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* **??**), 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 **??**).

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* (REFERENCIAR RELATÓRIOA DEPOIS), 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 **??**)

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* (REFERNCIAR IBM DEPOIS) 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 (REFERENCIAR SURVEY KDD).

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 (REFERENCIAR FAYYAD FROM DATA MINING TO KDD).

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 (REFERENCIAR SURVEY KDD).

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 **??**).

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 **??**).

É importante ressaltar a natureza iterativa dos modelos de KDP, podendo conter diversos laços de *feedback* e repetição entre quaisquer dois passos do processo (REFERENCIAR SURVEY KDD). 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 (REFERENCIAR KDD UNIFYING FRAMEWORK).

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 (REFERENCIAR FROM DATAMINING TO KDD).

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 (REFERENCIAR FROM DATAMINING TO KDD).

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* (REFERENCIAR ISTO ???). 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 (REFERENCIAR SURVEY KDD).

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 (REFERENCIAR SURVEY KDD).

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 (REFERENCIAR FROM DATA MINING TO KDD).

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 (REFERENCIAR 3 STEPS KDD) 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: REFERENCIAR DATA MINING A KNOWLEDGE DISCOVERY APPROACH

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 (REFERENCIAR 3 STEPS KDD). 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: REFERENCIAR DATA MINING A KNOWLEDGE DISCOVERY APPROACH

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: REFERENCIAR DATA MINING A KNOWLEDGE DISCOVERY APPROACH

3 REDES SOCIAIS

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 (REFERENCIAR DOSSIER STATISTA), 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.

Tentar definir em termos precisos o que seria uma rede social se prova difícil devido a grande variedade de serviços independentes e integrados de serviços de comunicação. Simplesmente definir como serviços que permitem aproximar pessoas de forma digital se torna uma definição muito ampla.

3.1 CONCEITOS FUNDAMENTAIS

Revisões de literatura permitiram extrair 4 características principais recorrentes nos serviços de *networking* social.

(REFERENCIAR SOCIAL MEDIA DEFINITION PEGAR ARTIGO NO SCIENCE DIRECT: https://www.sciencedirect.com/science/article/abs/pii/S0308596115001172)

1. social networking services are interactive Web 2.0 Internet-based applications,
2. user-generated content (UGC), such as user-submitted digital photos, text posts, "tagging", online comments, and diary-style "web logs"(blogs), is the lifeblood of the SNS organism,
3. users create service-specific profiles for the site or app that are designed and maintained by the SNS organization, and
4. social networking services facilitate the development of social networks online by connecting a user's profile with those of other individuals or groups.

Breve histórico de algumas redes sociais famosas.

Tipos de redes sociais brevemente explicados.

1. *Networking*
2. *Blogging*
3. *Microblogging*
4. *Photo Sharing*
5. *Video Sharing*

Usos diferentes de redes sociais

1. *Real time*
2. *Location based*
3. Mercados de nicho
4. Ciência
5. Educação
    a) Profissional

   b) Currículo
   c) Aprendizado
6. Procuro de emprego
7. Serviços de *hosting*
8. Trocas

3.2 IMPACTOS E RELEVÂNCIA

Alguns benefícios

Alguns problemas

Relevância dos estudos na área

3.3 *FACEBOOK*

Breve histórico *Facebook*.

Facebook utilizado em negócios para empresas.

Explicação das métricas de avaliação de desempenho de postagem pelo *Facebook*. Se encontram no quadro da metodologia

4 REVISÃO SISTEMATICA DA LITERATURA

Revisão dos trabalhos relacionados da área

4.1 MÉTODO DE REVISÃO SISTEMÁTICA

Explicação do método de revisão sistemática

4.2 APLICAÇÃO DO MÉTODO

Resultados da aplicação da revisão sistemática.
estou mudando o texto

Quadro 4 – Bases de dados pesquisadas.

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

Tabela 1 – Resultado das buscas.

Base de dados	**Resultados**
arXiv.org	21
Emerald Insight	4
IEEEXplore	752
Science Direct	266
Springer	902

5 METODOLOGIA

Metodologia, ferramenta e base utilizada na pesquisa.

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.

Quadro 5 – Descrição dos atributos 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*.
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. Esse número pode ser maior que o *Total Reach* pois um mesmo usuário pode ver a postagem diversas vezes.
Lifetime Engaged Users	Quantidade de usuários que clicaram na postagem de forma que geram ou não *Stories*. *Stories* são tipos de interações que fazem com que a postagem seja propagada para outros usuários, como, por exemplo, Compartilhar. Dessa forma essa estatística conta a quantidade total de usuários que clicaram na postagem de uma forma qualquer. (RETIRAR EXPLICAÇÃO DE STORIES DAQUI)
Lifetime Post Consumers	Quantidade de usuários que clicaram na postagem de forma que não geram *Stories*. Essa estatística é diferente da *Lifetime Engaged Users* pois só conta os usuários que clicaram na postagem de forma a não espalhar a postagem, dessa forma contando somente clique dentro do conteúdo em si, como tocar o vídeo ou ampliar a foto, por exemplo.
Lifetime Post Comsumptions	Quantidade total de cliques que não geram *Stories*. Essa estatística conta a quantidade de clique feitos pelos *Post Consumers*, tentando aproximar a quantidade de vezes que a postagem foi consumida, seja quantidade de vezes que o vídeo foi visto ou quantos clicaram no link compartilhado.
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	Quantidade de usuários únicos que a postagem alcançou, independente da forma com que a postagem chegou até o usuário, porém

6 ANÁLISE E DISCUSSÃO DOS RESULTADOS

6.1 RESULTADOS DO PROCESSO DE KDD

Conhecimento retirado do processo de kdd dos dados

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