Centro Universitário de Anápolis – UniEVANGÉLICA
Bacharelado em Engenharia de Computação

TATIANE GOMES DOS SANTOS

ANÁLISE DE OPINIÕES UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS EM REDES SOCIAIS.

ESTUDO DE CASO: *TWITTER*.

ANÁPOLIS - GO
2017

Tatiane Gomes Dos Santos

Análise de Opiniões Utilizando Técnicas de Mineração de Dados em Redes Sociais.

Estudo de Caso: *Twitter.*

Projeto de pesquisa apresentado ao Curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis– UniEVANGÉLICA como requisito parcial à aprovação na disciplina Trabalho de Conclusão de Curso II sob orientação da Prof.ª. Ms. Luciana Nishi.

ANÁPOLIS - GO

2017

Tatiane Gomes Dos Santos

Análise de Opiniões Utilizando Técnicas de Mineração de Dados em Redes Sociais.

Estudo de Caso: *Twitter.*

Projeto de pesquisa apresentado ao Curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA como requisito parcial à aprovação na disciplina Trabalho de Conclusão de Curso II sob orientação da Prof.ª. Ms. Luciana Nishi.

COMISSÃO EXAMINADORA

Profª. (Ms.) Luciana Nishi

Prof. (Dr.) Raphael de Aquino Gomes

Prof.(Ms) Marcelo de Castro Cardoso

Aprovado em: 05 de Dezembro de 2017

AGRADECIMENTO

Agradeço, primeiramente a Deus, por ter me permitido concluir esse trabalho, me dado força para superar as dificuldades.

Aos meus pais pelo incentivo aos estudos que sempre me proporcionaram.

A minha orientadora Profª (Ms) Luciana Nishi, por ter aceitado me orientar ao longo desse ano.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigada.

“Porque a sabedoria serve de defesa, como de defesa serve o dinheiro; mas a excelência do conhecimento é que a sabedoria dá vida ao seu possuidor”.

(Bíblia Sagrada, Eclesiastes, 7, 12).

RESUMO

As técnicas de Mineração de Dados, possibilitam classificar as opiniões extraídas das redes sociais, oportunizando na compreensão dos sentimentos emitidos nas menções online. Este estudo apresenta uma análise da execução de técnicas de Mineração de Dados na rede social *Twitter*, a preferência por essa mídia social, em razão de que ela fornece recursos que tornaram factível a coleta dos dados. Para a composição do trabalho utilizou-se as seguintes metodologias: levantamento bibliográfico, aplicação do processo KDD (*Knowledge Discovery in Databases*) e a execução das principais técnicas de *Data Mining* (*K–Nearest Neighbor*, Máquina de Vetor de Suporte e *Naives Bayes*), a realização da coleta dos dados relativo ao fato do presidente Michel Temer ser absolvido do crime de corrupção passiva e a classificação das opiniões contidas nos elementos extraído. Após a utilização desse procedimento, percebe-se que a técnica *K–Nearest Neighbor* apresentou melhores resultados pelas métricas de avaliações aplicadas.

Palavras-chave: *Twitter.* Mineração de Dados. Classificação das opiniões.

ABSTRACT

The Data Mining techniques allow to classify the opinions extracted from social networks, giving an opportunity to understand the feelings emitted in the online mentions. This study presents an analysis of the execution of data mining techniques in the social network Twitter, the preference for this social media, because it provides resources that made the data collection feasible. For the composition of the work, the following methodologies were used: bibliographical survey, KDD (Knowledge Discovery in Databases) process and execution of the main techniques of Data Mining (K-Nearest Neighbor, Support Vector Machine and Naives Bayes), the collection of data concerning the fact that President Michel Temer is acquitted of the crime of passive corruption and the classification of opinions contained in the extracted elements. After the use of this procedure, it is noticed that the K-Nearest Neighbor technique presented better results by the applied evaluation metrics.

Key-words: Twitter. Data Mining. Classification of opinions.

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

API	*Application Programming Interface*
ARFF	*Attribute Relation File Format*
CSV	Valores Separados por Vírgulas
HTTP	*Hypertext Transfer Protocol*
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
KNN	*K–Nearest Neighbor*
SVM	*Support Vector Machines*
URL	*Uniform Resource Locator*

SUMÁRIO

1. INTRODUÇÃO

As redes sociais denotam em uma interação e trocas sociais, nos acompanham desde o início da civilização, quando o homem se reunia em volta de uma fogueira para compartilhar as suas predileções e interesses comuns, elas surgiram quando houve a necessidade de compartilhar gostos e de criar laços, ligados por afinidades (SANTOS, 2017).

No momento em que as interações sociais atuam em ambiente online, dispomos das redes sociais online. Com o início da Web, as trocas de e-mail foram os primeiros modos de relacionamentos na Internet, com o decorrer dos anos ocorreu um aumento relevante do número de internautas (SANTOS, 2017).

O avanço da tecnologia da informação e a popularização dos computadores pessoais, *smartphones* e *tablets*, propiciou transformações na forma de se interagir, comunicar e expor as opiniões. As pessoas estão cada vez mais conectadas, no Brasil são mais de 94 milhões de usuários conectados na Internet (QUEEN, 2016).

Em consequência da popularidade e a facilidade de acesso à Internet, foram desenvolvidas várias redes sociais, Calazans e Lima (2013, p.12) destaca que: "essa maior velocidade de crescimento dos sites de redes sociais[...] é que a população conectada do mundo continua em expansão, impulsionada pela queda nos preços de serviços de telefone e de Internet banda larga".

> À medida que a Internet foi se tornando mais facilmente programável pelo usuário comum, seu uso foi se diversificando e se expandindo. Novos ambientes de expressão foram criados, permitindo a rápida produção e distribuição de conteúdo multimidiáticos e se popularizando de maneira veloz. Nesses espaços, usuários passaram a expressar sua individualidade, através da exposição de suas opiniões e gostos pessoais, saindo da posição passiva imposta pelas mídias tradicionais. A estrutura de rede se tornou cada vez mais popular, de maneira que diversos sites surgiram fundamentados nas conexões e laços sociais entre os atores que o utilizam. (CALAZANS; LIMA, 2013, p.11).

Tal advento viabilizou o crescimento do uso das redes sociais, um ambiente em que os usuários podem expressar as suas opiniões a respeito dos mais múltiplos assuntos: produtos, viagens, ideologias políticas e religiosas. Enfim, temáticas que estimulam os interesses de indivíduos de se conectarem em alguma rede social para emitirem as suas opiniões.

As opiniões expostas nas redes sociais contém emoções e sentimentos por trás da encadeação de palavras citadas, constituindo-se de dois componentes: um alvo tópico e um sentimento no alvo tópico.

Então, por exemplo na frase: “Eu amo essa empresa”, a palavra “essa empresa” é o tópico, e o sentimento é expressado pelo verbo “amor”, é uma palavra positiva. Se uma porção de conteúdo expressos contém mais palavras-chave positivas do que palavras-chave negativas, o conteúdo dessa informação é positivo, caso contrário, o conteúdo dessa informação é negativo (DONKOR, 2013).

A apuração das opiniões classificadas nas mídias sociais, ajudam a assimilar o que os usuários estão sentindo, sendo utilizada para obter uma compreensão das opiniões e emoções expressadas, propiciando obter uma visão geral e ampla da opinião pública por trás de certos tópicos (BANNISTER, 2015).

Porém, a intensa emissão de opiniões através das redes sociais disponibilizam informações difíceis de se interpretar devido à ambiguidade e o grande fluxo de dados. Com o propósito de descobrir padrões latentes inclusos nesses dados, as técnicas de Mineração de Dados, no campo da análise dos dados têm se revelado útil para a classificação das opiniões emitidas, uma vez que essas técnicas investigam por paradigmas nos dados.

No entanto, de que modo o procedimento da execução das principais técnicas de Mineração de Dados, impactariam no processo de classificação e análise de opiniões, em um domínio de dados específico extraído da rede social *Twitter?*

A escolha do *Twitter* para essa pesquisa, porque essa mídia social fornece recursos, que tornaram esse trabalho factível, na seção 3.2 desse estudo, há uma explicação pelos motivos e fatores que levaram a predileção dessa rede social.

Para responder a essa pergunta, tem-se como objetivo “apresentar os resultados obtidos da análise e validação do uso das principais técnicas de Mineração de Dados no *Twitter* para a classificação de opiniões para um domínio especifico definido durante o processo”.

Através das seguintes etapas: (a) averiguar as principais técnicas de mineração de dados; (b) definir quais técnicas de Mineração de Dados serão aplicadas; (c) determinar um domínio de dados para a coleta de dados através do parâmetro da disponibilidade dos dados; (d) coletar os dados do domínio delimitado; (e) executar técnicas de Mineração de Dados; (f) classificar as opiniões contidas nos dados coletados e analisar as informações classificadas.

Devido à expansão dos usuários expressando as suas opiniões nas mais diversas redes sociais, a respeito dos mais variados assuntos, as redes sociais oferecem um ambiente ideal para

investigar os dados publicados nelas, resultando na oportunidade de estudar e explorar as informações contidas nas mídias sociais.

> A internet e a interação via web mudaram a forma como os produtos e serviços estão sendo inventados, produzidos, comunicados e distribuídos. Estamos em uma era em que as pessoas participam da economia como nunca. Se antes o consumidor apenas passivamente recebia informações, produtos e serviços das marcas, hoje ele é ativo e sabe que tem poder de impacto sobre elas [...] (HIRANAKA, 2016, p.147).

Com esse cenário, surge o desafio de investigar os dados nas mídias sociais com o maior nível de exatidão possível, mas uma série de fatores dificultam a análise destes dados, pois os dados contidos nas redes sociais dispõem de vários formatos, fontes e estruturas.

Como por exemplo, uma mesma informação pode ser citada de várias formas (ambiguidade) devido a aspectos regionais e/ou culturais e através do uso de gírias e/ou abreviações.

> [...] o volume crescente de conteúdo subjetivo disponível diariamente, [...] nas redes sociais, motiva o crescimento da área[...]. Muitas são as aplicações centradas na sumarização e visualização do sentimento, ou na predição de comportamentos com base no sentimento existente. Empresas, eventos[...], personalidades, estão interessadas na compreensão de como são percebidas pelo público em geral em tempo real, e nas mais variadas mídias (BECKER; TUMITAN, 2013, p.23).

As mídias sociais viabilizam diversas fontes de informações, possibilitando na tomada de decisões estratégicas, podendo ser aplicadas em diversas áreas. Portanto, os dados das mídias sociais são claramente a maior e mais rica e dinâmica evidência da base do comportamento humano, trazendo novas oportunidades para entender indivíduos, grupos e sociedade (BATRINCA; TRELEAVEN, 2014).

Embora na Mineração de Dados (*Data Mining*), tenha técnicas para realizar a classificação de opiniões contidas em textos das mídias sociais, segundo Castro e Ferrari (2016, p.19) a aplicação de técnicas de mineração de dados: “possibilita extrair informações escondidas nos dados, [...] e muitas outras informações úteis e indispensáveis para a tomada de decisão estratégica”.

Contudo saber quais são as particularidades do manuseamento dessas técnicas, pode trazer vários benefícios no momento de se realizar a classificação das opiniões, pois se saberá quais são as implicações da utilização das principais técnicas de *Data Mining*.

Pretende-se demonstrar através da validação das principais técnicas de mineração de dados os seus aspectos e peculiaridades de aplicação. Assim sendo investigando qual técnica se

evidenciou a melhor execução para a classificação das opiniões nas mídias sociais e verificar os efeitos para a classificação, por conseguinte obtendo uma análise refinada das opiniões extraídas das redes sociais e uma maior assertividade na classificação dos dados.

O primeiro passo para realizar essa pesquisa, será a pesquisa bibliográfica, que fornecerá experiência teórica do conhecimentos necessários para produzir essa pesquisa, e habilitará a produção da escrita dessa monografia.

Lakatos e Marconi (2003, p.158) complementam que "a pesquisa bibliográfica é um apanhado geral sobre os principais trabalhos já realizados, revestidos de importância, por serem capazes de fornecer dados atuais e relevantes relacionados com o trabalho".

Posteriormente, serão investigadas as principais redes sociais utilizadas atualmente e os recursos e API que elas dispõem para a coleta dos dados, e em especial o *Twitter* (estudo de caso), para demonstrar que essa rede social é adequada para a efetivação das principais Técnicas de *Data Mining*.

Além disso haverá a necessidade de definir qual o domínio de dados poderá ser utilizado para esse estudo de caso, para esse fim será pesquisado o domínio que melhor se aplicará para a classificação de opiniões no *Twitter*, a escolha será através do parâmetro de disponibilidade de dados.

Subsequentemente será necessário executar o Processo de Descoberta de Conhecimento– KDD que constituem nas seguintes etapas: seleção de dados, identificação de conjuntos das bases de dados, limpeza dos dados, transformação dos dados para análise.

O próximo passo será executar técnicas de Mineração de Dados, para aplicar as técnicas será utilizado o Weka uma ferramenta construída em Java que auxilia no uso de classificadores de aprendizagem de máquinas (WEKA, 2017).

E a última etapa é a classificação das opiniões e a análise, que se constituem na identificação das opiniões contidas nos textos emitidos em uma rede social, Becker e Tumitan (2013, p.7) afirmam que a: "classificação de polaridade, é frequentemente um problema de classificação binário, isto é, que classifica um dado texto em uma de duas classes: positivo ou negativo".

Na seção 2 é demonstrado o referencial teórico no qual abordou os seguintes itens: Rede Social Online, Descoberta de Conhecimento em Bases de Dados, Mineração de Opiniões, Métricas

de Avaliação, a Tarefa de Mineração de Dados Classificação e as suas respectivas técnicas: Máquina de Vetor Suporte, *Naives Bayes* e *K-Nearest Neighbor*.

Na seção 3 é apresentado o Desenvolvimento no qual são explicados os tópicos: Trabalhos Relacionados, Rede social adequada para a coleta de dados, o Levantamento de domínios de dados para o *Twitter* e a escolha do domínio designado para essa pesquisa, a realização do processo KDD, a aplicação das técnicas de Mineração de Dados utilizando a ferramenta *Weka* e a análise dos resultados. Na seção 4 é demonstrado os possíveis trabalhos futuros para a continuidade desse trabalho e na seção 5 tem-se as considerações finais.

2. FUNDAMENTAÇÃO TEÓRICA

Esse tópico apresentará conceitos sobre a definição e as características das redes sociais online e como a sua expansão está proporcionando a geração de grandes quantidades de dados. Também será apresentado o processo KDD e quais são as etapas que constituem esse processo, o conceito de Mineração de Opiniões e as suas tarefas, as métricas de avaliação demonstrando conceitos de fórmulas matemática que auxiliam na classificação de opiniões, a tarefa *Data Mining* Classificação e suas principais técnicas de Mineração de Dados (SVM, KNN e *Naives Bayes*).

2.1 REDES SOCIAIS ONLINE

Redes sociais são aplicações que servem para manter os relacionamentos que não podem ser conservados por causa das distâncias e também para as pessoas fortalecerem os seus contatos profissionais e expandir o seu *networking*[1].

Segundo Recuero (2009, p. 24) uma rede social online: "é definida como um conjunto de dois elementos: atores (pessoas, instituições ou grupos; os nós da rede) e suas conexões (interações ou laços sociais)".

> Rede social é uma estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, que partilham valores e objetivos comuns. Muito embora um dos princípios da rede seja sua abertura, por ser uma ligação social, a conexão fundamental entre as pessoas se dá através da identidade. As redes sociais online podem operar em diferentes níveis, como, por exemplo, redes de relacionamentos [...], redes profissionais (*LinkedIn*) [...] Um ponto em comum dentre os diversos tipos de rede social é o compartilhamento de informações, conhecimentos, interesses e esforços em busca de objetivos comuns. As redes sociais costumam reunir uma motivação comum, porém podem se manifestar de diferentes formas (HALT, 2014, p.1).

Enfim, as redes sociais são formadas por usuários conectados nas redes de computadores por algum motivo, França et al. (2014. p.10) complementa que as redes sociais online: "são fruto do processo de socialização da informação nos últimos anos representado pela extensão do diálogo e do modo como as informações passaram a ser organizadas através da Web.".

Recuero (2009, p.24) afirma ainda que:

> O advento da Internet trouxe diversas mudanças para a sociedade. Entre essas mudanças, temos algumas fundamentais. [...] a possibilidade de expressão e sociabilização através das ferramentas de comunicação mediada pelo computador [...]. Essas ferramentas proporcionaram, assim, que atores pudessem construir-se, interagir e comunicar com outros atores, deixando, na rede de computadores, rastros que permitem o reconhecimento dos padrões de suas conexões e a visualização de suas redes sociais através desses rastros.

[1] Uma expressão que representa uma rede de contatos profissionais. (VUELMA, 2011).

As redes sociais são formadas por atores e conexões, os atores são os que compõem a rede social, por exemplo: pessoas, instituições e organizações, Recuero (2009, p.25) afirma que os atores: “são o primeiro elemento da rede social, representados pelos nós (ou nodos). [...] atuam de forma a moldar as estruturas sociais, através da interação e da constituição de laços sociais”.

E as conexões são a forma de comunicação e laços sociais que acontecem na rede social, Recuero (2009, p.30) diz que: “as conexões em uma rede social são constituídas dos laços sociais, que, por sua vez, são formados através da interação social entre os atores”.

Segundo Kietzmann et al. (2011) as caraterísticas das redes sociais online são:

a) Identidade: representa a maneira que os usuários revelam as suas identidades nas mídias sociais;
b) Conversas: caracteriza a comunicação entre os usuários em um ambiente virtual;
c) Compartilhamento: representa as redes sociais onde os seus usuários compartilham alguma informação entre si, por exemplo a *Wikipédia*;
d) Presença: demonstra quando outros usuários percebem que alguém está on-line na sua rede social;
e) Relações: constitui os relacionamentos mútuo entre os usuários;
f) Reputação: corresponde a possibilidade dos usuários verificar a sua popularidade nas redes sociais e também a popularidade de outros usuários, através de “curtidas”, *likes*, compartilhamento de fotos e vídeos;
g) Grupos: retrata as comunidades e grupo que podem ser formados nas mídias sociais.

Visto que essas características das mídias sociais proporcionaram a sua expansão, as redes sociais tornaram-se um lugar onde todos podem participar compartilhando ideias e opiniões e obtendo informações.

França et al. (2014. p. 10) afirma que as mídias sociais: “deram espaço para que usuários gerassem e compartilhassem conteúdo de forma expressiva, [...], caracterizando uma forma de democratização na geração de conteúdo.”.

Esse crescimento das mídias provocou uma mudança no paradigma das relações entre as pessoas. As redes sociais mais famosas normalmente exibem um grande número de contas de usuário ativas e um envolvimento forte dos seus usuários. Esse fato ocasiona enormes volumes de conteúdo e dados. França et al. (2014. p. 10) comenta que: “a popularidade dessas plataformas

pode ser evidenciada através da capacidade que possuem de produzir enormes volumes de conteúdo".

2.2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

O processo para transformar dados brutos em conhecimento é chamado de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases-KDD*), esse processo possibilita transformar esses dados brutos em algum conhecimento.

Galvão e Marin (2008, p. 687) afirmam que: "a descoberta de conhecimento em bases de dados pode ser definida como o processo de extração de informação a partir de dados registrados numa base de dados, um conhecimento implícito, previamente desconhecido, potencialmente útil e compreensível".

Outro aspecto levantado por Silva et al. (2016, p.11) que o objetivo do processo KDD é: "encontrar padrões intrínsecos aos dados [...], apresentando-os de forma a facilitar sua assimilação como conhecimento[...], tal descoberta está associada a um processo analítico, sistemático [...] onde possível, automatizado".

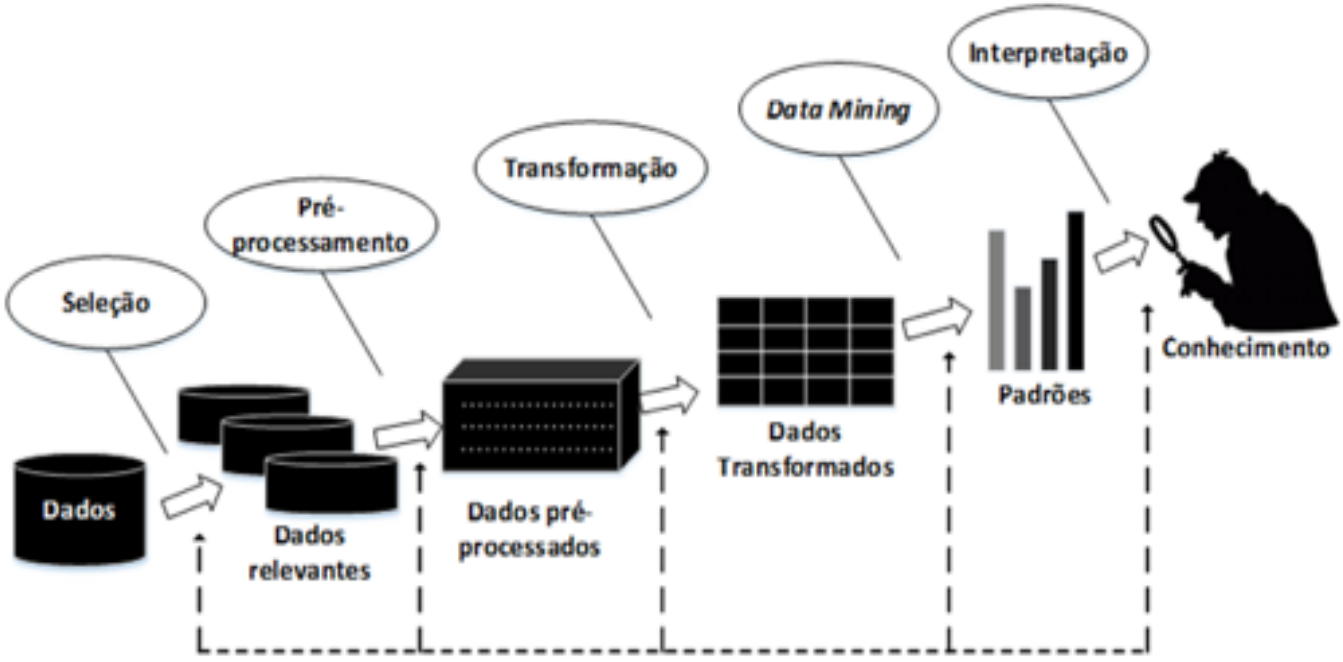
Porém esse processo não é uma atividade simples de se realizar, conforme Praz (2018, p.1) o processo KDD: "não é trivial já que alguma técnica de busca ou inferência é envolvida, ou seja, não é apenas um processo de computação direta".

Galvão e Marin (2008, p.688) complementa a ideia de que o processo KDD não é uma tarefa fácil de se realizar pois envolve vários conceitos da computação: "o processo de KDD utiliza conceitos de base de dados, métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial".

O processo de transformação dos dados em conhecimento contém uma série de etapas, Castanheira (2008, p.15) destaca que: "o processo KDD é um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados".

Essas atividades são: base de dados, seleção dos dados, preparação / pré-processamento e limpeza, transformação, mineração de dados *(Data Mining)* e interpretação / avaliação, conforme visualizado na Figura 01.

Figura 01 - Fases de um processo de descoberta de conhecimento em bases de dados



Fonte: Adaptado de Fayyad *et al.* (1996)

A primeira etapa é a Base de Dados, são nas Base de Dados que são retiradas as informações para obter o conhecimento desejado, Castro e Ferrari (2016, p.5), conceitua que "os dados podem ser entendidos como o nível mais básico de abstração a partir do qual a informação e, depois, os conhecimentos podem ser extraídos".

Segunda etapa desse processo é a fase de Seleção de Dados, é a etapa em que os conjuntos de dados são identificado e selecionados na base de dados, no dizer de Goldschmidt et al. (2015, p.23): "esta função, também denominada Redução de Dados, compreende, em essência, a identificação do subconjunto das bases de dados existentes que deve ser efetivamente considerado durante o processo de KDD".

A Preparação ou Pré Processamento de dados é o momento em que os dados são preparados para utilizar às técnicas de mineração de dados, Goldschmidt et al. (2015, p.23) exemplifica que "esta etapa tem como objetivo a preparação dos dados para os algoritmos que serão aplicados na etapa de mineração de dados".

Próxima etapa é a Limpeza dos Dados, em que é realizado o tratamento nos dados para ter corretude dos dados coletados, é a fase em que os dados não corretos são excluídos, Castro e Ferrari (2016, p36) explanam:" a limpeza dos dados atua no sentido de imputar valores ausentes, suavizar ruídos, identificar valores discrepantes (*outliers*) e corrigir inconsistências".

Transformação do Dados, os dados são formatados, para que os algoritmos de *Data Mining*, possam ser aplicados. Na opinião de Castanheira (2008, p.16) "para facilitar o uso de técnicas de mineração de dados, os dados ainda podem passar por uma transformação que os armazena adequadamente em arquivos para serem lidos pelos algoritmos".

Mineração de Dados, ocorre o processamento dos dados, para que se possa identificar as informações importantes nas Bases de Dados, é nesta etapa em que os algoritmos e as técnicas de mineração de dados são aplicados, Silva et al. (2016, p.11) reforça que a Mineração de Dados:

> [...]é definida em termos de esforços para descoberta de padrões em bases de dados. A partir dos padrões descobertos, têm – se condições de gerar conhecimento útil para um processo de tomada de decisão. Trata–se, portanto, da aplicação de técnicas, implementadas por meio de algoritmos computacionais, capazes de receber, como entrada, um conjunto de fatos ocorridos no mundo real e devolver, como saída, um padrão de comportamento.

Avaliação ou Validação de conhecimento, são os resultados encontrados no processo da Mineração de Dados, essa última etapa se preocupa em verificar se algum conhecimento útil foi descoberto, mostrando a importância dos dados descobertos. Essa etapa tem a finalidade de observar os conhecimentos obtidos são: verdadeiros, úteis e não triviais (CASTRO; FERRARI, 2016).

2.3 MINERAÇÃO DE OPINIÕES

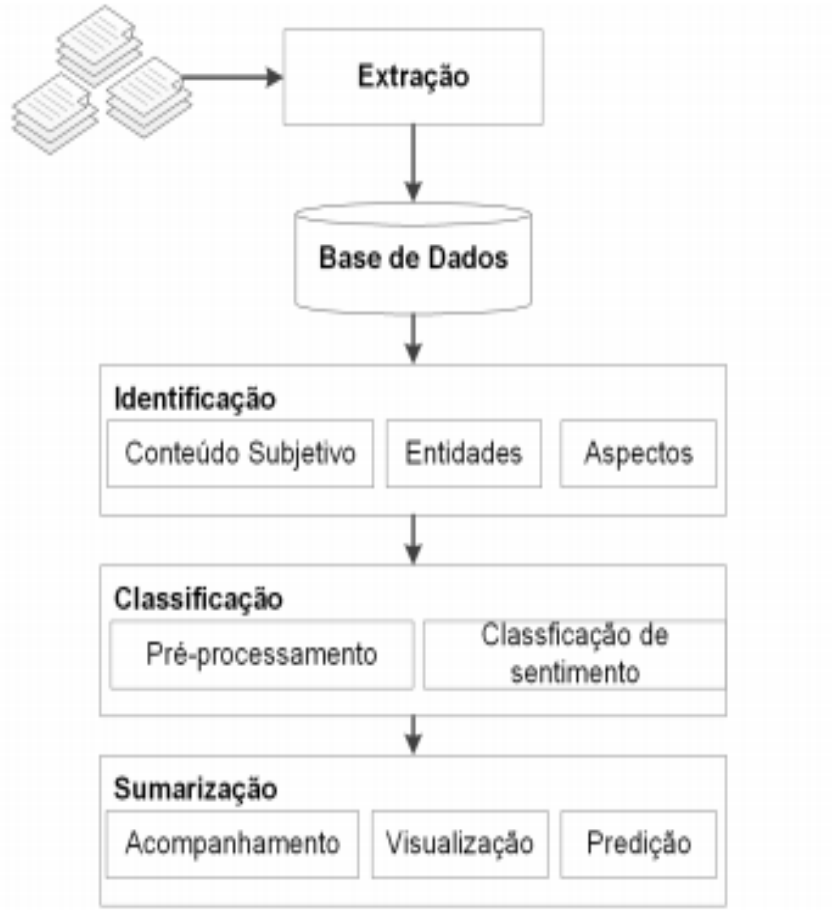
Opiniões são quando indivíduos demostram as suas ideias, crenças, avaliações sobre algum assunto específico, dispondo de um impacto considerável para orientações de pessoas no processo de tomada de decisão através das opiniões (KHAN, 2014).

A mineração de opiniões, também chamada de análise de sentimentos, se preocupa em identificar as opiniões expressadas, permitindo gerar padrões de dados que sejam interessantes e que agreguem algum conhecimento.

Santos (2012, p.30) enfatiza que: “a Análise de Sentimentos ou Mineração de Opinião, que está inserida no tópico de análise de subjetividade, corresponde ao problema de identificar (ou extrair) emoções, opiniões ou pontos de vista em textos”.

Segundo Becker e Tumitan (2013), a mineração de opinião pode ser caracterizada em termos de três tarefas, conforme ilustrado na Figura 02.

FIGURA 02 - TAREFAS DE MINERAÇÃO DE OPINIÃO



Fonte: (BECKER; TUMITAN,2013, p.7).

A Figura 02 apresenta as três tarefas de mineração de opinião que são: Identificação, Classificação e Sumarização. A tarefa de Identificação consiste em identificar conjuntos de dados extraído de alguma fonte de dados, a Classificação é a etapa que classifica a polaridade do sentimento em duas classes: positivos e negativos e a Sumarização é a criação de métricas e sumários para quantificar a diversidade de opiniões encontradas.

Uma abordagem que a classificação de opiniões manuseia é baseada em aprendizado de máquinas, que consiste em descobrir automaticamente as regras gerais em grandes conjuntos de dados. Segundo Silva et.al (2016, p.77) a abordagem de aprendizado de máquinas é: “um processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados [...]”.

Portanto, a abordagem fundamentada no aprendizado de máquina, permite procurar por padrões de informações dentro de um conjunto de dados e posteriormente, pode-se tomar decisões ou recomendar respostas com base no que se descobriu (GEITEY,2016).

2.4 MÉTRICAS DE AVALIAÇÃO

As métricas de avaliação são manuseadas para averiguar o modelo de dados utilizados para treinamento de dados e verificar se eles proporcionam bons resultados no momento da classificação dos dados (SANTANA, 2017).

Existem funções matemáticas que auxiliam a mensurar a capacidade de erro e acerto dos modelos de dados (FILHO, 2017), serão apresentadas as seguintes funções matemáticas: *Precision* (Precisão), *Recall*, F-*measure* e *Matthews Correlation Coefficient* (MCC).

Onde: *True Positives* (TP) são os valores classificados como verdadeiramente positivos, pelo classificador; *True Negatives* (TN): são instâncias negativas, que foram rotuladas corretamente pelo classificador; *False Positives* (FP): são os falsos positivos, são os dados classificados erroneamente como positivos pelo classificador; *Falses Negatives* (FN): são instâncias positivas, que foram classificadas incorretamente como negativas pelo classificador;

A *Precision* é o valor da predição positiva (número de casos positivos por total de instâncias):

$$Precision = \frac{\mathrm{TP}}{(TP + FP)}$$

E o *Recall* é considerado como uma medida de completude:

$$Recall = \frac{TP}{FN + TP}$$

O MCC leva em consideração os verdadeiros e falsos positivos e negativos e é geralmente considerado como uma medida equilibrada que pode ser usada mesmo em que as classes sejam de tamanhos muito diferentes:

$$\mathbf{MCC} = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}}$$

A F-*measure* é utilizada para medir o desempenho, combinando os valores de precisão e recall em um única fórmula:

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

2.5 TAREFA DE MINERAÇÃO DE DADOS

A principal tarefa de Mineração de Dados no domínio de análise das opiniões é a Classificação, segundo Becker e Tumitan (2013, p12): “na área de mineração de opiniões, nota-se um predomínio do uso de métodos supervisionados de aprendizagem, mais especificamente, classificação [...]”.

O método supervisionado é a técnica na qual o algoritmo de aprendizado recebe um conjunto de dados, que definem aquilo que deverá ser buscado pelo algoritmo, conforme Castro e Ferrari (2016, p.16) “aprendizado supervisionado é baseado em um conjunto de objetos para os quais as saídas desejadas são conhecidas, ou em algum outro tipo de informação que represente o comportamento que deve ser apresentado [...]”.

A classificação visa observar um conjunto de dados fornecidos, onde cada conjunto dos dados contém uma catalogação de qual classe ela pertence, por conseguinte a classificação utiliza dados já catalogados com a finalidade de aprender como classificar os novos conjuntos de dados apresentado para a tarefa.

Silva et. al (2016, p.79) enfatizam que a Classificação é: “o processo pelo qual se determina um mapeamento capaz de indicar a qual classe pertence qualquer exemplar de um domínio sob análise, com base em um conjunto de dados já classificados”.

2.6 TÉCNICAS DE MINERAÇÃO DE DADOS

As técnicas de aprendizado de máquina compreendem descobrir automaticamente as regras gerais em grandes conjuntos de dados, segundo Castro e Ferrari (2016, p.14) “aprendizado de máquina é a área que visa desenvolver programas computacionais capazes de automaticamente melhorar seu desempenho por meio de experiência”. Nos próximos tópicos serão apresentados as principais técnicas de Mineração de Dados.

2.6.1 Máquina de Vetor de Suporte (*Support Vector Machines*-SVM*)*

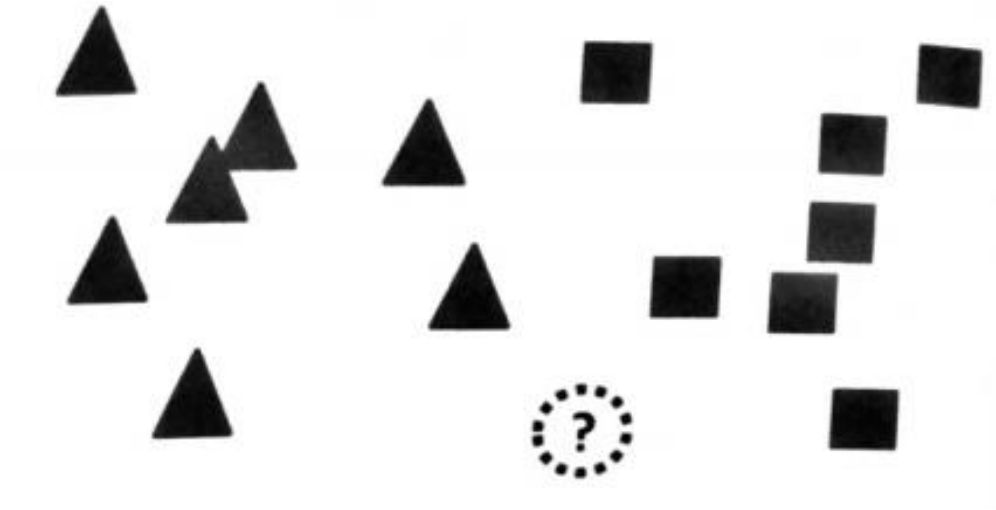
A técnica de Máquina de Vetor de Suporte é uma técnica de aprendizado, baseada na teoria de aprendizado estatístico, é aplicada para descrever a classificação através de vetores. Essa técnica funciona utilizando um conjunto de dados para treinamento, cada conjunto recebe uma categoria, a partir dessa categoria é possível categorizar os futuros conjuntos de dados, essa técnica permite analisar os dados e reconhecer padrões.

Lorena e Carvalho (2007, p.41) sustenta que: “essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu”.

A técnica de SVM encontra uma linha de separação, chamada de hiperplano (um conceito em geometria, ele é a generalização do plano em diferentes números de dimensões) entre os dados de duas classes, essa linha procura maximizar a distância entre os pontos mais próximos em relação a cada uma das classe (ZUBEN; ATTUX, 2010).

Um exemplo citado por Amaral (2016) sobre a técnica SVM, é demonstrado na Figura 03, essa ilustração apresenta triângulos e retângulos que são representados por duas classes de dados, enquanto a forma de uma interrogação no centro do plano é o dado alvo o qual se deseja classificar.

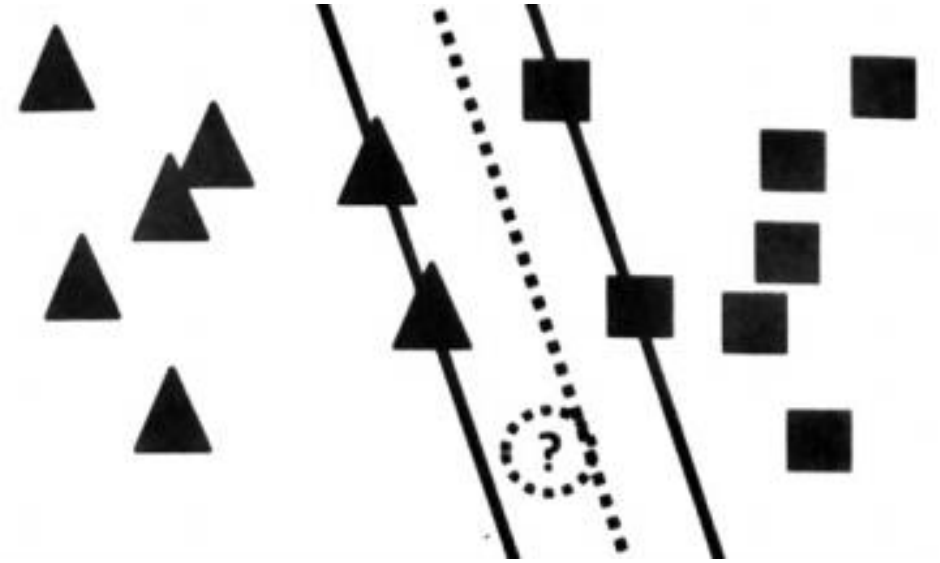
FIGURA 03 - ELEMENTOS DE UM PLANO



Fonte: Adaptado de Amaral (2016).

A Figura 04 são apresentados dois vetores não pontilhados, esses vetores são as margens otimizadas, as instâncias por onde as margens otimizadas percorrem são os vetores de suporte, e o vetor pontilhado é a referência para a classificar as novas instâncias, logo, a forma de uma interrogação é classificada como um triângulo.

FIGURA 04 – VETORES DE SUPORTE



Fonte: Adaptado de Amaral (2016).

2.6.2 Naive bayes

A técnica de *Naives Bayes* é um algoritmo de classificação que é baseado no Teorema de *Bayes*, Silva (2016) afirma que o Teorema de *Bayes* é uma ferramenta da estatística, esse teorema é uma fórmula matemática utilizada para cálculos de probabilidades condicionais, descrevendo a probabilidade de um evento com base no conhecimento prévio das condições que podem estar relacionadas com o evento:

$$P(c\,|x) \;=\; \frac{P(X\,|c)\;P(c)}{P(x)}$$

$$P(c\,|X) = \;P(x_1\,|c)\;\times P(x_2\,|c)\;\times\cdots\times P(x_n\,|c\,)\;\times P(c)$$

Onde: P (c | x) é a probabilidade posterior da classe alvo, P (c) é a probabilidade original da classe, P (X | c) é a possibilidade de que a probabilidade da classe preditora seja dada, P (x) é a probabilidade original do preditor.

O *Naives Bayes* é utilizado para a modelagem de previsões exploratórias, sendo uma técnica para construir classificadores, esses classificadores atribuem rótulos de classes para as instâncias de algum problema, esses rótulos são extraídos de algum conjunto de dados, Amaral (2016, p.41) complementa que:

> [...] é um algoritmo *bayesiano*, baseado na teoria das probabilidades e que supõe que os atributos vão influenciar a classe de forma independente. Na criação do modelo, este classificador vai construir uma tabela mostrando o quanto cada categoria de cada atributos contribui para cada classe. Uma vez montado o modelo, ao submetermos uma nova instância para o classificador, ele vai olhar os pesos nesta tabela, somá–los e ver qual classe teve um peso maior, que sairá como "vitorioso".

Portanto, *Naives Bayes* é um classificador baseado na estatística, que tem o propósito de classificar uma determinada classe, baseado na probabilidade do objeto de pertencer a determinada classe ou não.

2.6.3 K–*Nearest Neighbor*(KNN)

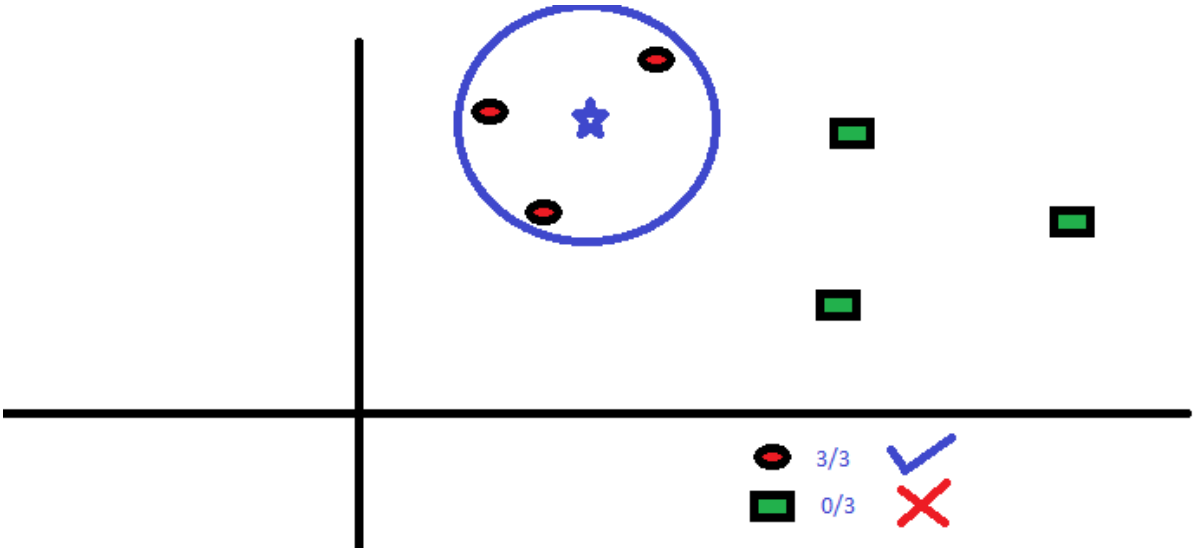
K–Nearest Neighbor (KNN) é uma técnica utilizada para classificar objetos em base de treinamento, essa classificação acontece quando os dados do treinamento estão mais próximos em caraterísticas, para trabalhar com o KNN é preciso: conjuntos de dados como exemplo para

treinamento, definir uma métrica para cálculo, definir o valor de K[2] (GOLDSCHMIDT et. al, 2015).

Essa técnica utiliza uma classe desconhecida para realizar a comparação dos exemplares, o fundamento dessa técnica é o de estocar um conjunto de treinamento e realizar comparações entre os exemplo de teste e o exemplares estocados (SILVA et. al, 2016).

Ray (2015) cita um exemplo que a técnica KNN pode ser mapeada em nossas vidas reais, por exemplo, se quisermos aprender sobre uma pessoa a qual não temos nenhuma informação, poderemos encontrar essa informação com amigos próximos dessas pessoas, a Figura 05 expõe essa técnica.

FIGURA 05 - VIZINHOS MAIS PRÓXIMOS (*K-NEAREST NEIGHBOR)*



Fonte: Ray (2015).

Na Figura 05 o valor de K=3, portanto é realizado um círculo onde a classe azul esteja no centro, no qual são incluídos três pontos. Os pontos mais próximos da classe azul são as classes com círculos vermelhos, assim pode-se dizer que essa classe pertence as classes de círculos vermelhos.

[2] O número de vizinhos próximos (GOLDSCHMIDT et. al, 2015).

3. DESENVOLVIMENTO

Nesse tópico irá apresentar os Trabalhos Relacionados, a rede social apropriada para a coleta de dados e os motivos que levaram a sua escolha, demonstrará o levantamento de domínios para a coleta de dados e o domínio elegido para a coleta das informações, também apresentará o processo KDD e todas as etapas que compõem esse processo e a aplicação das principais técnicas de mineração de dados (SVM, KNN e *Naives Bayes*) utilizando a ferramenta *Weka* e os resultados obtidos com a utilização das técnicas de *Data Mining*.

3.1 TRABALHOS RELACIONADOS

A partir de pesquisas realizadas, aferiu-se alguns trabalhos relacionados na análise de opiniões no *Twitter,* abrangendo problemas semelhantes com esse estudo, porém realizará uma análise das técnicas (SVM, KNN e *Naives Bayes*), nesse trabalho.

Santos (2014) teve-se como objetivo, classificar os *tweets* e minerar a opinião da viralidade de eventos educacionais, para o entendimento de como funciona esses tipos de dados postados nessa rede social.

Nessa pesquisa o algoritmo SVM se destacou em relação as métricas de avaliação para a classificação de opinião e como resultados pode-se observar que os seminários são os eventos mais propagados na rede social analisada.

Santos (2016) analisou o sentimento das opiniões sobre a *Black Friday* na rede social *Twitter*, quanto as suas respectivas polaridades foi utilizado as técnicas de Mineração de Texto, Análise de Sentimentos e o Processamento de Linguagem Natural para a extração das informações relevantes.

O trabalho utilizou o algoritmo de *Naïve Bayes* para a análise dos sentimentos, a solução encontrada por Santos (2016), foi capaz de identificar opiniões positivas e negativas sobre a *Black Friday*.

Filho (2014) demostrou como o processo de mineração de textos foi utilizado para coletar, estruturar o texto extraído do *Twitter* e como criar um modelo de classificação de texto utilizando o algoritmo *Naives Bayes*, possibilitando mapear a opinião da rede social dos usuários do *Twitter* sobre Copa do Mundo da FIFA Brasil 2014, no qual predominou o sentimento negativo.

3.2 REDE SOCIAL ADEQUADA PARA A COLETA DE DADOS

O portal de estatísticas *Statista* é um ambiente online de pesquisas estatísticas e *business intelligence*, que propicia o acesso aos dados de instituições de pesquisa de mercado e opiniões, bem como de organizações empresariais (*STATISTA*, 2017).

Esse portal realizou uma verificação para descobrir quais são as redes sociais mais famosas atualmente, a pesquisa elencou as principais redes sociais em números de contas ativas, que foram as seguintes: *Facebook*, *Instagram*, *Twitter* e *Linkedin* (*STATISTA*, 2017).

O *Facebook* não foi selecionado para essa pesquisa, pois apesar do grande número de contas ativas que essa rede social contém, as grandes menções que circulam nos perfis ficam exclusivos ao *Facebook Topic Data*, que é um serviço pago da ferramenta, o que restringe o acesso às publicações e demonstra apenas os dados estáveis (DATASIFT, 2017).

A rede social *Instagram* não foi estabelecida para essa pesquisa, em razão de que o serviço de coleta de dados dessa rede social possibilita o monitoramento de imagens, sendo dependente de legendas e *hashtags* [3](INSTAGRAM, 2017). E o propósito dessa pesquisa não é sobre a investigação de imagens e *hashtags* em uma rede social, mas de opiniões que os usuários mencionam em suas redes sociais.

O *Linkedin*, dissemelhante do *Instagram* e *Facebook* possibilita uma API para a extração dos seus dados (LINKEDIN, 2017), conquanto o *Linkedin* não se configura no intuito desse trabalho, pois o ideal é uma rede social em que os seus usuários expõem as suas opiniões sobre os mais variados assuntos e não somente sobre os seus *networking* profissional e o mundo dos negócios.

Apesar das contas ativas do *Twitter* serem menores que de outras redes sociais elencadas pelo Portal de Estatísticas *Statista*, foi possível compreender que essa rede social é factível para atingir os objetivos desse trabalho, em razão de que os recursos que o *Twitter* oferece para a coleta dos dados são *Open Source* (TWITTER, 2017).

E além disso os usuários dessa rede social se expressam claramente através de mensagens curtas e objetivas, no qual podemos perceber as opiniões nas mais diversas áreas, os dados do

[3] Um termo utilizado para pessoas identificar temas e assuntos nas redes sociais (DRUBSCKY, 2015)

Twitter são particularmente interessantes porque os *tweets* acontecem na velocidade em que as pessoas expressam os seus pensamentos e esses dados estão disponíveis para consumo, que ocorrem em tempo quase real, os *tweets* ligam as pessoas de uma variedade de maneiras, variando de diálogos conversacionais curtos, porém muitas das vezes significativos (RUSSELL, 2014).

Portanto, foi possível compreender que, para esse estudo, a rede social mais propícia para realizar a coleta de dados é o *Twitter*, por motivos de disponibilidades de recursos que essa mídia social oferece, além disso, essa rede social é uma fonte de dados muito ampla podendo realizar busca de dados sobre qualquer tópico.

3.3 LEVANTAMENTO DE DOMÍNIOS DE DADOS PARA O *TWITTER*

Segundo Rodrigues (2013), algumas das aplicações práticas para aplicação da mineração de opinião são sobre: produtos, empresas na bolsa de valores e análise de popularidade de pessoas, Drum (2016), menciona que pode-se utilizar da mineração de opinião na área política, subsequente Becker e Tumitan (2013) menciona o manuseio da análise de opinião no setor de filmes.

Em razão disso, ocorreu os seguintes levantamentos dos possíveis domínios para a coleta de dados no *Twitter*: produtos, empresas na bolsas de valores, investigação de popularidades de pessoas, filmes e na área política.

Todavia o domínio de pesquisa designado foi a temática política, pela motivação da crise política que está ocorrendo no Brasil. A temática para realizar a coleta dos dados se baseará no seguinte acontecimento: a Câmara dos Deputados rejeita a denúncia da Procuradoria Geral da República, por crime de corrupção passiva contra o presidente Michel Temer e o livra de responder ao processo no Supremo Tribunal Federal (SENADO NOTICIAS, 2017).

O objetivo será analisar as opiniões dos brasileiros, através das principais técnicas de *Data Mining* e analisar os sentimentos da população brasileira a respeito desse episódio e examinar quais são as palavras que se ressaltaram. Este tópico será capaz de fornecer uma base de dados robusta, para a investigação e a classificação das opiniões citadas no *Twitter* desse acontecimento político.

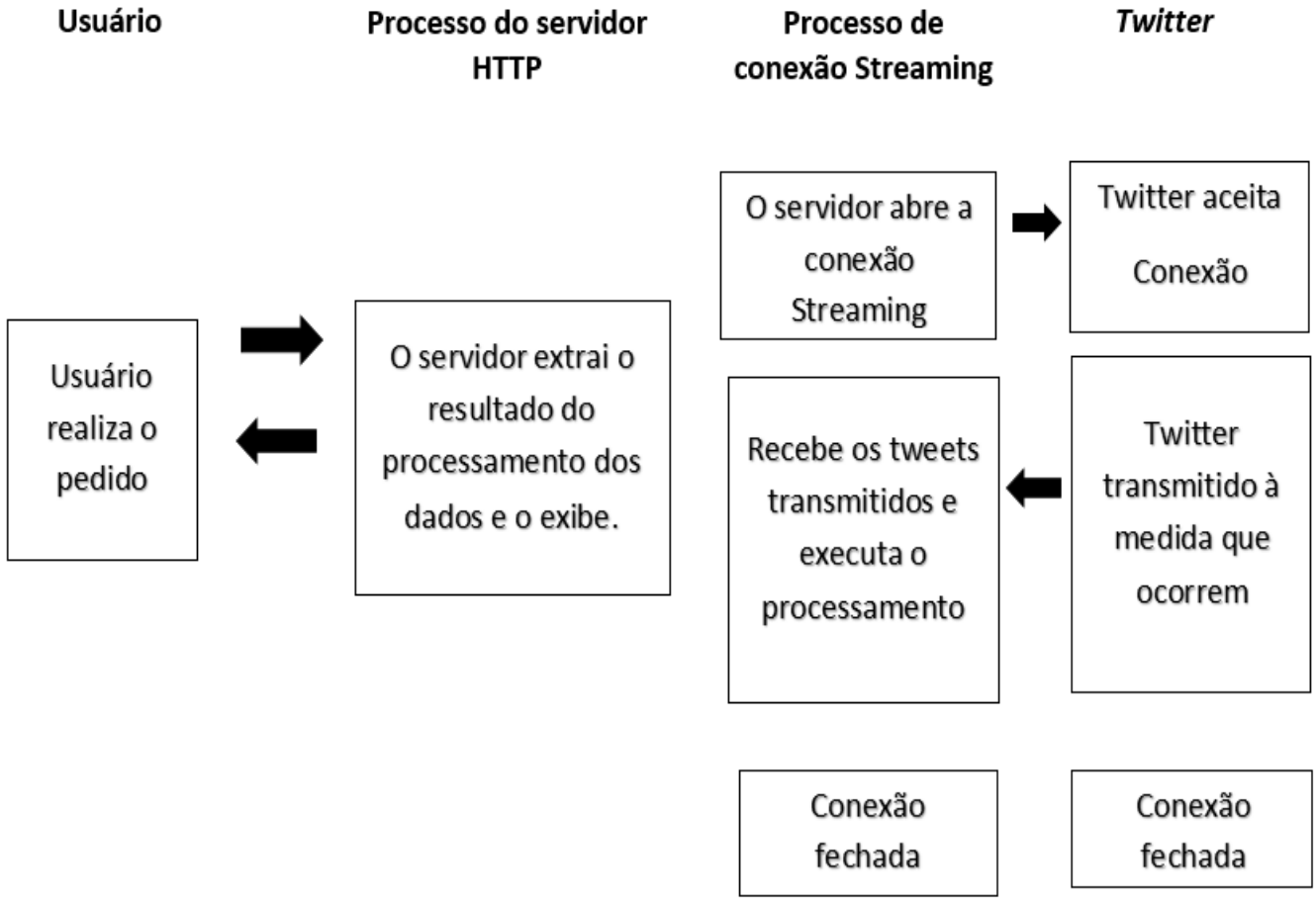
3.4 ETAPAS EXECUTADAS DO KDD

Segundo Praz (2017), o processo KDD apresenta toda a etapa que o dado percorre até se transformar em informação, nesse trabalho realizou-se a base de dados, a seleção dos dados que abrange a coleta dos dados do domínio elegido, a seleção dos dados que serão utilizados para esse estudo e a fase da limpeza dos dados que são removidos qualquer conteúdo indesejado.

A coleta dos dados do domínio delineado, foi realizada no período de 03 de Agosto de 2017 a 12 de Agosto de 2017, foram coletados cerca de 100.000 mil *tweets*, a extração foi efetuada por instrumento da API Streaming que a mídia social *Twitter* propicia.

A *Streaming* API fornece os tweets em tempo real e uma conexão por um longo período de tempo, o código fonte para manter a conexão do fluxo é normalmente executado em um processo separado do processo que trata as solicitações HTTP, conforme demonstrado na Figura 06.

Figura 06 - Fluxo dos Dados



Fonte: Elaborada pela autora (2017)

Conforme demonstrado na Figura 06 é necessário uma conexão ativa e persistente entre um servidor e o *Twitter*, no momento em que o tuites é realizado, o *Twitter* notifica o servidor em tempo real permitindo guardar os dados numa base de dados (TWITTER, 2017).

O processo que a API streaming recebe os tuites de entrada, executa qualquer filtragem necessária para o armazenamento dos dados, o processo de tratamento que o HTTP realiza é a consulta e o armazenamento de dados. Com isso possibilita obter os resultados das solicitações realizadas.

Para realizar a coleta desses dados foi fundamental cumprir-se as seguintes atividades: a criação de uma conta ativa no *Twitter*, a elaboração de um aplicativo no *Twitter Apps* para autenticação nessa rede social e o desenvolvimento de um script para se comunicar com API.

A criação desse aplicativo é necessária para ter acesso a base de dados dessa mídia social. O aplicativo será utilizado para transportar os *tweets*, uma vez que o aplicativo é criado, são fornecidos: a chave do cliente, a chave secreta do cliente[4], a chave de *token* de acesso e a chave secreta de acesso, essas chaves são utilizadas para autenticar o usuário quando o mesmo deseja acessar os dados do *Twitter*.

Posteriormente foi elaborado um script implementado na linguagem de programação *Python* para a coleta dos dados, que encontra-se no Anexo A, o script utiliza-se o *Tweepy* que é uma biblioteca *open source Python* que permite que o *Python* se comunique com o *Twitter* e utilize a sua API para coletar os dados (ROESSLEIN, 2017).

Neste script utiliza-se as chaves e os segredos que se obteve na criação do aplicativo, primeiro desenvolveu-se a classe ouvinte que é utilizada para carregar os dados do *twitter*, com a finalidade de coletar os dados, utilizou-se o protocolo *OAuth.*

Segundo Andrey (2017, p.1) o protocolo *OAuth*: “é um protocolo que permite aos usuários ter acesso limitado a recursos de um website sem precisar expor suas credenciais”. Permitindo que o usuário faça *login* em sites de terceiros usando qualquer conta do site da rede social sem expor as suas senhas, *OAuth* fornece segurança e autorização ao usuário.

[4] Por questão de privacidade as chaves fornecidas, por ser de uso pessoal não foram inseridas no script, que encontra-se no Apêndice A.

Os resultados da coleta de dados do *Twitter* são armazenados em uma matriz JSON de objetos contendo os campos dos resultados coletados, consistindo em uma lista de objetos que correspondem aos filtros fornecidos e o encadeamento realizado pela pesquisa.

A Figura 07 consiste na saída da chamada da API pela pesquisa GET onde os parâmetros especificam que a consulta é "Temer".

FIGURA 07 - RESULTADO DA COLETA


```
{
 "cell_type": "code",
 "execution_count": 9,
 "metadata": {
  "collapsed": true
 },
 "outputs": [],
 "source": [
  "auth = OAuthHandler(ckey, csecret)\n",
  "auth.set_access_token(atoken, asecret)"
 ]
},
{
 "cell_type": "code",
 "execution_count": null,
 "metadata": {},
 "outputs": [
  {
   "name": "stdout",
   "output_type": "stream",
   "text": [
    "Ileg\\u00edtimo?\\nVoc\\u00ea votou em Temer. CF \\u00e9 clara em caso de
    impeachment. \\nBolsa Fam\\u00edlia at\\u00e9 quem n\\u00e3o precisa
    recebe. D\\u00e1 f\\u2026 https:\\/\\/t.co\\/
    U5j3Zr1vGX\",\"display_text_range\":[0,140],\"source\":\"\\u003ca
    href=\\\"https:\\/\\/mobile.twitter.com\\\"
    rel=\\\"nofollow\\\"\\u003eTwitter Lite\\u003c\\/a\\u003e\",\"truncated\":t
    rue,\"in_reply_to_status_id\":null,\"in_reply_to_status_id_str\":null,\"in_
```


Fonte: Elaborada pela autora(2017).

Após a execução do script é gerado um arquivo .CSV (Valores Separados por Vírgulas), a utilização desse formato, sucedeu pois os arquivos CSV oportuniza um tempo de leitura e gravação melhor comparados com outros arquivos, os arquivos são armazenados no próprio disco rígido, no qual pode-se ser facilmente realizar a importação para a manipulação posterior.

O propósito da limpeza de dados é remover qualquer conteúdo indesejado dos dados coletados, foram excluídos da base de dados coletadas: pontuações, *hashtag*, nomes dos usuários, *retweets*[5] , Urls .

Foi elaborado um script em Python, que encontra-se no Anexo B, o seu propósito é ler cada *tweet* e limpar os dados indesejáveis, inicialmente na base de dados continha cerca de 100.000 mil *tweets*, com a execução do script, ocorreu uma redução para 31.423 mil *tweets*.

3.5 APLICAÇÃO DAS TÉCNICAS DE MINERAÇÃO DE DADOS

Para a continuidade do trabalho, utilizou-se o *software Weka*, para aplicar as técnicas de mineração de dados (Naives Bayes, KNN e SVM), a ferramenta ofereceu benefícios pela facilidade de uso, contém uma biblioteca de algoritmos de classificação, geração e visualização de relatórios no qual pode-se aplicar as técnicas que serão utilizados nesse trabalho (WEKA, 2017).

Sendo assim foi importado os dados na base do *Weka Explorer*, logo foi possível tratar as informações inseridas e convertê-las para o formato ARFF (Formato de Arquivo de Relação de Atributo), para que os dados possam ser lidos pela ferramenta.

O formato gerado pela ferramenta encontra-se na Figura 08, onde: *@ relation*: é o nome do arquivo definido como a primeira linha do arquivo ARFF; *@attribute*: cada atributo no conjunto de dados tem sua própria declaração que o define de forma exclusiva e o seu tipo de dados onde a especificação pode ser definidas como numéricos, *string* e data; *@data*: são os valores dos atributos que são delimitados por vírgulas.

5 São uma republicação de um *Tweet* (TWITTER,2017).

FIGURA 08 - FORMATO ARFF

```
@relation treinamento

@attribute Texto String
@attribute sentimento {positivo,negativo}

@data
'toda av paulista indignada contra corrupto michel temer gritavam',negativo
'corrupto  temer lavou alma povo incomodou gente poderosa ainda chamou janot corrupto
',negativo
'verdades ocultas congresso corrupto salva temer deixa brasil governado crime ',negativo
'eleitorado desse senhor defende afinco governo corrupto michel temer ',negativo
'corruptos ainda tentam justificar voto ',negativo
'fala sobre apoio corrupto temer',negativo
'falou corrupto temer hoje?! amigo?!',negativo
'acordar envergonhada temer fez brasil acordar impune corrupto',negativo
'politicos corruptos roubam continuam soltos',negativo
'deputados corruptos vendidos salvam temer ',negativo
'todos queremos corrupto temer preso',negativo
'quadrilha corruptos',negativo
```

Fonte: Autora da Pesquisa (2017)

Para prosseguir com este procedimento foi necessário converter os dados que estão no formato de uma *string* para um formato numérico, para esta etapa, utilizou o filtro '*StringToWordVector*'. Esse formato padrão é utilizado para organizar as bases de dados que serão inseridas na base de dados do *Weka*.

O filtro além disso, proporciona remover palavras que não tem um significado emocional, palavras que contém pronomes e artigos, para isso se utilizou um conjunto de *stopwords* [6], que encontra-se no Apêndice C.

Foi imprescindível rotular manualmente uma base de dados com 5000 instâncias, sendo 2500 positivas e 2500 negativos para o treinamento dos algoritmos de *Data Mining (Naives Bayes*, KNN e SVM) utilizados nesse trabalho.

O conjunto de dados que foram classificados fundamenta-se de classes na qual se sabe à qual pertence, por meio de análise manual da base de dados, um exemplo da especificação que foi efetuada é apresentada na Tabela 01.

Tabela 01 - Categorização das Instâncias

Twetts	**Classe**
'votam a favor de temer pela estabilidade e o crescimento'	Positivo
' congresso corrupto salva temer deixa brasil governado no crime’	Negativo

Fonte: Elaborada pela Autora (2017).

Na Tabela 01 evidência o twett com o teor de palavras positivas: ‘favor’, ‘crescimento’, por consequência esse texto foi classificado como positivo, contudo as palavras: ‘corrupto’, ‘crime’, são palavras que contém a acepção negativa, logo foram classificadas como negativas.

A ferramenta *Weka* proporciona métricas de avaliações, para os algoritmos que foram realizados o treinamento dos dados catalogados das 5000 instâncias, os resultados são apresentados na Tabela 02.

6 São palavras consideradas irrelevantes e insignificantes para o conjunto de decorrências a ser exibidos. (PÚBLIO, 2015).

TABELA 02 - MÉTRICAS DE AVALIAÇÃO

Algoritmos	**TP Rate**	**FP Rate**	**Precision**	**RECALL**	**F-Measure**	**MCC**
KNN	**0,927**	0,073	**0,928**	**0,927**	**0,927**	**0,855**
Naives Bayes	0,895	**0,105**	0,898	0,895	0,895	0,793
SVM	0,923	0,077	0,927	0,923	0,923	0,849

Fonte: Elabora pela Autora(2017).

Com essas métricas de avaliação, conforme apresentada na Tabela 02, percebe-se que o Algoritmo KNN e o SVM tiveram valores bem próximos, porém o KNN ressalta nas métricas de classificação das classes.

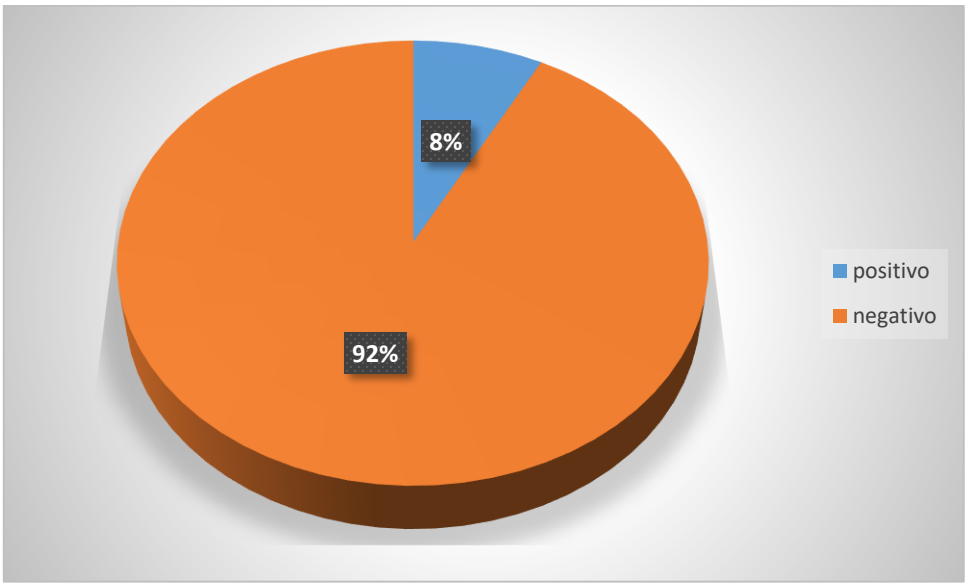
O algoritmo *Naives Bayes* apresentou um resultado inferior comparado com os outros dois algoritmos porém seu resultado também é bem próximo dos demais, isso demonstra que ambas as técnicas apresentou decorrências satisfatórias no processo de classificação de opiniões do domínio na área política.

Os algoritmos utilizarão esses conjuntos de dados de 5000 instâncias catalogadas, para aprender a mapear os exemplos de entrada dos dados desconhecidos (não catalogados) na base de dados do *Weka*, consequentemente os algoritmos irão classificar as novas informações, em positivos e negativos.

Posteriormente, realizou-se a avaliação do restantes dos dados que não foram catalogados, com base nos modelos de dados já treinados, foram apresentado o conjunto de dados testes para as técnicas (KNN, Naives Bayes e SVM), para os modelos de dados realizarem a classificação.

Para o algoritmo SVM foram classificados 92 % dos *twetts* em negativos e 8 % em positivos, conforme pode-se visualizar no gráfico 01.

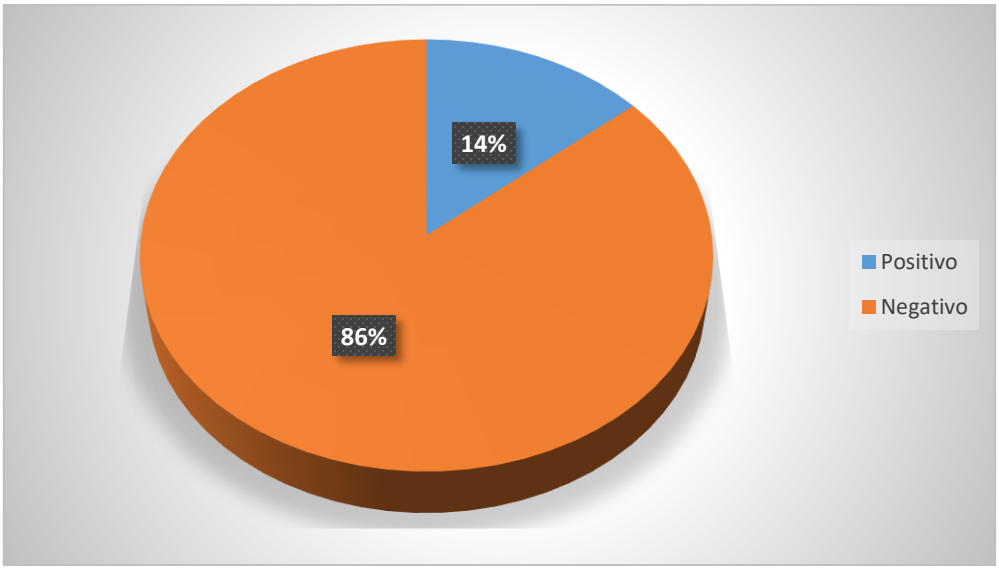
GRÁFICO 01 - SVM



Fonte: Elaborada pela autora(2017)

Para o classificador que utilizou o algoritmo KNN, mostrou-se um resultado de 86% das opiniões foram classificados em negativas e 14% em positivas, conforme pode-se visualizar no Gráfico 02.

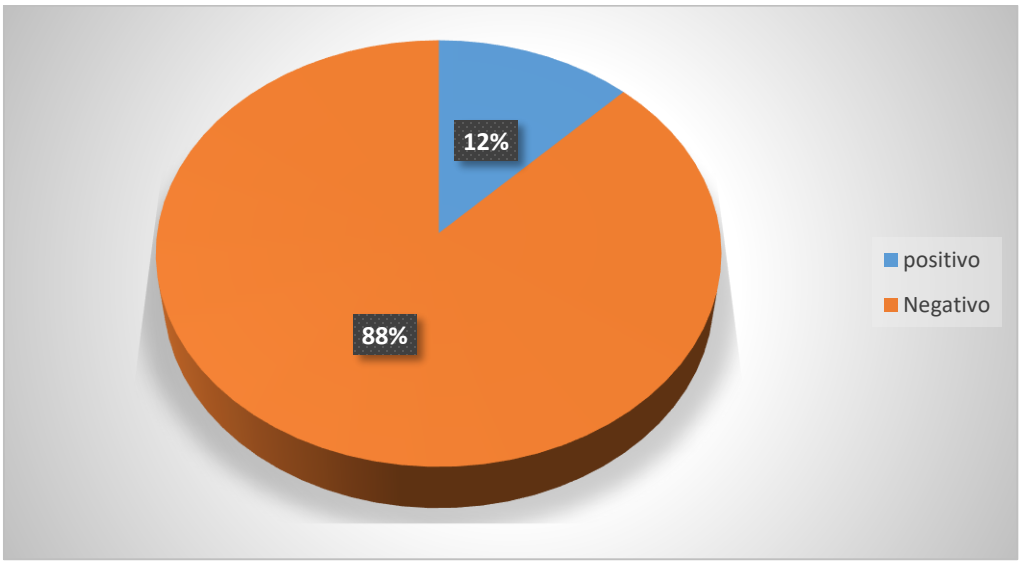
GRÁFICO 02 - KNN



Fonte: Elaborada pela Autora (2017).

O classificador que manuseou o algoritmo *Naives* Bayes, apresentou um resultado de 88% das opiniões classificadas sendo negativas e os 12 % em positivos, conforme ilustrado no Gráfico 03.

GRÁFICO 03 – NAIVES BAYES



Fonte: Elaborada pela Autora(2017).

Para ilustrar as palavras que mais se destacaram, foi desenvolvida a Nuvem de Palavras, segundo Silva (2013, p.1): "Nuvem de palavras, *word cloud* ou *tag cloud* são vários termos utilizados para um tipo de visualização[...] as nuvens de palavras escondem minúcias e fatos interessantes por trás de seu uso corriqueiro e apresentação de dados". Como pode-se perceber na Figura 09, as palavras que mais se destacaram foram palavras de teor negativo.

FIGURA 09 - NUVEM DE PALAVRAS

Fonte: Elaborada pela Autora (2017).

Portanto, percebe-se que a execução das três técnicas (KNN, SVM e *Naives Bayes*), apresentou resultados semelhantes em relação ao sentimento da população brasileira, sobre a absolvição do presidente, ambas demonstraram a carga de sentimento negativas em relação a esse acontecimento, em vista disso percebe-se que a utilização das três técnicas foi benéfico, para comprovar os sentimentos pesquisados.

4. TRABALHOS FUTUROS

Para a continuidade desse trabalho sugere-se algumas atividades: (a) percebe-se que nesse trabalho poderia ter utilizados outros filtros no momento da coleta de dados, não somente a palavra 'Temer', mas outras palavras chaves relacionadas com o tema político; (b) O tratamento de ambiguidade das palavras; (c) Escolha de outros domínios de dados para confrontar com os resultados obtidos nesse trabalho; (d) Aplicação de outros algoritmos de mineração de dados; (e) Aplicação das técnicas de mineração utilizando outras ferramentas de *Machine Learn*.

5. CONSIDERAÇÕES FINAIS

Foi possível avaliar que a análise de opiniões tem um mercado amplo e podem ser aplicadas em várias áreas do conhecimento, trazendo benefícios para empresas, pessoas e a sociedade em geral, pois com a análise pode-se gerar padrões de dados que sejam interessante, sendo possível determinar a emoção por trás das palavras citadas nas mídias sociais, possibilitando ter uma visão geral da opinião pública.

Efetuou-se a investigação sobre as API (*Application Programming Interface*) e os recursos, que as principais redes sociais disponibilizam para a coleta de seus dados, dessa forma pode-se averiguar que a API *Streaming* do *Twitter* é a mais propícia para a coleta de dados desse estudo, pois ela fornece recursos para desenvolvedores ter em acesso ao fluxo global de tuites.

Também aferiu-se que a abordagem em aprendizagem de máquina é uma boa alternativa para a classificação de opiniões, uma vez que, essa abordagem será capaz de descobrir automaticamente as regras gerais em grandes conjuntos de dados. Além disso ocorreu a averiguação das principais tarefas de Mineração de Dados que foi a Classificação e também a

verificação das principais técnicas de Mineração de Dados que são: Máquina de vetor de suporte (SVN), *Naive bayes* e *K–Nearest Neighbor(KNN).*

Foram realizadas as etapas do processo KDD, nesse método realizou a coleta dos dados, do domínio político com base na seguinte notícia: “Absolvição do presidente Michel Temer contra corrupção passiva”, a predileção desse tema foi por ser um tópico atual e ele proporcionou uma boa base de dados para realizar esse estudo.

A coleta foi realizada através de um script em *Python* para comunicar com a API *Streaming* do *Twitter*, a utilização dessa API facilitou muito esse procedimento, tornando factível a coleta dos dados, a etapa de limpeza dos dados foi a mais trabalhosa e que levou mais tempo para se cumprir essa etapa, pelo fato da grande quantidade de dados indesejados que a base de dados continha.

Portanto, com a execução desse trabalho a pergunta indagada: “de que modo o procedimento da execução das principais técnicas de Mineração de Dados, impactariam no processo de classificação e análise de opiniões, em um domínio de dados específico extraído da rede social *Twitter*?”, foi respondida, demonstrando que os impactos da execução das técnicas é benéfico no processo da classificação das polaridades das opiniões.

Com a execução das técnicas de mineração, percebeu-se que a técnica KNN, apresentou melhores resultados utilizando as métricas de avaliação, porém ambas as técnicas apresentaram porcentagens alta, identificando a carga de sentimentos dos brasileiros em relação ao episódio de absolvição do presidente Michel Temer como sendo negativas, desaprovando a decisão da Câmara dos Deputados, portanto compreende-se que a utilizou das três técnicas tornou a análise de opinião mais precisa e valida no processo de classificação das informações, pois ambas apresentaram resultados parecidos no processo de classificação de opinião.

REFERENCIAL BIBLIOGRÁFICO

AMARAL, Fernando. **Aprenda Mineração de Dados Teoria e Prática.** Rio de Janeiro: Alta Books, 2016.

ANDREY, Lucas. **Como funciona o protocolo OAuth 2.0.** Disponível em: < https://imasters.com.br/desenvolvimento/como-funciona-o-protocolo-oauth-2-0/?trace=1519021197&source=single>. Acesso em 21 set. 2017.

BANNISTER, Kristian. **Understanding Sentiment Analysis: What It Is & Why It's Used.** Disponível em: < https://www.brandwatch.com/blog/understanding-sentiment-analysis/ >. Acesso em: 16 jun. 2017.

BATRINCA, Bogdan; TRELEAVEN, Philip C. **Social media analytics: a survey of techniques, tools and platforms.** Disponível em: < https://link.springer.com/article/10.1007/s00146-014-0549-4>. Acesso em: 12 set. 2017.

BECKER, Karin; TUMITAN, Diego. **Introdução à Mineração de Opiniões:** Conceitos, Aplicações e Desafios. In: Simpósio Brasileiro de Banco de Dados, 28., 2013, Recife. Minicursos ... Recife: UFPE, 2013. p. 1-26. Disponível em: < http://inf.ufrgs.br/~kbecker/lib/exe/fetch.php?media=minicursosbbd_versaosubmetida.pdf >. Acesso em 19 jan. 2017.

CALAZANS, Janaina de Holanda Costa; LIMA, Cecília Almeida Rodrigues. **Sociabilidades virtuais: do nascimento da Internet à popularização dos sites de redes sociais online.** Disponível em: < http://www.ufrgs.br/alcar/encontros-nacionais-1/9o-encontro-2013/artigos/gt-historia-da-midia-digital/sociabilidades-virtuais-do-nascimento-da-internet-a-popularizacao-dos-sites-de-redes-sociais-online >. Acesso em: 15 maio.2017.

CASTANHEIRA, Luciana Gomes. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões**. Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal de Minas Gerais, Belo Horizonte, Programa de Pós-Graduação em Engenharia Elétrica da IFMG, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

CASTRO, Leandro Nunes De; FERRARI, Daniel Gomes. **Introdução a Mineração de Dados:** Conceitos Básico, Algoritmos e Aplicações. São Paulo: Saraiva, 2016.

DATASIFT. Disponível em: < http://datasift.com/products/pylon-for-facebook-topic-data/ >. Acesso em 17 mar. 2017.

DONKOR, Ben. **On Social Sentiment and Sentiment Analysis.** Disponível em: < http://brnrd.me/social-sentiment-sentiment-analysis/ >. Acesso em: 16 jun. 2017.

DRUBSCKY, Luiza. **Entenda o que é hashtag (#) para que elas servem e como utilizá-las.** Disponível em: https://marketingdeconteudo.com/o-que-e-hashtag/. Acesso em: 02 nov. 2017.

DRUM, Marluci. **A história do *Twitter*.** Disponível em: < https://www.oficinadanet.com.br/post/16210-a-historia-do-twitter >. Acesso em: 08 abr. 2017.

[Fayyad et al., 1996] Fayyad, U. M.; Piatesky-Shapiro, G.; Smith, P. **Advances in Knowlege Discovery and Data Mining**, MIT Press, Massachusetts, 1996.

FILHO, José Adair. **Mineração de textos: análise de sentimento Utilizando tweets referentes à copa do mundo 2014.** 2014, 44 f. Trabalho de Conclusão de Curso (Graduação) Universidade federal do ceará campus quixadá bacharelado em engenharia de software, QUIXADÁ, 2014.

FILHO, Mario. **As Métricas Mais Populares para Avaliar Modelos de Machine Learning.** Disponível em: < http://mariofilho.com/as-metricas-mais-populares-para-avaliar-modelos-de-machine-learning/ >. Acesso em: 03 nov. 2017.

FRANÇA, et al. **Big Social Data: Princípios sobre Coleta, Tratamento e Análise de Dados Sociais.** In: Simpósio Brasileiro de Banco de Dados, 11, 2014, Curitiba. Minicursos...Curitiba: SBSC.SBBD, 2014. P. 8-45. Disponível em: < http://www.inf.ufpr.br/sbbd-sbsc2014/sbbd/proceedings/artigos/pdfs/127.pdf >. Acesso em: 25 mar. 2017.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima**. Técnica de mineração de dados**: uma revisão da literatura. Revista Scientific Electronic Library Online. [On-line]. Edição 22. São Paulo: FAPESP–BIREME, 2009, setembro e outubro de 2009. Disponível em: < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-21002009000500014 > ISSN 1982-0194

GEITEY, Adam. ***Machine Learning is Fun!*** Disponível em: < https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471 >. Acesso em: 12 maio.2017.

GOLDSCHMIDT et al. ***Data Mining:*** Conceitos, técnicas, algoritmos, orientações e aplicações. 2° ed. Rio de Janeiro: Elsevier, 2015.

HALT, Glauber. **O que são redes sociais?** Disponível em: < https://www.campograndenews.com.br/marketing-pessoal/o-que-sao-redes-sociais >. Acesso em: 02 maio. 2017.

HIRANAKA, Andrea. **Comunidades Online Construindo conhecimento sobre o consumidor de forma ativa, interativa e colaborativa**. In: SILVA, Tarcízio; STABILE, Max. (Org.) Monitoramento e pesquisa em mídias sociais: metodologias, aplicações e inovações. São Paulo: Uva Limão, 2016, p. 145-161.

INSTAGRAM, Disponível em: < https://www.instagram.com/developer/ >. Acesso em: 09 mar.2017.

KHAN et. al. **Mining opinion components from unstructured reviews**: A review. Disponível em: < http://www.sciencedirect.com/science/article/pii/S131915781400010X >. Acesso em 30 abr.2017.

KIETZMANN, et. Al. **Social media? Get serious! Understanding the functional building blocks of social media.** Disponível em: < http://www.sciencedirect.com/science/article/pii/S0007681311000061 >. Acesso em 23 abr. 2017.

LAKATOS, Eva Maria, MARCONI, Marina De Andrade. **Fundamentos de metodologia científica**. 5º ed. São Paulo: Atlas, 2003.

LINKEDIN. Disponível em: < https://developer.linkedin.com/docs/rest-api >. Acesso em: 27 mar. 2017.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. **Uma Introdução às Support Vector Machines.** Disponível em: < http://www.seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67 >. Acesso em 23 abr. 2017.

PRAZ, Fernando Sarturi. **Um visão geral sobre as fases do Knowledge Discovery in Databases (KDD).** Disponível em: < http://fp2.com.br/blog/index.php/2012/um-visao-geral-sobre-fases-kdd/ >. Acesso em: 20 mar. 2017.

PÚBLIO Angelo. **Stop words: você não precisa delas!** Disponível em: < https://angelopublio.com.br/stop-words/ >. Acesso em: 26 out. 2017.

QUEEN, Jennifer. Descobrindo o Brasil que usa Internet. **Revista.br**, São Paulo, n.3, dez. 2016. Disponível em: < http://www.cgi.br/publicacoes/indice/periodicos/ >. Acesso em: 24 fev. 2017.

RAY, Sunil. **Essentials of Machine Learning Algorithms (com Python e R Codes).** Disponível em: < https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/ >. Acesso em: 03 maio. 2017.

RECUERO, Raquel. **Redes sociais na internet.** Porto Alegre: Sulina, 2009.

RODRIGUES et. Al. **Mineração de Opinião / Análise de Sentimentos.** Disponível em: < http://www.inf.ufsc.br/~luis.alvares/INE5644/MineracaoOpiniao.pdf>. Acesso em: 18 jun. 2017.

ROESSLEIN, Joshua. **tweepy Documentation.** Disponível em: < https://media.readthedocs.org/pdf/tweepy/v3.5.0/tweepy.pdf >. Acesso em: 21 set. 2017.

RUSSELL, Matthew A. **Mining the Social Web.** Mary Treseler: Sebastopol, 2014.

SANTANA, Rodrigo. **Café Com Código #09: Entendendo Métricas de Avaliação de Modelos.** Disponível em: < http://minerandodados.com.br/index.php/2017/10/10/cafe-com-codigo-09-metricas-de-avaliacao-de-modelos/>. Acesso em: 03 nov. 2017.

SANTOS, Camila. **Análise da viralidade em eventos Acadêmicos através das redes sociais.** 2014. 44 f. Trabalho de Conclusão de Curso(Graduação) – Universidade Federal Da Bahia Instituto de Matemática Departamento de Ciências da Computação, Salvador, 2014.

SANTOS, Marlon Vieira dos Santos. **Mineração De Opiniões Aplicada a Mídias Sociais**. Dissertação (Mestrado em Ciência da Computação), Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Programa de Pós-Graduação em Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

SANTOS, Natanael. **Redes Sociais História e Guia Completo.** Disponível em: < http://www.natanaeloliveira.com.br/a-historia-das-redes-sociais/ >. Acesso em: 15 maio. 2017.

SANTOS, Wilian Pereira da Silva. **Análise dos Tweets sobre a Black Friday através da Mineração de Texto e Análise de Sentimentos.** 2016. 51 f. Trabalho de Conclusão de

Curso(Graduação) - Universidade Federal do Estado do Rio de Janeiro Centro de Ciências Exatas e Tecnologia Escola de Informática Aplicada, Rio de Janeiro, 2016.

SENADO NOTÍCIAS. Disponível em: < http://www12.senado.leg.br/noticias/audios/2017/08/decisao-da-camara-de-rejeitar-denuncia-contra-temer-repercute-no-senado >. Acesso em 03 ago. 2017.

SILVA et. al. **Introdução à mineração de dados.** Rio de Janeiro: Elsevier, 2016.

SILVA, Tarcízio. **O que se esconde por trás de uma nuvem de palavras?** Disponível em: < http://tarciziosilva.com.br/blog/o-que-se-esconde-por-tras-de-uma-nuvem-de-palavras/>. Acesso em 03 nov. 2017.

STATISTA. Disponível em: < https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ >. Acesso em: 19 mar. 2017.

TWITTER. Disponível em: < https://dev.twitter.com/streaming/overview >. Acesso em: 22 fev. 2017.

VUELMA, Marta. **O que é networking profissional?** Disponível em: < https://martavuelma.wordpress.com/2011/07/22/o-que-e-networking-profissional/ >. Acesso em: 12 maio.2017.

WEKA. Disponível em: < https://www.cs.waikato.ac.nz/ml/weka/ >. Acesso em: 25 out.2017.

ZUBEN, Fernando J. Von; ATTUX, Romis R. F. **Máquinas de Vetores-Suporte.** Disponível em: < ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico8_IA004_1s2010.pdf >. Acesso em: 29 abr. 2017.