ANTONIO MENEGHETTI FACULDADE - AMF

CURSO DE SISTEMAS DE INFORMAÇÃO

JULIANA XAVIER

O USO DA ANÁLISE DE SENTIMENTOS NO TWITTER PARA AVALIAR A OPINIÃO DO PÚBLICO CONSUMIDOR A RESPEITO DO SISTEMA OPERACIONAL MOBILE ANDROID 10

RESTINGA SECA/RS

2019

JULIANA XAVIER

O USO DA ANÁLISE DE SENTIMENTOS NO TWITTER PARA AVALIAR A OPINIÃO DO PÚBLICO CONSUMIDOR A RESPEITO DO SISTEMA OPERACIONAL MOBILE ANDROID 10

Trabalho de Conclusão do Curso de Sistemas de Informação para obtenção do título em Bacharel de Sistemas de Informação.

Orientador: Profº. Ms. Fábio Sarturi Prass

Coorientador: Profº. Ms. João Otávio Cadó de Matos

RESTINGA SECA/RS

2019

FACULDADE ANTONIO MENEGHETTI

Juliana Xavier

O USO DA ANÁLISE DE SENTIMENTOS NO TWITTER PARA AVALIAR A OPINIÃO DO PÚBLICO CONSUMIDOR A RESPEITO DO SISTEMA OPERACIONAL MOBILE ANDROID 10

Trabalho de Conclusão de Curso-Monografia, apresentado como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação, Curso de Graduação em Sistemas de Informação, Faculdade AntonioMeneghetti - AMF.

Orientador: Prof. Ms. Fabio Sarturi Prass

Co-Orientador: Prof. Ms. João Otávio Cadó de Matos

Prof. Ms. Fabio Sarturi Prass
Orientador do Trabalho de Conclusão de Curso
Antonio Meneghetti Faculdade

João Otávio C. de Matos
Prof. Ms. João Otávio Cadó de Matos
Co-Orientador do Trabalho de Conclusão de Curso
Antonio Meneghetti Faculdade

Prof. Dr. Felipe Becker Nunes
Membro da Banca Examinadora
Antonio Meneghetti Faculdade

Prof. Ms. Pablo Chitolina
Membro da Banca Examinadora
Antonio Meneghetti Faculdade

Restinga Seca, RS, 10 de dezembro de 2019

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter-me concedido força ao longo do processo de minha graduação. Chegar ao término deste ciclo exigiu muito sacrifício e esforço pessoal, portanto, agradeço a minha tenacidade até o presente momento.

Agradeço também a minha família, pelo apoio incondicional e por acreditarem na minha capacidade, sobretudo à minha mãe, por estar presente me dando suporte, principalmente ao final desta árdua jornada.

Ao meu namorado, que acompanhou minha trajetória acadêmica desde o início e por ser meu grande incentivador no decorrer dela.

Aos meus amigos, que foram extremamente pacienciosos durante minha ausência no período de escrita do TCC e por me incentivarem constantemente.

Aos meus orientadores, que contribuíram e fundamentaram cada etapa de minha pesquisa.

A todos os professores que proporcionaram minha formação com excelência.

RESUMO

O crescente avanço do desenvolvimento da tecnologia e o consequente aumento da utilização das chamadas mídias sociais contribuem para uma ampla produção de dados na *Web*. Visto que os usuários da Internet utilizam tais mídias sociais como meio de divulgação para criar, compartilhar e expressar suas opiniões referentes a determinados produtos e/ou serviços, estas tornaram-se fonte valiosa de informação para as organizações. A partir da identificação de padrões de sentimentos por meio do emprego de técnicas computacionais aplicadas sobre o conteúdo, as organizações dispõem de uma oportunidade de investigar a opinião dos seus consumidores sobre sua marca e principalmente utilizar as métricas geradas como guia para tomadas de decisões futuras mais assertivas. Diante desse cenário, o presente trabalho apresenta a aplicação de técnicas e ferramentas da área de Processamento de Linguagem Natural (PLN) para extrair significado de textos publicados dentro da rede social Twitter. Baseou-se em técnicas de mineração de texto para executar o processo de extração de dados, juntamente com a prática do tratamento dos mesmos e por fim, a exploração de algoritmos de classificação de sentimentos, capazes de avaliar a polaridade dos textos Ao final desse processo, por meio da solução de *Business Intelligence* construída, apresentou-se a avaliação dos resultados, referente aos valores extraídos da análise de sentimentos do estudo de caso realizado. Neste caso, o produto escolhido foi o sistema operacional mobile Android 10. Obteve-se, a partir dos testes com a biblioteca *Pattern* utilizada em PLN, um nível de acurácia de 85% para o modelo de classificação de polaridade utilizado. Por fim, os resultados retornados a partir da análise de sentimento da opinião dos consumidores do Android 10 revelam um sentimento majoritariamente positivo do público em relação ao produto, visto que, dos *tweets* classificados, mais de 75% foram considerados positivos pelo modelo de classificação adotado. Ademais, são demonstradas com detalhamento as ferramentas que auxiliaram o desenvolvimento desta técnica através da utilização da linguagem de programação Python, bem como as bibliotecas empregadas para análise de dados.

Palavras-chave: Processamento de Linguagem Natural. Análise de Sentimentos. Business Intelligence. Twitter.

ABSTRACT

The increasing advancement of technology development and the consequent increase in the use of so-called social media contributes to a wide production of data on the Web. Since Internet users use such social media as a means of dissemination to create, share and express their opinions regarding certain products and/or services have become a valuable source of information for organizations. By identifying sentiment patterns using computational techniques applied to content, organizations have an opportunity to investigate their consumers' opinion about their brand and specially to use the generated metrics as a guide for more future decision making assertive. Given this scenario, this paper presents the application of techniques and tools from the Natural Language Processing (PLN) area to extract meaning from texts published within the Twitter social network. It was based on text mining techniques to perform the data extraction process, along with the practice of data processing and, finally, the exploration of sentiment classification algorithms, capable of evaluating text polarity. Using the built-in Business Intelligence solution, the results were evaluated, referring to the values extracted from the sentiment analysis of the case study. In this case, the product chosen was the mobile operating system Android 10. It was obtained from the tests with the Pattern library used in PLN, an accuracy level of 85% for the polarity classification model used. Finally, the results returned from the Android 10 consumer sentiment analysis reveal a mostly positive public sentiment regarding the product, since of the rated *tweets*, more than 75% were considered positive by the adopted rating model. In addition, the tools that aided the development of this technique using the Python programming language, as well as the libraries used for data analysis, are demonstrated in detail.

Keywords: Natural Language Processing. Sentiment Analysis. Business Intelligence. Twitter.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ANSI - American National Standard Institute

API – Application Programming Interface

BA – Business Analitycs

BI – Business Intelligence

BI&A - Business Intelligence& Analytics

CSV – Comma-Separated Values

IDE - Integrated Development Environment

IIoT –Industrial Internet of Things

IoT – Internet of Things

JSON - JavaScript Object Notation

NLTK – Natural Language Toolkit

PLN – Natural Language Processing

POST - Part-Of-Speech Tagging

RE – Regular Expressions

REST - Representational State Transfer

SO – Operating System

XML – Extensible Markup Language

SUMÁRIO

1. INTRODUÇÃO

Os recentes avanços tecnológicos, seja em nível de infraestrutura de rede, hardware ou software, corroboraram para que o acesso da população à tecnologia fosse facilitado, e esse fato constitui um dos principais promotores do aumento da quantidade de dados gerados ao longo da última década. Conforme Marquesone (2016, p.18), “a internet foi e continua sendo um dos fatores mais influentes no crescimento dos dados”, diante dessa explosão de quantidade de dados aumentados, o uso do termo Big Data ganhou força nos últimos anos.

Esse montante de dados disponíveis, por sua vez, pode se tornar informação valiosa para determinada organização. Entretanto, essa criação de valor a partir dos dados só ocorre se os mesmos forem “[...] tratados, analisados e usados para a tomada de decisões” (TAURION; 2013, p. 31). Para tanto, faz-se necessária uma solução tangível, de tal forma que haja aproveitamento dos resultados e agregação de valor ao negócio da empresa (PEREIRA, 2016, p. 18).

Devido a infinidade demídias sociais disponíveis, as pessoas estão constantemente produzindo diversos tipos de informações compartilhadas através destas. Essa popularidade das redes sociais online, faz-se estrategicamente útil para que as organizações possam entender a real necessidade de seus consumidores, descubram rumores acerca de determinados produtos, tracem os perfis dos clientes, dentre outras estratégias para estudar e conhecer melhor o seu público e ainda, prospectar novos clientes (COSTA et al., 2012).

Este cenário, demonstra que o vasto uso das redes sociais gera novas possibilidades para as empresas. Através da aplicabilidade de um modelo de análise de sentimentos dos textos, das postagens dos usuários de determinada rede social, obtém a capacidade de identificar se determinada afirmação é positiva, negativa ou neutra (COSTA et al., 2012), cujo objetivo desta é avaliar melhor a opinião do público consumidor.

Inúmeras são as organizações que implementaram esta abordagem, com intuito de transformar, uma parte, das muitas informações coletadas interna e externamente em ferramentas de inteligência competitiva, como é o caso da Netflix, que utiliza um algoritmo para mapear o comportamento dos seus clientes, e assim entendê-los para que possam oferecer, assertividade de conteúdo, conforme o perfil do cliente (SIMPLEZ, 2016).

Tais soluções, de Business Intelligence (BI), Analitycs (BA) e Análise de Sentimentos, quando bem incorporadas, contribuem fortemente na tomada de decisões, sejam relacionadas ao aperfeiçoamento de produtos, à mudança de modelos de negócios, à descoberta de novos nichos de mercado, à fidelização de clientes antigos ou à captura de novos clientes.

Uma das plataformas de redes sociais que possui razoável popularidade mundial é o Twitter. Este foi lançado no ano de 2006com uma premissa um tanto quanto diferenciada de outras redes sociais, aos usuários é permitido realizar postagens, chamados de *tweets*, com no máximo 140 caracteres. Devido a limitação da quantidade de caracteres, geralmente os *tweets* servem para expressar sentimentos, pensamentos e ideias em tempo real, porém apesar desta restrição, o aplicativo mantém uma média mensal de adeptos, até o primeiro trimestre de 2019 possuía em média, mais de 330 milhões de usuários ativos (STATISTA, 2019).

Por fim, Russel (2013, p.31, tradução livre para o Português) afirma que os dados do Twitter "[...] são particularmente interessantes pois estes acontecem na velocidade do pensamento e estão disponíveis para consumo imediatamente." Portanto dada rapidez da propagação do pensamento, fontes como o Twitter, geram uma volumetria de dados diariamente absurdas, e empresas capazes de implantar soluções combinadas de *Business Intelligence e Analitycs* (BI&A) com análise de sentimentos são capazes de realizar estudos mais aprofundados acerca destes, de forma que se tornam fonte valiosa para as organizações entenderem a visão do cliente sobre sua marca ou produto.

1.1 OBJETIVOS

1.1.1 Objetivo principal

O objetivo principal deste trabalho consiste em avaliar a opinião do público consumidor a respeito do sistema operacional *mobile* Android 10 por meio de postagens coletadas na rede social Twitter. Para isso, pretende-se desenvolver uma solução que proveja a análise de sentimentos por meio de técnicas de processamento de linguagem natural (PLN).

Trata-se, portanto, de um estudo de caso que, quando generalizado, encaixa-se como uma solução dentro das áreas de BI, Big Data e Analytics, visto que tal abordagem proporciona

às empresas subsídios para a tomada de decisões mais assertivas acerca de determinado produto e/ou serviço ofertado no mercado.

1.1.2 Objetivos específicos

Dentre os objetivos específicos deste trabalho, destacam-se:

- Coleta dos dados textuais (*tweets*) referente ao sistema operacional mobile Android 10 desde a data de seu lançamento e liberação para uso;
- Pré-processamento dos dados coletados;
- Aplicar a análise de sentimentos sobre os *tweets* coletados utilizando uma implementação da função de classificação de polaridade e subjetividade *Sentiment* da biblioteca *Pattern* da linguagem de programação Python;
- Análise e conclusão dos resultados alcançados com a análise de sentimentos.

1.2 JUSTIFICATIVA

O fluxo constante de uma massiva geração de dados, ao longo da última década, principalmente através de redes sociais, faz com que estas sejam visadas como objetos de estudos. Para as empresas, a aplicabilidade da análise de sentimentos do conteúdo que trafega pelas redes, torna-se mais relevante, pois nessas, os usuários expressam opiniões, ficam entusiasmados e até mesmo ocorrem decepções, sobre os mais diversos assuntos, produtos e serviços.

Portanto, no que diz respeito a técnica de análise de sentimentos, é possível extrair informação de valor, através dos dados coletados das redes sociais, dado que estes podem exprimir como as pessoas estão respondendo a uma empresa, produto ou determinado tópico. A análise de sentimentos procura associar “automaticamente”, uma determinada parte do texto a uma pontuação de sentimentos, pontuação esta, que pode ser positiva, negativa ou neutra (KUMAR et al., 2013).

Com intuito de tornar os dados supracitados, mais facilmente compreensíveis, principalmente em ambientes corporativos, técnicas de Business Analytics e Intelligence, combinadas com análise de sentimentos, além de darem sentido à dimensão do big data, cooperam para que seja possível identificar aspectos de expressões geradas por usuários,

atribuindo polaridade de sentimentos e extraindo resultados valiosos sobre determinados produtos ou serviços.

Com a evolução tecnológica e a necessidade de facilidade da comunicação entre a sociedade, os dispositivos móveis foram tornando-se gradativamente mais populares, entre o mais utilizados destacam-se os telefones celulares ou também conhecidos como smartphones, de acordo estatísticas de uso do telefone celular, as pessoas acabam “gastando”, em média, cerca de 2h e 51 minutos do seu dia, usando seus smartphones (BANK MY CELL, 2019).

Este uso constante, acaba frequentemente gerando a necessidade de expansão dos dispositivos eletrônicos, seja a nível de arquiteturas de software e hardware, como é o caso dos seus sistemas operacionais (SO). Assim, no mercado mundial, existe uma gigantesca concorrência, de empresas criadoras de aplicativos para dispositivos móveis, a fim de fornecer vantagens como usabilidade, performance, interface amigável e funcionalidades completas e robustas, tornando assim, a experiência para os usuários finais, o mais agradável, tanto quanto seja exequível. (COSTA E FILHO, 2013).

Diante das vastas opções disponíveis, o SO para dispositivos móveis que se destaca como empresa dominante de participação no mercado é o Android, onde, de acordo com a IDC (2019, tradução livre para o Português), até junho deste ano, a marca é detentora de mais de 85 (oitenta e cinco) por cento do mercado mundial.

Assim, a motivação para o presente trabalho, propõe classificar os *tweets* de acordo com o sentimento expresso no aplicativo Twitter, pelos usuários do novo sistema operacional lançado pela Google, no início do terceiro trimestre deste ano, o Android 10. Além disso, utilizando ferramentas de processamento de linguagem natural, analisar-se-á o conteúdo expresso sobre este sistema operacional mobile na rede social, reforçando assim, a importância da iniciativa estratégica da análise da opinião dos consumidores para as empresas, a fim de colaborar para as futuras tomadas de decisões desta.

2. REFERENCIAL TEÓRICO

Nesta seção, o objetivo é concretizar a fundamentação teórica do trabalho, descrevendo os principais temas que envolvem a solução apresentada. Dessa forma, a seção encontra-se dividida em subseções contendo cada uma das referidas temáticas.

Na subseção 2.1 são introduzidos os conceitos de Indústria 4.0, *Big Data*, IoT (do inglês Internet of Things) e *Business Intelligence&Analytics* (BI&A), bem como a relação existente entre esses campos de estudo. Em seguida, a subseção 2.2 traz a teoria acerca da análise de sentimentos, abordando o processamento de linguagem natural e os principais algoritmos adotados nessa abordagem.

Por fim, a última subseção descreve as ferramentas e tecnologias utilizadas durante o desenvolvimento do trabalho.

2.1 INDÚSTRIA 4.0

No processo industrial, o descobrimento e início de uso de motores a vapor no final do século XVIII inovou a produção, definindo assim, a Primeira Revolução Industrial. Já na segunda metade do século XIX, surgiram avanços tecnológicos ainda maiores e o aperfeiçoamento de tecnologias já existentes na primeira fase, a principal delas é a utilização da energia elétrica para produção em massa. Logo após a metade do século XX, o uso de eletrônicos e da tecnologia da informação instaurou-se nas indústrias, caracterizando assim, a terceira fase da Revolução Industrial, que ficou também conhecida por Revolução Tecnocientífica (PAMPLONA, 2018, p.13).

A Indústria 4.0 é considerada a Quarta Revolução Industrial. O conceito básico de Indústria 4.0 foi usado pela primeira vez, na Feira de Hannover (Alemanha), no ano de 2011. A ideia principal é instigar os potenciais de novas tecnologias, tais como: viabilidade e uso da Internet das Coisas (IoT), a integração dos processos técnicos e de negócios, mapeamento digital e virtualização do mundo real, instauração da fábrica “inteligente” através da combinação da tecnologia com a máquina, produzindo assim também, produtos “inteligentes” (ROJKO, 2018, p.80).

Por tratar-se de vários sistemas agrupados, que são caracterizados pela ligação entre os ativos físicos industriais e as tecnologias digitais, a Indústria 4.0 fornece, através de estruturas claras e arquiteturas de redes inteligentes, controle e conhecimento total das diversas ações que ocorrem na linha de produção, ou por assim dizer, no "chão da fábrica" (EXAME, 2018).

A partir desta comunicação entre a manufatura e o processo digital, as próprias máquinas, baseadas em dados coletados através dos dispositivos da Internet Industrial das Coisas (IIoT), possuem a capacidade de fundamentar tomadas de decisões, com intuito de aperfeiçoar o processo da cadeia produtiva da indústria, ou até mesmo reduzir os custos.

Esses são somente alguns benefícios do investimento na Indústria 4.0, Rojko (2018, p. 80, tradução livre para o Português) destaca que:

> "Há também uma série de outras vantagens e razões para a adoção deste conceito, incluindo: (1) menor tempo de colocação no mercado para os novos produtos; (2) melhor capacidade de resposta do cliente, (3) permitindo uma produção em massa personalizada sem aumento dos custos gerais de produção; (4) ambiente de trabalho mais flexível e amigável, e (5) uso mais eficiente de recursos naturais e energia."

Ou seja, no que diz respeito a manufatura das fábricas, inúmeros são os ganhos ao implementar a estruturação da Indústria 4.0, chamada revolução das máquinas e dos processos inteligentes, que combinam uma série de tecnologias digitais, tais como *machine learning*, IoT, inteligência artificial, *big data*, robótica, entre outras (EXAME, 2018, p.36).

> "Entretanto, este conceito de Indústria 4.0, não se limita somente ao processo de manufatura, mas diz respeito, também, a agregação de valor a uma cadeia de clientes, fornecedores, colaboradores e ainda, a todas as funções e serviços de negócios da empresa." Rojko (2018, p. 87, tradução livre para o Português).

Conforme mencionado anteriormente, uma nova possibilidade, um novo modelo de tomada de decisão surge com a Indústria 4.0: "a conexão completa e total do processo produtivo através da IoT e a IIoT, que permite a aquisição de dados em alto volume, alta velocidade e grande variedade" (AUTOMAÇÃO INDUSTRIAL, 2017).

Assim, como os conceitos do *Big Data* progressivamente expandem-se ao ambiente industrial, a evolução natural do BI, o *Business Analitycs*(BA) surge para se obter respostas

estatísticas avançadas e quantitativas, proporcionalmente ao volume das amostras coletadas, através de criação de modelos de análises preditivas, que irão precisar os resultados da indústria com base nas tendências do negócio para o futuro.(DELFINO, 2018).

2.1.1 Big Data

Gigantesca é a volumetria e a rapidez com que os dados são gerados ao longo das últimas duas décadas. Os amontoados de dados disponíveis digitalmente provêm de diversas fontes como sensores, biometria para identificação, sinais de GPS, mas, principalmente, por meio das chamadas mídias sociais como WhatsApp, Facebook, Twitter e Instagram.

O cenário descrito acima ocorre devido ao acesso constante e instantâneo de grande parte da população a diversos canais de comunicação, seja por meio de dispositivos móveis, como celulares e *tablets*, ou por meio dos tradicionais PCs, que podem ser utilizados e conectados de qualquer lugar do mundo (TAURION; 2013, p.8-19). Essa é uma das compreensões acerca do conceito de *Internet Of Things* (IoT), ou seja, a interconexão entre dispositivos por meio na rede mundial de computadores, possibilitando que estes sejam remotamente manuseados por humanos ou até mesmo por outros dispositivos.

Segundo reportagem publicada na revista Computação Brasil (2015, p.6, 7), essa integração de objetos (IoT), sejam eles físicos ou virtuais, conectados através da *internet* permite a coleta e a troca de informações e, ainda, o armazenamento em nuvem de um volume de dados em escala inimaginável, de tal forma que se estima 44 zetabytes (trilhões de gigabytes) de dados manipulados diariamente no mundo até o ano de 2020. Esse é um dos principais motivos que tornam os termos *big data* e *analytics* indispensáveis na discussão das organizações sobre como transformar dados em *insights*, ou seja, em informações relevantes que podem ser usufruídas eficientemente para a tomada de decisões.

Antes de tudo, o conceito de *big data* deve ser claramente explanado, pois este é o primórdio para as demais abordagens que serão descritas neste trabalho. É difícil obter uma definição única sobre o termo, mas este geralmente é associado ao armazenamento e consequente processamento de uma abundância de dados que são gerados diariamente pelas mais diversas fontes.

Porém, não é somente de volume que se fala. Hurwitz et al. (2013, p.16) afirma que *big data* é "[...] a capacidade de gerenciar um grande volume de diferentes dados, na velocidade e dentro do prazo certos, a fim de permitir a análise e reação em tempo real". Tipicamente, três são as características que foram formalizadas e associadas ao termo (IBM, 2017):

- Volume: trata-se da quantidade de dados produzidos. Atualmente, as pessoas estão mais conectadas do que já estiveram antes, esta interconexão gera ainda mais fontes de dados.
- Velocidade: este diz respeito a rápida geração de dados e tratamento em tempo real. A velocidade com que os dados acabam chegando a determinada empresa estão aumentando. Isso ocorre também devido aos avanços da tecnologia de rede, o que, na maioria das vezes, dificulta o acompanhamento para obtenção e extração de valor desses.
- Variedade: mais fontes de dados significa mais volume de dados em diferentes formatos. Essas informações são extraídas de diversas fontes, gerando assim dados estruturados, semiestruturados ou não estruturados.

É comum alguns autores e pesquisadores adotarem o conceito dos 5 "Vs", adicionando os termos veracidade e valor aos 3 "Vs" já apresentados. Segundo Erl et al. (2015, p. 31-32):

- Veracidade: refere-se à qualidade ou fidelidade dos dados, ou seja, à confiabilidade dos mesmos. Esta está relacionada à qualidade dos dados.
- Valor: está diretamente ligado à utilidade dos dados para uma empresa, ou seja, o quão significativa e valiosa uma informação pode ser para determinada solução. Neste caso, dados não tratados, que mantém uma baixa qualidade, geram a criação de informação sem valor, a qual pode conduzir a uma tomada de decisão incorreta.

As proporções de quantidade e acúmulo de dados que circulam a partir das atividades executadas na internet são colossais. Para explanar esta vasta evolução de dados, no infográfico apresentado na Figura 1, Lori Lewis e Chadd Callahan, da empresa Cumulus Media, ilustram o

que acontece em 60 (sessenta) segundos de atividade na internet de bilhões de pessoas globalmente.

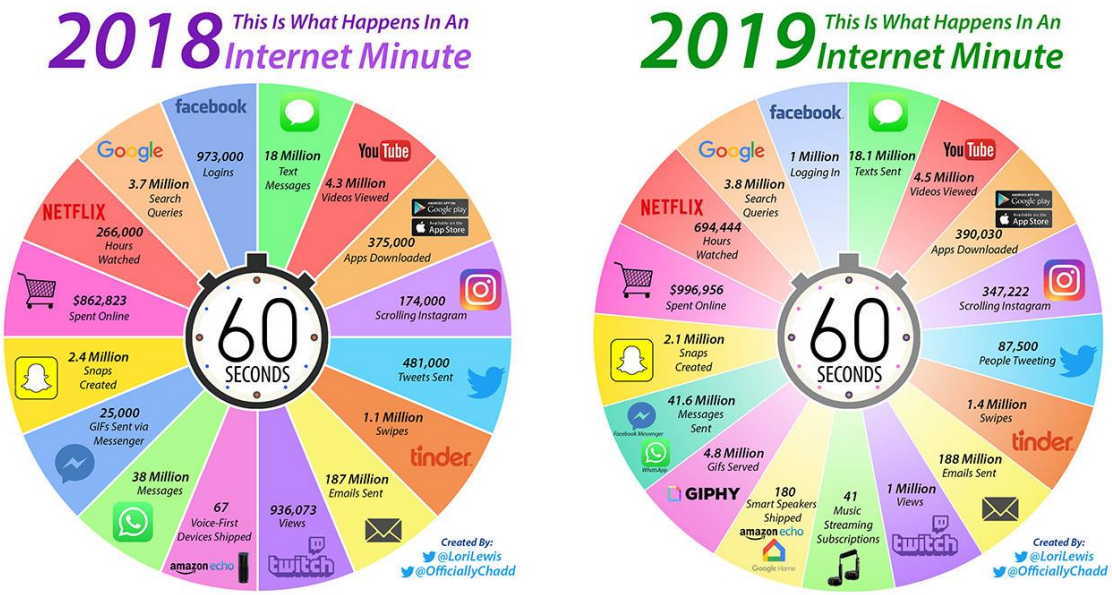

Figura 1. Comparação "Isto é o que acontece em um minuto na Internet".
Fonte: AllAcess (2019)

A análise da Figura 1 revela que o uso de determinadas plataformas cresce de maneira alucinante, como é o caso do Instagram e Netflix, onde o índice apresentado mostra que houve o dobro de rolagem no *feed* de notícias e quase o triplo de horas de *streaming* de vídeos assistidos, respectivamente. Diante desse contexto, pode-se afirmar que "a evolução da tecnologia proporciona que novas fontes de dados estruturados sejam produzidas - geralmente em tempo real e em grandes volumes." (HURTWITZ et al., 2013, p.50).

Apesar dessa enorme quantidade de dados gerados diariamente nas redes sociais, para as empresas, que necessitam constantemente relacionar a fim de tomar decisões gerenciais estratégicas para manterem-se no mercado concorrencial, buscar informações dessa "montanha" de dados pode tornar-se um desafio, visto que se trata de uma tarefa difícil e complexa.

De acordo com Turban et al. (2009, p.21):

> "Tomar essas decisões pode exigir quantidades consideráveis dedados oportunos e relevantes, além de informações e conhecimento. O processamento dessas informações,

> na estrutura das decisões necessárias, deve ser feito de forma rápida, com frequência em tempo real, e comumente exige algum apoio computadorizado.”

Diante desse cenário corporativo supracitado, *Business Intelligence&Analytics* (BI&A) surgem como uma abordagem para auxiliar na resolução de problemas de negócios, ou seja, como uma forma de apoio a melhores tomadas de decisões. O termo BI&A é associado a qualquer atividade, ferramenta ou processo que é utilizado para obter a melhor informação que dará suporte ao processo de tomada de decisões dentro de uma organização. (SCHEPS; 2008, p.35).

As redes sociais, por serem de grande utilização e apresentarem grande volume de tráfego de dados, expressões de opiniões e sentimentos sobre entidades, eventos e suas propriedades em tempo real, tornam-se as principais fontes de dados para a aplicação de BI&A por parte das organizações, a fim de que possam obter significativos *insights* a partir de seus dados operacionais (SCHEPS; 2008, p.41), identificando opiniões e até mesmo predizendo tendências, o que permite atuar diretamente na tomada de decisões.

O volume de dados sobre determinada empresa ou produto que pode ser coletado, organizado e visualizado através de várias técnicas de mineração de texto é imenso (CHEN et al., 2012, p. 1167). Após executada a coleta, as empresas focam na localização das informações de maior relevância dentre essa imensa quantidade, a fim de aplicar a análise de sentimentos e, assim, obter uma opinião geral das pessoas sobre seus produtos e serviços.

De acordo com Liu (2012), a análise de sentimentos é o campo de estudos mais ativo do ramo de processamento de linguagem natural (NLP). O autor ainda explica que esta área de estudo tem como foco principal identificar as opiniões, sentimentos, avaliações e atitudes das pessoas em relação a uma determinada entidade, analisá-los e classificá-los como positivo, negativo ou neutro.

Tal cenário contribui para que, cada vez mais, as organizações invistam em métodos e técnicas de análise de sentimentos, focadas nas redes sociais, que realizam o monitoramento dos seus produtos, marcas ou serviços, a fim de agregar vantagem competitiva à empresa, identificando pontos fracos e fortes, bem como potenciais produtos, através da opinião do seu público consumidor.

2.1.2 Business Intelligence

Decisões são o que movem as organizações. Uma decisão positiva, que seja tomada em um momento crítico em uma empresa, pode tornar uma operação mais eficiente, talvez um cliente mais satisfeito, ou até mesmo um negócio mais rentável (SCHEPS, 2008, p. 33). Neste ponto, de tomada de decisão, que o conceito de Business Intelligence (BI) se encaixa. De modo geral, pode-se pensar nesse termo como o uso de dados sobre o ontem e o hoje, que irão fazer com que melhores decisões sejam tomadas amanhã.

De modo mais específico, segundo Bentley (2017, p. 10, tradução livre para o Português), "BI pode ser descrito como um conjunto de técnicas e ferramentas para a aquisição e transformação de dados brutos em informações úteis e significativas para fins de análise de negócios", isto é, BI engloba desde intranets, sistemas distribuídos, processamento analítico online, até arquiteturas, *data warehouses*, metodologias e instrumentos que podem ser ambientados em uma única suíte de *software*.

Através deste conjunto de tecnologias, os analistas corporativos têm poder para cruzar informações e aprofundarem-se nos indicadores de performance de determinado negócio, contribuindo assim para próxima tomada de decisão. Trata-se das organizações utilizarem as ferramentas de transformação do BI, reunirem esse amontoado de dados disponíveis, e, assim, realizarem processamento em cima deste, capacitando-os a tomar as melhores decisões sobre determinado produto ou serviço tão rápido quanto seja possível.

Quer dizer, constantemente, as organizações estão em busca de resultados positivos que as mantenham competitivas no mercado, estas buscam aderência do consumidor a sua solução de negócio, para tanto, a coleta de informações de qualidade, inovação, otimização dos processos, assertividade nas decisões, *insights* valiosos, alcance das metas, entre outras ações, podem garantir este resultado para as empresas (SALDANHA, 2018, p.28), portanto, a aplicabilidade de uma estratégia de BI pode trazer diversas vantagens a uma organização, quando bem implementada.

Como é de conhecimento, toda alternativa de estratégia apresenta riscos e, também, possíveis recompensas. Nesse sentido, a alternativa de estratégia de BI deve estar alinhada com a estrutura e a cultura da organização, ou seja, deve-se considerar a forma como a atual política da empresa e a cultura corporativa poderão ser associadas a uma potencial solução de BI, a fim de

prevenir que a implantação de tal solução acabe gerando mais problemas do que benefícios. (SCHEPS, 2008, p. 182).”

2.1.3 Business Analytics

Desde a implementação dos exercícios de gerenciamento por Frederick Winslow Taylor no final do século XIX, o *analytcis* tem sido utilizado nas áreas de negócio, porém popularizou-se mais tarde, no final dos anos 60, quando os computadores passaram a ser usado em sistemas de suporte à decisão. (BENTLEY, 2017).

De acordo com Bentley (2017, p.31, tradução livre para o Português):

> *“Business Analytics* refere-se às habilidades, tecnologias, práticas para exploração iterativa contínua e investigação do desempenho comercial anterior para obter *insights* e impulsionar o planejamento comercial. BA concentra-se no desenvolvimento de novas ideias e no entendimento do desempenho dos negócios, com base em dados e métodos estatísticos.”

Business Analytics (BA) é implementado nas organizações a fim de melhorar o desempenho dos resultados desta Trata-se de um modelo proativo, altamente técnico e com foco no futuro, a fim de prever tendências, descobrir padrões e prescrever ações para os melhores resultados. Dito isto, BA geralmente responde a perguntas como:

- O que vai acontecer no futuro?
- O que acontece se...?
- O que vem depois?

Em outras palavras, BA trata-se de um conjunto de ferramentas e métodos utilizados por analistas para que, de maneira inteligente, sejam tomadas decisões sobre o futuro do negócio. Assim, diversos tipos de análise podem ser aplicados a dados corporativos para prever, descrever e melhorar o valor dos negócios. Ainda segundo Bentley (2017, p.34):

> "[...] as áreas de análise incluem análise preditiva, prescritiva, gerenciamento de decisões corporativas, análise de varejo, grande variedade e manutenção de estoque, otimização de unidades e modelagem de marketing, análise da web, otimização e dimensionamento da força das vendas."

Portanto, acultura de tomada de decisões baseada em BA, que seja implementada em toda uma organização, faz-se de extrema importância. E para que seja possível o apoio a essa cultura, os profissionais atuantes nesta análise, precisam não apenas saber como transformar os dados brutos e informações em significante conhecimento para as organizações, mas também transmitir, interagir e comunicar esse conhecimento a área de negócio e aos especialistas envolvidos (CHEN et al., 2012, p. 1183).

2.2 ANÁLISE DE SENTIMENTOS

2.2.1 Processamento de Linguagem Natural

De acordo com Jurafsky e Martin (2008, tradução livre para o Português), processamento de linguagem natural (PLN) implica em técnicas computacionais que têm por objetivo o processamento da fala e da escrita humana como forma de linguagem. Diante disso, PLN diz respeito ao desenvolvimento de softwares que sejam eficientes no processamento de informações em linguagem natural, com intuito de extrair informações significativas do texto interpretado por estes.

Com a dependência da compreensão automática e a crescente comunicação do ser humano com o computador surgiu o processamento de linguagem natural. Além de ser um mecanismo criado para captar informações textuais, foi criado também para intermediar a entrada de dados nos sistemas e estruturar os mesmos (BULEGON E MORO, 2010).

Dentro do PLN existe ainda uma série de ramificações que correspondem a diferentes áreas de pesquisa, cada uma delas com diferentes objetivos: algumas abordam o processamento de áudios, textos transcritos, outros tipos de dados relacionados à fala ou, ainda, imagens (SANCHES, 2017). A Figura 2 ilustra, de modo geral, um fluxo de trabalho tradicional dentro de aplicações de PLN, onde tem-se a entrada dos dados, o processamento da linguagem e,

posteriormente, como saída, tem-se a geração dos resultados (estruturas), que para pesquisas ou utilizações futuras podem ou não ser armazenados em um banco de dados.

Visto que na língua portuguesa existe uma grande variação morfológica e sintática, estas, juntamente com a ambiguidade, caracterizam-se como um obstáculo determinante no processamento de linguagem natural. Portanto, a interpretação de uma sentença por um sistema computacional trata-se de uma tarefa complexa, e, para que esta seja exequível, faz-se necessária a utilização das chamadas análises linguísticas computacionais (FIORIO, 2015).

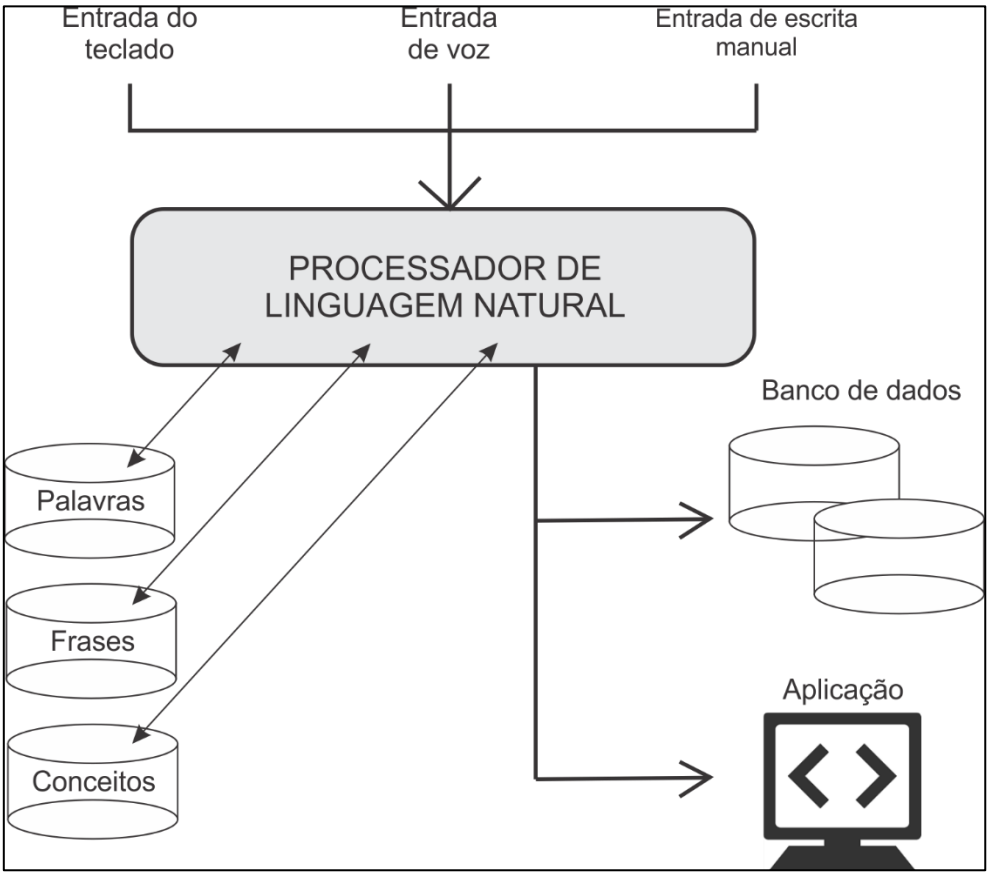

Figura 2. Etapas de PLN
Adaptado: NAGAO (2013)

2.2.1.1. Análise Léxica

O analisador Léxico é o responsável por identificar palavras e expressões isoladas em determinada sentença, ou seja, lida com a estrutura das palavras. Conforme Gonzales e Lima (2013, p.30) afirmam:

> "Na etapa da análise léxica, ocorre a conversão de uma cadeia de caracteres (o texto da consulta) em uma cadeia de palavras. Assim, o principal objetivo desta etapa é a identificação das palavras que constituem a consulta. Na fase seguinte, artigos, preposições e conjunções são candidatos naturais à lista de *stopwords*, ou seja, a serem eliminados. Então, pode ser executada a normalização lexical através de *stemming*, pois frequentemente o usuário especifica uma palavra na consulta, mas somente variações desta palavra estão presentes em um documento relevante."

O léxico mostra-se como um item de extrema importância para qualquer sistema de processamento de linguagem natural, dado que nesta fase de tratamento, os dicionários utilizados pelos sistemas de PLN devem ser apropriados para que não afetem negativamente todas as fases de processamento subsequentes (NETO et al., 2010).

É durante a análise léxica que ocorre o processo de *tokenização*, presente em grande parte das aplicações envolvendo PLN. De modo geral, a *tokenização* corresponde a uma técnica utilizada para isolar as palavras de um texto, onde cada palavra é denominada *token*. Após a explicitação dos *tokens*, pode-se prosseguir para a análise morfológica (FARINON, 2015).

2.2.1.2. Analise Morfológica

Crivelli (2011) conceitua que este tipo de análise lida com a união ou separação das ocorrências das palavras, quer dizer, a cisão do prefixo e do sufixo das mesmas, mantendo, assim, somente o seu radical, onde prefixos são elementos complementares que aparecem antes do radical e sufixos aparecem depois do radical, como por exemplo:

- Prefixo: **In**feliz
- Sufixo: Feliz**ardo**

Computacionalmente falando, o tratamento deste tipo de análise é relativamente básico, pois este fundamenta-se em regras que examinam as palavras, e assim, as classificam de acordo com as tabelas de afixos. Müller (2003, p.3) exemplifica:

> "Por exemplo, a entrada *zinho* de uma tabela de sufixos está associada a um diminutivo de um substantivo, portanto, a palavra *bonezinho* é o diminutivo da palavra *boné*, que é

> seu radical. Desta forma, são reconhecidas as palavras que não estão na sua forma padrão, já adequando-as para a fase posterior da análise sintática."

Em resumo, partes do texto podem ser simplificados através da permuta dessas variações de uma mesma palavra por sua raiz (*stem*). A remoção de afixos é comumente chamado de *stemming*. No pacote NLTK (do inglês *Natural Language Toolkit*) da linguagem de programação Python, o qual funciona como uma caixa de ferramentas computacionais voltadas para a análise de processamento de linguagem natural, existem diversas implementações para *stemmers*, onde as funções executam o processo de redução de palavras ou frases em sua forma raiz, de acordo com a necessidade do usuário, favorecendo assim o método para normalização da sentença.

2.2.1.3. Análise Sintática

Fiorio (2015, p. 22) afirma que:

> "A análise sintática faz a análise de como as regras gramaticais são combinadas de forma a gerar uma árvore, que represente a estrutura sintática que constitui a sentença avaliada, além disso, esse estilo de análise tem lugar de destaque sendo considerada uma interpretação da sintaxe."

Este tipo de análise tem por objetivo determinar a estrutura sintática ou gramatical de uma determinada frase. No contexto do processamento de linguagem natural, a definição de análise sintática está afiliada à infinidade de frases possíveis de serem modeladas, com intuito de representar certo domínio de análise (MÜLLER, 2013).

Como exemplificação desta análise, no contexto da linguística computacional, podemos citar diversas técnicas utilizadas na identificação e etiquetagem dos elementos sintáticos, bem como o arranjo da construção sintática:

1. <u>Etiquetagem (*tagging*</u>): Consiste na identificação das classes das palavras, onde determinada frase recebe uma etiqueta, ou seja, uma descrição abreviada, que representa a classe da palavra conforme exemplifica a Tabela 1.

Etiqueta	Descrição	Palavra
PPE	**P**ronome **PE**ssoal	eu
VP	**V**erbo no **P**assado	tropecei
PAF	**P**reposição + **A**rtigo **F**eminino	na
SSF	**S**ubstantivo **S**ingular **F**eminino	pedra

Tabela 1. Etiquetagem Sintática
Fonte: MÜLLER (2003)

Na linguagem de programação Python, diversas bibliotecas dispõem desta funcionalidade. No pacote *Pattern*, o módulo *Parse* identifica as sentenças, palavras e tipos de palavras em uma sequência de texto. Este processo envolve (CLIPS, 2019):

- Tokenização: quebras de frases em *tokens;*
- *Part-of-speech tagging* ou POST: anotação de palavras de acordo com a sua classe, por exemplo, verbo ou substantivo;
- *Chucking*, detecção para agrupamento de palavras consecutivas que são complementares, por exemplo: CARRO (N - Noun) esportivo (ADJ - Adjective).

Em situações onde ocorre ambiguidade, entretanto, a etiquetagem das palavras não é suficiente para realizar a análise sintática. Nesses casos, faz-se necessário recorrer a mais de um nível de sentença por meio da técnica conhecida como análise sintática profunda (*deep parsing*), a qual, conforme o próprio nome expressa, aprofunda-se na estruturação arbórea que é atribuída às sentenças. (MÜLLER, 2003; ALENCAR, 2011).

2. Árvores de Parser: Consiste da utilização de técnicas de busca em árvore a fim de definir o ajuste da construção da frase. De modo geral, realiza a comparação entre a frase em si e a estrutura da árvore.

Tomando novamente a biblioteca NLTK da linguagem de programação Python como exemplo, pode-se citar o módulo CoreNLPParser como uma ferramenta bastante eficiente para fornecer uma análise sintática completa da árvore de estrutura de uma frase. Esse módulo é escrito em Java e requer que o mesmo seja instalado no dispositivo, porém oferece suporte a diversas linguagens de programação (FREI, 2019).

2.2.2 Classificação de Polaridade

Ainda que diversos autores associam o termo análise de sentimentos a estudos mais complexos e aprofundados, este conceito geralmente diz respeito, resumidamente, à fragmentos textuais que são extraídos e submetidos a classificação com relação à sua polaridade (SANTOS L., 2015).

Análise de polaridade diz respeito a indicação da classificação da opinião do autor com relação ao objeto em discussão, isto é, consiste na verificação da opinião associada ao assunto, se esta é favorável, neutra ou negativa, nesta etapa ocorre a identificação do sentimento conectado ao texto em questão. Pela perspectiva funcional, a classificação de polaridade relaciona-se ao uso de técnicas de análise de texto, processamento de linguagem natural e linguística computacional, com objetivo de identificar, extrair e entender a subjetividade do conteúdo presente (SILVA, 2015).

Silva afirma que (2013, p.2):

> "As opiniões podem ser classificadas entre positivas, negativas ou neutras, indicando assim a polaridade do texto (a polaridade neutra ocorre quando o texto traz opiniões negativas e positivas na mesma proporção). Em geral, a polaridade de um texto é expressa por palavras opinativas (adjetivos –bom, ruim; advérbios –bem, rapidamente; e alguns substantivos –amigo, etc.)."

Já existem inúmeras ferramentas linguísticas que corroboram para o encargo automático da polaridade de determinada sentença, onde geralmente um dicionário contendo milhares palavras são pré-associados a escores numéricos, representando assim, a sua polaridade: *pos* (positivo), *neg* (negativo) ou *obj* (neutro) (SILVA, 2013).Como é o exemplo do *SentWordNet*, este trata-se de um recurso aprimorado, que foi desenvolvido com o fim de suportar aplicativos de classificação de sentimentos e de mineração de opiniões, esta ferramenta linguística, é derivada da base lexical WordNet (BACCIANELLA et al., 2010).

2.3 TECNOLOGIAS EMPREGADAS

2.3.1 Twitter

Esta rede social[1] foi criada no ano de 2006, e trata-se de um microblog, que permite aos usuários realizarem postagens em tempo real, de até no máximo 140 (cento e quarenta) caracteres, geralmente estas publicações, por serem bastante limitadas, costumam expressar ideias, pensamentos, sentimentos ou fatos que estão acontecendo na vida dos usuários em tempo real (BARBOSA, et al., 2012, p.2).

O uso do *hashtags*, ou seja, a famosa cerquilha (#) antecedendo determinada palavra-chave, serve para identificar o tema do conteúdo que as pessoas estão compartilhando nos seus *tweets* ou *retweets.* Vale destacar ainda, que diferentemente da maioria das redes sociais online, as conexões entre os usuários do Twitter são assimétricas, quer dizer, é bastante comum que um usuário siga determinado perfil, mas que este não seja seguido de volta (SANTOS L., 2015).

Para empresas, desenvolvedores e usuários, o Twitter oferece acesso aos dados através de uma Interface de Programação de Aplicativo (API), esta serve para que os programas se comuniquem e troquem informações entre si. O acesso é limitado a parte dos serviços através da API, porém, com este acesso, é permitido aos desenvolvedores, que estes criem softwares que sejam integrados ao Twitter, "[...] como, por exemplo, uma solução que ajude a empresa a medir opiniões dos clientes no Twitter" (TWITTER, 2019a).

Na REST API, fornecida pelo Twitter, os desenvolvedores são capazes de acessar atualizações de postagens, dados dos usuários, status, entre outros. Santos W. (2015, p.61) define REST API (Representational State Transfer) como:

> "[...] um estilo de arquitetura de software para sistemas hipermídia distribuído, como por exemplo, a Web do jeito que conhecemos atualmente, ou seja, onde utilizamos um navegador web para acessar recursos que estamos interessados, geralmente uma página HTML, ou um documento XML, mediante digitação de uma URL."

Além da REST API, o Twitter também oferece acesso a uma gigante massa de informações em tempo real, através da *Streaming* API, onde diferentemente da REST, em que as

[1]https://twitter.com/

solicitações são repetidas pelo aplicativo do cliente e as entregas ocorrem em lotes, uma única conexão é aberta entre o aplicativo e a API, e os resultados são enviados através desta, sempre que houver novas correspondências (TWITTER, 2019b).

Em diversos cenários, como marketing de produto, campanhas eleitorais e pesquisas, determinar os sentimentos das pessoas acerca dos mais diversificados tipos de assuntos, tornou-se uma jogada estratégica para as organizações, e uma fonte promissora de informação para este tipo de análise são as mídias sociais (BARBOSA, et al., 2012). Neste cenário, o Twitter provê dados extremamente relevantes, pois conforme mencionado anteriormente, os *tweets*, geralmente são compartilhados a fim de narrar experiências e expressar opiniões.

2.3.2 Sistemas Operacionais

De acordo com Tanenbaum e Woodhull (2008), um sistema operacional (SO) é um programa de sistema mais fundamental, que tem por tarefa controlar todos os recursos de um computador, e ainda, disponibilizar a base para os aplicativos poderem ser escritos, este dá suporte a inicialização do hardware, gerencia o escalonamento de tarefas e ainda controla os dispositivos.

A variedade de tipos de sistemas operacionais atualmente, que se encontram disponíveis, é bastante vasta, estes foram surgindo conforme a necessidade, para adaptar-se a progressão do *hardware* e as aplicações desenvolvidas especialmente para estes, de acordo com Machado e Maia (2013, p. 13):

> "Os tipos de sistemas operacionais e sua evolução estão relacionados diretamente com a evolução do hardware e das aplicações por ele suportadas. Muitos termos inicialmente introduzidos para definir conceitos e técnicas foram substituídos por outros, na tentativa de refletir uma nova maneira de interação ou processamento."

Dada esta gama de SO ofertados no mercado, os mais populares que podem se citar desenvolvidos para computadores são Windows da empresa Microsoft, MAC OS X da Apple e o Ubuntu da Canonical, este último, trata-se de código aberto, e foi construído baseado em Linux (UBUNTU, 2019). Para dispositivos móveis, os sistemas operacionais, também conhecidos como *firmwares*, mais populares englobam Android, este desenvolvido pela Google, e iOS,

desenvolvido pela empresa Apple, e executado somente em *hardware* próprio, assim como todos os SO desenvolvidos pela empresa (APPLE, 2019).

2.3.3 Sistema Operacional Android

Inicialmente, o sistema operacional Android foi desenvolvido por uma empresa denominada Android Inc., com fundação no ano de 2003, porém, a ausência de investidores acabou tornando-se um contratempo agravante para a organização, então, no ano de 2005, a Google demonstrou interesse pelo empreendimento e realizou a aquisição da companhia (SAKIS, 2015).

Após dois anos de mistério, em novembro de 2007, a Open Handset Alliance (OHA), um consórcio de empresas do ramo tecnológico liderados pelo Google, lança o projeto Android, um sistema operacional(SO) baseado em Linux e a primeira plataforma para aplicações móveis completamente livre e de código aberto (*open source*), ou seja, estrategicamente criou-se uma imensa vantagem competitiva para a companhia, visto que desenvolvedores e empresas mundiais, podem colaborar com a evolução da plataforma.(LECHETA, 2015).

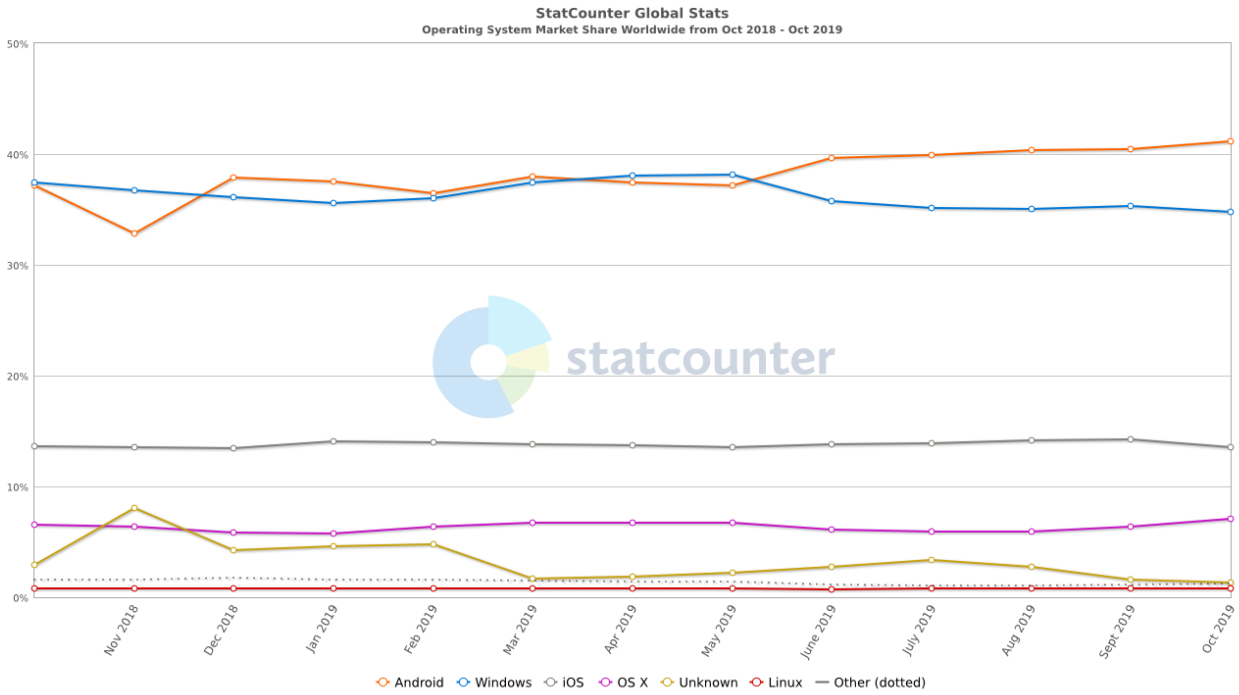

Figura 3. Participação de Sistemas Operacionais no Mercado Mundial
Fonte: StatCounter (2019)

Dito isto, um dos fatores fundamentais que contribuem para que Android comande a participação no mercado mundial, é que este SO se trata de uma plataforma de código aberto, que permite aos fabricantes de celulares usarem e adaptarem o sistema operacional para seus próprios dispositivos (IDC, 2019, tradução livre para o Português). O gráfico da Figura 3, ilustra claramente esta informação, a imagem foi extraída do site StatCounter, e exibe a porcentagem de atuação dos sistemas operacionais no mercado mundial, do período de outubro de 2018 a outubro de 2019.

É possível identificar que o Android é o sistema operacional mais apreciado mundialmente falando, este obteve um crescente aumento de fama desde outubro de 2018, onde era possuidor de 37.13% da parcela de uso rotineiro, e notoriamente cresceu para 41.19%, sendo assim o detentor do monopólio de popularidade, superando o SO *Windows* da Microsoft e tornando-se líder no mercado de sistemas operacionais.

De acordo com Morimoto (2009, p. 28):

> "Grande parte da estratégia em torno do Android é centrada no desenvolvimento de aplicativos por parte de outras empresas e programadores independentes. [...] Sendo a última empresa a entrar no mercado, o Google é quem mais está investindo pesado nessa frente, montando grande equipe de desenvolvimento, investindo em contatos com fabricante e na divulgação do sistema e incentivando a participação externa."

Uma vez que as empresas fabricantes não precisam comprar a licença para uso do Android, e não necessitam iniciar um sistema do zero pois apenas refatoram o código-fonte, reduzindo assim o tempo de desenvolvimento da solução para customização aos seus produtos, torna-se de extrema vantagem a estes utilizar esta solução, e assim, consequentemente colaborar para que a Google seja crescente potência do mercado móvel mundial (LECHETA, 2015).

Diante destas considerações, optou-se pela análise de opiniões sobre o sistema operacional Android, visto que possui um grau de popularidade elevado, e além do mais, recentemente houve o lançamento da última versão deste SO, o Android 10, quer tornou-se foco deste trabalho.

2.3.4 A Linguagem de Programação Python

Python é uma linguagem de programação dinâmica e de alto nível, que se popularizou rapidamente nos últimos anos. Por trata-se de uma linguagem de fácil aprendizado e sintaxe simples, apesar de sua robustez para realizar tarefas complexas, esta é utilizada por experientes programadores, ou por curiosos em busca de uma nova linguagem para explorar.

De acordo com Matthes (2016, p.30), *Python* geralmente é utilizada para um propósito em geral, tal como:

> "[...] criar jogos, construir aplicações web, resolver problemas de negócios e desenvolver ferramentas internas em todo tipo de empresas interessantes. Python também é intensamente usada em áreas científicas para pesquisa acadêmica e trabalhos aplicados."

Esta linguagem possui uma vasta biblioteca, que pode ser explorada a fim de realizar qualquer tarefa que o usuário desejar. Na área da análise de dados, também existem diversas bibliotecas que atendem ao objetivo do cientista de dados.

Estas fornecem diversas ferramentas para estruturas de dados, funções matemáticas de alto nível, métodos para aprendizado de *Machine Learning*, modelagem de matrizes multidimensionais e mineração de dados, este conjunto de bibliotecas, contribui para que Python seja uma linguagem relativamente sedutora para os analistas de dados.

2.3.5 PyCharm IDE

Pycharm trata-se de uma IDE (*Integrated Development Environment*) multiplataforma, desenvolvida pela companhia JetBrains. Além de fornecer ao desenvolvedor funcionalidades como análise de código, depurador gráfico, testador de unidade integrado, esta plataforma suporta desenvolvimento de web com o *framework* Django (JETBRAINS, 2019).

Este *software* é detentor de uma versão livre para estudantes, fator que corroborou para a utilização deste neste trabalho como auxílio a criação dos *scripts* referentes a coleta, pré-processamento e análise dos dados. Praticidade, facilidade de instalação de bibliotecas e uma documentação robusta das ferramentas JetBrains, também são fatores que justificam a escolha de

empregabilidade deste ambiente de desenvolvimento; bem como a já existente familiaridade da autora com este.

2.3.6 Biblioteca *Pattern* para Python

Trata-se de uma biblioteca compatível com a versão acima de 2.7 de Python, onde as funcionalidades possuem foco em *web mining* (Google, Twitter, Wikipedia, web spider, HTML DOM parser), PLN, *machine learning* e *network analytics*, sua essência principal, é disponibilizar ao usuário, uma ferramenta de fácil uso (SMEDT E DAELEMANS, 2012a).

Esta biblioteca é composta e organizada por diversos módulos que podem ser empregados em conjunto, como é o exemplo ilustrado na Figura 4 que exemplifica um fluxo de trabalho, onde inicialmente o texto é minerado da Wikipedia (*pattern.web)*, em seguida este pode ser analisado pelo part-of-speech tagging (ou popularmente conhecido por POST, módulo *pattern.text*), após sua sintaxe e semântica são consultadas (*pattern*.search), e por fim, os dados são utilizados para treinar um classificador (*pattern.vector*).

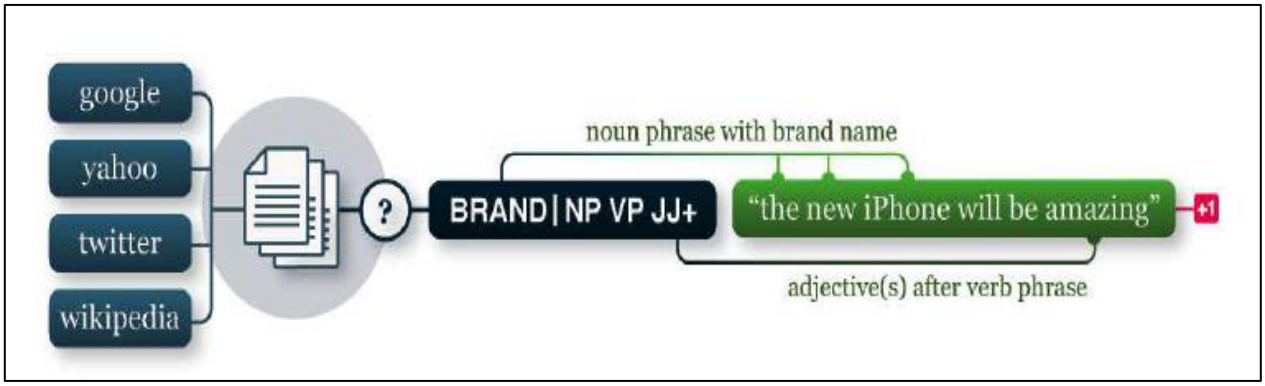

Figura 4. Fluxo de Trabalho da biblioteca *Pattern*
Fonte: SMEDT E DAELEMANS (2012a)

Diante disto, pode-se concluir que esta biblioteca faz-se uma das mais completas a serem exploradas no processamento de linguagem natural na linguagem de programação Python, dado sua funcionalidade multiuso para diversas tarefas, tais como Data Mining (disponibiliza APIs para minerar dados de mídias sociais e sites), Machine Learning (contém modelos de aprendizado máquina que podem ser utilizados para classificação, regressão e clusterização) e PLN (executando tarefas de tokenização, POST, análise de sentimentos, entre outros.) (MALIK, 2019)

2.3.7 Natural Language Toolkit (NLTK)

O foco deste framework é reforçar a criação de programas na linguagem de programação Python, que trabalhem com dados da linguagem humana2. Trata-se de uma ferramenta robusta para processamento de linguagem natura, pois compõe-se de módulos, conjuntos de dados, tutoriais e exercícios, além de abranger o processamento de linguagem simbólica e estatística (LOPER; BIRD, 2012).

Um conjunto de dados pré rotulados em suas respectivas categorias, é chamado de corpus. O NLTK disponibiliza alguns *corpus* para treinamento do classificador, Magano (2016) exemplifica:

> "Dentre os diferentes *corpus* existentes, o que mais se assemelha ao domínio deste trabalho é o movie_reviews que possui dois rótulos, positivo e negativo, para um conjunto de 2000 documentos de resenhas de filmes. Isso totaliza 7.786MB de conjunto de treinamento."

Neste sentido, o NLTK mostra-se um toolkit de ferramentas de linguagem natural com cobertura ampla, fornecendo ao usuário uma estrutura simples, extensível, porém uniforme para tarefas e projetos. Além disso, possui uma documentação bastante completa e de fácil entendimento, e por fim, as funcionalidades desta ferramenta são extremamente simples de serem utilizadas (LOPER; BIRD, 2012).

[2]https://www.nltk.org/

3. METODOLOGIA

O presente trabalho caracteriza-se como uma pesquisa prática, de natureza descritiva, e abordagem do tipo quali-quanti. Segundo Fonseca (2002), refere-se a uma pesquisa quantitativa: diferentemente da pesquisa qualitativa, os resultados da pesquisa quantitativa podem ser quantificados. Como as amostras geralmente são grandes e consideradas representativas da população, os resultados são tomados como se constituíssem um retrato real de toda a população alvo da pesquisa.

A pesquisa quantitativa caracteriza-se pela análise de uma amostra de dados brutos, recolhidos com o auxílio de instrumentos padronizados e neutros, decorrendo em resultados quantificados. Já a pesquisa qualitativa não se preocupa com a representatividade numérica, e sim com aspectos da realidade, com o aprofundamento da compreensão de um grupo social, uma organização, etc (GERHARDT E SILVEIRA, 2009).

Quanto aos objetivos, trata-se de uma pesquisa experimental, onde "determinamos um objeto de estudo, selecionamos as variáveis que seriam capazes de influenciá-lo, definimos as formas de controle e de observação dos efeitos que a variável produz no objeto". (PRODANOV; FREITAS, 2013).

A Figura 5 ilustra, por meio de um diagrama, o fluxo de trabalho empregado dentro da metodologia descrita.

Figura 5. Fases da Metodologia Aplicada

Durante a primeira etapa do trabalho, desenvolveu-se a pesquisa quantitativa, a partir da criação da base de dados extraídos do Twitter. Para aquisição dos dados, foi adotada uma ferramenta denominada *GetOldTweets*, escrita em linguagem Python e baseada na biblioteca *Twitter Search*, que permite ao usuário inserir parâmetros, como palavras-chaves da busca, e, assim, retornar os *tweets*, de acordo com as especificações.

Inicialmente, os dados foram armazenados em um arquivo de extensão CSV contendo: nome do usuário, data e horário de publicação, quantidade de *retweets*, o conteúdo das postagens, informação de geolocalização, *hashtags* e hiperlinks dos *tweets* recuperados. Posteriormente, o arquivo foi convertido em formato TXT para melhor manuseio no ato de pré-processamento dos mesmos.

A base de dados formada neste trabalho é composta por 7.303 *tweets* referentes ao sistema operacional móvel Android 10, no idioma inglês, coletados no período de 09 de setembro a 09 de outubro de 2019.

Após a fase de coleta, foi realizado o pré-processamento da base de dados textuais, removendo dos textos dos *tweets* coletados conteúdos não relevantes para a etapa de classificação, realizando, assim, uma limpeza dos dados. Para auxiliar nesta etapa de normalização dos dados, técnicas de PLN foram aplicadas através do uso das bibliotecas NLTK (Natural Language Toolkit) e RE (Regular Expressions), disponíveis na linguagem Python.

Uma vez realizada a coleta e o pré-processamento dos dados, cerca de 3% dos *tweets* normalizados foram submetidos ao algoritmo aplicado pela função *Sentiment* da biblioteca *Pattern* da linguagem de programação Python (versão 3.6.5), com intuito de efetivar a terceira etapa da metodologia, a validação de implementação da técnica, a fim de avaliar a acurácia obtida através do método de análise de sentimentos escolhido pelo autor.

Após resultado positivo da validação do método adotado, o restante dos *tweets* foi submetido ao algoritmo de classificação da Análise de Sentimentos. Este, por sua vez, compara o texto dos *tweets* a um conjunto de palavras já classificadas presentes no dicionário da biblioteca *Pattern* e, assim, classifica a polaridade e a subjetividade das frases. Este dicionário é composto por mais de 5.750 adjetivos que ainda dispõem de substantivos utilizados como recurso, ou seja, estes últimos servem de contexto de apoio à classificação daquela palavra. (SMEDT & DAELEMANS, 2012a).

Por meio da execução da etapa mencionada no parágrafo anterior, obteve-se, portanto, uma tupla de resultados, retornando assim, a subjetividade e principalmente a polaridade do *dataset* de *tweets*, permitindo assim, realizar a comparação, interpretação e por fim, a conclusão dos sentimentos dos consumidores a respeito do sistema operacional Android 10, recém lançado pela Google. Assim, obteve-se uma “primeira impressão” e opinião dos consumidores com relação à usabilidade e funcionalidade nos primeiros meses de lançamento deste produto.

4. ESTUDO DE CASO

Conforme Mortari e Santos (2016) afirmam, devido ao crescente uso da *Internet* e à facilidade de interação e busca de informações *online*, as empresas e/ou marcas estão suscetíveis a sérios danos de imagem que se dão, praticamente, junto da ocorrência do evento, como, por exemplo, o lançamento de um novo produto, dado que a opinião do público é relatada nas mídias sociais, em tese, instantaneamente.

É inexistente um padrão de estratégia organizacional para monitorar, extrair e analisar o conteúdo textual da opinião do público nas mídias sociais. Além do mais, torna-se inviável executar este tipo de análise sem suporte computacional, visto que o volume de dados a serem manipulados e analisados mostra-se extremamente grande (TEIXEIRA, 2018).

Diante disso, a fim de oferecer suporte a tomadas de decisões empresariais sobre determinado produto e/ou serviço ofertado, abordagens de *Business Intelligence* fundamentadas em extração e processamento de informações podem ser solidificadas para usufruto da área de negócio, com intenção de otimizar a análise dos dados extraídos das mídias sociais e, assim, estreitar o relacionamento com o público consumidor, possibilitando a tomada de decisões mais assertivas, agregando resultado positivo à imagem da organização.

Portanto, a fim de alcançar os objetivos propostos neste trabalho, desenvolveu-se uma solução computacional com o propósito de realizar a coleta de dados da rede social Twitter, e, posteriormente, após etapa de pré-processamento, submeter esses *tweets* a um método de análise de sentimentos em linguagem de programação Python, avaliando, assim, a polaridade dos textos com o propósito de extrair a opinião do público consumidor sobre determinado objeto de estudo.

Por fim, o objeto de estudo desta análise foi o sistema operacional *mobile* Android 10, escolhido por ser líder mundial no segmento (LECHETA, 2015) e, dessa forma, encontrar-se bastante popularizado entre os dispositivos móveis., Além disso, esse sistema operacional foi lançado em setembro deste ano, ou seja, refere-se a um produto novo no mercado.

4.1 DETALHAMENTO DO PROJETO

4.1.1 Coleta dos Dados

Inicialmente, seria utilizada a API oficial disponibilizada pela própria plataforma Twitter para os desenvolvedores, porém, após leitura da documentação, constatou-se que a ferramenta não se aplicaria ao estudo devido à recuperação dos *tweets* publicados ter limitação do tempo de busca em até sete dias antes da data da consulta. Além disso, o número de instâncias retornadas também é restrito (TWITTER, 2019c).

Assim, optou-se por utilizar a ferramenta *GetOldTweets*[3],escrita na linguagem de programação Python. Esta, por sua vez, oferece vantagens nas buscas de *tweets* publicados que superam as limitações da API do Twitter, visto que, através desta ferramenta, os *tweets* podem ser retornados de qualquer data conforme propósito do usuário. Através da interface de linha de comando, realizou-se consultas à rede social Twitter, a fim de extrair os *tweets* conforme os argumentos especificados.

Sobre os argumentos parametrizados nas consultas realizadas neste trabalho, seguem:

- *querysearch*: texto de consulta (palavra-chave) a ser considerado na busca pelos *tweets*, sequenciado pelo parâmetro *lang*, que define a limitação do idioma dos *tweets*. A palavra-chave utilizada como parâmetro foi “#Android10”.
- *since*: aponta a data inicial a ser considerada na consulta. Neste caso, a data utilizada foi “2019-09-03”, ou seja, a data de lançamento do sistema operacional.
- *until*: aponta a data limite a ser considerado na consulta. A data “2019-10-03” foi utilizada como parâmetro.

Optou-se pela busca dos *tweets* escritos em Inglês pois grande parte das ferramentas para Análise de Sentimentos em textos estão disponíveis neste idioma (REIS et. al, 2015), e, além disso, segundo pesquisa realizada e concedida pelo Instituto Cervantes[4], a língua inglesa é a mais utilizada mundialmente no Twitter, fato que corroborou para que a base de dados coletados possuísse uma quantidade significativa de *tweets*.

---

[3]https://github.com/Jefferson-Henrique/GetOldTweets-python
[4]http://bit.ly/353V6rA

Durante o período de coleta dos *tweets*, foram obtidos 7.303 textos distintos. Inicialmente, os *tweets* foram salvos em um arquivo do tipo .CSV, devido a padronização de *output* da ferramenta *GetOldTweets*. Além do textual dos *tweets* e das informações de data e horário das postagens, obtiveram-se retornos mais específicos, como nome do usuário, geolocalização, *hashtaghs*, menções e *links*. Na Tabela 2, pode-se visualizar informações sobre a etapa de coleta dos *tweets*.

Descrição	**Quantidade**
Número de dias	31
Tweets coletados	7.303
Tweets usados para testes	100
Tweets submetidos à análise final	7.203

Tabela 2. Informações sobre a etapa de coleta dos *tweets*

4.1.2 Pré-processamento dos *tweets* coletados

Após a primeira etapa de coleta dos dados, iniciou-se a fase de pré-processamento dos dados, na qual todo o conteúdo não relevante para a análise foi removido do conteúdo do tweet, como, por exemplo: links, nomes de usuário do Twitter, caracteres especiais e numéricos, *stopwords* e *hashtaghs*.

```
def str_rem_carac(text):
    prefixes = [redacted]
for separator in string.punctuation:
if separator not in prefixes:
            text = text.replace(separator,' ')
    words = []
for word in text.split():
        word = word.strip()
if word:
if word[0] not in prefixes:
words.append(word)
return ' '.join(words)
```

Figura 6. Limpeza dos Caracteres Especiais

As técnicas de pré-processamento que foram realizadas neste trabalho estão subscritas.

- Remoção de caracteres não alfabéticos, acentuação e pontuação. Com auxílio do módulo *string* da linguagem Python, criou-se uma função para remoção dos caracteres especiais.
- Remoção de *links* e URLs.
- Remoção de menção de usuários do Twitter (precedidos do símbolo "@").
- Padronização do texto e conversão em minúsculo. Utilizou-se o método de *string Lower*() da linguagem de programação Python, para retornar uma cópia de todos os caracteres da *string* passada como parâmetro, em minúsculos.
- Remoção de *stopwords*.

A Tabela 3 apresenta a exemplificação de alguns *tweets* em sua versão final, após passarem por todas as etapas de pré-processamento.

***Tweet* Original**	***Tweet* após limpeza**
Android finally got dark mode in the notifications and settings menu! #android #android10 #darkmode pic.twitter.com/jRFx0mTkdY	android finally got dark mode notifications settings menu darkmode
Oh Lordy Lord, I love #Android10, so freaking smooth and the gesture system is brilliant!	lordy lord love freaking smooth gesture system brilliant
I love this time of the year! New ui, new design, new features...Everything great!! #Android10	love time year new ui new design new features everything great

Tabela 3. Exemplificação dos *tweets* após etapa de pré-processamento dos dados.

4.1.2.1. Remoção de *Stopwords*

Dentre as técnicas do Processamento de Linguagem Natural (PLN), a remoção de *stopwords* geralmente é empregada na etapa de limpeza dos dados no pré-processamento (TEIXEIRA, 2018).

No âmbito de mineração de dados, as *stopwords* são palavras consideradas irrelevantes para este processo, ou seja, palavras que, de modo geral, não agregam significado a uma

sentença, como pronomes, artigos, advérbios, pronomes relativos, conjunções, entre outros, conforme exemplo ilustrado na Tabela 4.

a	an	and	are	as	at
be	by	for	from	has	he
in	is	it	its	on	that
the	to	was	were	will	with

Tabela 4. Exemplificação de stopwords
Fonte: FELIX (2016)

Dito isto, a não remoção das *stopwords* na etapa de limpeza dos dados pode afetar significativamente a eficiência da análise, visto que pode resultar em uma quantia gigantesca de processamento improdutivo (EL-KHAIR, 2016). O supracitado autor ainda afirma que (tradução livre para o Português):

> "Essas palavras irrelevantes têm impactos diferentes no processo de recuperação de informações. Eles podem afetar a eficácia da recuperação, porque possuem uma frequência muito alta e tendem a diminuir o impacto das diferenças de frequência entre palavras menos comuns, afetando o processo de análise. A remoção das palavras-chave também altera o tamanho do documento e subsequentemente afeta o processo de análise."

Para a remoção das *stopwords* deste trabalho, utilizou-se o método de mesmo nome, presente no *toolkit* NLTK da linguagem de programação Python, conforme exibido na Figura 7.

```
with codecs.open(r"cleaning.txt", 'r', encoding="utf8") as inFile,
codecs.open(r"stopwords.txt", 'w', encoding="utf8") as outFile:
stop_words = set(stopwords.words('english'))
for line in inFile.readlines():
[redacted]
filtered_words = " ".join(w for w in words if w not in stop_words)
outFile.write(filtered_words + '\n')
```

Figura 7. Exemplificação de remoção de *stopwords*

4.1.3 Validação da Função de Polaridade

Após o pré-processamento dos textos que foram coletados, a autora deste trabalho selecionou manualmente uma amostra de 100 (cem) *tweets* e classificou os mesmos de acordo

com suas polaridades, positiva e negativa, cuidando para obter um balanceamento de 50% entre as duas classes. Neste trabalho, a função *sentiment*, da biblioteca *Pattern*, foi o método utilizado para a classificação de polaridade dos textos. Portanto, fez-se imprescindível realizar a avaliação da acurácia deste método na execução da tarefa proposta.

Para realizar a classificação dos *tweets* em massa, fez-se necessário apenas escrever um código simples na linguagem Python, conforme exposto na Figura 8.

Essencialmente, após ser realizada a importação da biblioteca *sentiment*, cada linha do arquivo é percorrida através de um laço de repetição, sendo a *string* passada para a função *sentiment*, que realiza o cálculo de polaridade e subjetividade do texto, retornando um resultado por linha.

```
from pattern.text import sentiment
import codecs

with codecs.open('[redacted].txt', 'r', 'utf-8') as f:
for line in f:
        tweet = f.readlines()

        w = [sentiment(i) for iin tweet]
for iin w:
print(i)
```

Figura 8. Código de Leitura do arquivo e classificação utilizando a biblioteca *Pattern*

Imediatamente após a execução do código, uma tupla contendo os valores da polaridade (da esquerda, a primeira coluna) e da subjetividade (segunda coluna) é retornada para cada um dos *tweets* presentes no arquivo. Pode-se verificar parte do *output* na Figura 9 abaixo.

```
C:\[redacted]\Programs\Python\Python36\python.exe
(0.55, 0.75)
(0.375, 0.75)
(0.5375, 0.675)
(0.4375, 0.8194444444444444)
(0.7, 0.8)
(0.3181818181818182, 0.7272727272727273)
(0.26666666666666666, 0.6333333333333333)
(0.0, 0.0)
(0.5, 0.5)
(0.2, 0.5)
```

Figura 9. Retorno da Polaridade

Os valores obtidos variam dentro do intervalo [-1, 1], de tal forma que, quanto mais próximo o retorno for de “1”, traduz-se que maior é a força do sentimento positivo para determinado axioma, enquanto os valores mais próximos de “-1” propõem força de sentimento negativo (SMEDT E DAELEMANS, 2012b).

Os testes com esta biblioteca apresentaram uma acurácia de 85,07% na classificação do conjunto de dados de validação. De modo mais específico, o método apresentou uma melhor performance na classificação da polaridade positiva, alcançando uma acurácia de 96%, enquanto esta mesma métrica ficou em torno de 88% para a classificação de *tweets* negativos. Do total dos *tweets*, 16% não retornaram classificação. Diante destes números, pode-se afirmar que o modelo da biblioteca Pattern obteve uma acurácia elevada para a classificação dos *tweets* coletados.

4.1.4 Resultados da Análise de Sentimentos

Finalizada a etapa de validação do modelo, partiu-se para a classificação dos 7.203 *tweets* restantes. Estes, após serem pré-processados, foram submetidos ao modelo de classificação por meio da função *sentiment* da biblioteca Pattern, com o intuito de terem suas polaridades reveladas em relação ao recém lançado sistema operacional *mobile* Android 10.

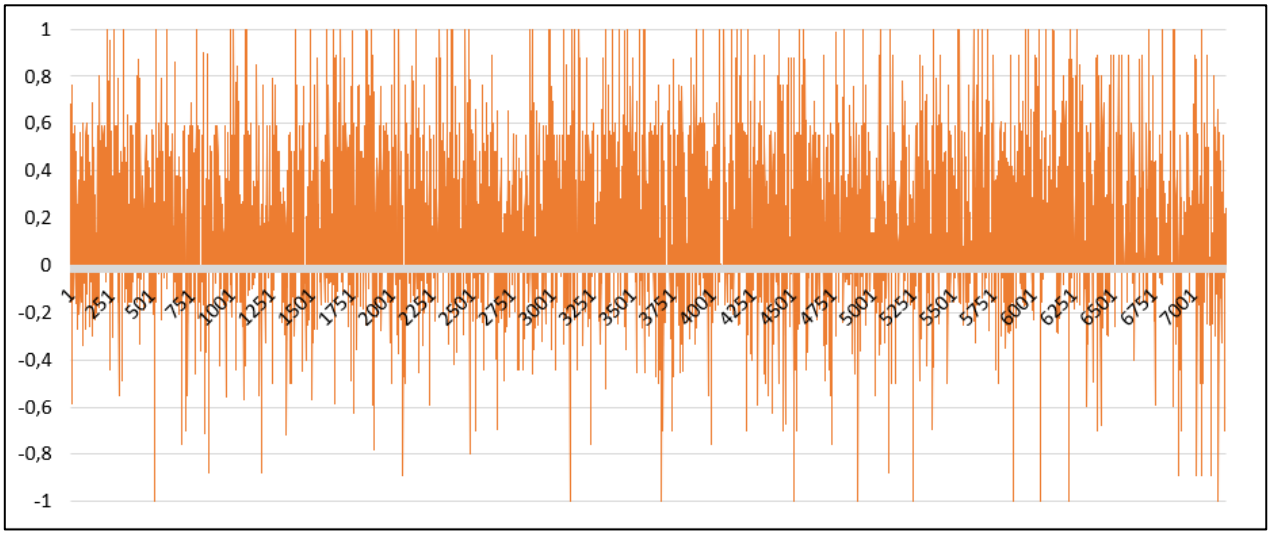

Figura 10. Classificação de Polaridade dos *Tweets* intervalo de 250

A Figura 10 ilustra a oscilação da classificação da polaridade dos 7.203 *tweets* que foram submetidos ao algoritmo.

Como resultado, foram obtidos os valores apresentados na Tabela 5, onde mais de 46,84% dos *tweets* que retornaram classificação são positivos e apenas 11,48% negativos. O método aplicado, por sua vez, não foi capaz de classificar 3.002 (aproximadamente 42%) do total de *tweets*.

Classificação	Quantidade Tweets	Porcentagem
Positivo	3374	46,84%
Negativo	827	11, 48%
Sem Classificação	3002	41,67%

Tabela 5. Sentimento dos Usuários do Twitter em relação ao Android 10

Esta impossibilidade de classificação deve-se, sobretudo, ao fato de o dicionário da biblioteca *Pattern* possuir uma quantidade limitada de adjetivos com escore de polaridade positivo ou negativo associado, conforme exemplificado na Figura 11. Portanto, quando o algoritmo realiza a varredura textual a cada linha, nenhum dos adjetivos são encontrados para que este execute a abordagem de classificação de polaridade do *tweet* em questão.

```
1  <word form="active" cornetto_synset_id="n_a-506270" wordnet_id="a
       -00031974" pos="JJ" sense="characterized by energetic activity"
       polarity="0.2" subjectivity="0.4" intensity="1.0" confidence="0
       .8" />
2  <word form="amazing" wordnet_id="a-01282510" pos="JJ" sense
       ="inspiring awe or admiration or wonder" polarity="0.8"
       subjectivity="1.0" intensity="1.0" confidence="0.9" />
3  <word form="angry" wordnet_id="a-00304144" pos="JJ" sense="(of the
       elements) as if showing violent anger" polarity="-0.5"
       subjectivity="1.0" intensity="1.0" confidence="0.9" />
4  <word form="awesome" wordnet_id="a-01282510" pos="JJ" sense
       ="inspiring awe or admiration or wonder" polarity="1.0"
       subjectivity="1.0" intensity="1.0" confidence="0.9" />
5  <word form="crap" wordnet_id="n-14854581" pos="NN" sense="obscene
       terms for feces" polarity="-0.8" subjectivity="0.8" intensity="1
       .0" label="profanity" confidence="0.9" />
6  <word form="cute" cornetto_synset_id="n_a-532620" wordnet_id="a
       -00167278" pos="JJ" sense="attractive especially by means of
       smallness or prettiness or quaintness" polarity="0.5"
       subjectivity="1.0" intensity="1.0" confidence="0.9" />
```

Figura 11. Dicionário de Classificação Léxica para Adjetivos da biblioteca *Pattern*
Fonte: GITHUB (2019)[5]

[5]https://github.com/clips/pattern/blob/master/pattern/text/en/en-sentiment.xml

A Figura 12 ilustra o cenário descrito acima, onde o eixo vertical trata-se do escore de classificação de polaridade e o eixo horizontal da quantidade de tweets analisados. Por meio da análise dessa figura pode-se observar que grande parte do intervalo do gráfico (a partir da 827ª posição) encontra-se alinhado ao eixo x, ou seja, com classificação igual a 0 (zero) de polaridade.

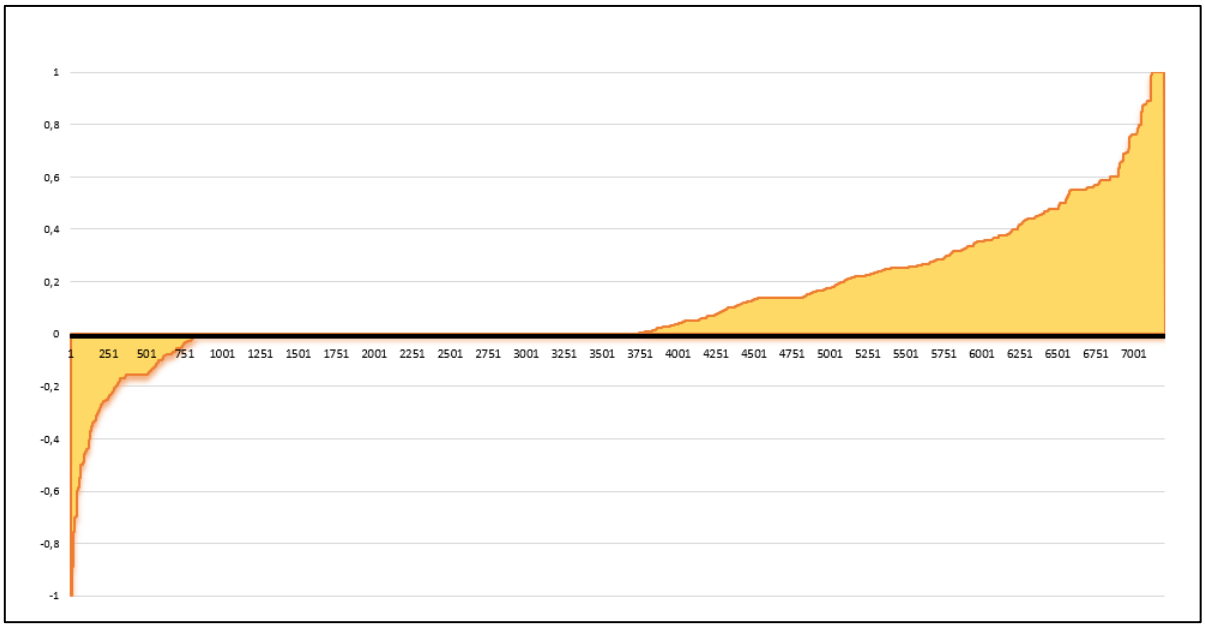

Figura 12. Decrescente de Polaridade Negativo e Positivo

Apesar do modelo de análise de sentimentos baseado em léxico utilizado possuir restrição em decorrência do dicionário de palavras supracitado, este trabalho apresentou, como uma das principais contribuições, através de aplicação de técnicas diversas de processamento de linguagem natural, a visão do público consumidor com relação ao objeto de estudo escolhido, o SO *mobile* Android 10 em seguida ao seu lançamento, e obteve uma assertividade positiva com relação a quantidade de *tweets* que foram classificados pelo modelo.

5. CONSIDERAÇÕES FINAIS

As opiniões expressas em mídias sociais transmitem sentimentos e avaliações das pessoas em relação ao mundo (SMEDT E DAELEMANS, 2012b) Assim, o uso crescente de tais mídias torna-se estratégico para a aplicação da abordagem de *Business Intelligence* por parte das empresas, as quais podem, a partir de tal estratégia, mapear a avaliação do público consumidor com relação a determinado produto ou serviço ofertado, e, assim, tomar decisões mais assertivas.

Dentro desse contexto, o presente trabalho detalhou o uso de mineração de opiniões para coletar, estruturar e normalizar o texto extraído do Twitter, e, por fim, testar um modelo de classificação de polaridade escrito na linguagem de programação Python. Tal modelo, por sua vez, permitiu realizar a análise de sentimentos e, assim, captar a opinião expressa pelos usuários do Twitter sobre o novo sistema operacional *mobile* lançado pela Google, no terceiro trimestre deste ano, o Android 10.

Entretanto, apesar de o resultado da análise de sentimentos sobre o produto Android 10 ter sido satisfatoriamente positivo, o modelo utilizado possui determinada limitação em função da quantidade restrita de adjetivos com escore de classificação de polaridade pré-definida existente dentro de seu dicionário. Isso acabou dificultando a classificação de aproximadamente 42% dos *tweets* avaliados, justamente pelo fato de estes não possuírem em seu contexto as palavras-chaves necessárias à classificação.

Ademais, a metodologia empregada neste estudo mostrou-se eficiente na análise de sentimentos, e, dessa forma, revela-se capaz de ser aplicada em outros contextos, por organizações que desejem mapear a opinião dos usuários do Twitter em relação a determinado produto e/ou serviço, a fim de obter embasamento para tomada de decisões.

6. TRABALHOS FUTUROS

Uma das limitações deste trabalho está associada às restrições que o método *Sentiment* da biblioteca *Pattern* apresenta ao tentar classificar determinados *tweets*, devido, sobretudo, à precariedade de adjetivos constantes no dicionário léxico do idioma inglês. Portanto, visto que esta biblioteca é de código aberto, uma das hipóteses a ser estudada seria a possibilidade de colaboração na expansão do dicionário da biblioteca.

Ainda como trabalho futuro, destaca-se a exploração de outros métodos de classificação para análise de sentimentos como *Naive Bayes*, *Bag of Words* ou Vader, e, assim sendo, construir um modelo classificador de acordo com a necessidade do objeto de estudo.

Por fim, na mesma sequência de metodologia deste trabalho, sugere-se ainda realizar uma separação dos usuários do Twitter conforme faixa etária e/ou geolocalização, permitindo segregar a opinião do público alvo de acordo com a idade e a localização.

7. REFERÊNCIAS

ALENCAR, Leonel Figueiredo de. Utilização de informações lexicais extraídas automaticamente de corpora na análise sintática computacional do português. In: Rev. Est. Ling., Belo Horizonte, v. 19, n. 1, p. 7-85, 2011. Disponível em: http://bit.ly/2NAWT1m. Acessado em 10 nov. 2019.

ANDROID. What Is Android. Disponível em: https://bit.ly/2pJrfp0. Acessado em 06 out. 2019.

APPLE. IOs 13. Disponível em: https://apple.co/2kW9EbH.Acessado em 06 out. 2019.

ARONSON, Jay E.; KING, David; SHARDA, Ramesh; TURBAN, Efraim. Business Intelligente: Um Enfoque Gerencial para a Inteligência do Negócio. 1 ed. Porto Alegre: Editora Bookman, 2009.

AUTOMAÇÃO INDUSTRIAL. Utilização de Big Data na Indústria 4.0. [S.I], 2017. Disponível em: https://bit.ly/2QAYVxh. Acessado em: 02 out. 2019.

BACCIANELLA, Stefano; ESULI, Andrea; SEBASTIANI, Fabrizio. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: 7th Conference on International Language Resources and Evaluation, p. 2200–2204. Valeta, 2010.

BANK MY CELL. Smartphone Addiction Facts & Phone Usage Statistics: The Definitive Guide (2019 Updated). Disponível em: https://bit.ly/2PQjs48. Acessado em 13 out. 2019.

BARBOSA, Glívia A.R.; MEIRA, Vagner Jr.; SILVA, Ismael S.; PRATES, Raquel O.; ZAKI, Mohammed J.; VELOSO, Adriano. Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In: PROCEEDING OF THE 2012 ACM ANNUAL CONFERENCE EXTENDED ABSTRACTS ON HUMAN FACTORS IN COMPUTING SYSTEMS EXTENDED ABSTRACTS. Austin, 2012.

BENTLEY, Drew. Business Intelligence e Analytics. 1 ed. New York: Press Library, 2017.

BULEGON, Hugo; MORO, Claudia Maria Cabral. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. Journal of Health Informatics, p. 51-56. São Paulo, 2010.

CHEN, H.; CHIANG, R. H. L.; STOREY, V. C. Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, v. 36, n. 4, p. 1165–1188, 2012.

CLIPS: Computacional Linguistics & Psycholinguistics Research Center. Pattern Developer Documentation. 2019. Disponível em: http://bit.ly/2Q2kslp. Acessado em: 10 nov. 2019.

COSTA, Norben P. O.; FILHO, Nemésio F. Duarte. Análise e Avaliação Funcional de Sistemas Operacionais Móveis: Vantagens e Desvantagens. Revista de Sistemas e Computação, Salvador, v.3, nº 1, p. 66-77, jan./jun. 2013.

COSTA, P. R. S.; SOUZA, F. F.; TIMES, V. C.; BENEVENUTO, F. "Towards Integrating Online Social Networks And Business Intelligence". Proceedings of the International Conferences Web Based Communities and Social Media, p. 21-32. Lisboa, 2012.

CRIVELLI, Ricardo Barbosa. Recuperação de Informação Por Meio de Processamento de Linguagem Natural. 47 f. Trabalho de Conclusão de Curso de Sistemas de Informação. Universidade Estadual do Norte do Paraná, Bandeirantes, PR, 2011.

DELFINO, Stephany. Analytics e BI Aplicado à Indústria 4.0. Blog Advantech, 2018. Dsiponível em: https://bit.ly/2LLEMVE. Acessado em 05 out. 2019.

EL-KHAIR, Ibrahim Abu. Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study. In: International Journal of Computing & Information Sciences, p. 119-133, v. 4, n. 3. 2006.

ERL, Thomas; KHATTAK, Wajid; BUHLER, Paul. Big Data Fundamentals: Concepts, Drivers & Techniques. 1 ed. Crawfordsville: Editora Prentice Hall, 2015.

FARINON, João Luís. Análise e Classificação de Conteúdo Textual. 77 f. Trabalho de Conclusão de Curso de Ciência da Computação. Universidade de Santa Cruz do Sul, Santa Cruz do Sul, RS, 2015.

FELIX Felipe da Silva, Nadia. Análise de Sentimentos em Textos Curtos. 138 f. Tese (Doutorado Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

FIORIO, Rosaine. Um Software De Apoio À Aprendizagem De Gramática E Estilo Literário Da Língua Portuguesa Brasileira. 86 f. Trabalho de Conclusão de Curso de Licenciatura em Informática. Universidade Tecnológica do Paraná, Francisco Beltrão, PR, 2015.

FREI, Lukas. Towards Data Science. 08/02/2019. Disponível em: http://bit.ly/2q0UAvD. Acessado em: 10 nov. 2019.

GERHARDT, Tatiana Engel; SILVEIRA, Denise Tolfo. Métodos de Pesquisa. 1 ed. Porto Alegre: Editora UFRGS, 2009.

HURWITZ, Judith; NUGENT, Alan; HALPER, Fern; KAUFMAN, Marcia. Big Data for Dummies. 1 ed. New Jersey: Editora John Wiley & Sons, Inc, 2013.

IDC. Smartphone Market Share. 18 Jun. 2019. Disponível em: https://bit.ly/2qNaCI2. Acessado em: 13 out. 2019.

JETBRAINS. PyCharm. 2019. Disponível em: http://bit.ly/2Xre09z. Acessado em: 08 nov. 2019.

JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2 ed. Nova Jersey: Prentice Hall, 2008.

KUMAR, Shamanth; MORSTATTER, Fred; LIU, Huan. Twitter Data Analytics. New York: Springer, 2013.

LECHETA, Ricardo R. Google Android: Aprenda a criar aplicações para dispositivos móveis com o Android SDK. 4.ed. São Paulo: Novatec Editora, 2015.

LEWIS, Lori. This Is What Happens In An Internet Minute. Allaccess.com. 05/03/2019. Disponível em: https://bit.ly/2Jc2nji. Acessado em: 27 abr. 2019.

LIU, Bing; Sentiment Analysis and Opinion Mining. Synthesis lectures on human language technologies, v. 5, n. 1, p. 1–168, 2012.

LOPER, Edward; BIRD, Steven. NLTK: The natural language toolkit. In: Proceedings of the ACL-02Workshop on Efective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics –v. 1, p. 63-70. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002.

MACHADO, Francis Berenger; MAIA, Luiz Paulo. Arquitetura De Sistemas Operacionais. 5.ed. Rio de Janeiro: Ltc Editora, 2013.

MAGANO, Fernada Camargo. Descoberta de Conhecimento em Redes Sociais e Bases de Dados Públicas. 55 f. Trabalho de Conclusão de Curso de Matemática e Estatística. Universidade de São Paulo, São Paulo, SP, 2016.

MALIK, Usman. Python for NLP: Introduction to the Pattern Library. Stackabuse, 2019. Disponível em: http://bit.ly/2XfXxog. Acessado em 15 nov. 2019.

MARQUESONE, Rosangela. Big Data: Técnicas e tecnologias para extração de valor de dados. 1 ed. São Paulo: Editora Casa do Código, 2016.

MATTHES, Eric. Curso Intensivo de Python: Uma Introdução Prática e Baseada em Projetos à Programação. 1. ed. São Paulo: Editora Novatec Ltda, 2016.

MORIMOTO, Carlos. Smartphones: Guia Prático. 1. ed. Porto Alegre: Editora Sulina, 2009.

MORTARI, Elisangela Carlosso Machado; SANTOS, Suzana Fernandes dos. Monitoramento de Redes Sociais Digitais como Estratégia Organizacional. Disponível em: http://bit.ly/2Qg09kN. Acessado em 14 nov. 2019. In: Intercom – RBCC, v.39, n.1, p.91-109. São Paulo, 2016.

MÜLLER, Daniel Nehme. Processamento de Linguagem Natural. Universidade Federal do Rio Grande do Sul. Porto Alegre/RS, 2003. Disponível em: http://bit.ly/32ywtSl. Acessado em 09 nov. 2019.

NAGAO, Makoto. 2013. “The Age of Content and Knowledge Processing”. National Institute of Information and Communications Technology, Japan. Disponível em https://bit.ly/32Gq9sZ. Acessado em 23 out. 2019.

NETO, João Mendes de Oliveira; TONIN, Sávio Duarte; PRIETCH, Soraia Silva. Processamento de Linguagem Natural e suas Aplicações Computacionais. 2010. Disponível em: https://bit.ly/2o7jhpu. Acessado em: Out. 2019.

PAMPLONA, Yuri Logatto. Indústria 4.0: Análise e simulação de uma nova era industrial. 49 f. Trabalho de Conclusão de Curso de Sistemas de Informação. Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, RJ, 2018.

PEREIRA, Vanessa Alves da Silva. Big Data: Um Estudo em Gestão Empresarial. 86 f. Trabalho de Conclusão de Curso de Biblioteconomia e Gestão de Unidades de Informação. Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 2016.

REIS, Júlio C.S.; GONÇALVES, Pollyanna; ARAÚJO, Matheus; PEREIRA, Adriano C. M.; BENEVENUTO, Fabrício. Uma Abordagem Multilíngue para Análise de Sentimentos. In: CBSC 2015 BraSNAM. Disponível em: http://bit.ly/3573bwg. Acessado em: 14 nov. 2019.

REVISTA COMPUTAÇÃO BRASIL. Internet das coisas - Nós, as cidades, os robôs, os carros: Tudo conectado!, nº 29, abr. 2015, 58 p. Disponível em: https://bit.ly/2mr274W. Acessado: 20 set. 2019.

REVISTA EXAME. Como Construir o Brasil 4.0. São Paulo: abril, edição 1162, ano 52, nº 10, 30 maio 2018.

ROJKO, Andreja. Industry 4.0 Concept: Background and Overview. Ijim. Nuremberga, p. 77-90. 2017. Disponível em: http://bit.ly/2PycAGs. Acessado em: 01 out. 2019.

SAKIS, Maria Augusta Santos. Aplicação De Heurísticas Para Avaliação De Usabilidade Em Dispositivos Móveis. 80 f. Trabalho de Graduação de Curso de Ciência da Computação. Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, 2015.

SALDANHA, Rodolfo Luis dos Santos. Business Intelligence: Análise sobre as Soluções de BI e Estudo de Caso Usando Pentaho. 48 f. Trabalho de Conclusão de Curso de Ciência da Computação. Universidade Estadual de Londrina, Londrina, 2018.

SANCHES, Matheus Ferraroni. Processamento E Entendimento de Linguagem Natural No Gerenciamento de Emergências Para Obtenção de Consciência Situacional. 78 f. Trabalho de Graduação de Curso de Ciência da Computação. Centro Universitário Eurípedes de Marília, São Paulo, 2017.

SANTOS, Leonardo José de Andrade. Business Intelligence e Análise de Sentimentos no Contexto de Redes Sociais Online. 60 f. Trabalho de Graduação de Curso de Ciência da Computação. Universidade Federal de Pernambuco, Recife, 2015.

SANTOS, Wagner Roberto dos. RESTful Web Services e a API JAX-RS. 2015, p. 59-73. Disponível em: https://bit.ly/2M9eLzP. Acessado em 13 out. 2019.

SCHEPS, Swain. Business Intelligene for Dummies. 1.ed. Hoboken: John Wiley & Sons, 2008.

SMEDT, Tom; DAELEMANS, Walter. Pattern for Python. Journal of Machine Learning Research: v. 13, p. 2063-2067, 2012a.

_____.“Vreselijkmooi!” (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), p. 3568-3572. 2012b.

SILVA, Josiane Rodrigues da. Detecção de Opiniões e Análise de Polaridade em Documentos Financeiros com Múltiplas Entidades. 62 f. Dissertação de Mestrado do Programa de Pós-Graduação em Informática. Universidade Federal do Amazonas, Manaus, AM, 2015.

SILVA, Nelson Gutemberg Rocha da. PairClassif -Um Método para Classificação de Sentimentos Baseado em Pares. Dissertação de Mestrado do Programa de Pós-Graduação em Ciência da Computação. 98 f. Universidade Federal de Pernambuco, PE, 2013.

SIMPLEZ. Inteligência de Negócio: Como o algoritmo da Netflix ajuda a entender seus usuários. [S.I], 2016. Disponível em: https://bit.ly/2Ls9OU2. Acessado em 05 out. 2019.

STATCOUNTER, Global Stats. Mobile Operating System Market Share Worldwide. 2019. Disponível em: http://bit.ly/2pMuVqJ. Acessado em 09 nov. 2019.

STATISTA. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions). [S.I], 2019. Disponível em: https://bit.ly/2dt7OI9. Acessado em: 05 out. 2019.

TANENBAUM, Andrew S.; WOODHULL, Albert S. Sistemas Operacionais: Projeto e Implantação. 3. ed. Porto Alegre: Editora Bookman, 2008.

TAURION, Cezar. Big Data. Editora Brasport, Rio de Janeiro, 2013. Disponível em: https://bit.ly/2JuZZmd. Acessado em: 06 abr. 2019.

TEIXEIRA, Gabriel Morais. ITeligence: Sistema de Apoio à Análise de Intenções. 2018. 96 f. Trabalho de Conclusão de Curso sem Sistemas de Informação. Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, RJ.

TWITTER. Consuming Streaming Data. Disponível em: https://bit.ly/2VHy8TY. Acessado em 13 out. 2019a.

TWITTER. Sobre as APIs do Twitter. Disponível em: https://bit.ly/2AShFCN. Acessado em 09 out. 2019b.

TWITTER. Standard Search API. Disponível em: http://bit.ly/33QdC6z. Acessado em: 14 nov. 2019c.

UBUNTU. About The Ubuntu Project. Disponível em: https://bit.ly/31TMxyn. Acessado em 06 out. 2019.

WAZLAWICK, Raul Sidnei. Metodologia de Pesquisa para ciência da computação. Rio de Janeiro: Elsevier, 2009. 6ª reimpressão.