
Në këtë punim, u bë një hyrje përmbledhëse e fushës së gjerë të Text Mining. U përdor një përkufizim më formal i termit dhe u prezantuan në formë përmbledhëse metodat e disponueshme aktualisht për TM, vetitë dhe aplikimet e tyre në probleme specifike. Gjithashtu është përshkruar koncepti Text Mining, përafrimet e tij dhe disa zbatime në disa fusha të rëndësishme.

Literatura

Amado A., Cortez P., Rita P., & Moro S. (2018): Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. In: European Research on Management and Business Economics, Vol. 24(1): 1-7

Chen K. C. (2009): Text Mining e-Complaints Data From e-Auction Store With Implications for Internet Marketing Research. In: Journal of Business & Economics Research. Vol. 7(5): 15-24

Fan W., Wallace L., Rich S., & Zhang Z. (2006): Tapping the Power of Text Mining. In: Communications of the ACM, 49(9): 76–82

Feldman R., & Dagan I. (1995): Knowledge Discovery in Textual Databases (KDT). In: International Conference on Knowledge Discovery & Data Mining: 112-117

Frawley W. J., Piatetsky-Shapiro G., & Matheus C. J. (1992): Knowledge Discovery in Databases : An Overview. In: AI Magazine, 13(3): 57–70

Giorgi J. M., & Bader G. D. (2018): Transfer learning for biomedical named entity recognition with neural networks. In: Bioinformatics, Vol. 34 (23): 4087–4094

Gonzalez G. H., Tahsin T., Goodale B. C., Greene A. C., & Greene C. S. (2016): Recent advances and emerging applications in text and data mining for biomedical discovery. In: Briefings in Bioinformatics, 17(1): 33-42

Grimes S. (2005): The developing text mining market. White paper. In: Text Mining Summit Alta Plana Corporation, Boston: 1-12

Gupta V., & Lehal G. (2009): A survey of text mining techniques and applications. In: Journal of Emerging Technologies in Web Intelligence, 1(1): 60–76

Hearst M. A. (1999): Untangling Text Data Mining. In: Proceedings of ACL '99: the 37th Annual Meeting of the Association for computational Linguistics, (New Jersey: Association for Computational Linguistics), 3-10.

Karanikas H. & Mavroudakis Th. (2005): Text Mining Software Survey. In: RANLP Text Mining Workshop No 1: 39-48

Liao S.-H., Chu P.-H., & Hsiao P.-Y. (2012): Data mining techniques & applications–a decade review from 2000 to 2011. In: Expert Systems with Applications, vol. 39, no. 12: 11 303–11 311

Ojo A. K., & Adeyemo A. B. (2017): Characterisation of Academic Journal Publications Using Text Mining Techniques. In: Journal of Computer Sciences and Applications. 2017, 5(2): 42-49

PubMed (2018): US National Library of Medicine National Institutes of Health, In: https://www.ncbi.nlm.nih.gov/pubmed/: E aksesuar me 14 Prill 2018

Sharp M. (2001): Text Mining. In: Seminar in Information Studies, Prof. Tefko Saracevic. 11 December 2001

Xu W., Liu X., & Gong Y. (2003): Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: 267-273

Yao K., Zhang L., Luo T., & Wu Y. (2018): Deep reinforcement learning for extractive document summarization. In: Neurocomputing 284: 52–62

Zanasi A. (2009): Virtual Weapons for Real Wars: Text Mining for National Security. In: Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08. Advances in Soft Computing: 53-60