revista GEDECON Gestão e Desenvolvimento em Contexto
UNIVERSIDADE DE CRUZ ALTA - UNICRUZ

# UM ESTUDO DE MAPEAMENTO SISTEMÁTICO DA MINERAÇÃO DE DADOS PARA CENÁRIOS DE BIG DATA

*A systematic mapping study of data mining for big data scenarios*

Patricia Mariotto Mozzaquatro Chicon[1]
Fabricia Roos-Frantz[2]
Rafael Zancan Frantz[3]
Sandro Sawicki[4]

**RESUMO**

O volume de dados produzidos tem crescido em larga escala nos últimos anos. Esses dados são de diferentes fontes e diversificados formatos, caracterizando as principais dimensões do Big Data: grande volume, alta velocidade de crescimento e grande variedade de dados. O maior desafio é como gerar informação de qualidade para inferir insights significativos de tais dados variados e grandes. A Mineração de Dados é o processo de identificar, em dados, padrões válidos, novos e potencialmente úteis. No entanto, a infraestrutura de tecnologia da informação tradicional não é capaz de atender as demandas deste novo cenário. O termo atualmente conhecido como Big Data Mining refere-se à extração de informação a partir de grandes bases de dados. Uma questão a ser respondida é como a comunidade científica está abordando o processo de Big Data Mining? Seria válido identificar quais tarefas, métodos e algoritmos vêm sendo aplicados para extrair conhecimento neste contexto. Este artigo tem como objetivo identificar na literatura os trabalhos de pesquisa já realizados no contexto do Big Data Mining. Buscou-se identificar as áreas mais abordadas, os tipos de problemas tratados, as tarefas aplicadas na extração de conhecimento, os métodos aplicados para a realização das tarefas, os algoritmos para a implementação dos métodos, os tipos de dados que vêm sendo minerados, fonte e estrutura dos mesmos. Um estudo de mapeamento sistemático foi conduzido, foram examinados 78 estudos primários. Os resultados obtidos apresentam uma compreensão panorâmica da área investigada, revelando as principais tarefas, métodos e algoritmos aplicados no Big Data Mining.

**Palavras-chave:** Big Data. Mineração de Dados. Mapeamento Sistemático.

**ABSTRACT**

The volume of data produced has grown on a large scale in recent years. These data are from different sources and diverse formats, characterizing the main dimensions of Big Data: large volume, high growth speed and wide variety of data. The biggest challenge is how to generate quality information to infer meaningful insights from such large and varied data. Data Mining is the process of identifying valid, new and potentially useful patterns in data. However, the traditional information technology infrastructure is not able to meet the demands of this new scenario. The term currently known as Big Data Mining refers to the extraction of information from large databases. One question to be answered is how is the scientific community approaching the Big Data Mining process? It would be valid to identify which tasks, methods and algorithms have been applied to extract knowledge in this context. This article aims to identify in the literature the research work already carried out in the context of Big Data Mining. We sought to identify the areas most approached, the types of problems treated, the tasks applied in the extraction of knowledge, the methods applied to perform the tasks, the algorithms for the implementation of the methods, the types of data that have been mined, source and their structure. A systematic mapping study was conducted, 78 primary studies were examined. The results obtained present a panoramic understanding of the investigated area, revealing the main tasks, methods and algorithms applied in Big Data Mining.

**Keywords**: Big data. Data Mining. Systematic Mapping.

[1] Mestrado em Computação Aplicada - Unicruz- Universidade de Cruz Alta- RS, Brasil. E-mail: patriciamozzaquatro@gmail.com. Orcid: https://orcid.org/ 0000-0002-0510-6643

[2] Doutora em Tecnologia e Engenharia de Software - Unijuí- Ijuí- RS, Brasil. Email: frzfrantz@unijui.edu.br. Orcid: https://orcid.org/0000-0001-9514-6560

[3] Doutor em Tecnologia e Engenharia de Software - Unijuí- Ijuí- RS, Brasil. Email: frfrantz@unijui.edu.br. Orcid: https://orcid.org/0000-0003-3740-7560

[4] Doutor em Ciência da Computação - Unijuí - Ijuí - RS, Brasil. Email: sawicki@unijui.edu.br. Orcid: https://orcid.org/ 0000-0002-7960-0775



Impresso 1982-3266
Online 2318-9150

## 1 INTRODUÇÃO

Ao longo dos últimos anos, um grande volume de dados tem sido gerado a partir de uma variedade de dispositivos digitais, serviços de computação em nuvem (cloud computing) e do crescente progresso de tecnologias e dispositivos de IOT (Internet of Things). In addition, a evolução de tecnologias associadas aos sistemas de bancos de dados possibilitou o surgimento de vários tipos de bases de dados interligadas e heterogêneas. A heterogeneidade dos dados refere-se a dois diferentes aspectos: sintático e semântico. Os dados podem ser heterogêneos quanto a sua natureza, sendo estruturados, semiestruturados ou não estruturados. E também podem ser heterogêneos quanto ao seu significado e interpretação, podendo ser do tipo relacionais, fluxo de dados, espaciais, hipertexto e multimídia, web, redes, sequenciais, data warehouse e transacionais. Neste cenário, as organizações necessitam processar e analisar rapidamente estes dados a fim de extrair valor para uma tomada de decisão. No entanto, a infraestrutura de tecnologia da informação tradicional simplesmente não é capaz de atender estas demandas.

Conforme os autores Oussousa *et.al,* (2017), as técnicas e modelos tradicionais de hardware e software não conseguem coletar, integrar, gerenciar e processar dados distribuídos provindos de fontes diversas de natureza não estruturada. Os seguintes desafios são citados: i) coletar, integrar e armazenar, com menos hardware e requisitos de software, enormes conjuntos de dados gerados a partir de fontes distribuídas (Chen *et al*., 2014); (Najafabadi *et al.*, 2015); ii) gerenciar a complexidade da natureza do Big Data (velocidade, volume e variedade) (Khan *et al*., 2014) e processá-lo de forma distribuída em um ambiente com uma mistura de aplicativos; iii) sincronizar fontes de dados externas e distribuir placas de Big Data (incluindo aplicativos, repositórios, sensores, redes) com as infra-estruturas internas de uma organização; iv) velocidades de E / S aumentam com a densidade lentamente) (Chen & Zhang, 2014). Consequentemente, as capacidades deste sistema desequilibrado podem diminuir o acesso aos dados e afetar o desempenho e a escalabilidade dos aplicativos de Big Data; v) análise em tempo real, como navegação, redes sociais, portanto, algoritmos avançados e métodos eficientes de mineração de dados são necessários. Desta forma, novos paradigmas de processamento são necessários para descobrir insights, tomar melhores decisões e otimizar o processo.

O processo conhecido como Descoberta de Conhecimento em Bases de Dados (KDD) vem sendo utilizado como uma forma de extrair informação de qualidade de um conjunto de dados por meio de algumas etapas fundamentais: Pré-processamento, Mineração de Dados e Pós-processamento ou Análise da Solução. O Pré-processamento compreende as ações para adequar os dados aos algoritmos de Mineração de Dados a serem aplicados. A Mineração de Dados é uma etapa não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e compreensíveis (HAN; KAMBER, 2011). O Pós-processamento refere-se à interpretação dos resultados e tradução de padrões úteis em termos inteligíveis pelos usuários.

Existem pesquisas que abordam as limitações e desafios na utilização do processo KDD tradicional quando aplicado ao contexto dos dados atuais. Particularmenete, no caso da

Impresso 1982-3266
Online 2318-9150

etapa de Mineração de Dados aplicada ao contexto do Big Data, apesar de existirem pesquisas relacionadas (Yan *et al*. 2016) e (Qiu *et al.* 2016), até onde sabemos, não existem trabalhos que sistematizem quais as tarefas, métodos e algoritmos vêm sendo empregados neste contexto. Neste trabalho, nos referiremos à etapa de mineração de dados aplicada a grandes volumes de dados pelo termo Big Data Mining, conforme (Sangeetha; Prakash, 2017) e (Oussousa et.al, 2017). Nosso principal objetivo é averiguar como a comunidade científica está abordando o processo de Big Data Mining; buscou-se identificar quais tarefas, métodos e algoritmos vêm sendo aplicados para extrair conhecimento neste contexto. A partir desse objetivo, revisou-se as áreas mais abordadas, os tipos de padrões tratados, as tarefas aplicadas na extração de conhecimento, os métodos aplicados para a realização das tarefas, os algoritmos para a implementação dos métodos, os tipos de dados minerados, fonte e estrutura dos dados.

Observou-se que as publicações são centradas em resolver problemas cuja característica relaciona-se a classificação, caracterização e discriminação aplicadas ao problema da Análise de Sentimentos no contexto de Business Intelligence, Educação e Saúde. Indentificou-se que a maioria dos estudos analisados utilizou as tarefas de Mineração Classificação e Clusterização. Os métodos Classificador Bayesiano, Classificador Vizinho mais Próximo, Suporte Vector Machine, Árvore de Decisão e Agrupamento Particional foram os mais citados. Os algoritmos mais citados nas pesquisas são os seguintes: Naive Bayes, K- Nearest Neighbor, SVM, J48 e K-Means. Quanto aos tipos de dados a serem minerados, a maioria tratou dados tipo Web e Relacional, originando-se das fontes como: Twitter, Comentários, Mensagens de Spam, Fórum de discussão, Resenhas de filmes, Comentários em Mídias Sociais, Dados de vendas e suporte, Mídias Sociais, Mensagens de texto, SMS, Facebook, Logs de auditoria, dentre outros. Quanto a sua estrutura a maioria classificou-se como Dados Estruturados e Semi Estruturados.

Este artigo está organizado da seguinte forma: A seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve o método adotado para o mapeamento sistemático. A Seção 4 descreve os resultados. Por fim, na seção 5, nós apresentamos as conclusões.

## 2 TRABALHOS RELACIONADOS

Nesta seção, discute-se os estudos de revisão identificados na literatura que abrangem temas relacionados com o Big Data Mining.

A pesquisa de Dutt *et al*. (2016), intitulada "A Systematic Review on Educational Data Mining" tem por objetivo desenvolver uma revisão sistemática da literatura ao longo de três décadas (1983-2016) sobre algoritmos de clustering e sua aplicabilidade e usabilidade no contexto da mineração de dados educacionais. A abordagem de agrupamento foi aplicada a diferentes variáveis, apresentando o algoritmo K-means e C-Means na maioria dos estudos.

Leena e Mansaf (2016) fizeram uma revisão sistemática intitulada" Educational Data Mining and its Role in Determining Factors Affecting Students Academic Performance: A Systematic Review". Os autores identificaram as ferramentas de mineração de dados disponíveis

Impresso 1982-3266
Online 2318-9150

na área educacional, tais como: R, RapidMiner, KEEL, WEKA, KNIME, ROSETTA, ANOVA, Clementine, Neuro Shell Classifier. Dentre as várias técnicas utilizadas para minerar dados, a tarefa de classificação é uma das mais comuns. Para implementar a tarefa citada, a revisão aponta os algoritmos: KNN, Decision Tree e Naïve Bayes.

O trabalho de Kaur (2017), intitulado "Systematic mapping study of big data mining tools and techniques" aborda um mapeamento sistemático sobre ferramentas, técnicas de mineração de dados e domínio de aplicação no contexto do Big Data. O software Hadoop foi indicado pela maioria dos estudos, seguido do software R. A tarefa para minerar os dados com maior indicação nos estudos analisados foi a tarefa de Clusterizadação, seguida da tarefa de Associação. O estudo aponta como principais áreas de aplicação: tomada de decisão, política e mineração sobre crimes. Ainda, grande parte dos estudos confirmou que a escolha da tarefa de mineração depende da situação, objetivo e do padrão a ser minerado.

A pesquisa de Bonidia *et al*. (2018) apresenta uma revisão sistemática sobre a aplicação da mineração de dados em dados esportivos. O estudo intitula-se "Data Mining in Sports: A Systematic Review". Na revisão se identificou as seguintes tarefas: Classificação, Clusterização, Associação e Regressão. Após a identificação das tarefas, os seguintes algoritmos e métodos foram detectados: Bat algorithm, Redes MLP, Algoritmo CART, Neural Network Ensemble, Classificador Bayesiano, Algoritmo SMO, Random Forest, Algoritmo C4.5, Logistic Regression, Algoritmo de Levenberg-Marquardt, Algoritmo de agrupamento K-means, Agrupamento hierárquico, Algoritmo Apriori e Backpropagation Algorithm.

Consideramos que os trabalhos relacionados descritos nesta seção se diferenciam desta proposta em diversos aspectos. A maioria dos trabalhos citados não tem como objetivo identificar as áreas de aplicação, nem os tipos de padrões tratados no contexto do Big Data Mining em geral. Embora o trabalho de Kaur (2017) identifique o domínio de aplicação, este estudo não se preocupa em identificar quais são os padrões tratados. Apenas o trabalho de Bonidia *et al*. (2018) aborda tarefas, métodos e algoritmos no contexto de Big data mining, porém o estudo trata apenas de propostas no domínio específico do esporte. Os trabalhos de Kaur (2017) e Leena e Mansaf (2016) buscam revisar apenas ferramentas e métodos aplicados no Big data mining. Dutt *et al.* (2016) revisaram apenas algoritmos de clustering aplicados a dados educacionais. Nenhum dos estudos analisados revisou tipos e fontes de dados a serem minerados, bem como sua classificação quanto a sua heterogeneidade (sintaxe e semântica).

## 3 MÉTODO DE PESQUISA

Nesta seção, apresenta-se a metodologia aplicada para este mapeamento sistemático (MS). O processo realizado segue, principalmente, as orientações descritas em (KITCHENHAM; CHARTERS, 2007) e está organizado nas seguintes atividades: definição do objetivo; definição do protocolo e condução.

Impresso 1982-3266
Online 2318-9150

### 3.1 Definição do Objetivo

A primeira atividade é a definição do objetivo do MS, que indica sua real necessidade. O objetivo deste MS é identificar e apresentar um panorama atual sobre Big Data Mining.

### 3.2 Definição do protocolo

A segunda atividade é a definição do protocolo da revisão, ou seja, os passos essenciais para sua execução. Esta atividade está dividida em seis passos: definição das questões de pesquisa, definição do método de busca, definição das fontes de busca, definição da string de busca, definição dos critérios para a seleção dos estudos e definição dos critérios de qualidade.

#### 3.2.1 Definição das Questões de Pesquisa

Foram definidas seis questões de pesquisa e duas sub-questões relacionadas. (QP1). Quando e onde os estudos têm sido publicados? (QP2). Quais problemas de mineração foram tratados no contexto do Big Data Mining e em que áreas? (QP3). Como classificar os dados que foram minerados nos estudos primários quanto a estrutura? (QP4). Como classificar os dados que foram minerados nos estudos primários quanto a Heterogeneidade semântica? (QP5). Qual a origem dos dados minerados? (QP6). Quais tarefas de Mineração de Dados foram aplicadas na extração do conhecimento no contexto de Big Data? (QP6.1). Quais as técnicas ou métodos foram aplicados para a realização das tarefas identificadas? (QP6.2). Quais são os algoritmos utilizados para implementar as técnicas?

#### 3.2.2 Definição do Método de Busca

O método de busca a ser utilizado é um ponto chave para o sucesso ou fracasso de um MS. Há três métodos principais que podem ser aplicados: busca automática em bases de dados digitais, busca manual e bola de neve (snowballing). Snowballing é realizado identificando trabalhos nas referências dos estudos primários selecionados que estão relacionados com o tema abordado (JALALI; WOHLIN, 2012), (WOHLIN, 2014). De maneira geral, para que um mapeamento ofereça uma visão ampla do tópico, é necessário realizar buscas automáticas. Assim, em MSs, busca manual e snowballing são considerados métodos complementares a serem usadas em conjunto com a busca automática. Neste MS foram utilizadas a busca automática em bases de dados digitais e a busca bola de neve (snowballing). Para dar suporte a organização e manipulação dos estudos primários, utilizou-se a ferramenta Parsif, uma ferramenta online projetada para auxiliar os pesquisadores na condução de mapeamentos e revisões sistemáticas de literatura, seguindo Keele *et al*. (2007).

#### 3.2.3 Definição das fontes de busca

A escolha de quais fontes de busca utilizar em um MS depende do esforço a ser

Impresso 1982-3266
Online 2318-9150

empreendido no mapeamento. No entanto, quanto maior a quantidade de fontes de pesquisa, maior é a chance de se obter uma cobertura abrangente dos trabalhos publicados. Dessa forma, buscando encontrar um equilíbrio, as bases de dados científicas selecionadas foram as seguintes: ACM Digital Library[5] , IEEE Xplore Digital Library[6] e Scopus[7] . Essas bases são atualizadas constantemente, indexam os principais journals, conferências e workshops da área e são utilizadas em diversos estudos secundários e terciários que mapeiam tópicos relacionados.

### 3.2.4 Definição da string de busca

Durante a definição da string de busca, o foco é a identificação de termos relacionados ao tópico de pesquisa, ou seja, as palavras – chave utilizadas nos estudos primários alvo do MS. Uma boa prática consiste em agrupar termos relativos a um mesmo aspecto, que podem ser considerados sinônimos, concatenando-os com o conectivo OR. Posteriormente, cada grupo de termos é concatenado com os demais por meio de conectivos AND. Vale ressaltar que os termos da string de busca devem estar alinhados ao objetivo e às questões de pesquisa do MS. As palavras-chave escolhidas foram as seguintes: Big Data, Data Mining, Techniques, Methods, Algorithms; termos relacionados a técnicas, métodos e algoritmos, todos no idioma inglês. Com base nas palavras-chave, a string de busca foi definida: (((“Big Data”) AND (“data mining”) AND ((techniques AND methods) OR (techniques AND algorithms))).

### 3.2.5 Definição dos critérios de seleção

A definição de critérios de seleção é fundamental para a qualidade do MS, estabelecendo características que um estudo deve conter para ser considerado relevante no contexto do MS (critérios de inclusão) e características que levam à exclusão de estudos que não obedecem aos critérios definidos (critérios de exclusão). Os critérios se baseiam nas questões de pesquisa e devem ser descritos previamente no protocolo e rigorosamente aplicados, podendo ser refinados durante o processo de busca. Os critérios de inclusão foram definidos com base nos objetivos do MS e devem ser aplicados em conjunto (ou seja, para um estudo ser incluído, ele deverá satisfazer todos os critérios de inclusão). Os critérios de inclusão foram os seguintes: CI- 1: O estudo primário foi publicado entre os anos de 2010 a 2018. CI- 2: O estudo primário aborda claramente a área de Mineração de dados aplicada a Big Data. CI- 3: O estudo primário apresenta pelo menos uma tarefa de Mineração de Dados; CI- 4: O estudo primário apresenta pelo menos um método de Mineração de Dados; CI- 5: O estudo primário indica qual o algoritmo usado na Mineração de Dados; CI- 6: O estudo primário está publicado em inglês. Os critérios de exclusão foram: CE-1: O estudo é um relatório técnico, um documento que está disponível no formato de resumo, é uma apresentação, uma chamada de artigo. CE-2: O estudo é um Livro ou um Capítulo de livro. CE-3: O estudo é um estudo secundário (survey, revisão de

[5] http://portal.acm.org
[6] http://ieeexplore.ieee.org
[7] http://www.scopus.com

Impresso 1982-3266
Online 2318-9150

literatura, mapeamento sistemático, revisão sistemática) ou estudo terciário. CE-4: O estudo é literatura cinzenta. CE-5: Artigos duplicados. CE-6: O Texto completo do estudo primário não está disponível. Foi definido o período compreendido entre 2010 e 2018.

#### 3.2.6 Definição dos critérios de qualidade

O critério de qualidade é utilizado para avaliar e julgar a qualidade dos estudos selecionados, podendo ser usado para excluir estudos abaixo de certo limiar de qualidade. Para analisar a qualidade desses 78 estudos primários, estabelecemos uma lista de verificação contendo quatro perguntas (ou critérios de qualidade), com base na avaliação da qualidade dos estudos primários propostos por Kitchenham e Charters (2007): Q1: Existe uma justificativa para o porquê do estudo ser realizado? Q2: É apresentada uma visão geral sobre o estado da arte da área em que o estudo é desenvolvido? Q3: É um artigo no contexto do Big Data Mining? Q4: O artigo descreve alguma tarefa, método e algoritmo? Para cada questão, foi aplicado o seguinte ponto de escala, conforme Garcés et. al (2016): (i) o estudo atende plenamente a um determinado critério de qualidade (1 ponto); (ii) o estudo atende ao critério de qualidade até certo ponto (0,5 ponto); e (iii) o estudo não atende a esse critério de qualidade (0 ponto). Estudos com pontuação acima ou igual a 0,5 foram considerados para a extração dos dados.

### 3.3 Condução

A atividade de condução do MS consiste em obter os resultados da aplicação do protocolo definido. Esta atividade é dividida em quatro passos: identificação dos estudos primários; seleção dos estudos primários, por meio da aplicação de critérios de seleção (critérios de inclusão e de exclusão), bem como os critérios de qualidade; extração e categorização dos dados contidos nos estudos selecionados; e síntese dos dados.

#### 3.3.1 Identificação dos estudos primários

Para a identificação dos estudos primários, utilizou-se o método de busca automática nas bases de dados, aplicando o critério de inclusão CI-1: O estudo primário foi publicado entre os anos de 2010 a 2018 e também o critério de exclusão CE:5: Artigos duplicados.

#### 3.3.2 Seleção dos estudos primários

Para a seleção dos estudos primários foram realizadas três etapas: seleção com base em pré-leitura, seleção com base em leitura completa e seleção com base em snowballing. Na seleção com base em pré-leitura, os critérios de inclusão e exclusão foram aplicados em todos os estudos identificados, por meio da avaliação de seus títulos, resumos e palavras-chave. Muitas vezes é difícil identificar se um estudo é ou não relevante apenas com a leitura do título, resumo e palavras-chave. Assim, na dúvida sobre a inclusão ou não de um estudo durante a primeira etapa, optou-se pela sua inclusão, sendo a decisão pela permanência tomada durante

Impresso 1982-3266
Online 2318-9150

a segunda etapa. Na seleção baseada em leitura completa, os critérios de inclusão e exclusão foram aplicados nos estudos por meio da avaliação de seus textos completos. Na seleção com base em snowballing, alguns trabalhos que não foram encontrados com a busca automática, mas que atendem os critérios de inclusão e exclusão, foram adicionados.

### 3.3.3 Extração e categorização dos dados

A atividade de extração e categorização dos dados busca, a partir dos estudos primários selecionados, definir esquemas de classificação de dados necessários para responder as questões de pesquisa do MS, e então extrair os dados e categorizá-los de acordo com estes esquemas. Para se obter um MS de qualidade é imprescindível ter um esquema de classificação confiável e bem definido (PETERSEN *et al*. 2015). Para derivar esquemas de classificação, três abordagens principais podem ser adotadas: (i) adotar esquemas existentes, (ii) definir um esquema previamente com base na literatura, ou (iii) deixar os esquemas emergirem dos próprios estudos selecionados (KITCHENHAM; CHARTERS, 2007).

Para derivar os esquemas de classificação utilizados neste MS, foram usadas essas três abordagens. Para a questão de pesquisa QP1 foi adotado um esquema de classificação existente. Para as questões de pesquisa QP2 e QP5 os esquemas de classificação emergiram dos próprios estudos selecionados. Finalmente, para as questões QP3, QP4 e QP6 os esquemas de classificação foram definidos com base na literatura.

### 3.3.4 Síntese dos dados

A síntese de dados envolve o resumo e a discussão dos resultados do mapeamento. De acordo com Kitchenham e Charters (2007), a síntese pode ser descritiva ou quantitativa.

## 4 RESULTADOS

Esta seção apresenta os resultados da Identificação dos estudos primários, da Seleção dos estudos primários e da Extração e categorização dos dados.

### 4.1 Identificação e seleção dos estudos primários

Para a identificação dos estudos primários, uma busca automática foi realizada nas bases de dados aplicando o critério de inclusão CI-1: O estudo primário foi publicado entre os anos de 2010 a 2018 e também o critério de exclusão CE:5: Artigos duplicados. Como resultado, foram obtidos um total de 3258 estudos primários: 1585 a partir da ACM Digital Library, 400 da IEEE Digital Library e 1273 da Scopus.

Uma vez identificados o conjunto de potenciais estudos primários, eles foram avaliados quanto à sua relevância e então selecionados para fazer parte do conjunto de trabalhos a ser analisados na condução do MS. A seleção foi realizada em três etapas: seleção baseada em pré-leitura, seteção baseada em leitura completa e seleção baseada em snowballing.

Impresso 1982-3266
Online 2318-9150

Na primeira etapa, os metadados dos artigos contendo dados como título, resumo, URL do artigo, palavras-chave e número DOI, foram exportados das bases citadas, em formato de arquivo ".BIB" e, posteriormente, os metadados foram importados na ferramenta Parsif[8] . A partir dos 3258 estudos primários identificados, realizou-se a pré-leitura de cada um deles e aplicou-se os demais critérios de inclusão e exclusão. Quando necessário, a leitura das seções de Introdução e de Conclusão foram consideradas. Como resultado da seleção baseada em pré-leitura, um conjunto de 277 estudos primários potencialmente relevantes foram selecionados.

Na segunda etapa, cada um dos 277 estudos foi lido integralmente e analisado considerando novamente os critérios de inclusão e exclusão, reduzindo o corpus final da pesquisa para 73 estudos. Após a aplicação do critério de qualidade, aqueles estudos que não tratavam de Big Data Mining, e não apresentavam explicitamente a aplicação de alguma tarefa de Data Mining, método e algoritmo foram excluídos. Portanto, como resultado da seleção baseada em leitura completa foram selecionados 73 estudos e excluídos 204.

Por último, na terceira etapa, o método de busca snowballing foi utilizado, resultando na inclusão de outros cinco (5) estudos. Portanto, como resultado da seleção foram selecionados 78 estudos. Uma lista de referências para o conjunto de artigos primários selecionados pode ser vista no Apêndice A.

### 4.2 Extração e categorização dos dados

Nesta seção, apresenta-se os esquemas para a extração dos dados conforme cada uma das questões de pesquisa, e os dados extraídos e categorizados seguindo estes esquemas.

**(QP1). Quando e onde os estudos têm sido publicados?**

Para responder essa questão de pesquisa, com respeito a quando os estudos foram publicados, a quantidade de estudos primários selecionados foram separados e classificados por ano. Foram encontrados os seguintes estudos no período de 2010 a 2018: 2010 (3 estudos), 2011 (1 estudo), 2012 (2 estudos), 2013 (6 estudos), 2014 (8 estudos), 2015 (7 estudos), 2016 (14 estudos), 2017 (15 estudos) e 2018 (21 estudos). Observa-se o crescimento de publicações nos últimos anos, ascentuando-se a partir de 2016. Apenas os estudos publicados até o início de outubro de 2018 foram considerados neste MS, uma vez que a atividade de seleção foi conduzida neste período. Para responder essa questão de pesquisa, com respeito a onde os estudos foram publicados, utilizamos uma classificação bastante adotada para tipos de veículos de publicação, que considera três categorias básicas: periódicos, conferências (incluindo simpósios) e workshops (eventos menores). Seguindo este esquema de classificação dos dados, identificamos o veículo de publicação de cada um dos 78 estudos e os categorizamos da seguinte maneira: 68 artigos publicados em Conferencia, 9 em Periódicos e 1 em Simpósio.

[8] https://parsif.al/

Impresso 1982-3266
Online 2318-9150

**(QP2). Quais problemas de mineração foram tratados no contexto do Big Data Mining e em que áreas?**

Para responder esta questão de pesquisa, os estudos foram analisados individualmente e classificados de acordo com o problema de mineração que pretendiam resolver. Todos os problemas de mineração de dados identificados nos estudos primários estão listados na Tabela 1, coluna 2. A primeira coluna da tabela relaciona os identificadores criados para cada um dos problemas, a segunda coluna lista esses problemas e a terceira mostra os estudos primários que tratam cada um deles.

Tabela 1- Problemas de mineração e estudos primários correspondentes

| ID | Problema | # Estudos |
|---|---|---|
| AS | Analisar sentimentos | [S1], [S4], [S12], [S13], [S18], [S19], [S20], [S22], [S24], [S27], [S34], [S54], [S55], [S61], [S62],[S63], [S65],[S69], [S71], [S73] |
| DS | Detectar spam em redes sociais | [S5], [S53] |
| RI | Recomendar informações | [S17], [S46], [S51] |
| DF | Detectar fraudes | [S15], [S59] |
| DD | Diagnósticar doenças | [S3],[S9],[S21],[S25],[S30], [S35],[S43],[S58],[S60],[S72] |
| DW | Detectar worms | [S23], [S29] |
| DDA | Detectar desempenho acadêmico e comportamento do aluno | [S7], [S14], [S56] |
| ADS | Agrupar dados semelhantes | [S2], [S6], [S10],[S49],[S57], [S64], [S66], [S76], [S78] |
| PC | Prever eventos climáticos | [S44] |
| CD | Classificar dados | [S11],[S26],[S33],[S45],[S48], [S67],[S68], [S75] |
| PED | Minimizar problemas de escalabilidade trabalhando com grandes conjuntos de dados | [S38], [S39], [S40], [S41], [S47], [S50], [S70] |
| DA | detectar outlier | [S52] |
| RIW | Recuperar informação na web | [S32] |
| ARC | Análisar riscos de crédito | [S28] |
| PIP | Predizer a identidade profissional de estudantes universitários | [S36] |
| DIA | Detectar a insegurança alimentar por meio de dados do tweets | [S37] |
| IIC | Identificar informações sócio demográficas do consumidor | [S74] |
| DST | Detectar sinal de tráfego e traçar rotas | [S31], [S77] |
| PS | Realizar previsão em sites de e-commerce | [S8] |
| DC | Detectar comportamento de usuário relacionado a segurança de informações | [S16] |
| DMTM | Detectar modelos para reduzir a toxidade de um medicamento | [S42] |

Fonte: Elaborada pelos Autores

Impresso 1982-3266
Online 2318-9150

Para responder a questão referente às áreas de aplicação dos estudos, cada estudo foi analisado individualmente para identificar em que área de aplicação se encontra o problema de mineração tratado. As seguintes áreas integram os estudos analisados: Business Intelligence ([S4],[S5],[S8],[S12],[S13],[S17],[S18],[S22],[S24],[S51],[S53],[S54],[S55],[S61],[S62],[S67],[S69],[S73], [S74]), Eduacação([S2],[S6],[S7],[S10],[S14],[S20],[S26],[S27],[S36],[S38], [S39],[S40],[S41],[S45],[S47],[S49],[S50],[S52],[S56],[S57],[S64],[S65],[S66],[S68],[S70], [S70],[S75],[S76]), Saúde([S3],[S9],[S11],[S21],[S25],[S30],[S32],[S33],[S34],[S35],[S37], [S42],[S43],[S58],[S60],[S72]), Segurança da informação ([S15],[S16],[S23],[S29], [S59]), Politica ([S48]), Ciências sociais ([S19], [S63]), Civil ([S31], [S77]), Metereologia ([S44]), Elétrica ([S48]), Financeira ([S28]), Computação em nuvem ([S46], [S78]) e Agricultura ([S71]).

**(QP3). Como classificar os dados que foram minerados nos estudos primários quanto a estrutura?**

O esquema de classificação utilizado para responder esta questão foi construído com base na literatura. De acordo com Han e Kamber (2011), os dados a serem minerados podem ser classificados, quanto a sua estrutra, em: estruturado, semiestruturado e não estruturado. Cada estudo foi analisado e categorizado pela estrutura dos dados originais que abordam, antes de qualquer tratamento. Pode-se observar que 43 estudos abordam dados não estruturados ([S1],[S3],[S4],[S5],[S7],[S8],[S12],[S13],[S14],[S15],[S17],[S18],[S20],[S22],[S23],[S24], [S27],[S29],[S30],[S31],[S36],[S37],[S41],[S43],[S44],[S46],[S47],[S51],[S53],[S54],[S55], [S56], [S60],[S61][S62],[S63],[S67],[S69],[S71],[S73],[S74],[S75]),[S77]) e 33 estudos dados estruturados([S2],[S6],[S9],[S10],[S11],[S16],[S19],[S21],[S25][S26],[S28],[S32],[S33],[S35] [S38],[S39],[S40],[S42],[S45],[S48],[S49],[S50],[S52],[S57],[S58],[S59],[S64],[S66],[S68],[S70],[S72],[S76],[S78]).Assim, constatou-se que a maioria dos dados de origem classificam-se quanto a estrutura como não estruturados.

**(QP4). Como classificar os dados que foram minerados nos estudos primários quanto a Heterogeneidade semântica? e (QP5). Qual a origem dos dados minerados?**

O esquema utilizado para responder a questão **QP4** foi baseado na literatura. De acordo com Han e Kamber (2011), os dados a serem minerados podem ser classificados, quanto a sua semântica, nos seguintes tipos: relacionais, fluxo de dados, espaciais, hipertexto e multimídia, web, redes, sequenciais, data warehouse e transacionais. Dessa forma, de cada estudo primário se extraiu a informação sobre o tipo de dado tratado na mineração, quanto a sua semântica, e então se categorizou cada um deles. Foram identificados dados de 5 diferentes tipos. A maioria dos estudos trataram dados do tipo Web([S1],[S4],[S5],[S8],[S10],[S12],[S13], [S16],[S17],[S18],[S20],[S22],[S23],[S24],[S26],[S27],[S28],[S29],[S30],[S32],[S34],[S36],[S37],[S41],[S43],[S46],[S51],[S53],[S54],[S56],[S57],[S61],[S62],[S63],[S64],[S65],[S67],[S68],[S69],[S70],[S73],[S74],[S75],[S76],[S78]), totalizando 45 e do tipo Relacional ([S2], [S3], [S6],[S7],[S9

Impresso 1982-3266
Online 2318-9150

],[S11],[S14],[S15],[S19],[S21],[S25],[S31],[S33],[S35],[S38],[S39],[S40],[S42],[S45],[S48],[S49],[S50],[S52],[S58],[S59],[S60],[S66],[S72]), totalizando 28.

O esquema utilizado para responder a questão **QP5** emergiu dos próprios estudos. Dessa forma, de cada estudo primário se extraiu a informação de onde proveem os dados utilizados nos experimentos/intervenção. Constatou-se que os dados se originaram das seguintes fontes: Twitter, Comentários, Mensagens de Spam, Fórum de discussão, Resenhas de filmes, Comentários em Mídias Sociais, Dados de vendas e suporte, Mensagens de texto, Facebook, Logs de auditoria, Dados de sensores captados sinal de trânsito, Tabelas, Documentos, Relatórios, Questionários, Data sets, Instagram, Blogs e Dados Metereológicos.

**(QP6). Quais tarefas de Mineração de Dados foram aplicadas na extração do conhecimento no contexto de Big Data?**

**(QP6.1). Quais as técnicas ou métodos foram aplicados para a realização das tarefas identificadas?**

**(QP6.2). Quais são os algoritmos utilizados para implementar as técnicas?**

Para responder esta questão de pesquisa, o primeiro esquema de classificação utilizado foi definido com base na literatura. Han e Kamber (2011) cita um conjunto de características que determinam os problemas de mineração, a saber: Caracterização e discriminação; Padrões a serem gerados; Associações e correlações entre os dados; Classificação de dados; e, Análise de cluster. Cada problema de mineração identificado na QP2 foi analisado e então categorizado de acordo com essas características, como pode ser observado na Tabela 2, primeira e segunda colunas. Posteriormente, cada estudo foi analisado para identificar qual tarefa de mineração foi aplicada para resolver os problemas identificados na QP2, conforme pode ser visto na terceira coluna. Por último, a quarta coluna lista os estudos primários correspondentes a cada categoria.

Tabela 2- Tarefas classificadas quanto a Característica do problema

| Característica do problema | Problema | Tarefa | Identificadores dos estudos primários |
|---|---|---|---|
| Caracterização e discriminação | CD, DIA, IIC, PC, PED, PIP | Classificação | [S14], [S36], [S37], [S38], [S39], [S40], [S41], [S44],[S47], [S50],[S52],[S67], [S68], [S70], [S74], [S75] |
| Padrões a serem gerados | CD, DC | Associação | [S16], [S26], [S33], [S45], [S48] |
| Associações e correlações entre os dados | DF, PS | Associação | [S8], [S11],[S15],[S59] |
| Classificação de dados | ARC, AS, CD, DD DF, DDA, DS, DMTM, DST, DW, PC, RI, DA | Classificação | [S1], [S4], [S7],,[S9], [S12], [S13],[S17],[S18], [S19],[S20],[S22], [S23][S24],[S25],[S27], [S28], [S29], [S31][S34], [S42],[S46],[S5], [S51],[S53], [S54], [S55],[S56],[S61],[S62],[S63],[S65],[S69], S71], [S73],[S77] |
| Análise de cluster | DD, DF, RIW, ADS | Clusterização | [S2],[S3],[S6],[S10],[S21],[S30],[S32],[S35],S43] [S49],[S57],[S58],[S60],[S64],[S66],[S72],[S76], [S78] |

Fonte: Elaborado pelos Autores

Impresso 1982-3266
Online 2318-9150

Para responder sobre os métodos e algoritmos utilizados nos estudos primários, um segundo esquema de classificação foi organizado, neste caso por tarefa de mineração. Com base na tarefa de mineração os estudos foram analisados para identificar quais métodos foram utilizados e consequentemente quais algoritmos foram aplicados para resolver o problema de mineração, conforme pode ser visto na Tabela 3. Esta tabela também mostra os estudos classificados pela estrutura dos dados utilizados na mineração. Esta classificação é diferente da apresentada na QP3 na qual 43 estudos mineraram problemas que tratam dados não estruturados. No entanto, desses estudos, aqueles que mineraram problemas com características de associação e correlação entre dados, análise de cluster e padrões a serem gerados, para que se pudesse aplicar determinados métodos, os dados precisaram ser estruturados. Dessa forma, de cada estudo primário se extraiu a informação sobre o tipo de dado tratado na mineração, quanto a sua estutura, antes e após o tratamento. Quanto a estrutura dos dados, antes do tratamento, 43 estudos integravam dados no formato não estruturado; 33 no formato estruturado e 2 no formato semiestruturado. Quanto a estrutura dos dados, depois do tratamento, 66 estudos integravam dados no formato estruturado; 7 no formato semiestruturados e 5 no formato não estruturado.

Tabela 3- Tarefas, Métodos, Algoritmos e Estrutura

| **Tarefa de Classificação** | | | | |
|---|---|---|---|---|
| Método | Algoritmo | Estrutura | | |
| | | E | SE | NE |
| Classificador Bayesiano | Naive Bayes | [S3],[S4], S8],[S9],[S13],[S14],[S15], [S18],[S22] [S24],[S25],[S27], [S29], [S30], [S31],[S35], [S37], [S44], [S52],[S55],[S56], [S58],[S60],[S61],[S67], [S72],[S73] | [S1], [S34], [S53], [S54], [S62], [S65], [S69] | [S5], [S7], [S20], [S51], [S71] |
| | Bag-of-Words | [S60] | [S65] | |
| | Bayesnet | [S56] | | |
| Classificador Vizinho Mais Próximo | K- Nearest Neighbor | [S3], [S12], [S13] [S23], [S27], [S28], [S37], [S40], [S41], [S45], [S46], [S50], [S60], [S75], [S77] | | |
| | Fuzzy- K- Nearest Neighbor | [S39], [S40] | | |
| | HSTF- K- Nearest Neighbor | [S39] | | |
| | Fuzzy Membership | [S46] | | |
| | Hybrid spill-tree | [S39] | | |
| Rede Neural Artificial | Support Vector Machines | [S1], [S12], [S28], [S43], [S52], [S55],[S74] | [S1], [S34], [S65], | [S20] |
| Suporte Vector Machine | Support Vector Machines | [S3], [S4] [S8], [S9], [S14], [S17], [S18], [S24], [S25], [S26], [S27], [S29], [S30], [S33], [S35], [S37], [S47], [S48], [S52], [S55], [S61], [S67], [S72], [S73],[S74] | [S53], [S54], [S62], [S65], [S69] | [S20] |

Impresso 1982-3266
Online 2318-9150

| | | | | |
|---|---|---|---|---|
| Árvore de decisão | C4.5 | [S21], [S23], [S25], [S52] | [S34], [S53] | |
| | ID3 | [S3],[S15], [S25], [S31], [S44], [S63] | | |
| | Rep Tree | [S72] | | |
| | J48 | [S9], [S14], [S28], [S29], [S56], [S58], [S60], [S67], [S72], [S73] | | |
| | Randow Forest | [S35], [S42], [S58],[S70] | | |
| | Randow Tree | [S31], [S42] | | |
| | XCS | [S70] | | |
| | Decision Tree | [S36], [S37] | | |
| Algoritmos genéticos | | [S11],[S33] | | |
| Classificador baseado em regras | PART | [S4] | | |
| | Jrip | [S56] | | |
| **Tarefa de Clusterização** | | | | |
| Agrupamento Particional | K-Means | [S2], [S3], [S6], [S10], [S14], [S19], [S27], [S32], [S49],[S57],[S59],[S64], [S66],[S68],[S76] | | |
| | K-Medoids | [S3], [S6] | | |
| | Bare Bones | [S68] | | |
| | Fuzzy k-means | [S32] | | |
| | C-Means | [S78] | | |
| Agrupamento Hierárquico | BIRCH | [S6] | | |
| Rede Neural Artificial | | [S49] | | |
| Agrupamento Baseado em Densidade | NK hybrid | [S66] | | |
| | DBSCAN | [S66], [S76] | | |
| Agrupamento Baseado em Grade | Clique | [S52] | | |
| Clustering baseado em Intervenção | Clustering baseado em Intervenção | [S35] | | |
| **Tarefa de Associação** | | | | |
| Padrões Frequentes | Apriori | [S11],[S14],[S15], [S16], [S38],[S59] | | |

Fonte: Elaborada pelos Autores

## 5 CONCLUSÃO

Neste artigo, apresentamos os resultados de um mapeamento sistemático sobre a aplicação de Tarefas, Métodos e Algoritmos de Mineração de Dados no contexto do Big Data. Nosso estudo corrobora com o crescente interesse em aplicar a Mineração de Dados no contexto do Big Data, conforme mostrado pelo número de publicações recentes.

O problema a ser minerado na maioria dos estudos referiu-se a análise de sentimentos. As áreas de aplicação mais comuns referem-se à Educação, Business Intelligence e Saúde.

Impresso 1982-3266
Online 2318-9150

As tarefas de Mineração Classificação e Clusterização aparecem nas pesquisas com maior proporção. Também se constatou que os métodos de mineração de dados mais utilizados no contexto citado foram: Classificador Bayesiano, Classificador Vizinho mais Próximo, Suporte Vector Machine, Árvore de Decisão e Agrupamento Particional. Os algoritmos mais citados nas pesquisas são os seguintes: Naive Bayes, K- Nearest Neighbor, SVM, J48 e K-Means. Quanto aos tipos de dados a serem minerados, a maioria tratou Web e Relacionais, originando-se das fontes como: Twitter, Comentários, Mensagens de Spam, Fórum de discussão, Resenhas de filmes, Comentários em Mídias Sociais, Dados de vendas e suporte, Mídias Sociais, Mensagens de texto, SMS, Facebook, Logs de auditoria, dentre outros. Quanto a sua estrutura dos dados a serem minerados, antes do tratamento, a maioria é do tipo não estruturado. Após o tratamento, a maioria classificou-se como Dados Estruturados e Semi Estruturados.

Nosso trabalho identificou a necessidade de melhorias e ajustes nos Métodos de Mineração de Dados a serem aplicados em bases Big Data. Métodos e modelos tradicionais não conseguem coletar, integrar, gerenciar e processar dados distribuídos provindos de fontes diversas de natureza não estruturada, ou seja, todos os objetivos diferentes de estudo precisam de métodos e algoritmos diferentes para alcançar o objetivo desejado. A solução de cada problema não pode ser limitada a apenas a aplicação de um método e um algoritmo. Neste sentido, ao se trabalhar com dados não estruturados ou semiestruturados, os mesmos não vêm em um formato de dados pré-definidos e não podem ser armazenados em tabelas relacionais. Dados não podem ser interpretados, adquiridos, gerenciados e processados pelos modelos tradicionais de hardware e software existentes. Torna-se necessário técnicas eficientes e escaláveis, iniciativas de tratamento, integração e análise de dados provenientes de diversas fontes em diferentes mídias e formatos relacionados ao contexto do Big Data Mining. Nosso trabalho também revelou pesquisas e possíveis oportunidades no estado da arte referente a Mineração de Dados aplicada a bases Big Data. Ainda proporcionou subsídios para o desdobramento de novas pesquisas nas temáticas de Data Science, Data Mining e Big Data.

Esperamos que este estudo de mapeamento não sirva apenas para destacar os principais tópicos de pesquisa na área do Big Data Mining, mas também sirva para atrair pesquisadores e profissionais para um corpo de conhecimento que identifique quais são as diferentes maneiras de extrair informações relevantes a uma tomada de decisão no contexto do Big Data.

## REFERÊNCIAS

BONIDIA, R. P.; BRANCHER, J. D.; BUSTO, R. M. **Data mining in sports: A systematic review.** IEEE Latin America Transactions, IEEE, v. 16, n. 1, p. 232–239, 2018.

CHEN, C. P., & ZHANG, C.-Y. **Data-intensive applications, challenges, techniques and technologies:** A survey on big data. Information Sciences, 275 , 314–347, 2014.

CHEN, M.; MAO, S.; LIU, Y. Big data: **A survey. Mobile Networks and Applications, Springe**r US, v. 19, n. 2, p. 171–209, 2014.

Impresso 1982-3266
Online 2318-9150

DUTT, A.; ISMAIL, M. A.; HERAWAN, T. **A systematic review on educational data mining.** IEEE Access, IEEE, v. 5, p. 15991–16005, 2017.

GARCÉS, R. *et al.* **A systematic mapping on quality attributes and quality models for ambient assisted living systems.** Instituto de Ciências Matemáticas e Computacão, Universidade de São Paulo (ICMC/USP), Brazil, 2016

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining concepts and techniques third edition.** The Morgan Kaufmann Series in Data Management Systems, v. 5, n. 4, p. 83-124, 2011.

.

JALALI. S; WOHLIN. C. S**ystematic literature studies: database searches vs. backward snowballing, in**: Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2012, pp. 29–38.

KAUR, G.; KAUR, E. H. **Prediction of the cause of accident and accident prone location on roads using data mining techniques**. In: IEEE. Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on. [S.l.], 2017. p. 1–7.

KEELE, S. *et al*. G**uidelines for performing systematic literature reviews in software engineering**. [S.l.], 2007.

KHAN, N., YAQOOB, I., HASHEM, I. A. T., INAYAT, Z., MAHMOUD Ali, W. K., ALAM, M., SHIRAZ, M., & GANI, **A. Big data: survey, technologies, opportunities, and challenges**. The Scientific World Journal , 2014 .

KHANNA, L.; SINGH, S. N.; ALAM, M. **Educational data mining and its role in determining factors affecting students academic performance: A systematic review.** In: IEEE. 2016 1st India International Conference on Information Processing (IICIP). [S.l.], 2016. p. 1–7.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for Performing Systematic Literature Reviews in Software Engineering,** Tech. rep., Technical report, EBSE Technical Report EBSE-2007-01, 2007

NAJAFABADI, M. M., VILLANUSTRE, F., KHOSHGOFTAAr, T. M., SELIYA, N., WALD, R., & MUHAREMAGIC, E. **Deep learning applications and challenges in big data analytics.** Journal of Big Data, 2, 1, 2015.

OUSSOUS, A. *et al.* B**ig data technologies:** A survey. Journal of King Saud UniversityComputer and Information Sciences, Elsevier, v. 30, n. 4, p. 431–448, 2018.

PETERSEN, Kai; VAKKALANKA, Sairam; KUZNIARZ, **Ludwik. Guidelines for conducting systematic mapping studies in software engineering:** An update. Information and Software Technology, v. 64, p. 1-18, 2015.

QIU, J. *et al.* **A survey of machine learning for big data processing.** EURASIP Journal on Advances in Signal Processing, Springer, v. 2016, n. 1, p. 67, 2016.

Impresso 1982-3266
Online 2318-9150

SANGEETHA, J.; PRAKASH, V. **Sinthu Janita. A survey on big data mining techniques.** International Journal of Computer Science and Information Security, v. 15, n. 1, p. 482, 2017.

WOHLIN. C. G**uidelines for snowballing in systematic literature studies and a replication in software engineering,** in: 18th International Conference on Evaluation and Assessment in Software Engineering, EASE '14, 2014, p. 38.

YAN, Y. *et al*. **A classifier ensemble framework for multimedia big data classification.** In: IEEE. 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI). [S.l.], 2016. p. 615–622.

## APÊNDICE A – ESTUDOS PRIMÁRIOS

[S1] ABDELHAMEED, H. J.; MUÑOZ-HERN'ANDEZ, S. **Emotion and opinion retrieval from social media in arabic language: Survey**. In: IEEE. Information and Communication Technologies for Education and Training and International Conference on Computing in Arabic (ICCA-TICET), 2017 Joint International Conference on. [S.l.], 2017. p. 1–8.

[S2] AKTHAR, N.; AHAMAD, M. V.; AHMAD, S. **Mapreduce model of improved k-means clustering algorithm using hadoop mapreduce**. In: IEEE. Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on. [S.l.], 2016. p. 192–198.

[S3] ALAMDARI, M. S. *et al*. **Disease detection in medical prescriptions using data mining tools.** In: IEEE. Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on. [S.l.], 2014. p. 159–164.

[S4] AL-AMRANI, Y.; LAZAAR, M.; ELKADIRI, K. E. S**entiment analysis using supervised classification algorithms.** In: ACM. Proceedings of the 2nd international Conference on Big Data, Cloud and Applications. [S.l.], 2017. p. 61.

[S5] ARIF, M. H.; LI, J.; IQBAL, M. **Solving social media text classification problems using code fragment-based xcsr.** In: IEEE. Tools with Artificial Intelligence (ICTAI), 2017 IEEE 29th International Conference on. [S.l.], 2017. p. 485–492.

[S6] ARORA, S.; CHANA, I. **A survey of clustering techniques for big data analysis**. In: IEEE. Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-. [S.l.], 2014. p. 59–65.

[S7] ATHANI, S. S. *et al*. **Student academic performance and social behavior predictor using data mining techniques**. In: IEEE. Computing, Communication and Automation (ICCCA), 2017 International Conference on. [S.l.], 2017. p. 170–174.

[S8] BACH, N. X.; PHUONG, T. M. **Leveraging user ratings for resource-poor sentiment classification**. Procedia Computer Science, Elsevier, v. 60, p. 322–331, 2015.

[S9] BASHIR, S.; QAMAR, U.; JAVED, M. Y. **An ensemble based decision support**

Impresso 1982-3266
Online 2318-9150

**framework for intelligent heart disease diagnosis**. In: IEEE. Information Society (i-Society), 2014 International Conference on. [S.l.], 2014. p. 259–264.
[S10] BELLO-ORGAZ, G.; CAMACHO, D. **Comparative study of text clustering techniques in virtual worlds**. In: ACM. Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics. [S.l.], 2013. p. 9.

[S11] BERKANI, L.; CHEBAHI, Y.; BETIT, L. **Using data mining techniques and genetic algorithm**. In: ACM. Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications. [S.l.], 2018. p. 25

[S12] BISIO, F. *et al*. **Data intensive review mining for sentiment classification across heterogeneous domains.** In: IEEE. Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. [S.l.], 2013. p. 1061–1067.

[S13] BLATNIK, A.; JARM, K.; MEZA, M. **Movie sentiment analysis based on public tweets.** ElektrotehniskiVestnik, v. 81, n. 4, p. 160–166, 2014.

[S14] CHATURVEDI, M. **Data mining and it's application in edm domain.** In: IEEE. Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on. [S.l.], 2017. p. 829–834.

[S15] CHAUHAN, C.; SEHGAL, S. **A review: Crime analysis using data mining techniques and algorithms.** In: IEEE. Computing, Communication and Automation (ICCCA), 2017 International Conference on. [S.l.], 2017. p. 21–25.

[S16] CHENG, M.; XU, K.; GONG, X. **Research on audit log association rule mining based on improved apriori algorithm.** In: IEEE. Big Data Analysis (ICBDA), 2016 IEEE International Conference on. [S.l.], 2016. p. 1–7.

[S17] CHEUNG, A.; SOLAR-LEZAMA, A.; MADDEN, S. **Using program synthesis for social recommendations.** In: ACM. Proceedings of the 21st ACM international conference on Information and knowledge management. [S.l.], 2012. p. 1732–1736.

[S18] CHONG, W. Y.; SELVARETNAM, B.; SOON, L.-K. **Natural language processing for sentiment analysis: an exploratory analysis on tweets**. In: IEEE. Artificial Intelligence with Applications in Engineering and Technology (ICAIET), 2014 4th International Conference on. [S.l.], 2014. p. 212–217.

[S19] DAS, A.; BANDYOPADHYAY, S. **Opinion summarization in bengali: A theme network model.** In: SocialCom/PASSAT. [S.l.: s.n.], 2010. p. 675–682.

[S20] DHANALAKSHMI, V.; BINO, D.; SARAVANAN, A. **Opinion mining from student feedback data using supervised learning algorithms.** In: IEEE. Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference on. [S.l.], 2016. p. 1–5.

[S21] DUAN, G. *et al*. **An improved medical decision model based on decision tree algorithms.** In: IEEE. Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (Sustain Com) (BDCloudSocialCom-SustainCom), 2016 IEEE International Conferences on. [S.l.], 2016. p. 151–156.

Impresso 1982-3266
Online 2318-9150

[S22] FIARNI, C.; MAHARANI, H.; PRATAMA, R. **Sentiment analysis system for indonesia online retail shop review using hierarchy naive bayes technique**. In: IEEE. Information and Communication Technology (ICoICT), 2016 4th International Conference on. [S.l.], 2016. p. 1–6.

[S23] FOROUSHANI, Z. A.; LI, Y. **Intrusion detection system by using hybrid algorithm of data mining technique.** In: ACM. Proceedings of the 2018 7th International Conference on Software and Computer Applications. [S.l.], 2018. p. 119–123

[S24] GARG, S.; SAINI, A.; KHANNA, N. **Is sentiment analysis an art or a science? impact of lexical richness in training corpus on machine learning.** In: IEEE. Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. [S.l.], 2016. p. 2729–2735.

[S25] GIRIJA, D.; SHASHIDHARA, M. **Data mining techniques used for uterus fibroid diagnosis and prognosis.** In: IEEE. Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on. [S.l.], 2013. p. 372–376.

[S26] GU, B. *et al*. **New incremental learning algorithm for semi-supervised support vector machine.** In: ACM. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. [S.l.], 2018. p. 1475–1484.

[S27] HAMZAH, A.; WIDYASTUTI, N. **Opinion extracting and classification from questionnaire comments using hmm-pos tagger and machine learning techniques.** In: IEEE. Data and Software Engineering (ICoDSE), 2016 International Conference on. [S.l.], 2016. p. 1–6.

[S28] HUANG, J.; CHEN, M. **Domain adaptation approach for credit risk analysis.** In: ACM. Proceedings of the 2018 International Conference on Software Engineering and Information Management. [S.l.], 2018. p. 104–107

[S29] ISMAIL, I.; MARSONO, M. N.; NOR, S. M. **Detecting worms using data mining techniques: learning in the presence of class noise**. In: IEEE. Signal-Image Technology and Internet-Based Systems (SITIS), 2010 Sixth International Conference on. [S.l.], 2010. p. 187–194.

[S30] JAIN, V. K.; KUMAR, S. **An effective approach to track levels of influenza-a (h1n1) pandemic in india using twitter.** Procedia Computer Science, Elsevier, v. 70, p. 801–807, 2015.

[S31] KAUR, G.; KAUR, E. H. **Prediction of the cause of accident and accident prone location on roads using data mining techniques.** In: IEEE. Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on. [S.l.], 2017.

[S32] KHENNAK, I.; DRIAS, H. **Data mining techniques and nature-inspired algorithms for query expansion**. In: ACM. Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications. [S.l.], 2018. p. 28

Impresso 1982-3266
Online 2318-9150

[S33] KOUKOUVINOS, C.; PARPOULA, C.; SIMOS, D. E. **Genetic algorithm and data mining techniques for design selection in databases**. In: IEEE. Availability, Reliability and Security (ARES), 2013 Eighth International Conference on. [S.l.], 2013. p. 743–746.

[S34] KROUSKA, A.; TROUSSAS, C.; VIRVOU, M. **The effect of preprocessing techniques on twitter sentiment analysis.** In: IEEE. Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on. [S.l.], 2016. p. 1–5.

[S35] KULEV, I.; PU, P.; FALTINGS, B. **A bayesian approach to intervention-based clustering.** ACM Transactions on Intelligent Systems and Technology (TIST), ACM, v. 9, n. 4, p. 44, 2018.

[S36] KURNAZ, S.; MAHMOOD, R. M. **Methodology preview on predicting students professional identity using data mining techniques**. In: ACM. Proceedings of the Fourth International Conference on Engineering & MIS 2018. [S.l.], 2018. p. 56.

[S37] LUKYAMUZI, A.; NGUBIRI, J.; OKORI, W. **Tracking food insecurity from tweets using data mining techniques.** In: ACM. Proceedings of the 2018 International Conference on Software Engineering in Africa. [S.l.], 2018. p. 27–34.

[S38] LUNA, J. M. *et al.* **Apriori versions based on mapreduce for mining frequent patterns on big data**. IEEE transactions on cybernetics, IEEE, v. 48, n. 10, p. 2851–2865, 2017.

[S39] MAILLO, J. *et al.* **A preliminary study on hybrid spill-tree fuzzy k-nearest neighbors for big data classification.** In: IEEE. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). [S.l.], 2018. p. 1–8

[S40] MAILLO, J. *et al*. **Exact fuzzy k-nearest neighbor classification for big datasets.** In: IEEE. 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). [S.l.], 2017. p. 1–6.

[S41] MAILLO, J.; TRIGUERO, I.; HERRERA, F. **A mapreduce-based k-nearest neighbor approach for big data classification.** In: IEEE. Trustcom/BigDataSE/ISPA, 2015 IEEE. [S.l.], 2015. v. 2, p. 167–172.

[S42] MISTRY, P. *et al*. **Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology.** Soft Computing, Springer, v. 20, n. 8, p. 2967–2979, 2016.

[S43] MOHD, N.; YAHYA, Y. A **data mining approach for prediction of students' depression using logistic regression and artificial neural network.** In: ACM. Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication. [S.l.], 2018. p. 52.

[S44] MORREALE, P.; HOLTZ, S.; GONCALVES, **A. Data mining and analysis of large scale time series network data.** In: IEEE. Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on. [S.l.], 2013. p.

Impresso 1982-3266
Online 2318-9150

39–43.

[S45] MUKAHAR, N.; ROSDI, B. **A. Performance comparison of prototype selection based on edition search for nearest neighbor classification**. In: ACM. Proceedings of the 2018 7th International Conference on Software and Computer Applications. [S.l.], 2018. p. 143–146

[S46] NADEEM, H. *et al.* **Knn-fuzzy classification for cloud service selection.** In: ACM. Proceedings of the 2nd International Conference on Future Networks and Distributed Systems. [S.l.], 2018. p. 66

[S47] NARAYANAN, S. *et al.* **Integration and automation of data preparation and data mining.** In: IEEE. Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. [S.l.], 2014. p. 1076–1085.

[S48] PARATE, M.; TAJANE, S.; INDI, B. **Assessment of system vulnerability for smart grid applications.** In: IEEE. Engineering and Technology (ICETECH), 2016 IEEE International Conference on. [S.l.], 2016. p. 1083–1088.

[S49] POOJITHA, V. *et al*. **A collocation of iris flower using neural network clustering tool in matlab.** In: IEEE. Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference. [S.l.], 2016. p. 53–58.

[S50] POUYANFAR, S. *et al.* **Multimedia big data analytics:** A survey. ACM Computing Surveys (CSUR), ACM, v. 51, n. 1, p. 10, 2018

[S51] PRIYANKA, K.; TEWARI, A. S.; BARMAN, A. G. **Personalised book recommendation system based on opinion mining technique.** In: IEEE. Communication Technologies (GCCT), 2015 Global Conference on. [S.l.], 2015. p. 285–289.

[S52] PURWAR, A.; SINGH, S. K. **Issues in data mining: A comprehensive survey.** In: IEEE. 2014 IEEE International Conference on Computational Intelligence and Computing Research. [S.l.], 2014. p. 1–6

[S53] RĂDULESCU, C.; DINSOREANU, M.; POTOLEA, R. **Identification of spam comments using natural language processing techniques.** In: IEEE. Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on. [S.l.], 2014. p. 29–35.

[S54] RADY, S. **A business intelligent technique for sentiment estimation by management sectors.** In: IEEE. Intelligent Computing and Information Systems (ICICIS), 2015 IEEE Seventh International Conference on. [S.l.], 2015. p. 370–376.

[S55] RAJALAKSHMI, S.; ASHA, S.; PAZHANIRAJA, N. **A comprehensive survey on sentiment analysis.** In: IEEE. Signal Processing, Communication and Networking (ICSCN), 2017 Fourth International Conference on. [S.l.], 2017. p. 1–5.

[S56] RAMAPHOSA, K. I. M.; ZUVA, T.; KWUIMI, R. **Educational data mining to improve learner performance in gauteng primary schools.** In: IEEE. 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). [S.l.], 2018. p. 1–6.

Impresso 1982-3266
Online 2318-9150

[S57] RATHORE, P.; SHUKLA, D. **Analysis and performance improvement of k-means clustering in big data environment.** In: IEEE. Communication Networks (ICCN), 2015 International Conference on. [S.l.], 2015. p. 43–46.

[S58] ROBU, R.; HORA, C. **Medical data mining with extended weka.** In: IEEE. Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on. [S.l.], 2012. p. 347–350.

[S59] SHESHASAYEE, A.; THOMAS, S. S. **Implementation of data mining techniques in upcoding fraud detection in the monetary domains**. In: IEEE. Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on. [S.l.], 2017. p. 730–734.

[S60] SHOUMAN, M.; TURNER, T.; STOCKER, R. **Using data mining techniques in heart disease diagnosis and treatment.** In: IEEE. Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on. [S.l.], 2012. p. 173–177.

[S61] SIDDIQUA, U. A.; AHSAN, T.; CHY, A. N. C**ombining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog.** In: IEEE. Computer and Information Technology (ICCIT), 2016 19th International Conference on. [S.l.], 2016. p. 304–309.

[S62] SINDHU, C.; VYAS, D. V.; PRADYOTH, K. **Sentiment analysis based product rating using textual reviews.** In: IEEE. Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of. [S.l.], 2017. v. 2, p. 727–731.

[S63] SONI, V. K.; PAWAR, S. **Emotion based social media text classification using optimized improved id3 classifier**. In: IEEE. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). [S.l.], 2017. p. 1500–1505.

[S64] SU, D. *et al.* **Differentially private k-means clustering.** In: ACM. Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy. [S.l.], 2016. p. 26–37.

[S65] THAPA, L. B. R.; BAL, B. K. **Classifying sentiments in nepali subjective texts**. In: IEEE. Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on. [S.l.], 2016. p. 1–6.

[S66] TINÓS, R. *et al*. **Nk hybrid genetic algorithm for clustering**. IEEE Transactions on Evolutionary Computation, IEEE, v. 22, n. 5, p. 748–761, 2018.

[S67] TU, Y.; YANG, Z.; BENSLIMANE, Y. **Towards an optimal classification model against imbalanced data for customer relationship management**. In: IEEE. Natural Computation (ICNC), 2011 Seventh International Conference on. [S.l.], 2011. v. 4, p. 2401–2405.

[S68] TUBA, E. *et al.* **Web intelligence data clustering by bare bone fireworks algorithm combined with k-means.** In: ACM. Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. [S.l.], 2018. p. 7.

Impresso 1982-3266
Online 2318-9150

[S69] TYAGI, E.; SHARMA, A. K. **An intelligent framework for sentiment analysis of text and emotions-a review**. In: IEEE. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). [S.l.], 2017. p. 3297–3302.

[S70] UWANO, F. *et al*. **Generalizing rules by random forest-based learning classifier systems for high-dimensional data mining**. In: ACM. Proceedings of the Genetic and Evolutionary Computation Conference Companion. [S.l.], 2018. p. 1465–1472

[S71] VALSAMIDIS, S. *et al*. **A framework for opinion mining in blogs for agriculture.** Procedia Technology, Elsevier, v. 8, p. 264–274, 2013.

[S72] VERMA, D.; MISHRA, N. **Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques**. In: IEEE. 2017 International Conference on Intelligent Sustainable Systems (ICISS). [S.l.], 2017. p. 533–538.

[S73] VIDYA, N. A.; FANANY, M. I.; BUDI, I. T**witter sentiment to analyze net brand reputation of mobile phone providers**. Procedia Computer Science, Elsevier, v. 72, p. 519–526, 2015.

[S74] WANG, Y. *et al*. **Deep learning-based socio-demographic information identification from smart meter data.** IEEE Transactions on Smart Grid, IEEE, v. 10, n. 3, p. 2593–2602, 2018.

[S75] YANG, H. *et al*. **Efficient and secure knn classification over encrypted data using vector homomorphic encryption**. In: IEEE. 2018 IEEE International Conference on Communications (ICC). [S.l.], 2018. p. 1–7

[S76] YANG, Y.; WANG, H. **Multi-view clustering: a survey.** Big Data Mining and Analytics, TUP, v. 1, n. 2, p. 83–107, 2018.

[S77] ZAMANI, Z.; POURMAND, M.; SARAEE, M. H. **Application of data mining in traffic management: case of city of isfahan.** In: IEEE. Electronic Computer Technology (ICECT), 2010 International Conference on. [S.l.], 2010. p. 102–106.

[S78] ZHANG, Q. *et al.* **Pphopcm: Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing**. IEEE Transactions on Big Data, IEEE, 2017

Recebido em: 10/06/2021
Aceito em: 19/07/2021
Publicado em: 08/2021

Impresso 1982-3266
Online 2318-9150