

**INSTITUTO FEDERAL DO ESPÍRITO SANTO
SISTEMAS DE INFORMAÇÃO**

MARIA LUIZA DE FREITAS IANNE

**ANÁLISE DE SENTIMENTOS E OPINIÕES POR MEIO DA REDE SOCIAL
TWITTER**

Serra
2019

MARIA LUIZA DE FREITAS IANNE

**ANÁLISE DE SENTIMENTOS E OPINIÕES POR MEIO DA REDE SOCIAL
TWITTER**

Trabalho de Conclusão de Curso apresentado à
Coordenadoria do Curso de Sistemas de
Informação do Instituto Federal do Espírito Santo
como requisito parcial para a obtenção do título de
Bacharel em Sistemas de Informação.

Orientadora: Prof.^a Dra. Kelly Assis de Souza
Gazolli.

Serra
2019

Dados Internacionais de Catalogação na Publicação (CIP)

I11a Ianne, Maria Luiza de Freitas
2019 Análise de sentimentos e opiniões por meio da rede social Twitter
 / Maria Luiza de Freitas Ianne. - 2019.
 52 f. ; il. ; 30 cm

Orientadora: Prof.^a Dra. Kelly Assis de Souza Gazolli.
Monografia (graduação) - Instituto Federal do Espírito Santo,
Coordenadoria de Informática, Curso de Bacharelado em Sistemas
de Informação, 2019.

1. Aprendizado do computador. 2. Sistemas de recuperação da
informação. 3. Mineração de dados (Computação). 4. Twitter (Rede
social on-line). I. Gazolli, Kelly Assis de Souza. II. Instituto Federal do
Espírito Santo. III. Título.

CDD 006

MARIA LUIZA DE FREITAS IANNE

“ANÁLISE DE SENTIMENTOS E OPINIÕES POR MEIO DA REDE SOCIAL TWITTER”

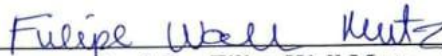
Trabalho de Conclusão de Curso apresentado como parte das atividades para obtenção do título de Bacharel em Sistemas de Informação, do curso de Bacharelado em Sistemas de Informação do Instituto Federal do Espírito Santo.

Aprovado em 16 de Dezembro de 2019.

COMISSÃO EXAMINADORA



Prof. Dr.^a Kelly Assis de Souza Gazolli
Instituto Federal do Espírito Santo
Campus Serra



Prof. Dr. Filipe Wall Mutz
Instituto Federal do Espírito Santo
Campus Serra



Prof. Msc. Jefferson Ribeiro Lima
Instituto Federal do Espírito Santo
Campus Serra

Dedico a todos aqueles que de alguma forma estiveram próximos de mim nesses 4 anos, fazendo esta vida valer cada vez mais a pena.

AGRADECIMENTOS.

Gostaria de agradecer a minha família, por acreditar em mim e não medir esforços para que eu chegasse até esta etapa de minha vida. Aos meus pais e irmãos, por investirem nos meus estudos desde pequena e sempre acreditarem em mim. Pai e mãe, obrigada por toda a paciência que tiveram durante os dias que eu estava ocupada com as tarefas da faculdade. Não posso deixar de agradecer meu irmão, por ter me dado meu primeiro computador e a minha irmã, por ter me dado meu primeiro celular, vocês me apresentaram a tecnologia.

Meus agradecimentos aos meus amigos, minha segunda família, pelas alegrias e tristezas compartilhadas, pelo incentivo e apoio constante. Vocês estavam presentes sempre que eu precisei. Sou grata especialmente à Alice e Ian pelo carinho e atenção, sem vocês, eu não teria chegado até aqui.

Agradeço à instituição de ensino, seu corpo docente e administração, pelo ambiente incrível que proporcionam e pela oportunidade de fazer esse curso. Se não fosse a gratuidade do IFES eu não teria concluído uma graduação.

Aos meus professores e colegas de faculdade, por todo conhecimento compartilhado nesses últimos anos e pela enorme contribuição no meu processo de formação profissional. Vocês terão meus eternos agradecimentos.

Um agradecimento especial à minha professora orientadora, pelo empenho dedicado ao ajudar no desenvolvimento deste trabalho, obrigada pela paciência e compreensão em todos os momentos. Kelly, você representa muito bem as mulheres da área de tecnologia. Sou muito grata por você ter me apresentado o *business intelligence* em suas aulas em 2017 e a mineração de opinião em nossa primeira reunião em 2018. Você me despertou bastante interesse nessas áreas de estudo e me ajudou a finalmente descobrir minha área de atuação profissional.

E agradeço ao universo, por mudar as coisas e nunca as fazer da mesma forma. Se não fosse assim, não teríamos o que pesquisar, o que descobrir e o que fazer.

“Lugar de mulher é onde ela quiser. Elas podem ser astronautas, programadoras, matemáticas ou qualquer outra coisa que sonharem.”

- *Autora*

RESUMO

A opinião é considerada muito importante para o comportamento das pessoas, principalmente para o processo de tomada de decisão em uma empresa. Esse trabalho apresenta a mineração de opinião, outros conceitos desta área da computação e propõe um estudo de caso onde o objeto de estudo é uma empresa chinesa de produtos eletrônicos. Para isso ser possível, foi extraída uma base de dados obtida através da rede social Twitter, visto que as redes sociais se tornaram importantes meios de comunicação com trocas de informação de maneira instantânea. Neste trabalho, foi utilizado um algoritmo de aprendizado de máquina, a máquina de vetores de suporte (SVM), para realizar uma análise de opiniões em textos escritos em português. O objetivo desse trabalho é classificar as opiniões obtidas realizando um processamento dos dados e apresentar sua polaridade de maneira resumida e estatística. Observou-se que as opiniões têm um papel importante na reputação das empresas, concluiu-se que as opiniões podem interferir na tomada de decisão.

Palavras-chave: Opinião. Mineração de opinião. Twitter. Aprendizado de máquina. Tomada de decisão.

ABSTRACT

Opinion is considered very important for people's behavior, especially for the decision-making process in a company. This paper presents opinion mining, other concepts in this area of computing, and proposes a case study where the object of study is a Chinese electronics company. To make this possible, a database was extracted through the social network Twitter, since the social network became an important media with instant information exchange. In this work, a machine learning algorithm was used, the support vector machine (SVM), to perform an analysis of opinions in texts written in Portuguese. The objective of this paper is to classify the opinions obtained by performing a data processing and present their polarity in a summarized and statistical way. It was observed that opinions have an important role in the corporate reputation, and it was concluded that opinions can interfere in the decision making.

Keywords: Opinion. Opinion mining. Twitter. Machine learning. Decision making.

LISTA DE ILUSTRAÇÕES

Figura 1 - Elementos formadores do conhecimento.....	15
Figura 2 - O processo de KDD.....	25
Figura 3 - Transformando uma frase em uma representação de <i>bag-of-words</i>	26
Figura 4 - Exemplo de uma SVM com duas classes.....	32
Figura 5 - Linha de código desenvolvida para seleção dos <i>tweets</i> utilizando a API do Twitter.....	34
Figura 6 - Processo de validação cruzada dividido em k-pastas.....	38
Figura 7 - Matriz de confusão com duas classes.....	39
Figura 8 - Gráfico com a sumarização das opiniões dos tweets sobre a Xiaomi.....	44
Figura 9 - Gráfico da participação da Xiaomi no mercado de fornecedores de smartphones no Brasil.....	44

LISTA DE TABELAS

Tabela 1 -	Sumarização dos dados contidos na base de <i>tweets</i> rotulados.....	35
Tabela 2 -	Alguns exemplos de <i>stopwords</i>	36
Tabela 3 -	Matriz de confusão com três classes obtido do modelo SVM.....	42

LISTA DE EQUAÇÕES

Equação 1 - Cálculo da acurácia do modelo de classificação.....	40
Equação 2 - Cálculo da precisão do modelo de classificação.....	40
Equação 3 - Cálculo da precisão para as classes positivo.....	41
Equação 4 - Cálculo da precisão para as classes negativo.....	41

LISTA DE SIGLAS

BI - Business Intelligence

KMS - Knowledge Management Systems

CI - Competitive Intelligence

PLN - Processamento de Linguagem Natural

KDD - Knowledge Discovery in Databases

SVM - Support vector Machine

API - Application Programming Interface

NLTK - Natural Language Toolkit

SUMÁRIO

1	INTRODUÇÃO	13
1.1	OBJETIVOS	19
1.2	ESTRUTURA DO TRABALHO	19
2	FUNDAMENTAÇÃO TEÓRICA	21
2.1	PROCESSAMENTO DE LINGUAGEM NATURAL	21
2.2	MINERAÇÃO DE OPINIÃO	23
2.3	DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS E MINERAÇÃO DE DADOS	24
2.4	CLASSIFICADORES DE POLARIDADE	27
2.4.1	Aprendizado de Máquina	28
2.4.1.1	Máquina de Vetores de Suporte	31
3	DESENVOLVIMENTO	33
3.1	TECNOLOGIAS UTILIZADAS	33
3.2	OBTENÇÃO DOS DADOS	34
3.3	CRIAÇÃO DA BASE DE DADOS ROTULADA	35
3.4	PRÉ-PROCESSAMENTO DE TWEETS	35
3.5	CLASSIFICAÇÃO DE TWEETS	37
3.6	AVALIAÇÃO DO CLASSIFICADOR	37
4	RESULTADOS	42
4.1	AVALIAÇÃO DOS CLASSIFICADORES	42
4.2	COMPARANDO A VENDA DOS PRODUTOS DA XIAOMI COM AS OPINIÕES DOS TWEETS	43
5	CONCLUSÃO	46
5.1	TRABALHOS FUTUROS	47
	REFERÊNCIAS	49
	ANEXO A - Repositório de código	52

1 INTRODUÇÃO

Business Intelligence (BI), no português Inteligência de Negócio, pode ser entendido como a utilização de variadas fontes de informação para definir estratégias de competitividade nos negócios de uma empresa. (BARBIERI, 2011). Para Turban et al. (2009), o BI é considerado um termo “guarda-chuva” o qual engloba ferramentas, arquitetura, bases de dados, data warehouse, gerenciamento de desempenho e metodologias, em uma suíte de software. Em seu livro, os autores afirmam:

O BI possui várias capacidades, o que inclui relatórios e perguntas, análise complicada, data mining, previsões e muito mais. Essas capacidades vieram de ferramentas e tecnologias nas quais o BI se baseia, e especialmente de sistemas de informação executiva (EIS), sistemas de apoio à decisão (DSS), perguntas, visualização, fluxo de trabalho, ciência de pesquisa/gerenciamento de operações e inteligência artificial aplicada.

Para os autores, o objetivo do BI se baseia em possibilitar que os gerentes de negócios e analistas acessem qualquer dado da empresa de maneira fácil e rápida, dessa forma, os tomadores de decisões obtêm valiosos insights que os ajudam a tomar melhores decisões. Dessa forma, segundo os autores, quase todas médias e grandes empresas hoje em dia estão utilizando BI de alguma forma para melhorar seu desempenho ou até mesmo sobreviver.

O BI é importante para ajudar as empresas a ficarem à frente da concorrência. Para Raisinghani (2004) o BI pode fornecer os meios para reunir e analisar dados, facilitando a geração de relatórios e contribuindo para uma tomada de decisão mais rápida, precisa e informada.

Em um processo de tomada de decisão, a informação é valiosa. Para Duarte (2013) “A informação é a componente chave para tomar decisões e escolhas com fundamento”. O autor também diz que:

[...] A opinião sobre uma empresa pode ajudar os investidores do mercado de ações e a opinião sobre marcas e produtos pode ajudar clientes indecisos, sobre o que devem comprar ou considerar, podendo também ajudar os profissionais de marketing a ver a eficácia de suas campanhas.

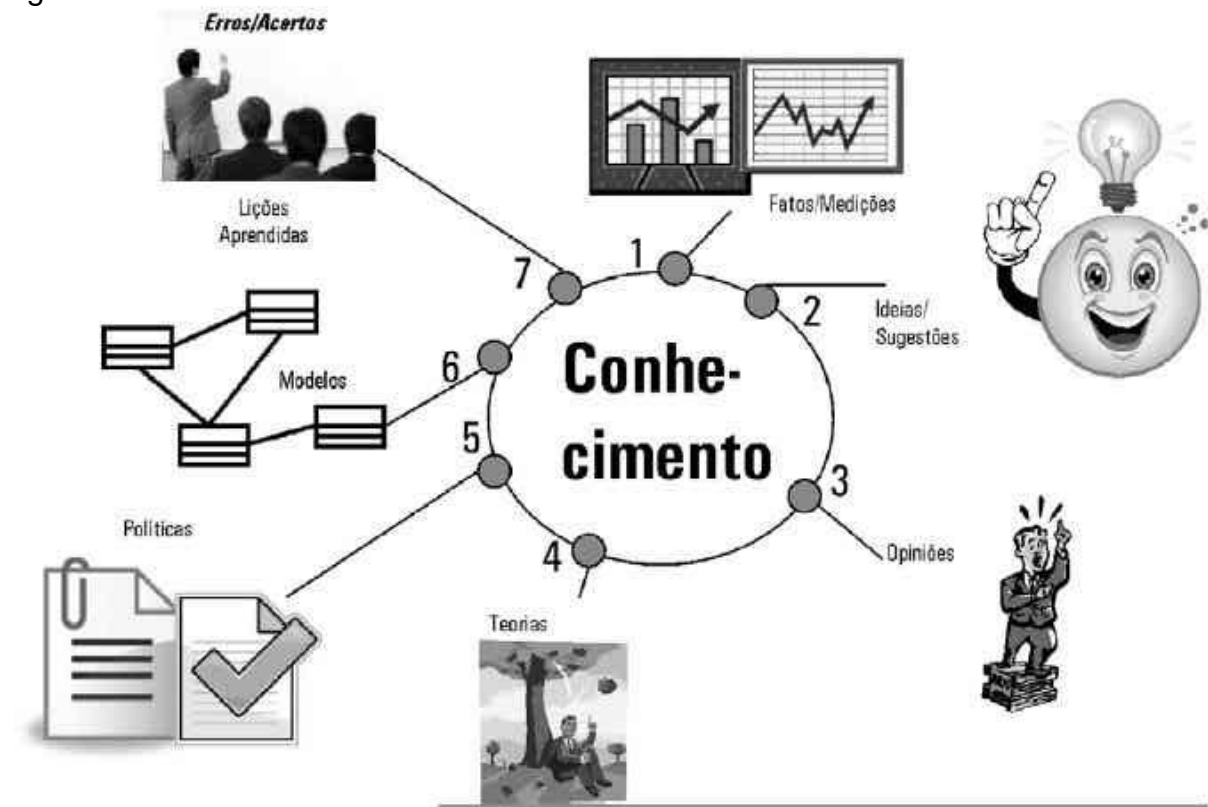
A informação é importante tanto dentro como fora da empresa. O conceito de gerência de conhecimento olha para dentro da empresa, enquanto a inteligência competitiva analisa o mundo exterior da empresa.

Sistemas de gestão do conhecimento (do inglês, KMS - *Knowledge Management Systems*), têm o objetivo de “estabelecer uma aproximação integrada e colaborativa para capturar, criar, organizar, disseminar e usar todos os ativos de informação de uma empresa” (BARBIERI, 2011). Dessa forma, acontece uma troca de conhecimento dentro da empresa, os colaboradores compartilham com a organização toda a sabedoria que possuem, e a organização, disponibiliza todo seu conhecimento para uso dos colaboradores. Em seu livro, o autor diferencia o BI do KMS:

[...] Se fizéssemos uma analogia dos conceitos de BI e KMS com um ambiente de condomínio, eu diria que os balancetes formais, estruturados, de onde se obtêm os gastos por dimensões de tempo, natureza, empregados etc. representam a abordagem de BI. O conjunto organizado de atas, circulares, orçamentos, notas fiscais, contratos de serviços, plantas de reformas, opiniões e experiências dos diversos moradores e síndicos, ao longo da vida do edifício, representaria a gerência/acervo de conhecimento.

Dentro da organização é interessante criar um repositório de conhecimento. As técnicas de BI podem ser aplicadas nesse acervo para transformar esse conhecimento em análise, que poderá ser utilizado para o desenvolvimento de projetos da empresa (BARBIERI, 2011). A Figura 1 ilustra algumas maneiras de obter conhecimento no ambiente corporativo.

Figura 1 - Elementos formadores do conhecimento.



Fonte: Barbieri (2011).

O conceito de sistemas de gestão do conhecimento olha para dentro da empresa, enquanto outra abordagem observa o mundo exterior da empresa. A inteligência competitiva (do inglês, CI - *Competitive Intelligence*), tem como ideia básica “explorar o outro lado da trincheira, obtendo informações detalhadas sobre os competidores e o mercado no qual se guerreia pela opção do cliente” (BARBIERE, 2011).

O mundo está cada vez mais competitivo, obter conhecimento da concorrência e dos clientes é fundamental. Para Barbierre (2011), a inteligência competitiva baseia-se no fato que ao conhecer melhor o ambiente externo, o diferencial competitivo em relação às outras empresas aumenta.

O CI pode ser entendido como um BI aplicado fora das fronteiras empresariais, identificando as oportunidades e ameaças. A inteligência competitiva permite que a empresa identifique seus pontos fortes e fracos, dessa forma os tomadores de decisão conseguem direcionar os negócios para o crescimento da empresa e para melhoria de sua reputação.

A reputação é a percepção ou sentimento que as pessoas expressam em forma de opinião definida sobre algo, alguém ou uma empresa (BARBIERRE, 2011), ela pode ser entendida como recurso de fonte de valor ou de vantagem competitiva (BARNEY, 1991).

Para Barbierri (2011), a reputação de uma empresa pode ser avaliada por:

- desempenho financeiro, de acordo com os resultados e à perspectiva de crescimento;
- liderança de mercado, com base na posição de destaque e visão de futuro;
- ambiente de trabalho, avaliando o clima organizacional;
- responsabilidade social, conforme as iniciativas de ações ambientais e sociais;
- capacidade de inovação, com colaboradores que sejam capazes de pensar diferente e apresentar ideias para melhorar processos, produtos e serviços;
- transparência, definida pela política de governança corporativa;
- qualidade de produtos e processos.

Para uma empresa, a opinião de seus clientes é fundamental para avaliar a qualidade de seus produtos e serviços, pois eles são os consumidores finais. É possível conhecer a opinião dos clientes, e compreender a repercussão da sua reputação, de algumas formas. As empresas podem obter opiniões de clientes por meio de pesquisa de satisfação feitas anualmente, preenchimento de pesquisas no momento da utilização de um serviço ou após um atendimento, pesquisas por e-mails etc.

Com uma grande quantidade de informação circulando diariamente pelo mundo em alta velocidade, a internet se apresenta como uma alternativa espontânea e imediata para obter opiniões. De acordo com o relatório *“Digital in 2019”* (KEMP, 2019), divulgado pelo serviço online *We Are Social*, 4.39 bilhões de pessoas no mundo estão conectadas com a internet, e dessas, 3.48 bilhões acessam as redes sociais compartilhando dados quase na velocidade da luz. Cobra (2009) afirma que a

internet com o passar do tempo está se tornando a principal mídia, estando presente em nossas vidas não só no nosso cotidiano, mas também no mundo dos negócios.

As redes sociais são o maior destaque do mundo virtual, elas se tornaram importantes meios de comunicação, onde usuários trocam informações diariamente sobre qualquer tema de maneira instantânea. Para Wasserman e Faust (1994), a rede social é um conjunto de dois elementos: atores, ou seja, nós (pessoas, instituições ou grupos) e suas conexões (interações ou laços sociais). A explosão das redes sociais, que possuem uma linguagem própria informal, vem modificando diversas áreas da atividade humana, disponibilizando opiniões diversificadas em grande volume de dados. Todas essas opiniões e sentimentos podem ser usados para identificar a posição das pessoas sobre determinados assuntos ou para verificar a reputação de uma empresa.

O Twitter, criado em 2006, é uma rede social onde as pessoas trocam mensagens em textos de até 280 caracteres, conhecidos como *tweets*, provenientes do mundo todo. No site da rede social o Twitter é definido como:

O Twitter é um serviço por meio do qual amigos, familiares e colegas de trabalho podem se comunicar e se manter conectados, trocando mensagens rápidas e frequentes. As pessoas publicam *tweets*, que podem conter fotos, vídeos, links e texto. Essas mensagens são publicadas em seu perfil e enviadas a seus seguidores, podendo ser encontradas por meio da busca do Twitter. (TWITTER, 2019)

As opiniões expostas nas redes sociais são importantes para verificar a reputação de uma empresa, por se tratar de opiniões espontâneas e imediatas. O grande volume e velocidade de produção de informação é um ponto essencial, porque isso aumenta a quantidade de opiniões disponíveis para análise, ao em vez da empresa se restringir à uma pesquisa de satisfação anual aplicada a um número restrito de clientes, por exemplo. Mas, esse grande volume e velocidade geram dificuldade em se obter esses resultados de forma manual. Desse modo, para ser possível utilizar a vasta quantidade de dados disponível nas redes sociais, se faz necessário a utilização de técnicas computacionais.

A análise de sentimentos ou a mineração de opiniões é “o estudo computacional de opiniões, sentimentos e emoções expressos em texto” (LIU, 2010). As opiniões são capazes de influenciar o comportamento das pessoas, seja para escolher um item a

ser comprado, um serviço a ser contratado, um filme para assistir, ou estratégias de mercado a serem utilizadas. Liu (2012) destaca que as opiniões são o centro de quase todas as atividades humanas, e quando precisamos tomar uma decisão, nós queremos saber a opinião das outras pessoas.

Segundo Freitas e Vieira (2015), pesquisas em análise de sentimentos geralmente são executadas em dados na língua inglesa, sendo outras línguas menos exploradas. No Brasil, essa área de pesquisa ainda está se estabelecendo e os recursos necessários para o processo só começaram a ser desenvolvidos na língua portuguesa em 2011.

Utilizar as redes sociais como fonte de dados para o BI pode ser uma alternativa de pesquisa de mercado, aumentando a quantidade e variação de informações disponíveis, além da rapidez na execução do processo. É possível aplicar técnicas de mineração de opiniões para obtenção da opinião das pessoas a respeito de uma empresa.

Xiaomi é uma empresa chinesa de produtos eletrônicos que foi fundada em 2010 que tem como lema “Tecnologia de qualidade acessível para todos” (XIAOMI, 2019), o que é possível observar devido aos preços baixos em dispositivos de qualidade conhecida. De acordo com a própria empresa:

O "Mi" em nosso logotipo significa "Mobile Internet". Ele também tem outros significados, incluindo "Missão Impossível", porque a Xiaomi enfrentou muitos desafios que pareciam impossíveis em nosso início.

A empresa chinesa é bastante popular ao redor do mundo e tem destaque no Brasil, principalmente pelos seus smartphones. Segundo Batista (2019), a Xiaomi saiu do Brasil em 2017 e este ano em maio de 2019 voltou oficialmente para o país, trazendo não apenas smartphones, mas também patinetes elétricos, relógios inteligentes, robô aspirador e outras novidades.

Sabendo da importância das opiniões para validação da reputação de uma empresa, a vasta quantidade de dados disponíveis nas redes sociais e as dificuldades de explorar esses dados de forma manual. Este trabalho propõe um estudo de caso utilizando as opiniões publicadas no Twitter, em português, a respeito da empresa Xiaomi, como uma alternativa de pesquisa de mercado mais moderna das existentes atualmente. A escolha da Xiaomi como objeto de estudo, foi devido a sua

popularidade e retorno ao Brasil em 2019, o que chamou atenção da autora. E a escolha do Twitter como fonte de dados, foi devido a quantidade curta de caracteres presentes em cada publicação e a forma espontânea e imediata que os usuários expressam os seus sentimentos nessa rede social.

1.1 OBJETIVOS

O objetivo deste trabalho é apresentar e aplicar técnicas de mineração de opinião através de um estudo de caso, utilizando publicações do Twitter em português acerca da empresa chinesa de produtos eletrônicos, Xiaomi. Sendo os objetivos específicos:

- a) extrair as informações presentes na rede social Twitter por meio dos *tweets*;
- b) analisar e classificar as opiniões e sentimentos a respeito da empresa Xiaomi, ou seja, se tendem a ser positiva, negativa ou neutra;
- c) apresentar os resultados de forma resumida e estatística;
- d) investigar se as opiniões interferem na tomada de decisão, em uma análise de correlação, comparando com o resultado de participação da marca no mercado.

1.2 ESTRUTURA DO TRABALHO

O restante do trabalho é organizado da seguinte maneira.

- Capítulo 2 - Fundamentação Teórica: é apresentada a explicação teórica necessária para o desenvolvimento e entendimento de estudo, incluindo uma visão geral sobre processamento de linguagem natural, uma apresentação da mineração de opinião, explicação das etapas de descoberta de conhecimento

em bases de dados, incluindo a mineração de dados, além de uma visão geral dos classificadores de polaridade, onde a abordagem de aprendizado de máquina e o algoritmo de máquina de vetores de suporte são melhores explorados.

- Capítulo 3 - Desenvolvimento: expõe as tecnologias e os métodos utilizados para o desenvolvimento do trabalho, descrevendo cada uma das etapas executadas durante o processo de mineração de opinião: a obtenção dos dados pelo Twitter, a criação de uma base de dados rotulada, o pré-processamento realizado para limpeza dos dados e os métodos para avaliar o desempenho do classificador utilizado.
- Capítulo 4 - Resultados: os dados obtidos a partir da classificação realizada são observados e analisados em comparação com a reputação da Xiaomi, de acordo com a sua participação e posição de vendas no mercado.
- Capítulo 5 - Conclusão: as principais conclusões deste trabalho são apresentadas, além de sugestões de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esse capítulo apresenta o conceito de processamento de linguagem natural e as dificuldades que a máquina possui para compreender a linguagem humana. Os conceitos relacionados à mineração de opinião são apresentados, e todas as etapas do processo de descoberta do conhecimento são explicadas. Também são abordadas as técnicas e a importância das etapas de pré-processamento e transformação dos dados. Em seguida os classificadores de polaridade mais conhecidos são brevemente apresentados. Por fim, é apresentado o algoritmo de máquina de vetores de suporte, um dos algoritmos de classificação mais utilizados.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

Processamento de linguagem natural (PLN) é uma área de pesquisa que explora como os computadores podem ser utilizados para entender e manipular uma linguagem natural, seja textual ou falada (CHOWDHURY, 2003). Segundo Jackson e Moulinier (2002) o termo PLN é usado para descrever a função de componentes de software ou hardware em um sistema de computador que analisa ou sintetiza a linguagem falada ou escrita. Ou seja, o processamento de linguagem natural tem como desafio utilizar os computadores para compreenderem a linguagem natural como um ser humano.

A linguagem humana, seja ela escrita ou falada, possui diversas regras e ambiguidades particulares de cada idioma. Não é tão simples para uma máquina compreender a linguagem natural, por exemplo, em português a palavra “banco” pode ser uma instituição financeira, um assento com ou sem encosto, um lugar onde são armazenadas amostras de sangue ou até mesmo um lugar onde são armazenados os dados de um sistema de computador. Uma mesma palavra pode ter vários significados, dependendo do contexto em que ela está inserida.

Sentenças inteiras também podem ser ambíguas, em relação à sua estrutura e ao seu significado. Na frase: "Viajar com os primos pode ser um incômodo.", pode significar um incômodo para os primos ou para o interlocutor da frase, dependendo da análise sintática (relação ou combinação entre as palavras).

Em seu livro, os autores Jackson e Moulinier (2002) apontam que um exemplo de manifestação comum da ambiguidade sintática é o apego à frase preposicionada (frase onde dois elementos estão ligados a uma preposição). A frase a seguir possui o tipo de preposição de instrumento, representada pela palavra "com": "John viu o homem no parque com o telescópio". Existe ambiguidade na frase, porque é possível questionar: com quem está o telescópio? O telescópio pode estar com John ou com o homem no parque, onde cada resposta sugere uma interpretação diferente, com base em um anexo diferente da frase preposicionada "com o telescópio".

De acordo com Jackson e Moulinier (2002) a linguagem está repleta de ambiguidade linguística, às vezes pode ser uma fonte de humor, mas muitas palavras e frases têm várias interpretações que passam despercebidas. Os seres humanos raramente confundem esses significados, devido aos diferentes contextos em que os símbolos dessa palavra podem estar e ao conhecimento do mundo real.

Frases que nenhum ser humano consideraria ambíguas podem também causar problemas aos programas de computador, por exemplo: "Ela entrou no carro de mochila" que parece semelhante à frase: "Ela entrou no carro de 4 rodas". Para uma pessoa fica claro que a mochila pertence a mulher e as rodas pertencem ao carro, mas um computador não tem esse conhecimento. A capacidade de entender as duas frases acima não é considerada evidência de inteligência superior, mas, segundo Jackson e Moulinier (2002), o desejo de lidar com esse tipo de ambiguidade alimenta várias teses de doutorado todos os anos.

Nas redes sociais, o grande desafio está na presença da linguagem informal, com as constantes gírias, abreviações, erros de português nas publicações dos usuários e as ambiguidades. Isso acontece, principalmente, nas publicações do Twitter, por serem restritas a um número pequeno de caracteres e por serem espontâneas, sem preocupação com a gramática da língua portuguesa. Não é tão simples para uma

máquina compreender a linguagem natural, e essa linguagem informal do Twitter ainda mais desafiadora para o processamento de linguagem natural, aumenta a dificuldade do computador em compreender a linguagem natural como um ser humano.

O desafio da tecnologia é conseguir desenvolver sistemas digitais que sejam capazes de entender a linguagem humana de acordo com os parâmetros de uma linguagem natural, ou o mais próximo dela. Lopes e Vieira (2010) afirmam que existem diversas técnicas computacionais que podem ser usadas para analisar e interpretar a linguagem natural. Por existir uma forte distinção entre a linguagem falada e a linguagem escrita, é necessário aplicar diferentes técnicas para cada tipo de aplicação, em diferentes tipos de textos, como por exemplo, textos científicos, jornalísticos, literários etc.

O entendimento do processamento de linguagem natural é importante para a aprendizagem automática, mineração de dados e mineração de opinião, que são campos que envolvem o aprendizado a partir de dados.

2.2 MINERAÇÃO DE OPINIÃO

Mineração de opinião, também chamada de análise de sentimentos, é um campo de estudo, que analisa as opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos (LIU, 2012). Na análise de sentimentos, são utilizadas técnicas para extrair e determinar automaticamente o sentimento expresso em uma linguagem natural, com o apoio de um computador. Segundo Duarte (2013) o que se deseja extrair das mensagens é o sentimento da emoção ou opinião expresso como positivo, negativo ou neutro.

Para Liu (2012) as opiniões são a relação entre pelo menos dois elementos: um objeto de estudo e um sentimento sobre esse objeto. Esse objeto de estudo pode ser uma organização, pessoa, evento, produto ou marca. De acordo com Tsytarau e Palpanas (2012) o sentimento é uma opinião ou emoção que o autor tem a

respeito do tema; e a polaridade do sentimento é representada em uma escala entre avaliação positiva, neutra ou negativa.

Para detectar opiniões é preciso identificar as palavras que são positivas ou negativas naquele contexto. Por exemplo, a frase “esse é um bom livro” é vista como uma opinião positiva, enquanto a sentença “esse filme não é nada bom” é vista como uma opinião negativa. Opiniões neutras acontecem quando na frase não há uma opinião positiva ou negativa, por exemplo, “Li esse livro ontem”.

Para o ser humano é simples classificar essas frases, mas para a máquina é um grande desafio. A utilização da mineração de opinião para identificar a opiniões expressas em publicações do Twitter de forma automática faz parte de um processo de descoberta de conhecimento, que tem como uma das etapas a mineração de dados, área na qual a mineração de opinião está inserida.

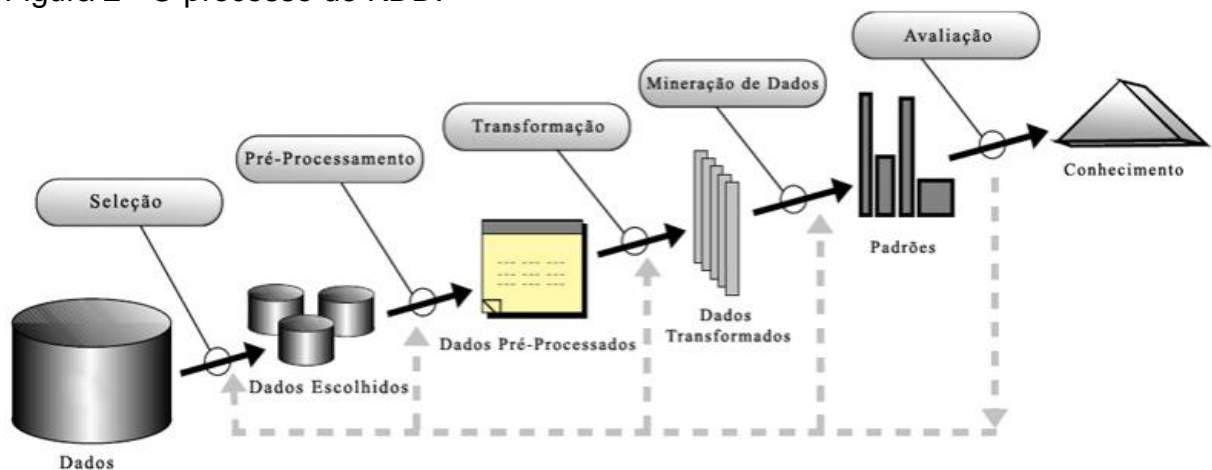
2.3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS E MINERAÇÃO DE DADOS

A descoberta de conhecimento em bases de dados (do inglês KDD - *Knowledge Discovery in Databases*) é uma das nomenclaturas utilizadas para noção de encontrar padrões nos dados. Segundo Fayyad *et al.* (1996):

Historicamente, a noção de encontrar padrões úteis nos dados recebeu vários nomes, incluindo mineração de dados, extração de conhecimento, descoberta de informações, coleta de informações, arqueologia de dados e processamento de padrões de dados. [...]

De acordo com os autores, o termo mineração de dados é utilizado em mais frequência por estatísticos, analistas de dados e comunidades de sistemas de informações gerenciais. Para eles, o KDD refere-se ao processo geral de descoberta de conhecimento a partir de dados, e a mineração de dados, apenas a uma etapa específica desse processo. Podemos observar uma representação do processo de KDD na Figura 2 abaixo.

Figura 2 - O processo de KDD.



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

A fase inicial do processo de KDD é a seleção dos dados. Existe uma variedade de possíveis fontes de dados, é possível utilizar diversos documentos, como obras literárias, livros acadêmicos, sites de notícias, blogs, notícias de jornal e até mesmo as redes sociais.

Depois que os dados foram selecionados, se inicia a fase de pré-processamento, onde os dados são transformados para melhorar sua qualidade. O objetivo principal dessa fase é “a identificação e remoção de problemas presentes nos dados antes que os métodos de extração de conhecimento sejam aplicados” (BATISTA, 2003).

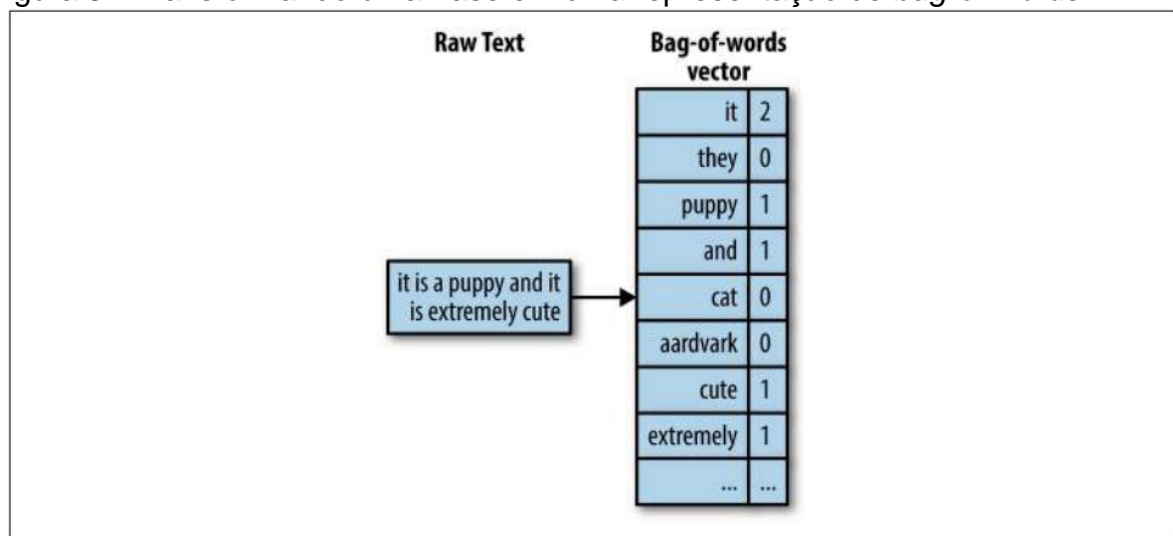
Conforme dito na seção 2.1, de acordo com Lopes e Vieira (2010), existem algumas técnicas computacionais que podem ser usadas para auxiliar na análise e interpretação da linguagem natural. Dessa forma, na fase de pré-processamento os dados são submetidos a algumas técnicas, como a remoção de palavras vazias (do inglês, *stopwords*), ou seja, limpar a base de dados de palavras que não agregam valores para a análise de sentimentos. Conforme Rajaraman e Ullman (2011) as *stopwords* são elementos como vírgulas, acentos, artigos (a, e, i, o, u), palavras comuns (em, um, na, como, para, de), entre outros.

De acordo com Zheng e Casari (2018) “Um problema com a análise simples é que variações diferentes da mesma palavra são contadas como palavras separadas”, por isso, ainda na fase de pré-processamento, os dados são submetidos ao processo de lematização (do inglês, *stemming*), onde considera-se cada palavra isoladamente,

tentando reduzi-la a sua palavra raiz. Palavras como “fazendo” e “fazer” viram apenas “faz”.

Após a eliminação das *stopwords* e da redução das palavras a sua forma raiz pelo processo de *stemming*, os dados passam pela fase de transformação. Uma das técnicas utilizadas é a conversão dos dados em tabelas, esse tipo de representação é conhecido como saco-de-palavras (do inglês, *bag-of-words*). De acordo com Zheng e Casari (2018) no processo de *bag-of-words* o texto é convertido em uma tabela de contagem de palavras. Se a palavra “amo” aparecer cinco vezes no texto, a tabela terá uma contagem de 5 na posição correspondente a essa palavra, caso a palavra não aparecer no texto, ela terá a contagem de 0. Por exemplo, a frase “é um filhote de cachorro e é extremamente fofo” (do inglês, “*it is a puppy and it is extremely cute*”) tem a representação de *bag-of-words* mostrada na Figura 3.

Figura 3 - Transformando uma frase em uma representação de *bag-of-words*.



Fonte: Zheng e Casari (2018).

Após a aplicação das transformações nos dados, inicia-se a fase de mineração de dados. Para Soares (2008):

É na etapa de mineração que ocorre a busca efetiva por conhecimentos novos e úteis a partir dos dados. Compreende a aplicação de algoritmos de aprendizado de máquina sobre os dados de forma a abstrair o conhecimento implícito presente nestes.

A fase de mineração de dados consiste em identificar padrões. Nessa etapa, são aplicados algoritmos de análise e descoberta de dados que produzem uma enumeração específica de padrões sobre os dados (ZHENG; CASARI, 2018).

Por fim, a fase de avaliação, ou pós-processamento, consiste em avaliar o desempenho do modelo aplicado na fase de mineração de dados utilizando medidas estatísticas. Nessa fase, os dados que foram gerados são interpretados e, dessa forma, o conhecimento do processo de KDD é descoberto.

Durante o processo de mineração de opinião, para a identificação de padrões é necessária a utilização de uma abordagem de classificação de polaridade.

2.4 CLASSIFICADORES DE POLARIDADE

Conforme dito na seção 2.2, a polaridade da opinião ou sentimento de uma mensagem pode ser representada em uma escala de avaliação positiva, negativa e neutra. Dessa forma, a classificação de polaridade, ou de sentimento, é responsável pela classificação do sentimento de uma mensagem, podendo ser um texto escrito.

Segundo Becker e Tumitan (2013) as abordagens de classificação de polaridade, podem ser divididas em quatro grandes grupos:

- a) Léxicas, também denominada de linguística ou abordagem baseada em dicionário, que tem como aspecto central o uso de dicionários de sentimentos, ou seja, compilações de palavras associadas à respectiva polaridade. Um dos métodos utilizados na abordagem léxica é o da co-ocorrência entre alvo e sentimento, dessa forma, para a classificação do sentimento em um texto, basta que exista uma palavra onde sua polaridade é dada por um léxico de sentimentos. Por exemplo, na frase “o celular Xiaomi Mi 9 é muito bom”, a polaridade positiva da palavra “bom” é associada à entidade “Xiaomi Mi 9”. Esse método apresenta bons resultados em documentos com poucos caracteres, pois a palavra detentora do sentimento pode estar próxima à entidade que qualifica;

- b) Aprendizado de máquina, com o objetivo de descobrir automaticamente regras gerais em grandes conjuntos de dados, permitindo extrair informações implicitamente representadas. As técnicas de aprendizado de máquina podem ser divididas em dois tipos: aprendizado supervisionado e aprendizado não supervisionado. A área de mineração de opinião tem o uso predominante de métodos supervisionado de aprendizado, como as técnicas de classificação e regressão;
- c) Estatísticas, que são denominadas como abordagens não supervisionadas, baseiam-se em técnicas para avaliar a co-ocorrência de termos. Nessa abordagem, se a palavra ocorre com mais frequência junto a palavras positivas (negativas) no mesmo contexto, então é provável que seja positiva (negativa). Já se a palavra ocorre em igual frequência, deve ser neutra. A polaridade de uma palavra pode ser determinada calculando a co-ocorrência com uma palavra positiva (negativa), tal como “excelente” ou “péssimo”;
- d) Semânticas, parecida com a abordagem estatística, porém a polaridade é calculada em função de sua proximidade semântica com outras palavras de polaridade conhecidas. A semântica de uma palavra é o sentido que ela pode tomar de acordo com o contexto. Nessa abordagem, palavras semanticamente próximas devem ter a mesma polaridade.

Essas técnicas podem ser combinadas, para obter melhores resultados.

A abordagem de aprendizado de máquina, de método supervisionado de aprendizado, é predominantemente aplicada na área de mineração de opinião, dessa forma, será explorada na seção a seguir.

2.4.1 Aprendizado de Máquina

O aprendizado de máquina (do inglês, *machine learning*), de acordo com Mitchell (1997) é um campo de estudo que se preocupa em como construir programas de

computador onde seu desempenho melhore com a experiência, ou seja, que possam “aprender” com a experiência.

Conforme dito na seção 2.4, para Becker e Tumitan (2013) as técnicas de aprendizado de máquina podem ser divididas em dois tipos: aprendizado supervisionado e aprendizado não supervisionado. De acordo com Jackson e Moulinier (2002) o aprendizado supervisionado é o aprendizado por exemplos. Dessa forma, quando são apresentados exemplos ao computador, ele está sendo ensinado a fazer as distinções corretas. Esse aprendizado é considerado mecânico, pois as regras de classificação são informadas à máquina. Já o aprendizado não supervisionado é o aprendizado por observação e descoberta, onde a máquina aprende sem receber exemplos, mas na verdade agrupando documentos semelhantes.

Segundo Duarte (2013) na classificação de aprendizado de máquina, para classificar objetos em classes específicas é utilizado um conjunto de algoritmos computacionais, que podem ser treinados com exemplos que já são classificados. Dessa forma, é possível encontrar a melhor combinação para classificar novas informações.

Em seu livro, Mitchell (1997) afirma que o entendimento detalhado a respeito dos algoritmos de aprendizado de máquina para o processamento de informação pode levar a um melhor entendimento da capacidade (e incapacidade) do aprendizado humano. Ainda não se sabe como fazer os computadores aprenderem tão bem como as pessoas, mas esses algoritmos são eficazes para certas tarefas de aprendizagem.

Segundo Mohri *et al.* (2012) os principais algoritmos de aprendizagem supervisionada criam modelos que permitem a classificação, para variáveis categóricas, e a regressão, para variáveis contínuas. A classificação é a aprendizagem de uma função que mapeie os dados em uma ou várias classes, que consiste em atribuir um rótulo para a saída a partir de determinada entrada. Dessa forma, pode-se dizer se uma “entrada” pertence a certa classe, por exemplo, positivo, negativo e neutro. Já a regressão é a aprendizagem de uma função que mapeie os dados em uma variável de previsão, onde ela retorna um valor que

pertence a um espectro contínuo de valores. Segundo Kutner *et al.* (2005), regressão é um método estatístico que utiliza a relação entre duas ou mais variáveis quantitativas para que uma variável de resultado possa ser prevista a partir da outra. Dessa forma, é realizada uma predição a partir de um exemplo quantitativo. Por exemplo, prever o valor da bolsa de valores amanhã de acordo com os valores de dias e meses anteriores.

A maioria dos problemas de processamento de linguagem natural em que se é utilizado o aprendizado de máquina podem ser tratados como problemas de classificação. De acordo com Becker e Tumitan (2013) na área de mineração de opinião é frequente o uso de métodos supervisionados. Dessa forma, o problema de classificação é dividido em dois passos:

1. Aprender um modelo de classificação em cima de uma coletânea de treinamento previamente rotulada com as classes consideradas, como positivo, negativo e neutro;
2. Prever a polaridade de novos textos com base no modelo que foi treinado.

Mas, Becker e Tumitan (2013) ressaltam que:

Uma das grandes limitações no uso de aprendizado supervisionado para definição de polaridade é a necessidade de dados rotulados para treino. O desempenho destes métodos é afetado não somente pela quantidade, mas igualmente pela qualidade dos dados de treino disponíveis.

Utilizar o aprendizado supervisionado para fazer um modelo de classificação de opinião se torna muito trabalhoso, devido a necessidade de rotular os dados para treino. Conforme dito por Mitchell (1997) ainda não se sabe como fazer as máquinas aprenderem tão bem como as pessoas, dessa forma, os algoritmos de aprendizado ainda não são 100% eficazes. À medida que o entendimento sobre os computadores cresce, é inevitável que o aprendizado de máquina tenha um papel cada vez mais central na ciência da computação e na tecnologia da informação.

Os algoritmos de classificação mais usados nesta área são: máquina de vetores de suporte, naive bayes, máxima entropia e algoritmos baseados em redes neurais (BECKER; TUMITAN, 2013). Dentre os algoritmos citados, a máquina de vetores de suporte entrou para a lista dos 10 melhores algoritmos para mineração de dados, de

acordo com WU *et al.* (2008). Em termos de performance, o algoritmo de SVM tende a ser melhor que o Naive Bayes (PANG *et al.*, 2002) e possui vantagem sobre as Redes Neurais Artificiais (LORENA; CARVALHO, 2007).

2.4.1.1 Máquina de Vetores de Suporte

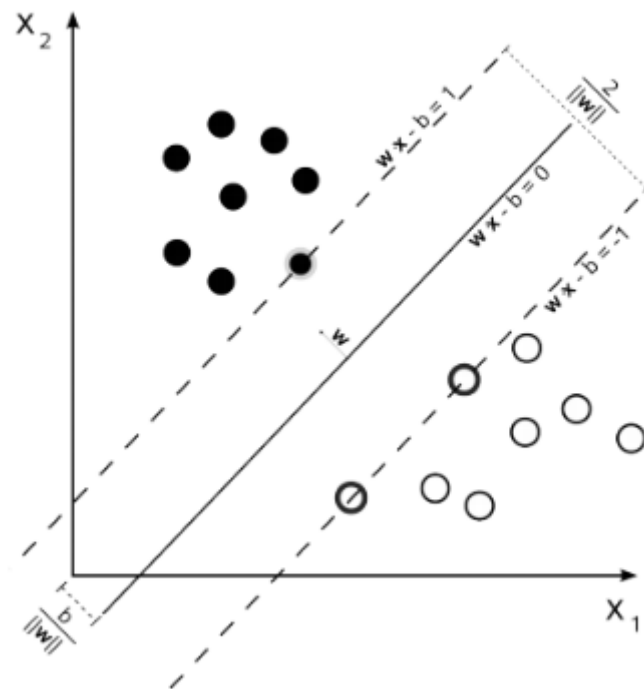
A Máquina de Vetor de Suporte (*support vector machine* ou SVM) é um padrão de classificação introduzido por Vapnik (1998), o qual afirma que o método SVM “pode ser aplicado para reconhecimento de padrões”. Dos algoritmos de aprendizado de máquina, a SVM é considerada uma ferramenta que oferece um dos métodos mais robustos e precisos entre todos os algoritmos conhecidos (WU *et al.*, 2008). Segundo Camilo e Silva (2009) as máquinas de vetores de suporte chamam muita atenção pelos seus resultados e assertividade, podendo ser utilizado para tarefas de classificação quanto de predição.

Segundo Duarte (2013) as SVM convertem os dados em pontos em um hiperplano e, de acordo, com Wu *et al.* (2008), em uma tarefa de aprendizado linear de duas classes. O objetivo da SVM é encontrar um hiperplano que separe as duas classes com a maior margem para oferecer a melhor capacidade de generalização. Os autores definem capacidade de generalização como:

[...] A capacidade de generalização refere-se ao fato de que um classificador não apenas possui um bom desempenho de classificação (por exemplo, precisão) nos dados de treinamento, mas também garante alta precisão preditiva para os dados futuros da mesma distribuição que os dados de treinamento.

A Figura 4 ilustra graficamente um hiperplano bidimensional.

Figura 4 - Exemplo de uma SVM com duas classes.



Fonte: Duarte (2013)

Segundo Duarte (2013):

na Figura 4 a posição do ponto dos recursos, estando mais próxima ou mais distante da área de uma classe no hiperplano, indicará a classe à qual esse recurso tem mais probabilidade de pertencer. Há também uma margem para separar as classes e os recursos contidos nessa margem são tratados como neutros.

Dessa forma, o objetivo do SVM é encontrar o hiperplano ótimo que tenha a margem de separação maximizada.

3 DESENVOLVIMENTO

Este capítulo expõe os métodos utilizados para o desenvolvimento do estudo de caso proposto neste trabalho, que tem como objetivo descobrir se há alguma relação entre a popularidade da empresa Xiaomi no Brasil e a opinião dos usuários do Twitter em relação aos produtos desenvolvidos por essa empresa. Além disso, é apresentado as tecnologias que foram utilizadas para o desenvolvimento deste trabalho.

3.1 TECNOLOGIAS UTILIZADAS

A autora desenvolveu o código deste trabalho utilizando a linguagem de programação Python como ferramenta, devido a sua familiaridade, o ambiente computacional utilizado foi o Jupyter Notebook e foram utilizadas algumas bibliotecas de código para apoio, como as bibliotecas:

- a) Pandas, para manipulação e análise de dados;
- b) NLTK (Natural Language Toolkit), para o pré-processamento dos dados, auxiliando no processamento de linguagem natural;
- c) Re, para auxiliar na etapa de pré-processamento ao remover os *links* presentes nos tweets;
- d) Scikit-learn, para aplicar o algoritmo de aprendizado de máquina escolhido, o SVM, e para transformação dos dados e validação do modelo;
- e) Além da utilização da biblioteca externa python-twitter, para realizar a conexão com o a API do Twitter e realizar a extração das publicações.

O computador utilizado para o desenvolvimento do código deste trabalho possui as seguintes configurações: processador de modelo Intel® Core™ i5-8265U, memória RAM de 8GB e memória de armazenamento HD de 1TB. E o tempo aproximado de execução do código foi de 8 segundos.

3.2 OBTENÇÃO DOS DADOS

Como fonte de dados deste trabalho foi utilizado o Twitter, por ser uma rede social que possibilita a seus usuários compartilharem ideias constantemente, gerando grande quantidade de informação, e devido a sua popularidade. A extração de dados foi realizada durante o período de 16 a 31 de outubro de 2019.

O Twitter fornece para os desenvolvedores uma Interface de Programação de Aplicativos (do inglês, *Application Programming Interface* ou *API*), que permite uma conexão entre os clientes e servidores para explorar os dados de maneira simples. Para coleta dos *tweets* foi utilizado a *Streaming API* (em português, API de fluxos do Twitter), que permite o acesso a um grande volume de dados em tempo real em nível global. É importante ressaltar que essa API possibilita obter todas as publicações abertas do Twitter, sem a necessidade de estar seguindo o usuário. Para ser possível realizar uma conexão com a API, foi necessário criar uma conta na plataforma de desenvolvedor do Twitter e foram disponibilizados dois tokens de acesso.

A seleção dos *tweets* foi feita pela escolha de uma palavra-chave, “Xiaomi” (nome da empresa de objeto de estudo), foram extraídos apenas *tweets* em português (informando à API por meio de um parâmetro para obter apenas *tweets* escritos nessa língua) e os *tweets* obtidos foram uma mistura das publicações mais recentes e das mais populares. A Figura 5 ilustra o processo de seleção dos dados com uma linha de código desenvolvida na linguagem de programação Python, onde foram extraídos 100 *tweets* aleatórios.

Figura 5 - Linha de código desenvolvida para seleção dos *tweets* utilizando a API do Twitter.

```
status_list = api.GetSearch(term="xiaomi",  
                             lang='pt',  
                             count=100,  
                             result_type='mixed')
```

Fonte: Elaborado pela autora.

3.3 CRIAÇÃO DA BASE DE DADOS ROTULADA

Após a seleção dos 100 *tweets*, as publicações foram rotuladas manualmente, pela autora e por uma outra pessoa, como positivas, neutras ou negativas. Nesse processo, as duas pessoas analisaram todos os *tweets* em conjunto e entraram em acordo a respeito de qual classe cada sentença pertencia. A Tabela 1 mostra quantos *tweets* cada classe possui depois da atribuição dos rótulos.

Tabela 1 - Sumarização dos dados contidos na base de *tweets* rotulados.

Sentimento	Quantidade de <i>tweets</i>
positivo	50
negativo	22
neutro	28
Total	100

Fonte: Elaborado pela autora.

De acordo com o objeto de estudo escolhido, foram consideradas opiniões positivas os *tweets* que falavam bem da marca ou dos produtos da Xiaomi. Consequentemente foram consideradas opiniões negativas os *tweets* que falavam mal ou continham reclamações a respeito da marca ou seus produtos. Por fim, foram consideradas opiniões neutras os *tweets* que não expressavam opiniões ou aqueles que eram propagandas de venda dos produtos da empresa.

3.4 PRÉ-PROCESSAMENTO DE TWEETS

Conforme dito na seção 2.2, o objetivo da etapa de pré-processamento é descartar o que não é considerado importante para a etapa de classificação. Dessa forma, foram aplicadas algumas operações para o tratamento dos dados:

- a) Conversão de letras maiúsculas em minúsculas, com o objetivo de padronizar o texto;
- b) Remoção de tweets repetidos. É importante ressaltar que após a remoção dos tweets repetidos, a base de dados ficou com 99 publicações;
- c) Remoção de *link*, uma vez que esses termos não possuem conteúdo semântico;
- d) Remoção de caracteres de pontuação, pois não agregam valor à classificação;
- e) Remoção de stopwords, retirando palavras que não agregam valores para a análise de sentimentos. A lista de stopwords foi retirada da biblioteca de código NLTK da linguagem de programação Python, alguns exemplos dessa lista podem ser observados na Tabela 2. Vale ressaltar que a única palavra que foi retirada da lista disponibilizada foi a palavra “não”, que é considerada pela biblioteca como uma stopwords, mas para a autora essa palavra é muito importante para identificação de frases negativas.

Tabela 2 - Alguns exemplos de *stopwords*.

Palavras		
o(a)	do(a)	no(a)
num(a)	nem	esse(a)
para	este(a)	aquilo
sou	houve	estão
seja	com	já
estive	quem	até

Fonte: Biblioteca NLTK - Python (2019).

3.5 CLASSIFICAÇÃO DE TWEETS

Depois que os dados foram extraídos, rotulados e pré processados, a etapa de classificação é executada, a qual consiste em aplicar o modelo de aprendizado de máquina nos vetores gerados por meio do *bag-of-words*. O algoritmo de aprendizado de máquina escolhido foi o SVM, devido aos autores apontarem o seu bom desempenho, e foi utilizado kernel linear.

O processo de *bag-of-words*, conforme dito na seção 2.3, é uma técnica da fase de transformação de dados que converte os dados em tabelas, onde o texto é convertido em uma tabela de contagem de palavras. Após aplicado o processo de *bag-of-words*, foi gerada uma matriz, com vetores dentro de um vetor, de tamanho 99 linhas e 444 colunas. As linhas representam a quantidade de *tweets* que foram obtidos e cada palavra da base de dados se tornou uma coluna, dessa forma, foi gerada uma matriz grande e esparsa.

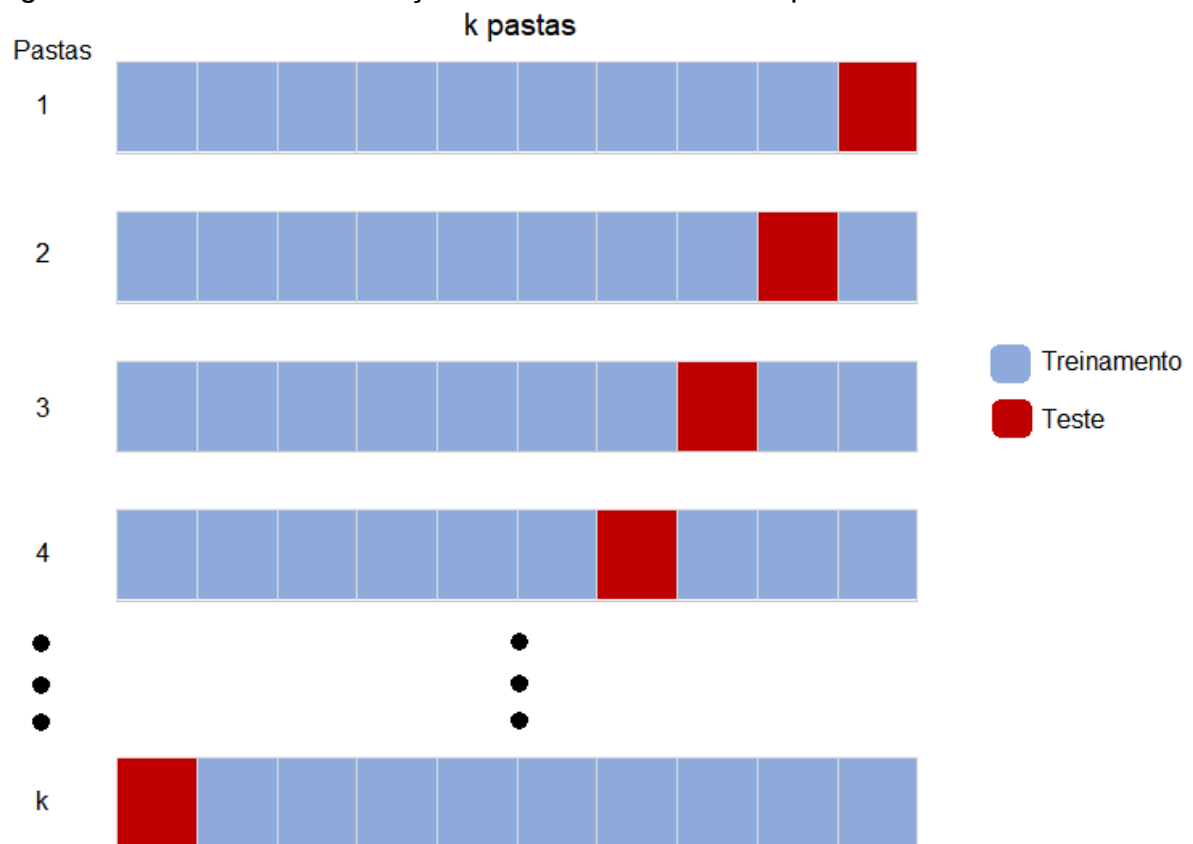
3.6 AVALIAÇÃO DO CLASSIFICADOR

O classificador SVM foi testado utilizando a base de *tweets* rotulada e a técnica validação cruzada. Segundo Mohri *et al.* (2012), a quantidade de dados rotulados disponíveis geralmente é muito pequena, o que deixa uma quantidade insuficiente de dados de treinamento. Para solucionar esse problema, o método conhecido como validação cruzada k-pastas é frequentemente adotado.

A técnica de validação cruzada k-pastas (do inglês, *cross validation k-fold*) consiste na divisão de dados em k partes aleatórias de tamanhos iguais, que são testadas individualmente, enquanto as partes restantes são utilizadas como conjunto de treinamento. O processo é repetido k vezes até todas as partes tenham sido usadas para teste apenas uma vez (TAN, STEINBACH; KUMAR, 2006).

A Figura 6 ilustra como funciona essa divisão dos dados em k partes aleatórias. Nesse processo, uma parte dos dados é separada para teste e o restante é utilizado para treinamento do modelo. Então quando a primeira pasta está em execução, o modelo é treinado e depois testado apenas com a parte que foi separada para isso. Esse processo é repetido k vezes até todas as partes serem utilizadas para teste apenas uma vez.

Figura 6 - Processo de validação cruzada dividido em k -pastas.



Fonte: Elaborado pela autora.

Neste trabalho os 99 vetores da matriz que foram gerados a partir do processo de *bag-of-words*, foram divididos em 48 partes. Dessa forma, 1 parte foi utilizada para teste e as outras 47 para treinamento, contendo no total 94 exemplos, garantindo que os dados de teste não fossem utilizados na fase de treinamento.

De acordo com Tan, Steinbach e Kumar (2006), a performance de um modelo de classificação está relacionada com a quantidade de predições corretas e incorretas.

Para obter essa informação é elaborada uma matriz de confusão, uma tabela que mostra as frequências de classificação para cada classe do modelo. A Figura 7 ilustra uma matriz de confusão de duas classes

Figura 7 - Matriz de confusão com duas classes.

		Classificação prevista	
		positivo	negativo
Classificação rotulada	positivo	Número de verdadeiros positivos (VP)	Número de falsos negativos (FN)
	negativo	Número de falsos positivos (FP)	Número de verdadeiros negativos (VN)

Fonte: Olson e Delen (2008)

Explicando os elementos contidos na matriz de confusão, temos:

- a) Verdadeiros positivos (VP) é a quantidade de dados que foram classificados corretamente como positivo, da mesma forma que os dados rotulados;
- b) Falsos negativos (FN) é a quantidade de dados que foram classificados incorretamente como negativo. Esses dados estariam corretos se fossem classificados como positivo de acordo com os dados rotulado, mas o modelo entendeu como negativo;
- c) Falsos positivos (FP) é a quantidade de dados que foram classificados incorretamente como positivo. Esses dados deveriam ter sido classificados como negativo, mas o modelo entendeu como positivo;
- d) Verdadeiros negativos (VN) é a quantidade de dados que foram classificados corretamente como negativo, da mesma forma que os dados rotulados.

De acordo com Tan, Steinbach e Kumar (2006), a matriz de confusão fornece uma informação importante para determinar a performance do modelo de classificação, a acurácia. Segundo Becker e Tumitan (2012), a qualidade do modelo preditivo pode

ser medida em termos de métricas pela acurácia, que é a capacidade do modelo de classificar corretamente, de acordo com as classificações apresentadas para aprendizado. Logo, o cálculo da acurácia resulta em um valor que representa quanto o modelo acertou das previsões possíveis. Quanto maior o valor da acurácia, melhor a performance do modelo.

A Equação 1 mostra como é realizado o cálculo da acurácia, onde $VP + VN$ é a soma de previsões corretas e $VP + VN + FP + FN$ é a soma do total de previsões do modelo.

Equação 1 - Cálculo da acurácia do modelo de classificação.

$$Acurácia = \frac{\text{Número de previsões corretas}}{\text{Número total de previsões}} = \frac{VP + VN}{VP + VN + FP + FN}$$

Fonte: Olson e Delen (2008)

Neste trabalho, foi elaborado uma matriz de confusão de três classes, divididas em positivo, negativo e neutro. Para determinar a performance do modelo, o cálculo da acurácia foi similar à Equação 1.

Segundo Becker e Tumitan (2012), a qualidade do modelo preditivo também pode ser medida em termos de métricas pela precisão, que é o número de vezes que uma classe foi prevista corretamente. A Equação 2 mostra como é realizado o cálculo da precisão:

Equação 2 - Cálculo da precisão do modelo de classificação.

$$Precisão = \frac{\text{Número de previsões corretas da classe}}{\text{Número de vezes que a classe foi predita}}$$

Fonte: Elaborado pela autora.

Para calcular a precisão da classe positivo, por exemplo, o número de vezes que a classe foi classificada corretamente é dividido pelo número total de vezes que essa

classe foi prevista pelo modelo. Onde VP é o número de vezes que a classe foi prevista corretamente e $VP + FP$ é o número total de vezes que essa classe foi prevista. É possível calcular a precisão de todas as classes do modelo, dessa forma, as fórmulas do cálculo da precisão das classes positivo e negativo podem ser observadas na Equação 3 e 4:

Equação 3 - Cálculo da precisão para as classes positivo.

$$\textit{Precisão da classe positivo} = \frac{VP}{VP + FP}$$

Fonte: Olson e Delen (2008)

Equação 4 - Cálculo da precisão para as classes negativo.

$$\textit{Precisão da classe negativo} = \frac{VN}{VN + FN}$$

Fonte: Olson e Delen (2008)

4 RESULTADO

Este capítulo tem como objetivo analisar os dados obtidos pela classificação automática dos *tweets* e sua relação com a popularidade da empresa Xiaomi.

4.1 AVALIAÇÃO DO CLASSIFICADOR

Conforme dito na seção 3.5, a performance de um modelo de classificação pode ser medida elaborando uma matriz de confusão. A Tabela 3 mostra a matriz de confusão obtida a partir do modelo gerado.

Tabela 3 - Matriz de confusão com três classes obtido do modelo SVM.

		Classificação pelo algoritmo SVM			
		positivo	negativo	neutro	Total
Rotulado pela autora	positivo	43	2	4	49
	negativo	12	7	3	22
	neutro	14	3	11	28
Total		69	12	18	99

Fonte: Elaborado pela autora.

Conforme dito na seção 3.5, sabendo que a performance do modelo de classificação está relacionada com a quantidade de predições corretas, a acurácia pode ser calculada a partir dos dados obtidos. A acurácia obtida nos experimentos realizados foi de aproximadamente 62%, isso significa que o modelo acertou 62% dos testes.

O valor de predições corretas não foi tão alto, isso pode ter acontecido pela base de dados não ser muito grande, o que dá ao modelo poucos exemplos para

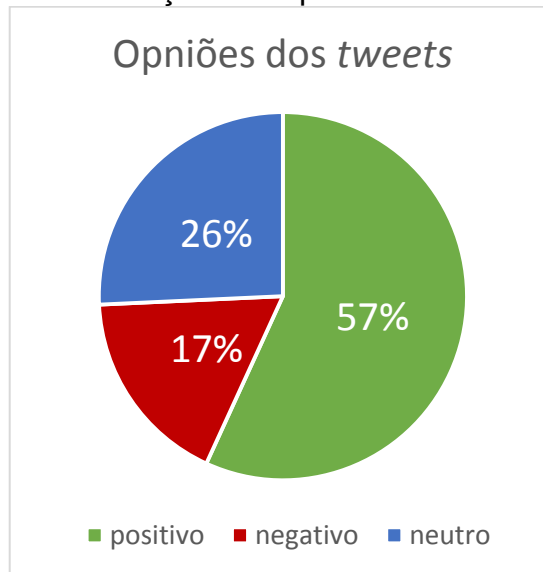
treinamento. Como foi explicado na seção 2.1, é um grande desafio para a máquina compreender a linguagem natural do ser humano, ainda mais com a linguagem informal das redes sociais, o que pode ter contribuído para o resultado alcançado não ser tão alto. E de acordo com o que foi exposto na seção 2.3.1, o desempenho dos métodos de aprendizado supervisionado é afetado não somente pela quantidade, mas também pela qualidade dos dados de treino.

Também foi calculada a precisão de cada classe do modelo. Os valores obtidos foram: 62% de precisão para a classe positivo, 58% para classe negativo e 61% para classe neutro. Pode-se observar que a classe positivo teve melhor desempenho em relação as outras classes, já que ela teve o maior índice de previsões corretas. Isso pode ter acontecido devido a maior quantidade de exemplos de *tweets* rotulados com opinião positiva, a máquina teve mais exemplos dessa classe para tentar compreender o que era cada opinião.

4.2 COMPARANDO A VENDA DOS PRODUTOS DA XIAOMI COM AS OPINIÕES DOS TWEETS

A fim de investigar se há uma correlação entre as opiniões emitidas pelos usuários e o desempenho de vendas da empresa Xiaomi, foram extraídos 1000 *tweets* para serem classificados pelo modelo de aprendizado de máquina gerado. Essas publicações passaram por todas as etapas de pré-processamento e transformação de dados, se tornando uma base de dados com 819 *tweets*. A partir desses dados foi possível observar que a opinião do público do Twitter em relação a Xiaomi é positiva. A Figura 8 ilustra a sumarização dos resultados obtidos no modelo SVM, onde 57% das opiniões expressas foram positivas, 17% negativas e 26% neutras. O que leva a compreender que a maiorias dos usuários do Twitter estão satisfeitos com a empresa e falam bem dos produtos, aumentando a reputação da organização.

Figura 8 - Gráfico com a sumarização das opiniões dos tweets sobre a Xiaomi.

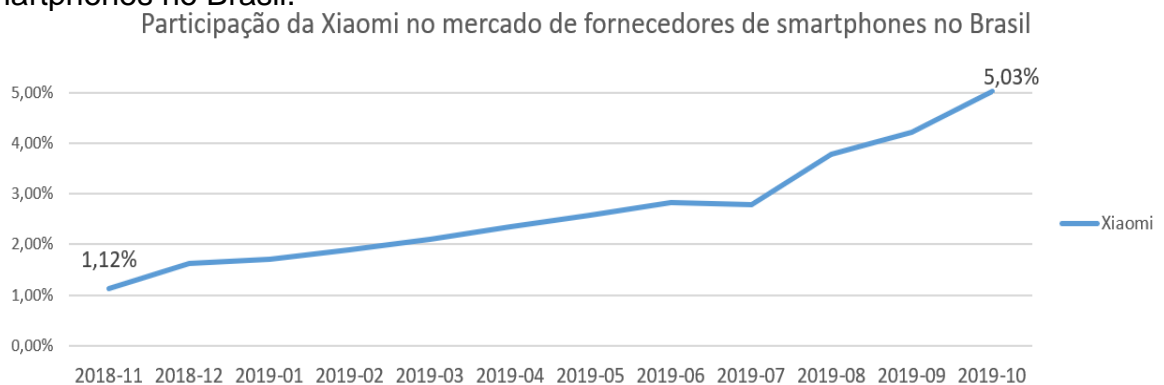


Fonte: Elaborado pela autora

Segundo a International Data Corporation (2019), ou IDC, a Xiaomi aparece em 4 lugar no ranking mundial de vendas de smartphones. E de acordo com a empresa Google os modelos de celulares mais procurados pelos brasileiros para as compras de Black Friday 2019 foi da Xiaomi, ocupando espaço apenas com a Apple no top 5 de smartphones mais buscados.

O relatório da Statcounter (2019) mostra que a participação da Xiaomi no mercado de fornecedores móveis no Brasil teve crescimento de 4 vezes mais com aumento de 3,91% de participação. A Figura 9 ilustra a presença da Xiaomi no mercado brasileiro.

Figura 9 - Gráfico da participação da Xiaomi no mercado de fornecedores de smartphones no Brasil.



Fonte: Statcounter (2019)

A partir dos resultados obtidos é possível concluir que a maioria dos usuários do *Twitter* tem uma visão positiva da empresa Xiaomi, que está crescendo no mercado brasileiro conforme dados analisados. Dessa forma, é possível que exista uma relação entre a opinião das pessoas e as vendas da marca Xiaomi, já que as pessoas opinam sobre os produtos e acabam influenciando a tomada de decisão de outros nas escolhas de qual marca optar em suas compras.

5 CONCLUSÃO

Neste trabalho, foram apresentados os conceitos de business intelligence, enfatizando a importância da reputação de uma empresa, que pode ser medida a partir da opinião de seus clientes. As redes sociais foram abordadas como fonte de dados relevantes para obter opiniões e o Twitter recebeu foco, devido a sua velocidade e disponibilidade de informação instantânea, o que torna uma ferramenta interessante a ser explorada na área de mineração de opinião.

As opiniões da internet são importantes para verificar a reputação de uma empresa, devido à grande quantidade de usuários ativos na internet e por possuírem opiniões espontâneas e imediatas. Essa pode ser uma alternativa para as empresas obterem opiniões, em vez de realizar uma pesquisa de satisfação com uma pequena quantidade de clientes.

Dessa forma, o objetivo deste trabalho foi aplicar as técnicas de mineração de opinião através de um estudo de caso, para isso foram utilizadas as publicações do Twitter na língua portuguesa e comparado as opiniões com a empresa chinesa Xiaomi.

A avaliação dos resultados das classificações, obtidas a partir do modelo de aprendizado de máquina, foi feita por meio do cálculo da acurácia e da precisão de cada classe. O valor de acurácia do modelo foi de 62%, e os valores obtidos no cálculo da precisão foram: 62% de precisão para a classe positivo, 58% para classe negativo e 61% para classe neutro.

Foi observado que o nível de acurácia do classificador não foi tão alto e, sabendo que o desempenho dos métodos de aprendizado supervisionado é afetado pela quantidade e pela qualidade dos dados de treino, pode ser, que o fato de a base de dados não ser muito grande tenha contribuído para esse resultado, pois o classificador recebeu poucos exemplos para aprender o que são opiniões positivas, negativas e neutras. Outro fator que pode ter contribuído, é que as publicações do Twitter apresentam o uso de uma linguagem informal contendo alguns erros de português, gírias e ambiguidade.

O resultado da sumarização das opiniões a respeito da Xiaomi mostrou que 70% dos usuários do Twitter possuem uma opinião positiva a respeito da empresa e apenas 12% possuem uma opinião negativa, sendo os outros 18% opiniões neutras e propagandas de venda dos produtos.

A partir de um levantamento feito sobre a marca Xiaomi no Brasil, que foi escolhida como objeto de estudo devido a sua popularidade e retorno ao país em 2019, foi possível observar que a empresa tem boa posição no mercado, ocupando o 4º lugar no ranking mundial de vendas de smartphones e ganhando destaque nas buscas de *black friday* de 2019. Além do seu aumento em 3,91% de participação no mercado de fornecedores de smartphones no Brasil, no período de novembro de 2018 a outubro de 2019.

A partir dos resultados obtidos e da análise dos *tweets*, foi possível perceber que a maioria dos usuários do Twitter possui uma visão positiva a respeito da empresa Xiaomi, que está crescendo no mercado. Esses dados sugerem que existe uma relação entre as vendas da marca e as opiniões das pessoas. Sendo assim, a opinião das pessoas pode influenciar na tomada de decisão de outros nas escolhas de qual marca optar em suas compras.

5.1 TRABALHOS FUTUROS

Devido aos resultados que este trabalho apresentou, em um trabalho futuro seria interessante obter uma maior quantidade de dados e verificar se os resultados melhoram, além disso, utilizar de outros algoritmos e classificadores de polaridade para comparar os resultados obtidos e verificar qual é o melhor a ser utilizado nesse tipo de aplicação.

Outra abordagem seria considerar apenas as *hashtags* (são utilizadas nas redes sociais palavras chaves após o símbolo #, por exemplo, #amei) contidas em cada *tweet* e verificar se elas contêm algum sentimento que reflete no respectivo *tweet*.

Vale ressaltar que a metodologia aplicada neste estudo de caso também pode ser aplicada para a análise das opiniões a respeito de outras empresas. Uma outra forma de abordagem seria realizar um estudo utilizando uma base de dados composta por tweets publicados antes e após o lançamento de algum produto, e analisar o sentimento dos usuários da rede social em relação as expectativas e se elas foram alcançadas.

Como foi apresentado neste trabalho, é possível explorar as técnicas de mineração de opinião para descobrir o sentimento expresso em um texto. E é um desafio desenvolver sistemas que compreendam a linguagem humana de acordo com os parâmetros de uma linguagem natural.

REFERÊNCIAS

BARBIERI, C. **BI2 - Business Intelligence**: modelagem & qualidade. Rio de Janeiro: Elsevier, 2011.

BARNEY, J. Firm resources and sustained competitive advantage. **Journal of Management**, v. 17, n. 1, p. 99-120, 1991.

BATISTA, A. **Xiaomi volta ao Brasil e lança aparelhos inteligentes para casa conectada**. TECHTUDO; 2019. Disponível em: <<https://www.techtudo.com.br/noticias/2019/05/xiaomi-volta-ao-brasil-e-lanca-lampada-de-cabeceira-e-sensor-inteligentes.ghhtml>>. Acesso em: 28 nov. 2019.

BECKER, K., TUMITAN, D. Introdução à mineração de opiniões: conceitos, aplicações e desafios. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS. 2013. [Anais...]. [S.l.:s.n.], 2013.

CAMILO, C. O.; SILVA, J. C. **Mineração de dados**: conceitos, tarefas, métodos e ferramentas: relatório técnico. Goiânia, 2009.

CHOWDHURY, G. G. Natural language processing. Annual review of information science and technology. **Wiley Online Library**, v. 37, n. 1, p. 51-89, 2003.

COBRA, M.; **Administração de marketing no Brasil**. 3. ed. Rio de Janeiro: Elsevier, 2009.

DUARTE, E. S. **Sentiment analysis on twitter for the portuguese language**. 2013. Dissertação (Mestrado em Engenharia Informática) - Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, 2013.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P.; **From Data Mining to Knowledge Discovery in Databases**. [S.l.]: American Association for Artificial Intelligence, 1996.

FREITAS, L.; VIEIRA, R. **Exploring resources for sentiment analysis in portuguese language**. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS). 4-7 nov. 2015, Natal. **Proceedings...** [S.l.]: IEEE, 2015.

INTERNATIONAL DATA CORPORATION. **Worldwide smartphone shipments rise by 0.8% in the third quarter as huawei went full steam in China**. 2019. Disponível em: <<https://www.idc.com/getdoc.jsp?containerId=prUS45636719>>. Acesso em: 28 nov. 2019.

JACKSON, P.; MOULINIER, I. **Natural language processing for online applications**: text retrieval, extraction, and categorization. [S.l.]: John Benjamins Publishing Company, 2002.

KEMP, S. **The state of digital in april 2019**: all the numbers you need to know. WeAreSocial. 2019. Disponível em: <<https://wearesocial.com/uk/blog/2019/04/the-state-of-digital-in-april-2019-all-the-numbers-you-need-to-know>>. Acesso em: 28 maio 2019.

KEMP, S.; **Digital in 2019**: global internet use accelerates. WeAreSocial. 2019, Disponível em: <<https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates>>. Acesso em: 14 nov. 2019.

KUTNER, M. H. *et al.* **Applied linear statistical models**. [S.l.]: McGraw-Hill Irwin, 2005.

LIU, B. **Sentiment** analysis and opinion mining. [S.l.]: Morgan & Claypool Publishers, 2012.

LIU, B. Sentiment analysis and subjectivity. In: INDURKHYA, N.; DAMERAU, F. J. (ed.). **Handbook of natural language processing**. 2. ed. [S.l.:s.n.], 2010.

LOPES, L.; VIEIRA, R. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. IN: PERNA, Cristina Lopes *et al.* (org.). **Linguagens especializadas em corpora**: modos de dizer e interfaces de pesquisa. Porto Alegre: EDIPUCRS, 2010.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de. Uma introdução às support vector machines. **RITA**, v. 14, n. 2, 2007.

MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: Massachusetts Institute of Technology, 2012.

OLSON, L. D.; DELEN, D. **Advanced data mining techniques**. [S.l.]: Springer, 2008.

PANG B., LEE L. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, v. 2, 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment classification using machine learning techniques. In: HAJIC, J.; MATSUMOTO (Ed.). **Language 10**. [S.l.:s.n.], 2002.

RAISINGHANI, M. **Business Intelligence in the Digital Economy**. Hershey PA: The Idea Group, 2004.

RAJARAMAN, A.; ULLMAN, J.; **Data mining in mining of massive datasets**. Cambridge: Cambridge University Press, 2011.

SOARES, A. F. **Mineração de textos na coleta inteligente de dados na web**. 2008. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

STATCOUNTER. **Mobile vendor market share Brazil**. Disponível em: <<https://gs.statcounter.com/vendor-market-share/mobile/brazil/#monthly-201901-201911>>. Acesso em: 28 nov. 2019.

TAN, N. P.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Boston: Pearson Addison Wesley, 2006.

TSYTSARAU, M.; PALPANAS, T. **Survey on mining subjective data on the web**. Data Mining and Knowledge Discovery. 2012.

TURBAN, E. *et al.* **Business Intelligence**: um enfoque gerencial para a inteligência do negócio. Porto Alegre: Bookman, 2009.

TWITTER. **Glossário**. Disponível em: <<https://help.twitter.com/pt/glossary>>. Acesso em: 20 jun. 2019.

TWITTER. **Perguntas frequentes de novos usuários**. Disponível em: <<https://help.twitter.com/pt/new-user-faq>>. Acesso em: 20 jun. 2019.

TWITTER. **Plataforma de desenvolvedor do Twitter para acesso da API**. Disponível em: <<https://developer.twitter.com/>>. Acesso em: 15 ago. 2018.

TWITTER. **Streaming API Documentation**. Disponível em: <http://dev.twitter.com/pages/streaming_api>. Acesso em: 19 ago. 2019.

VAPNIK, V. The support vector method of function estimation. In: SUYKENS, J.; VANDEWALLE, J. (Ed.). **Nonlinear Modeling**. [S.l.]: Springer US, 1998.

VELOSO, T. **iPhone e Xiaomi lideram buscas por celulares na Black Friday**. TechTudo. 2019. Disponível em: <<https://www.techtudo.com.br/noticias/2019/11/iphone-e-xiaomi-lideram-buscas-por-celulares-na-black-friday.ghtml>>. Acesso em: 28 nov. 2019.

WASSERMAN, S.; FAUST, K. **Social network analysis**: methods and applications. Cambridge, UK: Cambridge University Press, 1994.

WU, X. *et al.* **Top 10 algorithms in data mining**. Knowledge Information Systems, 2008.

XIAOMI. **About**. Disponível em: <<https://www.mi.com/br/about/>>. Acesso em: 28 nov. 2019.

ZHENG, A.; CASARI, A. **Feature engineering for machine learning**: principles and techniques for data scientists. [S.l.]: O'Reilly Media, 2018.

ANEXO A - Repositório de código

Esse anexo possui o endereço do *github* que contém o repositório de código realizado para elaboração deste trabalho.

Endereço: <https://github.com/malufreitas/mineracao-de-opinioao>.