UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DAINF - DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO

GABRIEL LEVIS ZAWALSKI

APLICAÇÃO DO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM UMA BASE DE DADOS DE REDE SOCIAL

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA
2018

GABRIEL LEVIS ZAWALSKI

APLICAÇÃO DO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM UMA BASE DE DADOS DE REDE SOCIAL

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Ciências da Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel.

Orientadora: Helyane Bronoski Borges
UTFPR - Ponta Grossa

Coorientadora: Thissiany Beatriz Almeida
UTFPR - Ponta Grossa

PONTA GROSSA
2018

Altere este texto inserindo a dedicatória do seu trabalho.

AGRADECIMENTOS

Edite e coloque aqui os agradecimentos às pessoas e/ou instituições que contribuíram para a realização do trabalho.

É obrigatório o agradecimento às instituições de fomento à pesquisa que financiaram total ou parcialmente o trabalho, inclusive no que diz respeito à concessão de bolsas.

Eu denomino meu campo de Gestão do Conhecimento, mas você não pode gerenciar conhecimento. Ninguém pode. O que pode fazer - o que a empresa pode fazer - é gerenciar o ambiente que otimize o conhecimento. (PRUSAK, Laurence, 1997).

RESUMO

ZAWALSKI, Gabriel. Aplicação do processo de descoberta de conhecimento em uma base de dados de rede social. 2018. 30 f. Trabalho de Conclusão de Curso – Curso de Bacharelado em Ciências da Computação, Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2018.

O Resumo é um elemento obrigatório em tese, dissertação, monografia e TCC, constituído de uma seqüência de frases concisas e objetivas, fornecendo uma visão rápida e clara do conteúdo do estudo. O texto deverá conter no máximo 500 palavras e ser antecedido pela referência do estudo. Também, não deve conter citações. O resumo deve ser redigido em parágrafo único, espaçamento simples e seguido das palavras representativas do conteúdo do estudo, isto é, palavras-chave, em número de três a cinco, separadas entre si por ponto e finalizadas também por ponto. Usar o verbo na terceira pessoa do singular, com linguagem impessoal, bem como fazer uso, preferencialmente, da voz ativa. Texto contendo um único parágrafo.

Palavras-chave: Palavra. Segunda Palavra. Outra palavra.

ABSTRACT

ZAWALSKI, Gabriel. Appliance of knowledge discovery in databases process in a social media database. 2018. 30 f. Trabalho de Conclusão de Curso – Curso de Bacharelado em Ciências da Computação, Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2018.

Elemento obrigatório em tese, dissertação, monografia e TCC. É a versão do resumo em português para o idioma de divulgação internacional. Deve ser antecedido pela referência do estudo. Deve aparecer em folha distinta do resumo em língua portuguesa e seguido das palavras representativas do conteúdo do estudo, isto é, das palavras-chave. Sugere-se a elaboração do resumo (Abstract) e das palavras-chave (Keywords) em inglês; para resumos em outras línguas, que não o inglês, consultar o departamento / curso de origem.

Keywords: Word. Second Word. Another word.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KDP	*Knowledge Discovery Process*
KDD	*Knowledge Discovery in Databases*

LISTA DE SÍMBOLOS

Γ	Letra grega Gama
λ	Comprimento de onda
$\in$	Pertence

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

Desde o seu surgimento, a Internet têm movimentado o mundo de uma forma que seus idealizadores na década de 60 jamais poderiam imaginar. Segundo dados estimados pelo *site* Group (), 54,4% dos 7,6 bilhões de pessoas no planeta estão conectados a rede mundial de computadores, somando assim um total de mais de 4,1 bilhões de usuários *online*. Desse total, 2,62 bilhões fazem uso de algum tipo de rede social e a previsão é que esse número aumente para 3,02 bilhões até o fim de 2021, segundo pesquisas da Statista Dossier ().

Dentro desse universo de redes sociais, o *Facebook* possui a maior base de usuários ativos. Em abril de 2017, segundo o relatório do segundo quadrimestre *GLOBAL DIGITAL STATSHOT* Social e Hootsuite (), publicado através da parceita entre a *We Are Social* e o *Hootsuite*, o *Facebook* atingiu a marca de 1,9 bilhão de usuários ativos mensais e no primeiro quadrimestre de 2018 esse número já passava de 2,2 bilhões segunda pesquisas da Statista Dossier ()

Somente esses números são suficientes para colocar o *Facebook* como principal rede social utilizada, em muitro superando o segundo colocado, o *YouTube* com 1,5 bilhão de usuários ativos mensais.

Além do crescimento do uso da *Internet* como meio de consumo, os avanços das tecnologias móveis e redes sociais estão possibilitando a geração de quantidades de dados cada vez maiores. Um estudo da IBM *Marketing Cloud* Cloud () descreve que são criados, aproximadamente, 2,5 quintilhões de *bytes* de dados todos os dias e que 90% de todo o montante de dados presente no mundo hoje foi criado a partir de 2016.

1.1 DESCRIÇÃO DO PROBLEMA E MOTIVAÇÃO

Tecnologias atuais possibilitam o armazenamento e acesso a essa grande quantidade de dados há um custo muito baixo. Porém, o principal problema associado a um mundo centrado em informações é utilizar os dados brutos coletados Kurgand e Musilek (2006).

Neste cenário que nos encontramos, cada vez mais se mostra necessário o uso de ferramentas computacionais para auxiliar na extração de conhecimentos desses volumes de dados, uma vez que somente acumular dados não necessariamente se traduz em informações úteis e aplicáveis Fayyad, Piatetsky-Shapiro e Smyth (1996a).

1.2 OBJETIVOS

Esta Seção apresenta o obejtivo geral e os objetivos específicos deste trabalho. Na Subseção 1.2.1 encontra-se o objetivo geral e na Subseção 1.2.2 encontram-se os objetivos específicos.

1.2.1 Objetivos gerais

O objetivo geral deste trabalho é aplicar o KDD numa base de dados retirada do *Facebook* com o intuito de extrair informações a respeito da relação entre os metadados das postagens e as métricas de avaliação geradas pelos algoritmos da rede social em questão.

1.2.2 Objetivos específicos

Como objetivos específicos deste trabalho têm-se:

- compreender o funcionamento do processo de KDD;
- analisar as etapas do KDD, identificado técnicas que podem ser aplicadas;
- compreender o funcionamento das métricas geradas pelo *Facebook*
- aplicar o processo de KDD na base de dados de postagens;
- realizar experimentos;
- analisar os resultados obtidos por meio de comparação estatística com outros trabalhos da área.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho encontra-se dividido nos seguintes sete capítulos:

- Capítulo 1: Capítulo introdutório de contextualização do trabalho, apresentando em linhas gerais a situação atual, a motivação que levou a idealização deste trabalho e os objetivos a serem alcançados.
- Capítulo 2: Este capítulo aborda os conceitos necessários para compreender o processo de KDD. Contém uma descrição das etapas que o constituem como um todo e entra mais a fundo em conceitos importantes da parte de mineração de dados.
- Capítulo 3: O capítulo apresenta conceitos e definições a respeito de redes sociais, classificando-as e descrevendo sua relevância fora e dentro da área acadêmica. Também contém uma explicação a respeito de termos da rede social Facebook, sendo esta o local onde os dados utilizados neste trabalho foram retirados.
- Capítulo 4: Neste capítulo encontra-se uma explicação a respeito do método de revisão metodológica, ressaltando os pontos importantes do processo e a aplicação deste em bases de artigos acadêmicos, permitindo a identificação de trabalhos correlatos que serviram de referência para a elaboração deste.
- Capítulo 5: O capítulo descreve a base de dados utilizada para a realização dos experimentos deste trabalho, bem como as técnicas e métodos a serem aplicados na mesma, seguindo o processo definido no capítulo 2.
- Capítulo 6: Este capítulo contém a descrição e análise dos resultados obtidos após a aplicação dos métodos e técnicas descritos no capítulo 5 sobre a base de dados deste trabalho.

- Capítulo 7: O capítulo contém possibilidades de continuações deste trabalho, bem como as considerações finais a respeito dos resultados obtidos e do cumprimento dos objetivos propostos.

2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Este capítulo aborda o processo de Descoberta de Conhecimento em Bases de dados, abreviado pela sua silga em inglês KDD. A Seção 2.1 apresenta o histórico e informações básicas a respeito do processo como um todo, oferecendo uma visão geral de suas partes principais e de sua importância. A Seção 2.2 contém uma explicação mais detalhada das tarefas principais do processo, cada uma explicada em detalhes nas Seções secundárias 2.2.1, 2.2.2 e 2.2.3, além de suas respectivas subetapas internas. Dentro da Seção secundária 2.2.2, uma Seção terciária 2.2.2.1 entra em mais detalhes a respeito das tarefas associadas em especfício à etapa mineração de dados.

2.1 CONCEITOS FUNDAMENTAIS DO KDD

Antes de qualquer tentativa de realizar esta extração de informações, se faz necessário estabelecer um método a ser seguido. O principal objetivo de estabeler uma padrão é de ajudar a compreender o processo de descoberta de conhecimento, oferecendo um roteiro a ser seguido no decorrer do projeto e, em consequência, reduzir custos de tempo e recursos Kurgand e Musilek (2006).

Desde 1990, diversas abrodagens para a criação de modelos a esse processo foram desenvolvidas, primeiramente por academicos e posteriormente pela indústria, porém, a fim de formalizar esses modelos concorrentes de KDP, se faz necessários colocar todos dentro de um *framework* comum Cios et al. (2007).

De forma geral, os diversos modelos propostos consistem de um conjunto de passos de processamento executados de forma sequencial e dependente dos resultados gerados pelo passo anterior como entrada. Esses passos englobam uma grande variedade de tarefas, abrangendo desde a análise e preparação dos dados brutos até a compreensão e aplicação dos resultados gerados ao final do processo Cios et al. (2007).

É importante ressaltar a natureza iterativa dos modelos de KDP, podendo conter diversos laços de *feedback* e repetição entre quaisquer dois passos do processo Kurgand e Musilek (2006). Por fim, o KDP é considerado não-trivial por conter um certo grau de inferência durante algumas de suas etapas, significando que algumas delas podem não ser diretas como de valores pré-definidos Fayyad, Piatetsky-Shapiro e Smyth (1996b).

O objetivo final do KDP é identificar padrões dos dados em que o processo foi aplicado. Por padrões entende-se por ajustar um modelo aplicável aos dados brutos, encontrar estruturas que os dados seguem ou até mesmo uma descrição abstrata de alto nível dos conjuntos de dados Fayyad, Piatetsky-Shapiro e Smyth (1996a).

Estes padrões produzidos ao final da aplicação do KDP precisam atingir certos critérios mínimos. Precisam ser, até certo grau de certeza, válidos para o conjunto de dados de onde

o padrão foi inferido, além de novos, tanto no escopo do sistema quanto, preferencialmente, para o usuário. Os padrões precisam ser potencialmente úteis, oferecendo algum benefício para a tarefa a qual foi necessário extrair conhecimento. E por fim, precisam ser compreensíveis em alguma linguagem, tanto imediatamente quanto após alguma etapa de pós-processamento Fayyad, Piatetsky-Shapiro e Smyth (1996a).

Seguir um processo estabelecido como o KDP, além de redução de custos como já citado no começo do Capítulo, garante que os dados possam ser verificados, reutilizados e replicados de maneira consistente.

2.2 ETAPAS DO KDD

Os fundamentos básicos para o KDP foram propostos por Fayyad no lançamento do livro *Advances in Knowledge Discovery in Databases 1996*. A pesquisa do livro apresentava dois tipos de modelos: o *human-centric*, que se concentra no papel ativo do analista durante o processo; e o *data-centric*, que foca na natureza iterativa e interativa da tarefa de análise de dados Kurgand e Musilek (2006).

O modelo *human-centric* consiste em uma série de tarefas com interações complexas ao decorrer do tempo do processo entre um humano e uma base de dados, possivelmente auxiliados por uma variedade de ferramentas. Sua estrutura é dividida em três tarefas principais: seleção de modelo e execução (pré-processamento); análise de dados (mineração de dados); e geração da saída (pós-processamento). Cada uma dessas etapas foi dividida em outras subetapas totalizando nove passos ao todo Kurgand e Musilek (2006).

Um grande número trabalhos acadêmicos se focam exclusivamente no passo de mineração de dados do KDP, porém todos os passos do processo são importantes.Os passos adicionais de descoberta de conhecimento, como preparação e limpeza dos dados e interpretação apropiada dos resultados são essenciais para garantir que o conhecimento derivado da aplicação está correto. Aplicação de métodos de mineração de dados de maneira não criteriosa é uma atividade que facilmente pode levar a descoberta de padrões sem sentido e muitas vezes não válidos Fayyad, Piatetsky-Shapiro e Smyth (1996a).

2.2.1 Pré-processamento

O principal objetivo desta tarefa é a garantir que os dados estejam prontos para o processo de mineração. É um dos principais componentes do KDP, sendo todo o sucesso das etapas subsequentes dependetes da identificação e iteração correta desta tarefa Bagga e Singh (2011) Suas subetapas podem ser compreendidas no Quadro 1.

2.2.2 Mineração de dados

É a única etapa do processo de KDD que se preocupa na aplicação de técnicas computacionais. Têm o papel de encontrar padrões no conjunto de dados previamente preparado

Quadro 1 – Etapas do pré-processamento do KDD.

Etapas	**Descrição**
1. Desenvolver e compreender o domínio da aplicação	Este passo diz respeito ao aprendizado de quaisquer conhecimentos anteriores ao domínio da aplicação além dos objetivos do usuário para o novo conhecimento descoberto. É um passo preparatório para o trabalho com a base de dados em sim.
2. Criar o conjunto de dados alvo	Envolve a seleção dos atributos e instâncias em que serão aplicadas as tarefas de descoberta de conhecimento, geralmente percorrendo a base a fim de selecionar os dados do subconjunto.
3. Limpeza dos dados	Este passo consiste em na remoção de *outliers*, limpeza de ruído e inputação de valores faltantes.
4 Redução de dados e projeção	Consiste na aplicação de métodos de transformação a fim de reduzir as dimensões dos dados, continuam o processo somente com os atributos relevantes ou encontrar representação não variantes dos dados a serem tratados

Fonte: (CIOS et al., 2007)

e transformado para permitir com que o processo ocorra sem grande problemas. Caso a tarefa de pré-processamento não tenha sido realizada com sucesso, em consequência tanto esta tarefa como o processo todo também não terão êxito Bagga e Singh (2011). Podemos ver a descrição de suas subetapas no Quadro 2.

Quadro 2 – Etapas da mineração do KDD.

Etapas	**Descrição**
5. Escolha da tarefa de mineração	Escolha da tarefa de mineração em sincronia com os objetivos levantados no passo 1 do processo: e.g. classificação, clusterização, regressão, etc.
6. Escolha do algoritmo de mineração	Inclui tanto a seleção de métodos de busca por padrões quanto quais modelos e parâmetros são mais apropriados para os critérios do conhecimento a ser extraído.
7. Mineração de dados	Geração dos padrões de interesse em determinada forma de representação.

Fonte: (CIOS et al., 2007)

2.2.2.1 Tarefas da mineração de dados

Como dito anteriormente, por ser a única etapa dentre todas as etapas do processo de KDD por se preocupar com aplicação prática de técnicas computacionais, a mineração de dados é uma das etapas que possui maior ênfase nas áreas acadêmicas.

O processo de mineração de dados envolve a descoberta de padrões a partir de dados e a adaptação de modelos para melhor acomodar os dados existentes.

2.2.3 Pós-processamento

Ultima parte do processo de KDD, envolve a visualização e interpretação do conhecimento extraído dos padrões encontrados. Esta etapa tem grande importância no sentido de permitir que o conhecimento gerado seja compreensível para o usuário final. No Quadro 3 encontram-se as subetapas do pós-processamento.

Quadro 3 – Etapas do pós-processamento do KDD.

Etapas	**Descrição**
8. Interpretação dos padrões minerados	O analista realiza a visualização dos padrões e modelos extraídos. É possível um retorno a qualquer um dos passos até agora realizados a fim de corrigir erros numa próxima iteração.
9. Consolidação do conhecimento descoberto	Passo final do processo, consiste na incorporação do novo conhecimento descoberto nas métricas de performance do sistema, além da documentação, relatório, checagem e resolução de conflitos com conhecimentos previamente adquiridos.

Fonte: (CIOS et al., 2007)

3 REDES SOCIAIS

Este capítulo aborda os conceitos fundamentais a respeito de redes sociais e em particular às características do *Facebook* necessárias para o desenvolvimento deste trabalho. Na Seção 3.1 se encontra um panorama de características em comum a maioria das redes sociais atuais, seguido um histórico breve do surgimento desses tipos de *sites* nos últimos anos dentro da Seção 3.2. A Seção 3.3 contém explicações a respeito de alguns tipos diferentes de mídias e os diversos usos para as plataformas de redes sociais. Na última Seção 3.4 encontramos as definições providas pelo *Facebook* para os dados retirados e utilizados na base de dados deste trabalho.

3.1 CONCEITOS FUNDAMENTAIS

Desde seu surgimento, *sites* de redes sociais, como *Facebook* e *Twitter*, vêm atraindo milhões de usários ao redor do mundo. Segundo estatísticas do site Dossier Statista Dossier (), em 2017 haviam 2.46 bilhões de usários de redes sociais ao redor do mundo e é estimado que em 2019 esse número suba para 2.77 bilhões.

Definir um conceito de redes sociais exibe dois problemas distintos: a velocidade com que a tecnologia se expande, dificultando a nossa habilidade de definir com clareza limites claros para o conceitos; e, se as redes sociais servem para facilitar a comunicação de pessoas, deveriamos considerar telefone, *e-mail* e *fax* como redes sociais também? Para tentar endereçar estes problemas, Obar e Wildman (2015) resume as deinições de redes sociais da literatura em 4 pontos comuns:

1. serviços de rede sociais são aplicação interativas da *Web* 2.0;
2. conteúdo gerado pelos usuários é vital para a longevidade das redes sociais;
3. indiviuos ou grupos podem criar perfis específicos para um *site* ou *app* mantido pela rede social;
4. desenvolvem e facilitam o desenvolvimento de novas conexões entre perfis de um usuário com os de outros indivídos ou grupos.

(EXPLICAÇÃO MAIS DETALHADA DOS PONTOS)

3.2 HISTÓRICO DAS REDES SOCIAIS

Baseado nessas definições, podemos considerar o *site SixDegress.com*, lançado em 1997 como a primeira rede social lançada. No começo, o serviço apenas permitia criar um perfil e assinalar seus amigos, porém no ano seguinte ao lançamento foi adicionado a função de navegar pelas listas de seus amigos. Muitos desses recursos já existiam em outros tipos de serviços, como perfis em *sites* de relacionamento e listas de amigos em serviçso de comunicação

IRC, porém nenhum deles integrou essas funcionalidades para compor o conceito estabelecido de rede social Boyd e Ellison (2008).

Mesmo com o sucesso inicial, atingindo milhões de usuários, o *site SixDegress.com* teve seus servidores fechados em 2000. Seu fundador acredita que o serviço estava muito a frente de seu tempo, uma vez que, mesmo com a grande quantidade de pessoas entrando no mundo digital, muitos ainda não tinham uma extensa rede de amigos *online* e menos ainda estavam interessados em conhecer estranhos através da plataforma. Apesar de seu sucesso não ter sido duradouro, *SixDegress.com* serviu de inspiração para um grande número de novos *sites* entre 1997 e 2001 Boyd e Ellison (2008).

Em 2001 fora lançado o *site* de *business network Ryze.com* e no ano seguinte um complemento focado em relacionamentos *Friendster*, que viria a ser conhecido "uma das maiores decepções na história da *Internet*"Boyd e Ellison (2008).

Friendster passou por dificuldades técnicas durante seu primeiro ano de crescimento, atingindo trezentos mil usuários nesse período somente através de *marketing* de referência. Os servidores e centros de dados do *site* não estavam devidamente preparados para lidar com o rápido crescimento, causando quedas de servidores constante.

Outras decisões de negócios, como restringir de atividades de usuários muito ativos, permitir a descoberta de desconhecidos somente até quatro graus de distância, o que gerou a criação de perfis conhecidos como *Fakesters*, utilizados para coletar a maior quantidade de amigos possíveis a fim de circular a restrição de quatro graus de separação Boyd e Ellison (2008).

A decisão do *Friendster* de excluir as contas de perfis *fakes* indicava um rompimento entre os interesses do usuários e os desenvolvedores da plataforma e após rumores de que o serviço adotaria um modelo de inscrição paga, seus usuários começaram a encorajar o uso de redes sociais alternativas que surgiam na mesma época tentando capitalizar nos usuários que se decepcionavam com as decisões do *Friendster* Boyd e Ellison (2008).

A partir do ano de 2003, um grande número de serviços de redes sociais surgiram, ao ponto de analistas cunharem o termo *YASNS*: *"Yeat Another Social Networking Service"*. Apesar do grande sucesso de várias dessas plataformas em diversos países, como a "invasão Brasileira"do *Orkut*, poucas pessoas prestavam atenção em serviços de sucesso fora do mercado interno dos Estados Unidos da América. Uma dessas redes que capitalizou em cima da queda do *Friendster* foi o *MySpace*.

Um dos grupos de usuários que foram acolhidos pelo *MySpace*, bandas de músicas independentes, tiveram grande impacto para a popularização do serviço além dos antigos usuários da concorrência, fazendo com que o *site* ganhasse tração para crescer organicamente. O *MySpace* também se diferenciava de outras plataformas por regularmente adicionar funcionalidades requisitados por seus usuários , também "permitia"a personalização de perfis uma vez que o *site* não restringia a adição de código *HTML* dentro das páginas dos perfis Boyd e Ellison (2008).

Ao invés de restringir a presença de menores dentro do *site*, *MySpace* remodelou suas

políticas de usuários e adolescentes começara a utilizar o serviço em massa. Nesse período, os usuários do serviço se dividiam em três grandes grupos: musicos e artistas; adolescentes; e os antigos usuários do *Friendster*, estes já mais velhos. O único elo entre os 2 últimos grupos sendo os gostos por artistas em comum Boyd e Ellison (2008).

Em julho de 2005, após ganhar atenção em massa devido a compra do *site* por 580 milhões de dólares pela *News Corporation*, vários problemas de segurança começaram a surgir dentro do *MySpace*. Séries de alegações de interações sexuais entre adultos e menores começaram a surgir e aos poucos o pânico causado, apesar de pesquisas demonstrarem haver um exagero, foi responsável pela tomada de ações legais contra o serviço Boyd e Ellison (2008).

Em paralelo a esses serviços abertos, diversas redes sociais menores e focados em demográficos específicos começaram a ser lançadas, dentre estas, focada em somente alunos da Universidade de Harvard, foi criado e lançado em 2004 o *Facebook*. Aos poucos, o serviço foi aceitando novos usuários de outros círculos além de Harvard.

Em 2005, ano seguinte ao seu lançamento, o *Facebook* começou a aceitar estudantes de ensino médio e corporações, cada uma delas ainda precisando de um endereço de *e-mail* correto para acessar as redes privadas dentro do serviço. Eventualmente o serviço se tornou aberto a todos, permitindo que todo mundo pudesse criar uma conta dentro do *site*. Duas características distintas do *Facebook* das outras redes sociais são a inabilidade de seus usários de tornar todas as informações do seu perfil públicas para todos os usuários e a facilidade com que desenvolvedores externos podem criar aplicações com as informaçõs disponíveis dos perfis Boyd e Ellison (2008).

3.3 TIPOS E USOS DE REDES SOCIAIS

Apesar de não existir uma separação clara entre todos os serviços de redes sociais, é possível distinguir os serviços entre três categorias ou tipos em relação as suas funcionalidades centrais:

1. *networking*, são serviços focados no gerenciamento de círculos sociais e nas interações entre seu perfil e estes vários circulos, e.g. *Facebook*, *LinkedIn*;
2. *blogging*, onde as relações se dão ao redor da discussão de determinado conteúdo, e.g. *Blogger*, *Twitter*, *Reddit*.
3. *media sharing*, consistem de *sites* onde o enfoque é na distribuição de conteúdo gerado pelos usuários em diversos formatos, e.g. *YouTube*, *Vimeo*, *Instagram*.

É importante notar que essas classificações não significam que o serviço não possa possuir funcionalidade dos outros tipos. Várias funcionalidades podem ser encontradas em diversas redes sociais, como divulgação de vídeos dentro do *Facebook* ou *Twitter*, porém mesmo com essa possibilidades, a divulgação de vídeos não é o enfoque principal de nenhum desses dois seviços citados.

Ainda dentro de cada uma dessas classificações, encontramos diversos mercados diferentes para as funcionalidades de cada rede social, tanto de nicho quanto mais abrangentes.

Podemos tomar como alguns exemplos:

1. *real time*, focado na atualização de conteúdo no menor espaço de tempo possível entre o acontecimento real e digital, e.g. *Twitter*, *Snapchat*, *Periscope*;
2. *location based*, que se utiliza das tecnologias de localização por GPS na geração de conteúdo, e.g. *Swarm*.

3.4 *FACEBOOK*

Com o grande número de usuáriosna plataforma, não é de surpreender que o *Facebook* receba atenção de empresas como meio de comunicação e de marketing. A forma mais comum como negócios podem se utilizar da plataforma em seu benefício é através de páginas próprias, onde uma pessoa responsável pode controlar a publicação de conteúdo, espalhar esse conteúdo pela rede e permitir com que demais usuários possam interagir com essas publicações.

Dentro de áreas específicas dessas páginas, restritas a administradores,podemos encontrar uma grande quantidade de estatísticas a respeito de todas as postagens da página. A base de dados utilizada para este trabalho contém estatísticas retiradas das postagens da página de uma empresa renomada de cosméticos durante o ano de 2014 Moro, Rita e Vala (2016).

A base de dados contém 500 instâncias das 790 originais devido a questões de confidencialidade da empresa. Todas as instâncias contam com 19 atributos e, para a realização da modelagem das relações entre os atributos, foram separados em atributos de entrada do modelo e atributos de saída a serem modelados.

Os atributos de entrada estão descritos no Quadro 4 enquanto os atributos de saída estão descritos no Quadro 5.

Quadro 4 – Descrição dos atributos de entrada da base de dados utilizada no trabalho.

Atributo	**Descrição**
Page total likes	Quantidade de *likes* totais da página quando feita a postagem.
Type	Tipo de postagem, entre Fotos, *Link*, Status e Vídeo.
Category	Categoria de tipo de propaganda utilizada internamente pela empresa, adicionado de forma manual na base de dados.
Post Month	Mês em que a postagem foi feita, retirado da data da postagem.
Post Weekday	Dia da semana em que a postagem foi feita, retirado da data da postagem.
Post Hour	Hora do dia em que a postagem foi feita, retirado da data da postagem.
Paid	Se o post usou os serviços de anúncios pagos do *Facebook*.

Fonte: (MORO; RITA; VALA, 2016)

Quadro 5 – Descrição dos atributos de saída a serem modelados no trabalho.

Lifetime Post Total Reach	Quantidade de usuários únicos que a postagem alcançou, independente da forma com que a postagem chegou até o usuário.
Lifetime Post Total Impressions	Quantidade total de vezes que uma postagem foi vista [a].
Lifetime Engaged Users	Quantidade de usuários que clicaram na postagem de forma que geram ou não *Stories* [b].
Lifetime Post Consumers	Quantidade de usuários que clicaram na postagem de forma que não geram *Stories* [c].
Lifetime Post Comsumptions	Quantidade total de cliques que não geram *Stories* [d].
Lifetime Post Impressions by people who have liked your Page	Quantidade total de vezes que uma postagem foi vista por usuários que deram *like* na página.
Lifetime Post reach by people who like your Page	Parecido com *Lifetime Post Total Reach*, porém somente conta usuários que deram *like* na página.
Lifetime People who have liked your Page and engaged with your post	Quantidade de usuários que deram *like* na página e interagiram com a postagem de alguma forma que gera ou não *Stories*.
Comment	Quantidade de comentários da postagem.
Like	Quantidade de *likes* da postagem.
Share	Quantidade de compartilhamentos da postagem.
Total Interactions	Quantidade total de interações com a postagem, ou seja, a soma do total de comentários, *likes* e compartilhamentos.

Fonte: (MORO; RITA; VALA, 2016)

[a] Esse número pode ser maior que o *Total Reach* pois um mesmo usuário pode ver a postagem diversas vezes.
[b] *Stories* são tipos de interações que fazem com que a postagem seja propagada para outros usuários.
[c] Diferente da *Lifetime Engaged Users* pois conta os cliques dentro do conteúdo em si.
[d] Conta a quantidade de clique feitos pelos *Post Consumers*, aproximando quantas vezes que a postagem foi consumida

4 REVISÃO SISTEMATICA DA LITERATURA

Neste Capítulo encontra-se um levantamento bibliográfico a fim de obter um panorama abrangente dos estudos a cerca de mineração de dados em bases retiradas de redes sociais e os métodos e técnicas que são utilizados sobre esses dados. Na Seção 4.1 está descrito o método de revisão sistemáticas utilizado neste trabalho. A Seção 4.2 contém a aplicação do método e os resultados obtidos do mesmo. Por fim, a Seção 4.4 apresenta as considerações finais a cerca do Capítulo.

4.1 MÉTODO DE REVISÃO SISTEMÁTICA

Explicação do método de revisão sistemática. Baseado em PAGANI 2015 dividido em algumas etapas: definição da intenção da pesquisa e características dos tipos de trabalhos; definição palavras chaves e busca preliminar exploratória; busca efetiva e gerenciamento de bibliografia; primeiras filtragens; ordenação dos artigos filtrados; leitura e análise integral.

4.2 APLICAÇÃO DO MÉTODO

Nesta Seção está a aplicação do método utilizada neste trabalho.

4.2.1 Estabelecimento da inteção da pesquisa

4.2.2 Definição de palavras chaves e busca preliminar exploratória

Abaixo encontram-se as bases que foram usadas para a pesquisa de artigos da área de descoberta de conhecimento e mineração de dados

Quadro 6 – Bases de dados pesquisadas.

Base de dados	**Sites**
arXiv	<https://arxiv.org/>
Emerald Insight	<https://www.emeraldinsight.com/>
IEEE Xplore	<https://ieeexplore.ieee.org>
Science Direct	<https://www.sciencedirect.com/>
Springer	<https://link.springer.com/>

Fonte: Autoria própria

Foi aplicado um filtro de busca.

arXiv: query "social media and data mining" no abstract entre os anos 2000 e 2018 na área de ciência da computação

Emerald Insight: query "social media and data mining" no abstract entre os anos 2000 e 2018

IEEE Xplore: query "social media and data mining" no abstract entre os anos 2000 e 2019, filtrando por journals & magazines e conferences

ScienceDirect: query "social media data mining" no abstract entre os anos 2000 e 2018, filtrando research e review articles

Springer: query "social media data mining" entre os anos 2000 e 2018, filtrados por english, computer science, artificial intelligence, data mining and knowledge discovery, article

4.2.3 Busca efetiva e gerenciamento de bibliografia

Foi utilizado o *software* gerenciador de bibliografia Zotero para facilitar a vizualição e processo de filtragem dos artigos. Após a busca nas bases citadas na Tabela REFERENCIA PARA AS TABELAS E QUADROS foram encontrados o seguinte numero de artigos

Tabela 1 – Resultado das buscas.

Base de dados	**Resultados**
arXiv.org	75
Emerald Insight	30
IEEEXplore	457
Science Direct	236
Springer	320
Total:	1118

4.2.4 Procedimento de filtragem

Após os processos de deleção de artigos repetidos e seleção dos artigos relevantes ficamos com a quantidade de artigos mostrados an tabela abaixo.

Tabela 2 – Resultado dos filtos.

Base de dados	**Resultados**
arXiv.org	0
Emerald Insight	4
IEEEXplore	8
Science Direct	3
Springer	3
Total:	18

4.2.5 Ordenação dos artigos

4.2.6 Leitura e análise integral dos artigos

O artigo AN INTELLIGENT APPROACH FOR PREDICTING SOCIAL MEDIA IMPACT ON BRAND BUILDING buscou propor um modelo para previsão de influência de

postagens baseados em suas características, avaliar o modelo gerado com medidas de dissimilaridade entre os valores reais e preditos e fazer uma exploração das técnicas de mineração de dados que podem ser aplicadas na previsão de influência de postagens em redes sociais. O modelo proposto utilizou os algoritmos de *General Linear Regression* para associar as variaveis dependes de forma linear as suas variáveis independentes, *Normal Regression*, *Support Vector Machine* para classificar os dados em dois grandes grupos sem *overfitting*, *Neural Network* utilizado para clusterizar os dados em diferentes grupos menores e, por fim, *CHAID Decision Tree* para criar árvores não binárias que se utiliza da estatística *chi-squared* para determinar os melhores pontos de corte. O *dataset* foi dividido em 80% para treinamento e 20% para testes, sendo selecionado como atributos de entrada para o treinamento do modelo todos os atributos da instância exceto o atributo a ser previsto. Foram escolhidos como alvos para previsões os seguintes atributos: *Lifetime post total reach*, *Lifetime post total impressions*, *Lifetime post consumers* e *Lifetime post impressions by people who have liked your page*.

O artigo ANALYSIS OF DATA USING MACHINE LEARNING APPROACHES IN SOCIAL NETWORKS compara as técnicas de Aprendizagem de Máquina *Logistic Regression*, *Random Forest* e *Adaboost* utilizando métricas de análise de performance como *precision*, *recall* e *F1 score*. Os dados passaram pelo processo de teste e treinamento utilizando o método *10-fold crossover*. O atributo *Type* foi utilizado como classe, sendo os valores *Status*, *Video*, *Link* e *Photo*.

4.3 ARTIGOS ADICIONAIS

O artigo (ULTSCH; VETTER; VETTER, 1995) buscava comparar os métodos de clusterização estatística clássicos com o método *Self-Organizing Feature Map* (SOFM ou SOM), desenvolvido por Kohonen. Ao final do artigo foi provado como os grupos formados pelo método SOM eram melhores se comparados aos métodos clássicos de agrupamento.

4.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Considerações finais.

5 METODOLOGIA

Metodologia, ferramenta e base utilizada na pesquisa.

SVM:

Inventado originalmente por Vladimir Vapnik e Alexey Chervonenkis em 1963 com modificações para permitir classificações não-lineares em 1992 por Vapnik, os SVM são modelos de aprendizagem supervisionada utilizados nas tarefas de classificação e regressão. Na SVM, os dados são representados por pontos num espaço n-dimensional, sendo n o número de atributos dos seus dados. SVM são classificadores binários lineares pois tenta buscar um hiperplano que separe linearmente os dados em dois grupos distintos. Podem existir diversos hiperplanos de separação, porém a melhor escolha é o plano que possui maior margem, ou seja, o plano que contém a maior separação entre os dados mais próximos, dessa forma diminuindo o erro de generalização do classificador. Novos elementos que forem classificados serão julgados de acordo com qual lado do hiperplano eles se encontram. Casos que não podem ser linearmente separados podem ser modificados através das funções de kernel: funções que convertem os dados em espaços de maiores dimensões onde a análise pode ser feita e os dados separados linearmente por um hiperplano. Este hiperplano encontrado pode não ser uma reta nos planos originais dos dados. São algoritmos que funcionam bem em espaços de muitas dimensões, até mesmo maior do que o número de exemplos da base de dados, porém não são eficientes quando os dados possuem muito ruído. Em problemas de regressão as técnicas principais do SVM ainda podem ser usadas. Um hiperplano pode ser encontrado que divide o conjunto de dados em 2, porém a margem de separação passa a ser uma margem de erro aceitável, sendo que os dados que caiam dentro dessa margem de erro podem ser calculados seus valores a partir do hiperplano de separação e os dados que caem fora das margens são descartados como pontos fora da curva.

Dados:

A base de dados contava originalmente com 790 instâncias, porém apenas 500 foram disponibilizadas de forma pública devido a confidencialidades da página do Facebook em questão. Todas as instâncias contam com 19 atributos, após a criação de um modelo de previsão de performance da postagem, 7 desses atributos foram considerados como entrada e os outros 12 como medidas de performance. Os 7 atributos de entrada são: a quantidade de likes totais da página no momento da postagem em questão, o tipo de postagem (foto, vídeo, link ou status), a categoria (atributo interno da empresa adicionado manualmente), mês, dia da semana e hora da postagem, cada um em um atributo separado e se o post foi pago para ter mais publicidade. Os outros 12 atributos de performance são: quantas pessoas o post atingiu, quantas vezes a postagem foi vista, quantos usuários interagiram com a postagem ao todo, quantidade de pessoas que interagiram com a postagem sem "compartilhar com outras pessoas" através de histórias, quando vezes o conteúdo da postagem foi clicado sem "compartilhar com outras

pessoas" através de histórias, quantas vezes a postagem foi vista por pessoas que curtem a página em questão, quantas pessoas que curtem a página viram a postagem, quantas pessoas que curtem a página e interagiram com a postagem, quantidade de comentários, curtidas, compartilhamentos e o total desses últimos 3 atributos. Todos esses atributos, com exceção da categoria que foi adicionada manualmente por um funcionário da empresa, foram retirados das estatísticas da página da empresa de cosméticos diretamente do Facebook. Após o processo de mineração utilizando SVM, foi gerado um modelo para cada um dos 12 atributos de performance utilizando os 7 atributos de entrada. Os resultados dos modelos foram comparados com os valores reais retirados dos dados através de diferença absoluta e percentual, além da média de erro absoluto percentual para cada um dos 12 atributos. Para facilitar a análise dos atributos de saída, estes foram divididos em visualização e interação, sendo que os de interação tiveram como melhor resultado 27% de erro e o melhor de visualização com 37% de erro. Especulasse que os erros maiores com relação a previsão de visualização se devem a maiores fatores fora do controle dos dados e dos modelos, como a aleatoriedade do algoritmo de feed do Facebook.

SOM:

(ULTSCH; VETTER; VETTER, 1995) SOFM são usados para classificar dados de altas dimensões mapeando os dados para dimensões menores e quanto combinado com o método de U-Matrix ajuda a identificar clusteres de dados de forma mais fácil que métodos tradicionais de clusterização. Apesar de ser comparado com o método k-means por causa do seu método de particionamento, o SOFM não é idêntico e em alguns casos tem melhor performance que o seu comparativo. Kohonen's self organizing map representa um modelo de rede neural utilizado em aprendizagem não supervisionada, DEFINIÇÃO DE NÃO SUPERVISIONADO: o que significa que os resultados são provenientes das propriedades inerentes dos dados em si, não tendo um mestre que diz o que é certo ou errado e que não é preciso conhecer a estrutura dos dados anteriormente à aplicação do algoritmo. O algoritmo adapta dados em altas dimensões numa grade de duas dimensões e preservando a vizinhança na topologia do mapa as estruturas dos dados podem ser descobertas explorando o mapa, e se a propriedade de preservação de vizinhança estiver correta então dados próximos terão uma correspondência na mesma latitude no mapa de Kohonen, ou seja, os clusteres de dados de altas dimensões estaram também em clusteres na projeção em menor dimensão. Como as distâncias entre os pontos do mapa estão igualmente distribuídos, somente o método de SOFM não nos indica a presença ou não de clusteres. O método de U-Matrix tem como ideia visualizar a topologia dos mapas de atributos. Ao analisar os pesos em cada ponto da grade com seu vizinho e mostrando a distância entre os dois vizinhos como altura, temos uma figura em tres dimensões do mapa de Kohonen. U-Matrix, curto para unified distance matrix, contém uma aproximação geométrica do vetor distribuição da rede de Kohonen. Os vales representam dados que estão próximos enquanto colinas representam uma separação e maior distância entre os vizinhos.

(KOLEHMAINEN et al., 2004) seja $\mathbf{X}$ uma matriz de dados com p atributos e n amostras, um mapa auto-organizado consiste de M neurônios organizados numa grade de duas

dimensões, cada uma das células contendo seu vetor de peso para cada uma das variáveis. FORMULA w_m =(w_m1,..w_mp) , (m=1...M). Os pesos são inicializados para valores aleatórios. Para cada vetor **x**_i é encontrado o melhor neurônio (BEST MATCHING UNIT), ou seja o neurônio com a maior proximidade com o vetor de entrada, ou seja o BMU é o neurônio que tenha a menor distância Euclidiana do vetor de entrada e dos pesos da matriz. Os pesos dos vizinhos (que são escolhidos de acordo com uma função) são corrigidos junto com o peso do próprio BMU se utilizando do contador de interações, o index do BMU e o index do neurônio. Processo básico do SOM: 1. econtre o BMU para um vetor de entrada de acordo com a menor distância euclidiana, 2. mova o vetor peso do BMU em direção ao vetor de entrada de acordo com a formula apropriada, 3. mova o vetor de peso dos neurônios vizinhos em direção ao vetor de entrada, 4. repetir até que todosos vetores de entrada tenham sido utilizados, reptia tudo até que o mapa seja totalmente coberto, 6. encontre o BMU final para cada uma das entradas de acordo com a distância euclidiana, ou seja, o neurônio a qual o vetor pertence. FÓRMULAS NESTE ARTIGOS DE COMO OS CÁLCULOS SÃO FEITOS.

(WENDEL; BUTTENFIELD, 2010) segundo a pesquisa, não foi encontrado nenhuma fonte única de regras e boas práticas na criação de SOMs efetivas, então este trabalho tenta estabelecer *guidelines* para a criação de SOMs através de teste empíricos, variando os parâmetros de inicialização, os níveis de treinamento, o tamanho da rede e dos nós, comparando saída, medidas de incerteza e interpretações, tudo isso em conjunto com ´boas práticas´ da área. As guidelines foram agrupadas em 6 categorias: inicialização (linear ou aleatória, sem necessidade de normalização de dados binários, porém Skupin 2008 recomenda a inicialização aleatória para preservar o processo como completamente auto-governado), tamanho do mapa (as recomendações iniciais se referem ao propósito da construção e objetivos do mapa a ser gerado seungo Ultsch e Simon 1990, Vesanto 2005 recomenda uma fórmula de tamanho ótimo RETIRAR FÓRMULA DE TAMANHO. Buscar um tamanho ótimo evita que a computação do mapa fique muito pesada por lidar com muitos espaços vazios, mas um pouco precisa ser preservado a fim de facilitar na interpretação dos clusteres), forma do mapa (as recomendações por Kohonen sugeram uma geometria assimétrica a fim de evitar efeito de bordas, sendo que o lado mais curto deveria ser no máximo metade do lado mais comprido, isso é necessário pois clusteres localizados no centro do mapa são mais fáceis de interpretar como separados de grupos de clusteres que se encontram em bordas), geometria e tamanho da vizinhança (esse fator influencia na hora do recalculo dos pesos pois é necessário levar em conta e atualizar os pesos dos vizinhos do dado a ser conferido na interação, para estabelecer o melhor tamanho a primeira interação o tamanho da vizinhança abragia o mapa todo e a medida que as interações iam passando o tamanho foi sendo reduzido pela metade), duração de treinamento e ajuste de matrix (training lenght = 10000 e matrix tunning rate = 2000), qualificação da incerteza (utiliza o erro de quantização e U-Matix, erro de quantização é dado por uma fórmula que retorna valores entre 0 e 1 de acordo com o quanto o mapa se adequa aos dados, um retorno de 0 pode indicar overfitting e essa medida permite a comparação quantitativa entre mapas

diferentes, já a U-Matrix indica qualitativamente e a clareza com que os clusteres são divididos, contém 2 dobro de espaços da SOM e é definida pela medida de similaridade entre uma célula e suas vizinhas, bordas entre clusteres muito forte podem indicar overfitting dos dados também. ISSO É UM RESUMO DE UMA PALESTRA DE CONFERÊNCIA, VERIFICAR A VALIDAD DE USAR ESSAS INFORMAÇÕES.

(ULTSCH, 1990) desde sua introdução em 1982, SOMs vem sendo usados em diversas áreas por sua característica de a estrutura de dados em altas dimensões, porém a aplicação exclusiva dos mapas auto-organizáveis não necessariamente indica alguma informação a respeito de clusteres e análises de dados devem ser conduzidos sobre os mesmos para que sejam tiradas conclusões a respeito de clusterização. Explicação de como o algoritmo de Kohonen para SOM funciona e como U-Matrix funcionam em detalhes e fórmulas.

5.1 PRÉ-PROCESSAMENTO

Tudo que foi aplicado na base de dados.

5.2 MINERAÇÃO DE DADOS

Tudo que foi aplicado para mineração, svm e regressão.

5.3 PÓS-PROCESSAMENTO

Tudo que foi feito após o proc de datamining.

6 ANÁLISE E DISCUSSÃO DOS RESULTADOS

6.1 RESULTADOS DO PROCESSO DE KDD

Diversos processos de SOM foram aplicados na base a fim de definir quais seriam bons parâmetros de geração de mapas. Foram feitos mapas com grids 16x7 e 16x16 tanto hexagonais quanto retangulares. Nesses 4 grids foram variados os valores de aprendizagem entre 0.05, 0.25 e 0.5 e números de iterações entre 100, 500 e 1000 interações. Foram gerados 36 gráficos de progressão do treinamento dos mapas para dois grupos distintos, os atributos de entrada e saída da base estudada.

INPUT

Algumas características interessantes podem ser retiradas destes gráficos: Apesar de 100 iterações serem na maioria das vezes suficientes para encontrar os mesmos valores de distâncias média entre unidades da SOM mais próximos, todos em torno do 0,01, em alguns dos casos não se pode determinar com precisão se os dados chegaram em um platô onde não é mais perceptível variações, podendo indicar que mais interações poderiam resultar em distâncias melhores. Para esse propósito, os testes com 500 interações são mais confiáveis onde podemos claramente ver que alguns dos teste geraram distâncias mais estáveis a medida de mais iterações. De 500 iterações para 1000 iterações não produzem resultados visivelmente melhores para o treinamento do mapa, mas aumentam bastante o tempo de treinamento dos mesmos, então foi feito a preferência pelos testes de 500 iterações.

Podemos notar também que os mapas retangulares demonstram decaimentos em degrau mais acentuados que os hexagonais na maioria dos casos, exceto naqueles com baixas iterações ou taxas de aprendizagem.

padrões de destaque input:

1. 16x7h 0.5 500 it
2. 16x7r 0.5 500 it muito próximo de 16x7r .25 500 it
3. 16x7r é sempre bem mais definida do que 16x7r 100
4. 16x16h .5 e .25 tanto 500 quanto 1000 it atingiram os valores mais baixos de distânica ($<$0.010) porém com maior custo de treinamento
5. 100 it não presta, 500 e 1000 não tem muita diferença, 16x16 atingiu menores distâncias do que 16x7, r tem caimento em degrau mais acentuado que h que possui caimento mais acentuado somente nas iterações finais, 0.5 de aprendizagem se mostrou mais eficiente do que o 0.25 porém os dois são bem mais eficientes do que somente 0.05.

mapa escolhido inputs: 16x16r 0.5 500 it

OUTPUT

De forma geral, provavelmente pela maior dispersão dos dados, os dados mostraram clareza de convergência somente com mais de 500 iterações e muitas vezes somente quando

realizados o treinamento com 1000 iterações. Todos os resultados de outputs geraram distâncias menores do que os inputs.

padrões de destaque output:

1. Para 16x7h 0.05 não é possível perceber nenhum padrão ou convergência clara de menor distância, comente nas 16x7h com 1000 iterações podemos ver uma queda com indicação de estabilidade tanto em 0.5 quanto 0.25 de aprendizagem.
2. para 16x7r as 1000 iterações se provaram melhores do que somente 500 iterações com os melhores resultados em 0.5 e 0.25 de aprendizagem
3. 16x16h 0.5 500 it é suficiente, porém os outrs casos com 1000 it são bons também
4. 16x16h 0.25 1000 it único nessa faixa de aprendizagem
5. 16x16r 0.5 é a unica taxa de aprendizagem que mostrou alguma estabilidade das distâncias

mapa esccolhido para outputs: 16x16r 0.5 500 it

6.2 ANÁLISE DOS RESULTADOS

Comparação dos resultados com os dados reais da base e de oturos trabalhos

6.2.1 Comparativo entre dados reais coletados

6.2.2 Comparativo entre outros trabalhos relacionados

7 CONCLUSÃO

Resultados obtidos da regressão

7.1 TRABALHOS FUTUROS

Continuações do trabalho

7.2 CONSIDERAÇÕES FINAIS

Considerações finais

Referências

BAGGA, S.; SINGH, G. N. Three phase iterative model of kdd. **International Journal of Information Technology and Knowledge Management**, v. 4, n. 2, p. 695–697, 2011. Citado 2 vezes nas páginas 5 e 6.

BOYD, D. D.; ELLISON, N. B. Social network sites: Definition, history and scholarship. **Journal of Computer-Mediated Communication**, v. 13, p. 210–230, 2008. Citado 2 vezes nas páginas 9 e 10.

CIOS, K. J. et al. **Data Mining, A Knowledge Discovery Approach**. [S.l.]: Springerl, 2007. Citado 3 vezes nas páginas 4, 6 e 7.

CLOUD, I. M. **10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations**. Disponível em: <https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN>. Citado na página 1.

DOSSIER, S. **Social media & user-generated content - Number of global social network users 2010 - 2018**. Disponível em: <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>. Citado 2 vezes nas páginas 1 e 8.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, 1996. Citado 3 vezes nas páginas 1, 4 e 5.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unyfing framework. **KDD-96 Proceedings**, 1996. Citado na página 4.

GROUP, M. M. **Internet World Stats**. Disponível em: <https://www.internetworldstats.com/stats.htm>. Citado na página 1.

KOLEHMAINEN, M. T. et al. **Data exploration with self-organizing maps in environmental informatics and bioinformatics**. [S.l.]: Helsinki University of Technology, 2004. Citado na página 17.

KURGAND, L. A.; MUSILEK, P. A survey of knowledge discovery and data mining process models. **The Knowledge Engineering Review**, v. 21, n. 1, p. 1–24, 2006. Citado 3 vezes nas páginas 1, 4 e 5.

MORO, S.; RITA, P.; VALA, B. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. **Journal of Business Research**, v. 69, 2016. Citado 2 vezes nas páginas 11 e 12.

OBAR, J. A.; WILDMAN, S. Social media definition and the governance challenge: An introduction to the special issue. **Telecommunications Policy**, v. 39, 2015. Citado na página 8.

SOCIAL, W. A.; HOOTSUITE. **Global Digital Statshot**. Disponível em: <https://wearesocial.com/special-reports/digital-in-2017-global-overview>. Citado na página 1.

ULTSCH, A. Kohonen's self organizing feature maps for exploratory data analysis. **Proc. INNC90**, p. 305–308, 1990. Citado na página 19.

ULTSCH, A.; VETTER, C.; VETTER, C. **Self-Organizing-Feature-Maps versus statistical clustering methods: a benchmark**. [S.l.]: Fachbereich Mathematik, 1995. Citado 2 vezes nas páginas 15 e 17.

WENDEL, J.; BUTTENFIELD, B. P. Formalizing guidelines for building meaningful self-organizing maps. **GIScience 2010 Short Paper Proceedings, Zurich, Switzerland, September**, 2010. Citado na página 18.

Apêndices

APÊNDICE A – Nome do apêndice

Lembre-se que a diferença entre apêndice e anexo diz respeito à autoria do texto e/ou material ali colocado.

Caso o material ou texto suplementar ou complementar seja de sua autoria, então ele deverá ser colocado como um apêndice. Porém, caso a autoria seja de terceiros, então o material ou texto deverá ser colocado como anexo.

Caso seja conveniente, podem ser criados outros apêndices para o seu trabalho acadêmico. Basta recortar e colar este trecho neste mesmo documento. Lembre-se de alterar o "label" do apêndice.

Não é aconselhável colocar tudo que é complementar em um único apêndice. Organize os apêndices de modo que, em cada um deles, haja um único tipo de conteúdo. Isso facilita a leitura e compreensão para o leitor do trabalho.

APÊNDICE B – Nome do outro apêndice

conteúdo do novo apêndice

Anexos

ANEXO A – Nome do anexo

Lembre-se que a diferença entre apêndice e anexo diz respeito à autoria do texto e/ou material ali colocado.

Caso o material ou texto suplementar ou complementar seja de sua autoria, então ele deverá ser colocado como um apêndice. Porém, caso a autoria seja de terceiros, então o material ou texto deverá ser colocado como anexo.

Caso seja conveniente, podem ser criados outros anexos para o seu trabalho acadêmico. Basta recortar e colar este trecho neste mesmo documento. Lembre-se de alterar o "label" do anexo.

Organize seus anexos de modo a que, em cada um deles, haja um único tipo de conteúdo. Isso facilita a leitura e compreensão para o leitor do trabalho. É para ele que você escreve.

ANEXO B – Nome do outro anexo

conteúdo do outro anexo