UNIVERSIDADE DE TAUBATÉ

Debora Vanessa Gobbo

UMA ABORDAGEM BASEADA EM *SMALL DATA* PARA COMPARAR O RESULTADO DA APLICAÇÃO DAS TÉCNICAS DE ANÁLISE DE SENTIMENTOS DOS CLIENTES DE UMA PEQUENA EMPRESA.

TAUBATÉ - SP

2019

Debora Vanessa Gobbo

UMA ABORDAGEM BASEADA EM *SMALL DATA* PARA COMPARAR O RESULTADO DA APLICAÇÃO DAS TÉCNICAS DE ANÁLISE DE SENTIMENTOS DOS CLIENTES DE UMA PEQUENA EMPRESA.

Trabalho de Pós-Graduação apresentado como requisito parcial para a conclusão do curso de Gestão de projetos em *Business Intelligence* do Departamento de Informática da Universidade de Taubaté.

Orientador: Fernando Gama da Mata

TAUBATÉ - SP

2019

Debora Vanessa Gobbo

UMA ABORDAGEM BASEADA EM *SMALL DATA* PARA COMPARAR O RESULTADO DA APLICAÇÃO DAS TÉCNICAS DE ANÁLISE DE SENTIMENTOS DOS CLIENTES DE UMA PEQUENA EMPRESA.

Trabalho de Pós-Graduação apresentado como requisito parcial para a conclusão do curso de Gestão de projetos em *Business Intelligence* do Departamento de Informática da Universidade de Taubaté.

Data: _____

Resultado: _____

Banca Examinadora

Profº. Me. Fernando Fabio Dias Gama da Mata

Assinatura

_____

Profº. Dr. José Carlos Lombardi

Assinatura

_____

Profª.Me. Dawilmar Guimarães de Araújo

Assinatura

_____

Profº. Esp. Fábio Rosindo Daher de Barros

Assinatura _____

DEDICATÓRIA

.

A todas as pessoas que acreditaram neste trabalho.

AGRADECIMENTOS

Primeiramente agradeço a DEUS por ter concedido a oportunidade de apresentar esta monografia.

Agradeço aos familiares por sempre estarem apoiando durante todo o período do curso.

Agradeço a todos os professores pela colaboração e ajuda durante todo o período de aprendizagem.

Agradeço aos colegas e amigos pelo apoio durante todo esse tempo.

“O impossível não é um fato, é uma opinião. O impossível não é uma declaração, é um desafio.”

Muhammad Ali

Resumo

Atualmente as empresas estão investindo em tecnologia que gerem subsídios para tomada de decisão, a fim de fidelizar seus clientes, necessitando para isso conhecer o perfil dos mesmos, bem como seus gostos, expectativas e frustações, podendo recorrer para isto, a utilização do conceito de *Small Data,* sendo necessário para armazenar estas informações observadas. Para a base de dados neste caso foi escolhida a plataforma no *Sqlite.* Para verificar a satisfação dos clientes com relação aos serviços de suporte em tecnologia da informação prestados, houve o desenvolvimento de um classificador de emoção predominante em comentários da plataforma de chamados em formato *tickets* de uma pequena empresa, sendo necessário para isso o estudo e implementação de um modelo baseado em aprendizado de máquina, o *Naive Bayes* , por ser um excelente estimador probabilístico. O algoritmo desenvolvido trabalha com duas classes para classificação do atendimento, sendo as classes: positivo e negativo, se ainda houver dúvidas quanto esta classificação, deverá ser recorrida a análise de uma base de Small Data para validação. Por fim, obteve-se um classificador de ótimo desempenho, com 91% de acurácia nos testes, que será utilizado em futuras aplicações reais na plataforma, automatizando análises qualitativas, por exemplo, sendo a base construída em *Small Data* utilizada em poucos casos.

Palavras-chave: Análise de Sentimentos, Mineração de Dados, *Small Data*.

Abstract

Companies are currently investing in technology that enables them to have decision-making, in order to build customer loyalty, thus needing to know their profile, as well as their tastes, expectations and frustrations of *Small Data*, being necessary to store this observed information, as a database, in this case, was chosen the platform in *SqLite*. In order to verify the satisfaction of its customers regarding the information technology support services provided, a predominant emotion classifier was developed in the comments of the ticket format platform of a small company, which required the study and implementation. of a machine learning model, *Naive Bayes*, for being an excellent probabilistic estimator. The developed algorithm works two classes for service classification, being the classes: positive and negative, if there are still doubts about this classification, the analysis of a *Small Data* base for validation should be resorted to. Finally, was obtained a great performance classifier, with 91% accuracy in the tests, which will be used in future real applications on the platform, automating qualitative analysis, for example, a small data base used in a few cases.

Keywords: Sentiment analysis, Data Mining, Small Data.

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

LISTA DE ABREVIATURAS E SIGLAS

BI – Business Intelligence

CRM - Customer Relationship Management

CWI - *Centrum Wiskunde & Informatica*

CSV - Valores Separados por Vírgulas

KDD – Knowledge Discovery in Database

MPE – Micro e pequenas empresas

NL - *Natural Language*

NLTK - *Natural Language Toolkit*

PLN - Processamento de Linguagem Natural

SQL - Structured Query Language

TI – Tecnologia da informação

SUMÁRIO

1.INTRODUÇÃO

A frequente mudança do mercado torna-se presente no cotidiano das pequenas empresas, que acabam sendo influenciadas por decisões tecnológicas, governamentais, mercado consumidor, aumento de renda, entre outros fatores. Muitas vezes as organizações não obtêm êxito em filtrar e trabalhar essas informações de mercado e tampouco criar um conhecimento de planejamento de médio e longo prazo. As micro e pequenas empresas correspondem ao maior segmento empresarial do país, segundo os dados de uma pesquisa realizada pelo Sebrae (2018), no Brasil existem 6,4 milhões de estabelecimentos, sendo 99% composto por micro e pequenas empresas (MPE), sendo elas responsáveis por 52% dos empregos com carteira assinada no setor privado (16,1 milhões).

A fim de atrair e reter clientes, bem como parceiros de negócios, as organizações necessitam prestar serviços consistentes e com excelência de qualidade, consequentemente, fazendo-se necessária uma perspectiva de automação de processos, pois os mesmos devem ser corretamente projetados e sua execução deve ser apoiada por um sistema que possa atender aos requisitos do trabalho, para isso, a adoção de *Business Intelligence* é fundamental.

O conceito de *Business Intelligence* mais amplamente aceito vem do Gartner: “um termo abrangente que inclui os aplicativos, infraestrutura e ferramentas e as melhores práticas que permitem o acesso e análise de informações para melhorar e otimizar decisões e desempenho.” (FORBES INSIGHTS, 2016). O seu maior propósito é permitir o acesso aos dados, viabilizar a manipulação, e fornecer para os gestores de negócios a competência de realizar a análise adequada para a escolha da melhor decisão a ser tomada. Esse processo baseia-se na transformação dos dados em informações visíveis, depois em decisões e por fim em ações. O *BI* tem por desígnio trabalhar as informações, ou seja, facilitar as consultas, geração de relatórios, gráficos e análises gerenciais. O autor afirma que um dos grandes benefícios que *BI* oferece para uma empresa é a capacidade que tem de prover informações fundamentais, quando necessárias, incluindo uma visão rápida do desempenho corporativo, sendo essas informações fundamentais para todos os tipos de decisões da organização. Esses benefícios podem ser compreendidos como: economia de tempo, versão única

da verdade, melhores estratégias e planos, melhores decisões táticas, processos mais eficientes, economia de custos definição dos tipos de clientes.

Para entender o comportamento do cliente, a utilização do conceito de *Small Data*, adquirido através de informações de anotações, questionários, vídeos e fotografias (MITI, 2016) possibilita complementar os *insights* não captados por ferramentas tradicionais de análise de dados, fornecendo uma visão mais próxima das reais necessidades ou desejos do cliente.

O *Small Data* entende as necessidades do cliente de forma mais precisa, não sendo preciso um grande investimento para isso, pois há baixo custo de implementação, e os dados são fáceis de serem acessados.

1.1. Justificativa

Diante do cenário exposto, este trabalho tem por finalidade implantar uma solução de *BI* de baixo custo em uma pequena empresa da área de suporte e infraestrutura de TI da cidade de Taubaté -SP, que por motivos de sigilo não terá seu nome divulgado. No apêndice A, está o acordo de utilização e confidencialização dos dados. A empresa visa conhecer melhor quem são e os perfis de seus clientes, opinião, hábitos, expectativas e frustrações dos mesmos, sobre os serviços realizados, a fim de otimizar o tempo de atendimento com criação de rotinas de atendimento, a partir de outros chamados semelhantes já realizados anteriormente, melhorar a performance do atendente com base nas considerações feitas pelos clientes, através dos indicadores gerados com pesos e medidas, que terão a missão de comunicar, de forma simples e por meio da quantificação, os resultados diretos de cada processo ou operação, havendo desta forma uma maior satisfação de seus clientes, aumentando a possiblidade dos mesmos indicarem os serviços da empresa a outras organizações.

Com o auxílio do *Small Data*, a empresa poderá resolver também as seguintes questões:

- Como os dados pequenos são estruturados e acionáveis, a empresa pode tomar decisões operacionais e administrativas rápidas, em vez de dedicar tempo para entender o conjunto de dados primeiro.

- Devido às expectativas e à crescente necessidade de produtos e serviços personalizados, os clientes estão mais dispostos a revelar informações sobre suas preferências, resultando em armazenamentos de dados pessoais.
- Formação de uma base sólida de informações para personalização. Com *insights* bem organizados sobre os clientes, ele pode ser usado para personalização de atendimento.
- Ao simplificar os dados, a empresa pode capacitar seus funcionários para entender melhor seus clientes e fornecer os melhores serviços de forma personalizada para eles.

Estas questões poderão ser solucionadas através da análise realizada das respostas dos chamados abertos no sistema de *Ticket* e juntamente com um acompanhamento pessoal ao cliente da seguinte maneira:

- Acessando o histórico de chamados do usuário;
- Estilo linguístico utilizado no atendimento;
- Uso de pronomes pessoais;
- Proporção de palavras funcionais e não funcionais;
- Índice de variação qualitativa;
- Padrão na avaliação;
- Identificando o comportamento/ perfil do cliente;
- Detalhes pessoais do usuário observados;
- Detalhes do como ele está;
- Ligação para confirmar satisfação ou frustração com o fechamento do chamado;
- Visita no local de trabalho do cliente pelo funcionário periodicamente;
- Pesquisa de campo com os colegas de trabalho do cliente alvo;
- Registros de indicação da empresa por parte do cliente.

Acredita-se que a abordagem utilizando *Small Data* será fundamental para o sucesso na implantação da solução de *BI* na empresa em questão, pois há baixo volume de dados, coletando o *feedback* do cliente, quanto a sua satisfação com o serviço, tempo de atendimento, solução do problema, tudo realizado de maneira *online*, *Small Data* permitirá o “real conhecimento do cliente”, pois será possível

conhecer suas reais necessidades e confrontar essas informações com os dados da empresa para que uma fidelização entre o cliente e a empresa possa ser consolidada.

1.2. Objetivo Geral

O objetivo deste trabalho é a criação de um classificador de análise de sentimentos.

1.3. Objetivos Específicos

- Explorar e aplicar algoritmo de análise de sentimentos para identificação da bipolaridade do atendimento (positivo/negativo). EX. *Naive Bayes.*
- Confrontar os resultados em que a há dúvidas quanto a qualificação do atendimento com *Small Data*.

1.4. Estrutura do trabalho

Este trabalho está dividido da seguinte forma:

Capítulo 1 – Seção introdutória apresentada, expondo o contexto e a justificativa pela dedicação em análise de sentimentos, além dos objetivos gerais e específicos aos quais este trabalho é direcionado.

Capítulo 2 – Fundamentação teórica: estrutura necessária para que houvesse o conhecimento e posterior desenvolvimento deste, bem como conceitos relacionados a Análise de Sentimentos.

Capítulo 3 – Desenvolvimento: é descrita a metodologia adotada para este desenvolvimento, bem como cada etapa executada, para analisar os comentários dos clientes com relação ao atendimento que a empresa vem prestando.

Capítulo 4 – Resultados: são relatados os dados obtidos com a realização da classificação, realizada com o Algoritmo *Naive Bayes.*

Capítulo 5 – Conclusão: é apresentada aqui a contribuição deste, além de sugestões de trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Na literatura é possível se deparar com muitos trabalhos que utilizam à análise de sentimentos de frases extraídas de diversos meios. A finalidade deste capítulo é apresentar os principais aspectos conceituais, que foram imprescindíveis para a produção deste trabalho.

2.1. *Small Data*

O conceito *Small Data* pode ser definido pelo uso de dados simples e pequenos, que em análise conjunta podem gerar ideias inovadoras, oportunidades (MITI, 2016), não havendo a necessidade da utilização de ferramentas robustas e pessoas altamente qualificadas, pois os dados estão prontos para consumo, diferentemente do *Big Data*, como aprofundado no próximo item.

Nas empresas, possuir informação é indispensável para os negócios, sendo informação compreendida como a base para a construção do conhecimento, pois uma vez processadas são de grande importância para a tomada de decisão, por este motivo as empresas estão investindo constantemente em dados.

Para se alcançar o objetivo de real transformação do negócio, é preciso considerar a utilização de *Small Data*, que segundo Mario Hime "é uma coleção de informações que podem ser encontradas por exemplo em um CRM (*software* voltado para o relacionamento com o consumidor) de uma pequena empresa ou pesquisas de mercado, o *Small Data* há anos suporta as tomadas de decisões das empresas, muito antes do *Big Data*, pois as empresas já utilizavam o histórico de compras para as tomadas de decisão" (HIME , 2018). O termo se refere a uma parte dos dados, que oferece uma visão mais profunda de quem é o cliente, quais são os seus gostos, opiniões, expectativas, frustrações e de como ele se comporta, pois consumidor é a peça-chave do negócio, sendo preciso desenvolver para ele estratégias direcionadas, havendo uma diferenciação do cliente, personalização do seu atendimento e uma maior interação entre o consumidor e a empresa, para uma fidelização entre ele e a organização, sendo possível descobrir as causas e motivos nas pequenas correlações do volume de dados em análise, além de permitir uma decisão em curto prazo, sem grandes ruídos de informações, permitindo-se encontrar respostas para decisões assertivas.

Para Lindstrom (2016) *Small Data* é essencial no estudo do comportamento humano, muito usado em pesquisas que visam explicar causas de acontecimentos pontuais. Dessa maneira, o autor desenvolveu a metodologia 7C que consiste em sete etapas, que são:

a) coleta: momento de estabelecer uma busca de informações.

b) comunicação: nessa etapa se devem buscar pistas.

c) conexão: buscar relações entre os dados coletados;

d) correlação: nessa etapa deve-se buscar fatos pessoais que podem ter influenciado as mudanças de comportamento e hábitos.

e) causalidade: identificar quais sentimentos estão por trás da mudança;

f) compensação: identificação das lacunas.

g) conceito: momento de juntar as informações e buscar solução ou decisão.

2.1.1 Diferença entre *Big Data* e *Small Data*

A fim de tornar clara a diferença entre estes dois conceitos é preciso definir *Big Data* como conjunto de dados com grande volume, variedade e velocidade que estão em constante atualização (MCKINSEY GLOBAL INSTITUE, 2011), sendo mais complexa a extração de *insights*, no que diz respeito ao comportamento humano, enquanto, *Small Data* são dados prontos para análise e sem forma definida, podendo ser fotos, depoimentos de clientes entre outros (LINDSTROM, 2016), por este motivo não há a necessidade prévia de qualificar o conteúdo.

Para auxílio da tomada de decisão com relação a análise de sentimento expressa pelo cliente, o *Small Data* se destaca por buscar causas nos dados, tornando a busca mais específica e direcionada.

2.2. Sentimentos, Polaridade e Emoções

Sentimento e outros conceitos relacionados, como avaliação, postura e humor, estão vinculados a crenças e sensações das pessoas (LIU, 2012), não havendo consenso na área de mineração de opinião sobre uma forma padrão para medir sentimentos.

Como não existe uma única definição quanto às categorias básicas de emoção, cada autor define o seu conjunto de emoções que mais se adéque aos seus objetivos, por este motivo sendo recorrente a utilização também da polaridade.

Quando se usa emoção, o objetivo é classificar o sentimento em categorias como tristeza, alegria, surpresa, entre outras. Há a caracterização de emoções sobre três componentes: valência, excitação e dominância (WARRINER; KUPERMAN; BRYSBAERT, 2013). Valência corresponde ao quanto agradável um estímulo é, variando de infeliz a feliz. Excitação é relacionada à intensidade. Dominância denota o poder de um estímulo, variando de fraco/submisso a forte/dominante. Cada estímulo tem uma pontuação atribuída aos seus componentes, a qual varia de 0 (baixo) a 10 (alto).

O uso de polaridade é bem mais simples, já que tende a posicionar o sentimento em uma escala cujos sentimentos variam do negativo ao positivo, o que justifica sua popularidade (TSYTSARAU; PALPANAS, 2012).

2.3. Mineração de dados

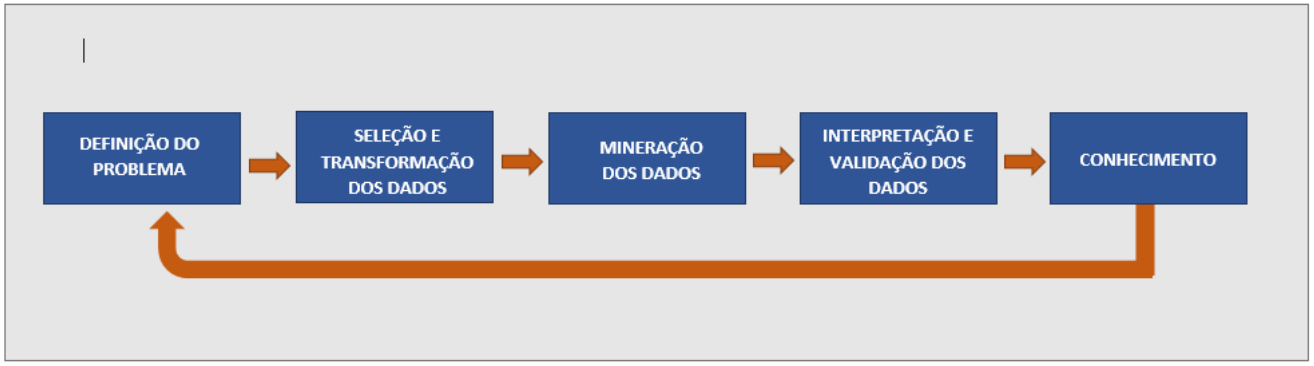
A Mineração de Dados consiste na etapa principal de uma maior atividade denominada descoberta de conhecimento em base de dados (*Knowledge Discovery in Database -KDD*), no qual, conforme Baker et al. (2012), é importante estabelecer metas para que se obtenha conhecimento relevante, onde há um foco maior na descoberta de padrões entre os dados.

É válido ressaltar que estes resultados gerados, necessitam ser compreensíveis, principalmente para os usuários finais do processo, que geralmente são os responsáveis pela tomada de decisão nas empresas. A principal finalidade deste processo é permitir a melhora dos resultados das explorações feitas utilizando ferramentas tradicionais de exploração de dados (Morais e Ambrósio 2007).

Este processo é executado de forma muito semelhante ao de Descoberta de Conhecimento em Textos, sendo o foco deste trabalho. Sendo necessária desde a definição do problema, até a extração de conhecimentos. Na definição do problema, os objetivos a serem alcançados devem ser definidos, para que em seguida dê continuidade a fase de seleção e pré processamento dos dados, na qual os dados normalmente não estão em um formato adequado para extração de conhecimento, por isso faz-se necessária a aplicação de métodos para extração, integração, transformação, limpeza, seleção e redução de volume destes dados, antes da etapa de mineração (MORAIS; AMBRÓSIO, 2007). Na etapa seguinte de Mineração de dados, busca-se alcançar os objetivos delimitados ao início do processo, nesta fase são configurados e executados algoritmos em busca de padrões relevantes, podendo

executá-los diversas vezes, até que sejam alcançados os resultados esperados. Por fim, na última fase do processo, a de Conhecimento, diversos padrões podem ser identificados, alguns não relevantes, outros interessantes para o domínio do problema.

Figura 1 - Processos da Mineração de Dados



Fonte: Adapatado de Zaidan, 1996

Na Figura 1, ilustra-se os principais resultados gerado pelo processo que devem ser apresentados. Este processo no total é parecido com o de Mineração de Textos, porém, ele lida com dados estruturados enquanto a Mineração de Textos é aplicada sobre dados não estruturados. O conceito de Mineração de Textos é explicado no tópico seguinte.

2.3.1. Mineração de Textos

A Mineração de Textos, trata do processo de extração de padrões ou conhecimento em textos, que se encontram no formato não estruturado, nela são aplicadas as mesmas funções analíticas da Mineração de Dados (GOMES, 2013), porém para dados textuais que possuem uma grande fonte de informação, mesmo em um formato que seja difícil de extrair de maneira automatizada.

Este número grande de informação não pode ser utilizado por computadores, pois os mesmos tratariam apenas como uma sequência de caracteres, fazendo-se necessária a aplicação de diferentes métodos e algoritmos para dar estruturação aos dados textuais, visando facilitar a extração de conhecimento dos respectivos dados.

Aranha (2007) propõe um modelo de Mineração de Textos contendo cinco fases distintas: coleta, pré-processamento, indexação, mineração e análise.

Figura 2 - Etapas do processo de Mineração de Textos

Fonte: Aranha, 2007

Na primeira etapa que é a de coleta, utiliza-se ferramentas ou outros meios para realizar a extração das informações, para extração dos textos que serão utilizados para a extração de conhecimento. No pré-processamento, são utilizadas técnicas como o Processamento de Linguagem Natural para estruturar os textos que serão analisados. A indexação é a etapa onde são extraídos conceitos dos documentos através da análise de seu conteúdo e traduzidos em termos da linguagem de indexação. Esta representação identifica o documento e define seus pontos de acesso para consultas (GOMES, 2006). Na etapa de mineração, são aplicados métodos e algoritmos para a identificação de padrões interessantes e extração de conhecimento. Na parte de análise, os resultados são avaliados e validados.

Para a realização deste trabalho, os passos seguidos foram os definidos por Aranha (2007).

2.4. Análise de sentimentos

A Análise de sentimentos, que pode ser encontrada também como “Mineração de Opiniões”, pertence a um âmbito que estuda opiniões, sentimentos, avaliações, atitudes e emoções direcionadas a entidades, que podem ser produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos em relação aos seus atributos e as suas características (LIU, 2012). O conceito de opinião ainda é muito vago,

todavia a análise de sentimentos valida principalmente opiniões que expressam ou implicam sentimentos em duas classes principais, sendo elas a positiva ou negativa.

Umas das principais etapas no processo de análise de sentimento é a classificação (BECKER; TUMITAN, 2013), que se refere aos sentimentos que constam em um documento, podendo o mesmo ser analisado dentro de um grupo, ou de forma individual. Para uma melhor validação, se faz necessário um pré-processamento, via técnicas de linguagem natural (PLN) no texto original. Para classificação, existem diferentes abordagens, como a léxica, a de aprendizado de máquina e a estatística (TSYTSARAU; PALPANAS, 2012).

A principal finalidade da análise de sentimentos é poder “definir técnicas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser empregado por um sistema de apoio ou tomador de decisão” (BENEVENUTO, RIBEIRO e ARAÚJO, 2018).

Benevenuto, Ribeiro e Araújo (2018) citam duas principais técnicas utilizadas para extrair sentimentos em textos: a supervisionada e a não-supervisionada. Enquanto a primeira exige uma etapa de treinamento de um modelo com uma porcentagem de amostras previamente classificadas, a segunda não realiza treinamento de modelos de aprendizado de máquina e faz uso de um dicionário de termos. Na técnica não-supervisionada, cada termo está associado a um sentimento, que possui um significado qualitativo ou quantitativo, ou seja, um valor numérico que varia em uma escala de -1 a 1, onde -1 é o valor sentimental mais negativo e 1 o mais positivo.

O objetivo desde trabalho é analisar tíquetes escritos por usuários em um determinado sistema de suporte à área de TI, utilizando a técnica supervisionada.

2.5. Processamento de Linguagem Natural para Análise de Sentimentos

O Processamento de Linguagem Natural (PLN) possui uma vasta coleção de técnicas que permitem a manipulação e transformação de textos e também processar a linguagem natural humana para várias atividades, as quais são essenciais ao pré-processamento de textos para classificação de sentimentos na mineração de opiniões (FELIX, 2016). O PLN está relacionado à compreensão da linguagem, oral ou escrita, e pode se dar em diferentes níveis sendo os principais:

- Morfológico: conhecimento das construções e dos componentes de palavras;

• Sintático: conhecimento das relações estruturais entre palavras;

• Semântico: conhecimento do significado de palavras em sentenças.

Na etapa de pré-processamento, as técnicas de PLN podem ser utilizadas para a remoção de *stopwords*, remoção de *Stemming*, dentro outras. O termo *stopword*, pode ser definido como palavras nos textos que não possuem importância ao assunto (EL-KHAIR, 2006), enquanto *Stemming*, pode ser caracterizado pela separação de cada termo em radical e terminação, removendo-se a terminação (FELIX, 2016) (OLIVEIRA, 2013) (RIBEIRO, 2015).

Para a realização deste trabalho, as técnicas fornecidas pela PLN foram fundamentais, para toda a etapa de estruturação dos dados, auxiliando na melhora da classificação de texto.

2.6. *Pandas*

A biblioteca *Pandas*, foi desenvolvida em linguagem *Python,* para possibilitar a observação de dados de forma mais simples e com grande performance, possuindo para isso estruturas rápidas e flexíveis.

Para McKinney (2011), o *Pandas* dispõe de ferramentas para trabalhar com conjuntos de dados estruturados comuns a estatísticos. Tal biblioteca proporciona meios integrados para executar, manipular e analisar dados

Uma das estruturas dessa biblioteca mais utilizada é a de *data-frame*, que trás uma estrutura de tabelas com indexação integrada, onde cada coluna possui um índice, associado a um conjunto de valores, sendo cada linha com vários valores, um deles referente a cada coluna.

2.7.*Scikit-Learn*

O *Scikit-Learn* é uma biblioteca de aprendizado de máquina de código aberto para *Python*, possuindo uma ampla gama de ferramentas para mineração e análise de dados, constituindo uma mistura de pacotes como *NumPy*, *SciPy* e matplotlib.

Seus recursos principais são algoritmos para:

• Clusterização: agrupamento de objetos com características similares.

• Regressão: para predição de atributos de valores contínuos para objetos a eles associados.

• Seleção de modelos: módulos para comparação e validação de parâmetros e modelos.

• Redução de dimensionamento: ou diminuição do número de variáveis a serem utilizadas.

• Pré-processamento: preparação dos dados.

• Classificação: métodos de atribuição de um dado a um conjunto específico.

2.8. Tokenização

Tokenizar texto é o primeiro passo no pré-processamento de dados, caracterizado pela ação de fragmentar uma sequência textual em palavras ou elementos significativos, chamados *tokens*, para uso posterior em análise e mineração de dados. A palavra *Token* é utilizada ao se referir a somente uma palavra do texto, porém em alguns casos, não podem ser consideradas palavras ou apresentam, mais de uma palavra: “SP”, “R$2,00” e “super-homem”. Estes *tokens* de 1..n, são denominados n-gram.

Este processo de fragmentação do texto, consiste na separação de palavras, tendo como referência os espaços entre elas ou os elementos de pontuação (FELDMAN & SANGER, 2007). A tokenização deve ser ajustada à necessidade de cada problema. O objetivo deste processo é transformar frases num conjunto de *tokens*, de forma a poder trabalhar os dados.

Nos casos em que há a presença de um dicionário de dados, o mesmo poderá verificar as sequências de caracteres do termo, a fim de validação do termo, corrigindo alguns erros ortográficos.

2.9. Stemização

O processo de *Stemming* consiste na redução das palavras ao seu radical, ou seja sua raiz morfológica. Em alguns casos ela não é validada, pois uma raiz pode conter várias derivações, até mesmo com classes gramaticais diferentes (Moral, de Antonio, Imbert e Ramirez, 2014).

Um exemplo dessa não validação, pode ser com relação a aplicação da função na palavra “livro” que é um substantivo, que ao final do processo de *Stemming,* ela volta assumir seu radical como “livr”, o mesmo do advérbio “livremente”, ou seja mesmo radical e significados diferentes.

O objetivo principal ao se utilizar esta função é a possibilidade da redução da grande dimensionalidade das aplicações de Mineração de Textos, pois com as

palavras sendo reduzidas ao seu radical, obtém-se um único token para representar todas elas, e não mais vários, devidos as suas alterações.

2.10. *Stopwords*

Com o processo de Tokenização realizado, é preciso remover aqueles “*tokens*” que não possuem valor semântico para o entendimento da sentença. Esses *tokens* são classificados como *stopwords* em um sistema de mineração de textos. As *stopwords* mais frequentes são as preposições, artigos e pronomes, como por exemplo podemos citar as palavras “um”, “uma”, “a”, “e” e “para”, que tornam-se irrelevantes para o contexto (Sedbrook e Lightoot, 2010). A sua remoção da sentença, agrega uma melhora na velocidade de processamento de dados. Normalmente, 40 a 50% do total de palavras de um texto são removidas com este processo.

2.11. Aprendizado de máquina

O aprendizado de máquina, que pode ser tratado também como reconhecimento de padrões ou mineração de dados (SCHERER; MEULEMAN, 2013), é uma segmentação da área da inteligência artificial, na qual é estudado e desenvolvido algoritmos em que o que é implementado, de fato, é a maneira com a qual o algoritmo melhora seu desempenho para executar determinada função contrastando assim com os algoritmos "tradicionais" em relação ao que de fato é tarefa do programador. O principal objetivo é prever com precisão uma classe ou atributo contendo valores reais, com base em outros atributos e um modelo com parâmetros pré-estabelecidos.

Nos algoritmos tradicionais, é necessário definir e explicitar a ordem com as quais os dados serão processados para adquirir as respostas. Já em aprendizado de máquina é possível escolher um modelo que possua suas próprias regras com as quais conseguirá aperfeiçoar as capacidades de obter a resposta após uma fase de treino.

Os algoritmos de aprendizado de máquina, podem ser classificados em duas classes principais: a supervisionada e não supervisionada. Segundo Baeza-Yates e Ribeiro-Neto (2013), um algoritmo de classificação é considerado supervisionado quando uma coleção de treinamento é usada para treinar o classificador.

Para os algoritmos não supervisionados, não é necessário que sejam fornecidos exemplos de treinamento (BAEZA-YATES; RIBEIRO-NETO, 2013). Os

algoritmos supervisionados, necessita dos dados e suas respectivas respostas, e os algoritmos não supervisionados, necessitam somente dos dados para criar regras que os façam modelar alguma função relevante em relação aos dados. Ambos precisam de uma etapa de treino, que é quando os dados (e resposta, quando supervisionado) são fornecidos ao algoritmo para que o mesmo, de acordo com cada algoritmo, consiga obter uma função que, após o treino, consiga predizer a resposta correta para dados não vistos.

O aprendizado supervisionado pode apresentar a inconveniência no uso, pois precisa de vários registros de exemplos para uma boa performance e confiabilidade no treinamento (SILVA; LIMA; BARROS, 2012), (QIU et al., 2009), sendo importante algumas vezes a construção de exemplos manuais para uma melhor detecção dos sentimentos e polaridade.

A análise de sentimentos realizada pela técnica supervisionada emprega textos já classificados que servem como base de treinamento. Com o uso de modelos já treinados, tenta-se prever o conteúdo emocional de um texto desconhecido.

Neste trabalho utiliza-se um algoritmo de aprendizado de máquina supervisionado, por consequência, os dados foram divididos em duas classes, sendo uma delas de treino, na qual já possuía as respostas associadas a cada instância.

2.12. Algoritmo *Naive Bayes*

O algoritmo *Naive Bayes* é excelente classificador probabilístico baseado na aplicação do teorema de *Bayes*, (RIBEIRO, 2015) que consiste na probabilidade de um evento ocorrer, baseado em um conhecimento, a priori que poder estar de alguma maneira relacionado ao evento, além de permitir a alteração das probabilidades, quando há novas evidências, sendo o nome *naive* (ingênuo) por assumir algumas fortes hipóteses de independência entre as variáveis e por também ser chamado de *bag of words* ou saco de palavras, onde supõe-se que a ordem das palavras dentro de um documento não tem importância, sendo claramente falso na prática, mas não causa complicações nos resultados e é algo simples para a implementação..

A utilização deste é muito recorrente em classificação de textos, por ser ágil e simples para implementação. Gomes (2013) defende a ideia de que este classificador é o mais eficiente em questões relacionadas com o processamento e precisão na classificação de novas amostras.

O classificador *Naive Bayes* é fundamentado no Teorema de *Bayes*, como representado pela equação abaixo, na qual a variável B representa um evento que ocorreu previamente e A um evento que depende de B, para que seja calculada a probabilidade de A ocorrer dado o evento B, o algoritmo verifica o número de casos em que A e B ocorrem juntos e dividir pelo número de casos em que B ocorre sozinho.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Como é um método de aprendizado de máquina supervisionado, ele possui uma etapa de treino, onde são calculados estimadores ótimos para os parâmetros do modelo.

2.13. *Python*

Python é uma linguagem de programação surgida no ano de 1989, desenvolvida por Guido van Rossum (PYTHON, 2019) dentro do instituto nacional de pesquisa em Matemática e Ciência da Computação, CWI (Centrum Wiskunde & Informatica), na Holanda.

Esta linguagem surgiu para corrigir algumas falhas que a linguagem C apresentava, por isso sua semelhança de código, tornando -se uma linguagem interpretada, orientada a objetos e interativa, podendo ser utilizadas para diferentes finalidades, como a análise de dados.

Por todas as funcionalidades desta linguagem, ela permite uma fácil aprendizagem, podendo ser utilizada por profissionais que não pertencem ao setor de informática, mas que precisam dela para automatizar tarefas distintas.

Segundo Jeff Hale (HALE, 2018), por conta de seu alto crescimento em usabilidade e sua alta popularidade entre os profissionais dentro e fora do setor de informática, ela está entre as 20 habilidades mais requeridas para os profissionais que trabalham com análise de dados.

2.14. *Tickets*

Um sistema de *Tickets,* tem a finalidade de auxiliar a comunicação entre o cliente e a empresa, com relação a um chamado a ser resolvido. Isto nada mais é que “um registro único eletrônico feito por um usuário autorizado de uma solicitação

de serviço iniciada através de uma central de atendimento, uma interface *Web* ou outros meios" (KARLYN, 2013). Através dele o cliente por intermédio de um formulário eletrônico descreve uma situação, que deverá ser recebida pela empresa, que designará um responsável para atender a este *ticket*.

Neste sistema de atendimento, o solicitante insere os detalhes do problema que ocorre, ou melhorias a serem implantadas, da melhoria que necessita ser desenvolvida no sistema para atender um cenário de utilização do mesmo. Em contrapartida, a empresa que o recebe, o atrela a um funcionário para resolver aquela questão. Por fim, depois da solução desenvolvida o cliente, avalia aquele chamado, que é utilizado para a empresa como medida de satisfação pelo atendimento.

A adoção de *tickets* para o recebimento de demandas de clientes é uma prática adotada em diversas empresas, devido a facilidade no treinamento dos funcionários para utilização do processo, distinção entre tipos de usuário, controle de status, entre outras vantagens.

No próximo capítulo são descritos os métodos utilizados para o desenvolvimento da solução.

3. METODOLOGIA

Para resolver o problema da empresa com relação ao comportamento dos clientes, era exigida uma abordagem colaborativa ente pesquisador/organização, visto que a solução seria de grande valia para ambas as partes, necessitando um estudo minucioso para conhecimento implícito nos dados.

Com este intuito, a metodologia de pesquisa-ação, foi escolhida pois ela engloba pesquisa e ação em um único processo, no qual as partes envolvidas participam, junto com os pesquisadores, para chegarem interativamente a elucidar a realidade em que estão inseridos, identificando problemas coletivos, buscando e experimentando soluções em situação real. Simultaneamente, há produção e uso de conhecimento (THIOLLENT, 1997).

McKay e Marshall (2001) realizaram um modelo com cada etapa a ser desenvolvida para um projeto envolvendo pesquisa-ação, que é subdivido em 8 fases, descritas abaixo.

Fase1: Identificação do problema, consiste na tarefa do pesquisador em identificar o problema que tenha interesse em resolver.

Fase 2: O pesquisador deve se empenhar em promover uma ampla revisão de literatura em busca de teorias que possam estar alinhadas com fatos relevantes sobre o problema e sirvam para dar suporte à solução do problema.

Fase 3: Desenvolvimento de um plano de ações para a solução do problema.

Fase 4: Inserção do plano de ação desenvolvido em prática.

Fase 5:Acompanhamento das ações implementadas para observação dos resultados.

Fase 6: Avaliação do efeito das ações, ponto de decisão

Fase 7:Ocorre enquanto os resultados obtidos na Etapa 6 não forem satisfatórios.

Fase 8:Etapa final, o problema deverá estar resolvido e os objetivos da pesquisa atingidos com sucesso.

3.1. Criação de uma base para *Small Data*

Antes de se iniciar a fase da coleta de dados, é criada uma base no banco de dados *Sqlite*, para armazenamento das observações realizadas sobre o

comportamento dos clientes, e posteriormente coleta, tratamento e análise dos dados. Os experimentos são detalhados abaixo.

Durante o período de análise foram realizadas ligações telefônicas, envios de e-mails, visitas locais e acompanhamento de redes sociais associadas a empresa para colhimento de informações (*Small Data*), o qual para seu armazenamento foi implementado um banco de dados no *Sqlite* para armazenar as informações observadas com relação ao comportamento do cliente, durante e após o atendimento.

3.2. Extração dos Dados

Para a realização deste trabalho foi utilizado o banco de dados do sistema de tickets da empresa que é relacional, utilizando hoje o *MySql*, o qual é único não havendo separação por departamentos existentes.

Uma consulta SQL foi realizada no banco de dados da empresa a fim de coletar os registros para análise. Os dados coletados e exportados no formato “CSV” possuem 350 *tickets*, sendo que cada um contém registros de avaliação enviados por clientes após serviços prestados pela empresa categorizados como o rótulo “Avaliação do chamado”. Os dados são referentes ao ano de 2019 entre os meses de janeiro e junho, no qual o sistema foi implantado pela empresa.

As variáveis para análise são:

- id: identificador de uma mensagem (e-mail) enviada por um cliente;
- empresa: identifica o nome do cliente (empresa);
- nome: nome do representante do cliente (empresa);
- prioridade: prioridade de atendimento de um ticket;
- avaliação: resposta ao atendimento realizado;
- técnico: indica o responsável pelo atendimento;

As variáveis “id” e “prioridade” são quantitativas discretas. Todas as demais variáveis são qualitativas nominais.

3.3. Definições de polaridade

Com o intuito de auxiliar a classificação manual dos comentários, foi criada esta guia com as definições de polaridade de sentimento dentro do domínio do trabalho, para que as classificações sejam consistentes e não exista um ruído de opiniões subjetivas quanto aos sentimentos presentes nos dados classificados. Para cada

categoria (positivo ou negativo), foram listados os principais critérios para pertencimento com exemplos da própria base de dados.

Polaridade positiva: a emoção associada a esse tipo de comentário é predominantemente positiva.

• Parabenização - "Meu Problema foi resolvido, ótimo trabalho. Grato."

• Agradecimento – “Serviço bem feito!"

• Elogios - "Muito obrigado!!!! Show.”

Polaridade negativa: emoção associada a esse tipo de comentário é predominantemente negativa.

• Xingamentos a entidades ou usuários - "parem de reclamar Zé povinho"

• Insatisfação - "Serviço mal feito, muito mal executado”

• Cobranças - "E aí pessoal, vamos resolver não?"

• Aguardando solução - "Aguardando providências."

3.4. Tratamento dos Dados Extraídos

Todas implementações relacionadas ao processamento e visualização dos dados foram realizadas utilizando a linguagem *Python* 3.6, pois permitia que esse tratamento pudesse ser feito de forma mais rápida e eficaz.

3.5. Pré-processamento dos *tickets*

Após a coleta de dados, foram aplicadas técnicas de mineração de texto na variável em linguagem *Python*, pois os dados precisam ser tratados, com o propósito de descartar o irrelevante para a etapa de classificação (FELIX, 2016). Para as etapas descritas abaixo foi utilizada a plataforma *NLTK* que já possuía os recursos para Tokenização, remoção *Stemming,* e remoção de *Stopwords*, como descritas abaixo.

• Conversão de letras maiúsculas em minúsculas, com o objetivo de padronizar o texto.

• Tokenização do texto, que consistiu na remoção de caracteres não alfabéticos e pontuação, pois não agregam valor à classificação.

A Tabela 1 evidencia alguns exemplos de caracteres removidos do texto.

Tabela 1. Exemplos de caracteres especiais

Descrição	***Token***
Acentos	´ ~ ^
Pontuação	‘ ’ , . ; : ? !
Especiais	@ # * ( ) &
Emoticons	:) ;) :D :(:(;(
HTML	<p>

Fonte: Autoria própria, 2019.

• Remoção de *Stemming*, para remover apenas o sufixo de uma palavra. Ex. A palavra “meninas” se reduziria a “menin”, assim como “meninos” e “menininhos”.

• Remoção de *stopwords*, para retirar do texto as palavras que não trazem sentido ao texto.

A Tabela 2 mostra exemplos de *stopwords*, palavras que são bastante comuns em um idioma e, portanto, não possuem muito valor semântico. Por isso, são removidas durante o pré-processamento (FILHO, 2014). Foi utilizada a própria lista de *stopwords* desta biblioteca nesta etapa do pré-processamento.

Tabela 2. Exemplos de *stopwords* removidas

.

a	de	deu	e	esta
meu	muito	mesmo	nossa	para
talvez	tem	tendo	pelo	com

Fonte: Autoria própria, 2019.

3.6. Desenvolvimento e validação do modelo de classificação

Os Algoritmos de Classificação pertencem a subcategoria de Aprendizagem Supervisionada, na qual Classificação diz respeito ao processo de tomar algum tipo de entrada e atribuir-lhe um rótulo. Sistemas de classificação são usados geralmente quando as previsões são de natureza distinta, como por exemplo "positivo" e "negativo".

Para este trabalho, foi escolhido um algoritmo classificador baseado no modelo *Bayes* com relação a outros algoritmos de classificação mais conhecidos, como por exemplo *IBK*, *Forest* e *Random Committee*, por este utilizar a probabilidade condicional para criar o modelo de dados a ser trabalhado. O algoritmo *Naive Bayes* determina uma classe da sentença não apenas pelas palavras existentes, mas pelas frequências que elas acorrem no texto (Wittern et al., 2016). Ademais, França e

Oliveira [2014] utilizam o algoritmo *Naive Bayes* no idioma português brasileiro e apresentam resultados de até 92% de acurácia ao classificar polaridades expressas.

A biblioteca *NLTK* além dos recursos para tratamento do texto, teve como principal objetivo o auxílio ao desenvolvimento e implementação do algoritmo *Naive Bayes*, pois possui treino de palavras. Porém, como ela não possui um algoritmo de validação cruzada que consiste no particionamento de um conjunto de dados em subconjuntos exclusivos, para posterior utilização para estimação dos parâmetros do modelo (dados treinamento), sendo o restante para validação (dados teste), foi utilizada a biblioteca *Scikit-Learn* para trabalhar em conjunto.

Para a realização deste, foi necessária a construção de uma base de dados com classificações de sentenças em positivas e negativas, para que assim fossem realizadas as inferências necessárias ao algoritmo de treinamento e após realizar a aplicação destas etapas de validação cruzada e teste.

A base original para treinamento e teste foi fornecida no *Kaggle*, na sua versão original há aproximadamente 500 mil linhas, de avaliações de empresas de cinema, como o processamento de uma grande quantia como essa pode demorar bastante, foi utilizada uma amostra menor, com 3.959 registros, sendo 1.999 classificados de forma negativa e 1.960 de forma positiva. A mesma foi traduzida por uma ferramenta de tradução automática, com a classificação das sentenças em positivas ou negativas. A figura abaixo, apresenta as informações contidas na base.

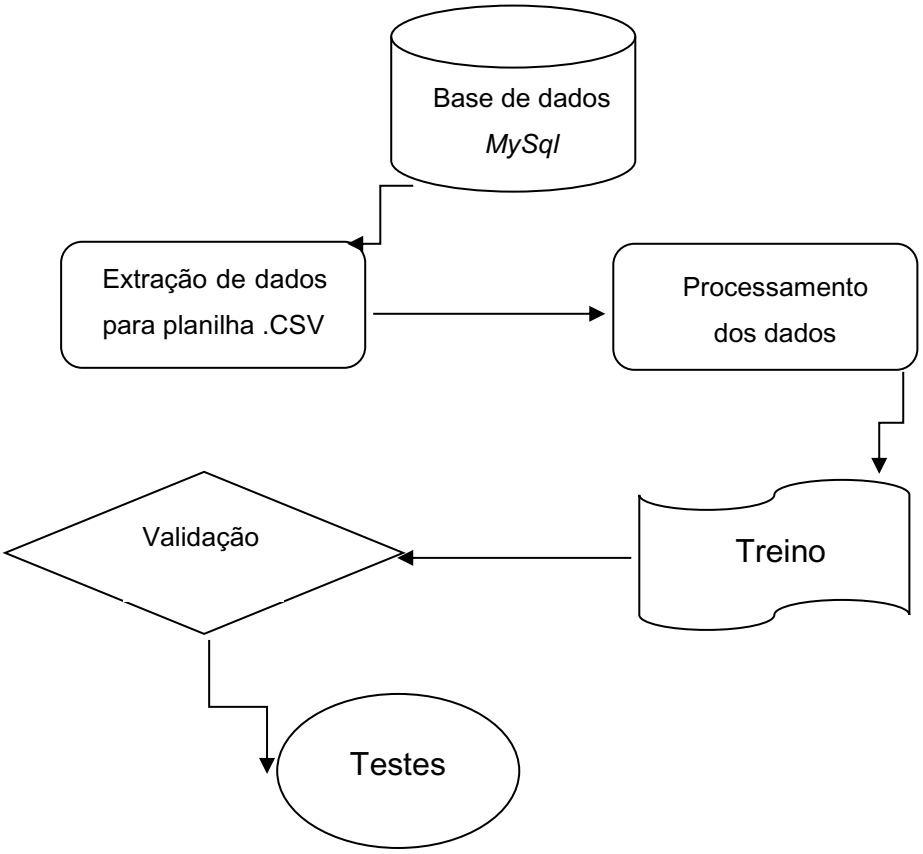
Figura 3 – Base para treinamento e teste

	text_pt	sentiment	classification
1	este é um exemplo do motivo pelo qual a maiori...	neg	0

Fonte: Autoria própria, 2019

Foram seguidas algumas etapas para tornar possível a obtenção da base da empresa de forma íntegra, conforme é visto na figura 4.

Figura 4 - Fluxo de análise



Fonte: Autoria própria, 2019.

Os dados coletados e utilizados nos procedimentos acima foram distribuídos segundo a seguinte ordem: 66% para treino e 34% para testes. Com isto foi possível validar a ferramenta de forma a dar continuidade no seu desenvolvimento.

As analises verificam a qualidade do classificador, ou seja, a eficácia do classificador utilizado. Para tanto, depois da etapa de treino, o classificador foi avaliado com a base de teste. Foi verificada a acurácia e a matriz de confusão gerada.

O capítulo a seguir, descreve os resultados obtidos.

4.RESULTADOS

Este capítulo tem como objetivo analisar os dados obtidos pelo pré-processamento dos *tickets,* os resultados obtidos e análise da base criada para a avalição pertencente ao *Small Data.*

4.1. Criação da Base *Small Data*

Para armazenar as observações coletadas com relação ao comportamento dos clientes, que foram realizadas através de ligações, visitas ao local, entre outros meios, foi implementado um banco de dados no *Sqlite* para que fosse possível registrar estes fatos.

Na figura abaixo, é possível identificar os campos criados para estas informações adquiridas.

Figura 5. Campos criados no banco de dados para *Small Data*.

```
CREATE TABLE clientes (
    codigo_cliente                  INTEGER       PRIMARY KEY AUTOINCREMENT
                                                  NOT NULL,
    forma_contato_cliente           VARCHAR (30)  NOT NULL,
    solicitante_cliente             VARCHAR (30),
    reclamacao_cliente              VARCHAR (100),
    tecnico_responsavel_cliente     VARCHAR (30),
    atendente_chamado_cliente       VARCHAR (30),
    chamado_anterior_aberto_cliente VARCHAR,
    humor_cliente                   VARCHAR (30),
    fato_relevante_observado_cliente VARCHAR (100),
    problema_resolvido_cliente      VARCHAR (100),
    data_contato_cliente            VARCHAR (12)  NOT NULL,
    observacao_final_cliente        VARCHAR (100),
    cliente                         VARCHAR (40)  NOT NULL
);
```

Fonte: Própria autoria, 2019

Com esta figura é possível averiguar que observações pertinentes ao comportamento dos clientes foram registradas na data em que o chamado foi aberto pelo mesmo junto ao sistema.

4.2. Verificação do Resultado do Classificador *Naive Bayes*

Nesta seção são apresentados os resultados obtidos após a fase de treino e teste realizados sobre a base rotulada. Após treinar o algoritmo fornecendo ao *NLTK*

66% dos *tickets* classificados e rotulados como positivos ou negativos, utilizou-se os 34% restantes para analisar a acurácia. A Tabela 3 apresenta a matriz de confusão contendo as informações dos *tickets* classificados corretamente ou de maneira equivocada.

Tabela 3: Matriz de confusão do modelo

PREDITO/REAL	POSITIVO	NEGATIVO	TOTAL
POSITIVO	**705**	29	734
NEGATIVO	21	**592**	613
TOTAL	726	621	**1347**

Fonte: Própria autoria, 2019

A diagonal principal da matriz, representa os acertos do modelo, e os demais registros os erros de classificação do mesmo. A coluna representa a classificação real do dado e linha representa a predição do classificador. É possível verificar que os erros mais comuns do modelo envolvem classificar como negativo, tickets positivos.

A acurácia do modelo obteve um percentual de acerto de 91% com os dados de teste, como é evidenciado pela figura a seguir.

Figura 6. Acurácia do modelo desenvolvido

```
# Divindo no dataset em treino e teste
X_train, X_test, y_train, y_test = train_test_split(text_counts, df['classificati
                                                    test_size=0.34, random_state=
                                                    shuffle=True)
# Criar modelo e treinar
clf = MultinomialNB().fit(X_train, y_train)

# Fazendo  predict do valor de X para teste de acuracidade
y_predicted= clf.predict(X_test)
print("MultinomialNB Accuracy:",metrics.accuracy_score(y_test, y_predicted).round
```

```
MultinomialNB Accuracy: 0.912
```

Fonte: Própria autoria, 2019

4.3. Criação das Bases rotuladas

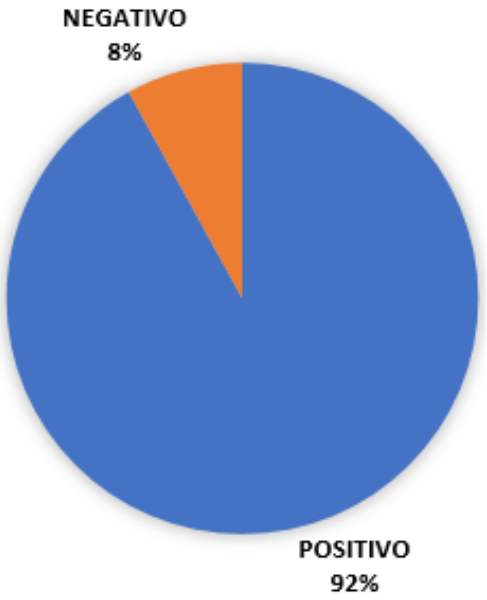
Os *tickets* da base foram exportados do banco de dados *MySql*, para uma planilha no formato .CSV para posterior utilização no ambiente *Python* . Depois da extração apenas dos campos de interesse para a análise, a remoção de *stopwords*, tokenização e o *stemming* foram implementados usando a linguagem *Python* e o módulo *NLTK*. A base rotulada foi construída após essa etapa.

4.4. Validando a base coletada

Nesta seção é apresentado o resultado obtido com o carregamento da base coletada do sistema de *tickets* da empresa.

Após o seu carregamento, foi aplicado todo o processo de pré-processamento detalhado no capítulo anterior na base, para pós classificação pelo modelo gerado. O gráfico 1, evidencia a porcentagem de avaliações dos clientes, que foram classificadas de forma positiva e negativa.

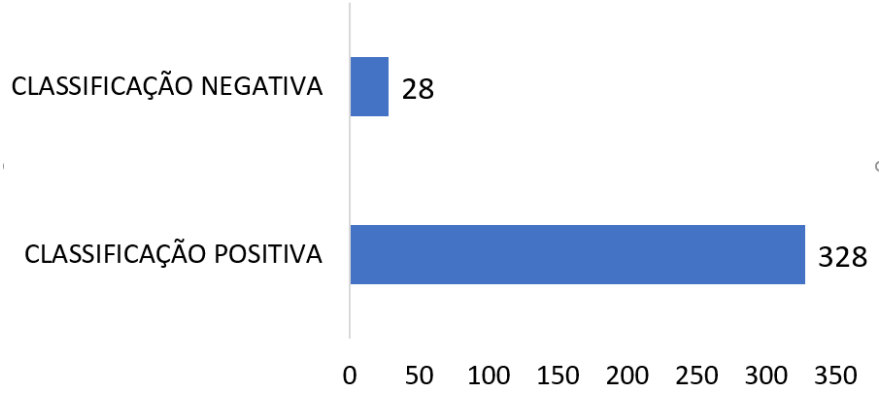
Gráfico 1: Resultado classificação de sentimentos.



Fonte: Própria autoria, 2019.

No gráfico abaixo é evidenciado o resultado em quantidades de registros.

Gráfico 2: Resultado em quantidades de registros.



Fonte: Própria autoria, 2019.

A partir dos gráficos acima, é possível verificar que uma porcentagem muito pequena dos atendimentos realizados pela empresa, foram classificados de forma negativa.

Com a finalidade de identificar o porque da avaliação negativa, é recorrida a base criada para *Small Data,* para identificar possíveis causas, para que a empresa possa saber se esta avaliação negativa, foi realmente pelo atendimento prestado, ou algum motivo pessoal do cliente naquele momento que o levou a gerar aquela avaliação. Na figura 7 é possível verificar parte dos resultados armazenados na plataforma para *Small Data*, que contém as possíveis causas desses 8% avaliados de forma negativa.

Figura 7: Possíveis causas da classificação negativa

fato_relevante_observado_cliente	problema_resolvido_cliente	data_contato_cliente	observacao_final_cliente	cliente
queria que o tecnico leandro atendesse	sim	29/01/2019	desapontamento	[redacted]
pressa no atendimento	sim	12/02/2019	desapontamento	
apreensiva pela voz	nao	28/01/2019	insatisfeita	
voz apreensiva	sim	30/01/2019	insatisfeita	
abatido pela voz	nao	15/02/2019	insatisfeita	
urgente	sim	08/02/2019	insatisfeita	
empresa parada até finalização do chamado	nao	12/01/2019	insatisfeito	
mal humorado	nao	26/01/2019	insatisfeito	
decepção com o computador comprado	nao	14/01/2019	insatisfeito	
perguntou pelo tecnico washington	nao	04/03/2019	insatisfeito	
queria falar com o técnico diego	sim	26/02/2019	insatisfeito	
cansada, clube exigindo que ele resolva logo	sim	24/02/2019	insatisfeito	
perguntou pelo tecnico diego	sim	25/02/2019	insatisfeito	
decepção com o computador comprado	nao	18/02/2019	insatisfeito	
irritada, com queda de energia varios equipamentos eletronicos queimaram	sim	08/02/2019	insatisfeito	
mal humorado	nao	07/02/2019	insatisfeito	
irritada, cliente parado com os pedidos a serem impressos	nao	04/02/2019	insatisfeito	

Fonte: Própria autoria, 2019.

Analisando a figura acima, é possível verificar os fatos relevantes observados com relação ao atendimento ao cliente, como por exemplo: mal humor, irritação e decepção, sendo esses registros separados por data, funcionário da empresa responsável pelo atendimento realizado e quem foi o cliente. Desta maneira a empresa observa essas possíveis causas da avaliação negativa por parte do cliente, e pode adotar medidas para reverter este cenário.

Na figura 8 é possível verificar parte dos resultados armazenados, para poder analisar se os atendimentos que foram classificados como positivos, foram classificados de forma correta.

Figura 8: Possíveis causas da classificação positiva

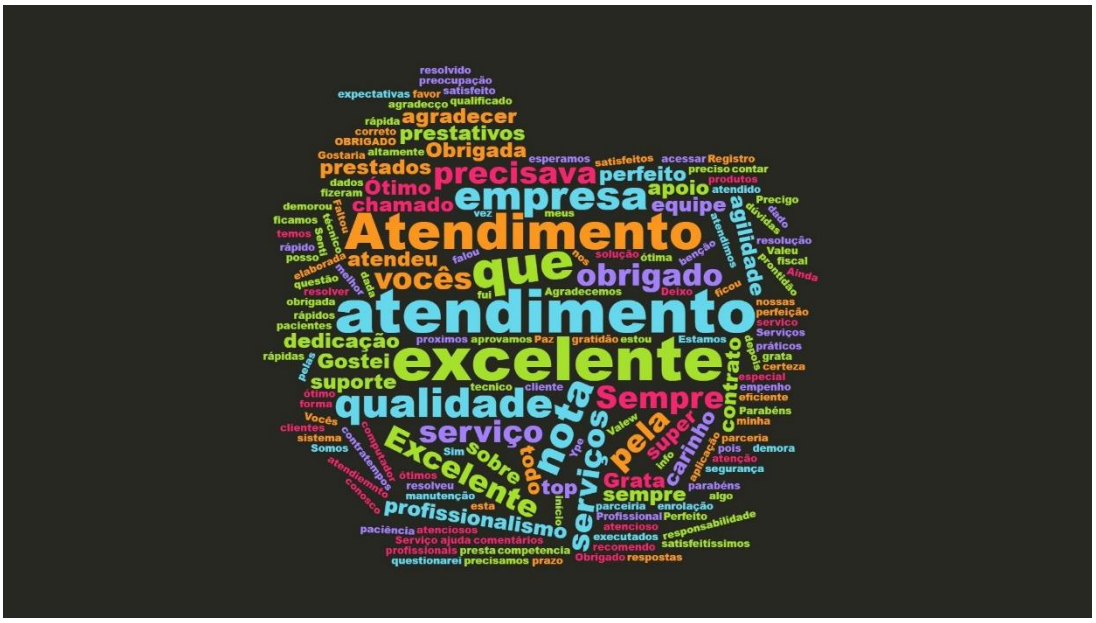
debora	não	tranquilo	pressa pra levar uma maquina para reunião	sim	26/03/2019	satisfeito
debora	não	pressa	impaciente, pra conseguir utilizar o sistema logo	sim	27/02/2019	satisfeito
washington	não	alegre	promovido a gerente neste dia	sim	24/02/2019	satisfeito
debora	não	pressa	pressa para resolver o problema para emitir nota ...	sim	28/03/2019	satisfeito
debora	não	feliz	entusiasmado com novo servidor da empresa	sim	27/03/2019	satisfeito

Fonte: Própria autoria, 2019.

A figura acima, mostra que embora o resultado final seja o cliente satisfeito, foi observado que alguns apresentavam pressa para conclusão do mesmo, preferência por funcionário a realizar o atendimento, com isso a empresa pode rever se esta classificação positiva, tem que ser melhorada.

Na figura 9, é possível verificar uma Nuvem de Palavras, que é uma lista hierarquizada visualmente, para evidenciar as palavras mais recorrentes na base extraída.

Figura 9. Palavras mais recorrentes na base



Fonte: Própria autoria, 2019.

Como possível verificar, em evidência se encontram apenas palavras positivas com relação aos atendimentos realizados pela empresa.

Os resultados obtidos são satisfatórios para esta finalidade. Os experimentos demonstram que o objetivo foi atingido o que demonstra que o método escolhido é eficaz na classificação de sentimentos de clientes com relação aos atendimentos realizados. Caso a classificação precise de um complemento a base de *Small Data*

deverá ser recorrida, para auxílio de tomada de decisão final, sendo trivial para o sucesso da decisão, como foi observado.

A próxima seção apresenta as considerações finais e projetos futuros.

5. CONCLUSÃO

Neste estudo de caso, foi apresentada a justificativa para a realização do mesmo, destacando-se o sistema de chamados da empresa envolvida, no caso o *tickets* a ser explorada nas áreas de Mineração de Dados e Análise de Sentimentos.

Foi realizada a análise dos sentimentos dos *tickets* relacionados a avaliação dos serviços prestados pelos funcionários da empresa, conforme o objetivo deste trabalho. Para isso, foram realizadas as etapas da tarefa de Análise de Sentimentos voltada para o sistema envolvido: extração dos *tickets*, pré-processamento dos dados, construção de uma base de dados rotulada, classificação dos textos e validação dos resultados, utilizando-se da abordagem de aprendizado supervisionado com o algoritmo *Naive Bayes,* que para análise textual, apresenta os melhores resultados, com relação a acurácia.

A base de *Small Data* produzida, foi utilizada em poucos casos, em que a classificação como positivo ou negativo, não respondia o que a empresa precisava saber, sendo utilizada então, para responder esta questão de forma mais assertiva, sendo responsável pela tomada correta de decisão da empresa.

Em resumo, a empresa obteve a resposta com relação a satisfação dos seus clientes com o trabalho que vem sendo prestado a eles, os mesmos na sua maioria estão satisfeitos com seus serviços, necessitando talvez de alguns reajustes mínimos, como por exemplo, a questão do tempo para finalização do chamado, como foi analisado, que pode ser apontado como um dos possíveis motivos para classificação negativa, com base nas observações presentes na base de *Small Data*.

Como projeto futuro, poderá ser acrescentada à análise, as outras categorias de classificação do sentimento, tornando o prognóstico mais completo, além de realizar uma análise em uma base mais rica, composta não só por *tickets*, mas também por comentários publicados nas redes sociais relacionadas a empresa.

Uma linha do tempo poderia ser construída, demonstrando como os usuários estão se expressando durante cada época escolhida.

Ademais, a metodologia aplicada neste estudo também pode ser aplicada para futuras situações em que se deseja obter um panorama sobre a opinião dos clientes da empresa em relação a certo(s) tópico(s).

6.BIBLIOGRAFIA

Abellón, Marcos. **Business Intelligence para pequenas empresas** Disponível em: <https://www.profissionaisti.com.br/2013/04/business-intelligence-para-pequenas-empresas/>. Acesso em: 04 out. 2018.

AFFELDT, F. S., JUNIOR, S. D. S. Information architecture analysis using business intelligence tools based on the information needs of executives. JISTEM - Journal of Information Systems and Technology Management . vol.10 no.2. Versão on-line. São Paulo maio/ago. 2013.

ARANHA, C.N. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) – Pontífica Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2007.

BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proc. of the International Conference on Language Resources and Evaluation (LREC), Valletta, Malta. [s.n.], 2010. Disponível em: . http://www.lrec-conf.org/proceedings/lrec2010/summaries/769.html. Acesso 24/07/2019.

BAEZA-YATES, R.; RIBEIRO-NETO, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. 2. ed. Porto Alegre: Bookman Editora, 2013. 590 p.

BAKER, Ryan S. J.; COSTA, Evandro; AMORIM, Lucas; MAGALHÃES, Jonathas; MARINHO, Tarsis. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. In: Jornada de Atualização em Informática na Educação, 2012. Anais eletrônicos disponíveis em: < http://www.br-ie.org/pub/index.php/pie/article/view/2341>. Acesso em: 13 abril 2019.

BECKER, K.; TUMITAN, D. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. In: Lectures of the 28th Brazilian Symposium on Databases. [S.l.: s.n.], 2013.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para Análise de Sentimentos em Mídias Sociais. In: WebMedia2015 (minicurso). Disponível em: . Acesso em novenbro de 2018.

BOLLEGALA, D.; MU, T.; GOULERMAS, J. Y. Cross-domain sentiment classification using sentiment sensitive embeddings. IEEE Trans. on Knowl. and Data Eng., IEEE Educational Activities Department, Piscataway, NJ, USA, v. 28, n. 2, p. 398–410, fev. 2016. ISSN 1041-4347. Disponível em: http://dx.doi.org/10.1109/TKDE.2015.2475761 Acesso em novenbro de 2018.

Computerworld. Mercado mundial de BI e analytics movimentará US$ 18,3 bi neste ano. Disponível em:< https://computerworld.com.br/2017/04/12/mercado-mundial-de-bi-e-analytics-movimentara-us-183-bi-neste-ano/ >.Acesso em: 08 out. 2018.

EL-KHAIR, Ibrahim Abu. Effects of stop words elimination for Arabic information retrieval: a comparative study. International Journal of Computing & Information Sciences, v. 4, n. 3, p. 119-133, 2006.

FAYYAD, U.; SHAPIRO, G.P.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases, *American Association for Artificial Intelligence*, 1996

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

FELIX, N. Análise de sentimentos em textos curtos provenientes de redes sociais. 138 p. Tese (Doutorado) Ů Universidade de São Paulo - São Carlos, 2016.

FILHO, J. A. C. Mineração De Textos: Análise de Sentimento Utilizando Tweets Referentes à Copa Do Mundo 2014. 2014.

FORBES INSIGHTS (2016). Breakthrough Business Intelligence - How Stronger Governance Becomes a Force For Enablement.

França, Tiago C. de; Oliveira, Jonice. (2014). Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013 . In Procs. of Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), p. 128 - 139, Brasília, Brasil.

GARTNER GROUP. **Key Issues for Analytics, Business Intelligence and Performance Management, 2011.** Disponível em: < http://www.gartner.com/technology/it-glossary/businessintelligence.jsp>. Acesso em: 10/06/2014.

GIGLIOTTI, W. Em Busca de Más Notícias. 2012.

GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. Large-scale sentiment analysis for news and blogs. In: Proc. of the First International AAAI Conference on Weblogs and Social Media (ICWSM). [S.l.: s.n.], 2007. v. 2.

GRIGORI, D.; CASATI, F.; CASTELLANOS, M.; DAYAL, U.; SAYAL, M.; SHAN, M. Business Process Intelligence. Computers in Industry. Elsevier. Vol. 53, Issue 3, April, P. 321–343. 2004.

GOMES, G. R. R. Integração de Repositórios de Sistemas de Bibliotecas Digitais e de Sistemas de Aprendizagem. 2006. 143 f. Tese (Doutorado em Informática) – Pontífica Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2006.

GOMES, Helder Joaquim Carvalheira. Text Mining: análise de sentimentos na classificação de notícias. Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on. Lisboa. 2013.

HALE, Jeff. https://towardsdatascience.com/the-most-in-demand-skills-for-datascientists-4a4a8db896db. <Acesso em 24 novembro 2018>.

HIME. Não deixe de desvendar o universo do *Small Data*. Disponível em: < https://cio.com.br/desvende-o-universo-do-small-data/> Acesso em 15 de novembro de 2018.

KARLYN, M. R. O. e M. A. A Guide to IT Contracting. Boca Raton, Florida, USA: CRC Press, 2013.

LINDSTROM, M. Small data: Como poucas pistas indicam grandes tendências. Rio de Janeiro: HarperCollins Brasil, 2016.

LIU, Bing. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, v. 5, n. 1, p. 1-167, 2012.

McKAY, J.; MARSHALL, P. The Dual Imperatives of Action Research. Information Technology & People, v. 14, n. 1, p. 46-59, 2001.

MCKINSEY. Marketing Analytics: funciona, então por que não são mais empresas que o utilizam? Disponível em: https://rockcontent.com/blog/web-analytics/ Acesso em 19 de Novembro de 2018

MITI. Big Data você conhece. E Small Data, você já ouviu falar? Disponível em: < http://miti.com.br/blog/big-data-voce-conhece-e-do-small-data-voce-ja-ouviu-falar/ > Acesso em 15 de novembro de 2018.

McKinney, Wes (2011), 'pandas: a foundational python library for data analysis and statistics', Python for High Performance and Scientific Computing pp. 1–9.

MOHAMMAD, S. M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In: MEISELMAN, H. (Ed.). Emotion Measurement. [S.l.]: Elsevier, 2016.

MOHAMMAD, S. M.; TURNEY, P. D. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: Proc. of the NAACL

HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. [S.l.: s.n.], 2010. (CAAGET '10), p. 26–34.

MORAIS, Edilson Andrade Martins; AMBRÓSIO, Ana Paula L. Mineração de Textos. Goiânia: UFG. 2007. (Série Texto Técnico, INF_005/07)

Moral, C., de Antonio, A., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. Information Research: An International Electronic Journal, 19(1), n1.,

OLIVEIRA, F. W. C. de. Análise de sentimentos de comentários em português utilizando SentiWordNet. 2013.

PETRINI, M.; POZZEBON, M.; FREITAS, M. T. Qual é o Papel da Inteligência de Negócios (BI) nos Países em Desenvolvimento? Um Panorama das Empresas Brasileiras. In: Anais do 28º ENANPAD, Curitiba – PN, 2004.

PYTHON. https://docs.python.org. <Acesso em 18 abril 2019>.

PRIMAK, Fábio Vinícius da Silva. **Decisões com B.I. (Business Intelligence).** Rio de Janeiro: Ciência Moderna Ltda., 2008.

QIU, G.; ZHANG, F.; BU, J.; CHEN, C. Domain specific opinion retrieval. In: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology. Sapporo, Japan: Springer-Verlag, 2009. p. 318–329. ISBN 978-3-642-04768-8.

RIBEIRO, L. B. Análise de sentimento em comentários sobre aplicativos para dispositivos móveis: Estudo do impacto do pré-processamento. 2015.

SAP.O que é Small Data e como obter grandes resultados com essa tecnologia?

Disponível em: <https://news.sap.com/brazil/2016/05/o-que-e-small-data-e-como-obter-grandes-resultados-com-essa-tecnologia/>.Acesso em: 08 out. 2018

Scherer KR, Meuleman B (2013) Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. PLoS ONE 8(3): e58166. https://doi.org/10.1371/journal.pone.0058166

Sebrae. Pequenos negócios em números Disponível em:<http://www.sebrae.com.br/sites/PortalSebrae/ufs/sp/sebraeaz/pequenos-negocios-em-numeros,12e8794363447510VgnVCM1000004c00210aRCRD>.Acesso em: 21 out. 2018.

Sedbrook, T., & Lightfoot, J. M. (2010). Dear: a new technique for information extraction and context-dependent text mining. Communications of the IIMA, 10(3), 3.

Setti, Alexandre.**CRIAÇÃO DE FERRAMENTA DE BUSINESS INTELLIGENCE VOLTADA PARA PROCESSOS GERENCIAIS.** 2014**.** Trabalho de conclusão de curso (especialização) - Universidade Federal do Paraná, Curitiba, 2009.

SILVA, M.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. Computational Processing of the Portuguese Language, Springer, p. 218–228, 2012.

SILVA, N. R.; LIMA, D.; BARROS, F. Sapair: Um processo de análise de sentimento no nível de característica. In: 4nd International Workshop on Web and Text Intelligence (WTI'12), Curitiba. [S.l.: s.n.], 2012.

SMITH, T. C.; FRANK, E. Statistical Genomics: Methods and Protocols. New York, NY: Springer, 2016. 353Ű378 p. Disponível em: . https://link.springer.com/protocol/10.1007%2F978-1-4939-3578-9_17. acesso 09 de setembro de 2019

STEINBERGER, J. et al. Creating sentiment dictionaries via triangulation. In: Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (WASSA '11), p. 28–36. ISBN 9781937284060. Disponível em: https://dl.acm.org/citation.cfm?id=2107653.2107657 .acesso 24 de julho de 2019

EXAME. O que faz um gestor de Marketing? Disponível em: < https://exame.abril.com.br/pme/o-que-faz-um-gestor-de-marketing/ > Acesso em 29 de julho de 2019

THIOLLENT, M. Pesquisa-Ação nas Organizações. São Paulo: Atlas, 1997.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, Springer, v. 24, n. 3, p. 478–514, 2012.

TUMITAN, D.; BECKER, K. Sentiment-based features for predicting election polls: A case study on the brazilian scenario. In: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 02. Washington, DC, USA: IEEE Computer Society, 2014. (WI-IAT '14), p. 126–133. ISBN 978-1-4799-4143-8. Disponível em: https://ieeexplore.ieee.org/document/6927616/ .

TURBAN, Efrainet
Al. **Business Intelligence: Um Enfoque Gerencial para a Inteligência do Negócio .** Porto Alegre: Bookman, 2009.
http://www.meioemensagem.com.br/home/marketing/2018/08/13/o-protagonismo-do-small-data-na-era-dos-resultados-urgentes.html.

VISAGIO. Small Data. Disponível em < http://visagio.com/pt/insights/small-data > Acesso em 15 de novembro de 2017

WARRINER, A. B.; KUPERMAN, V.; BRYSBAERT, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior research methods, v. 45, n. 4, p. 1191–207, dez. 2013.

Witten I. H.; Frank, E; Hall, M. A.; Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 4th edition. ISBN: 978-0128042915.

ANEXO A – Acordo de confidencialidade

ACORDO DE CONFIDENCIALIDADE

O presente Acordo é celebrado entre

Debora Vanessa Gobbo, portadora do CPF [redacted] residente na [redacted] **("COMPROMITENTE")**; e [redacted], inscrita no CNPJ sob o nº [redacted], com sede na [redacted] - **("EMPRESA")**, sendo **COMPROMITENTE** e **EMPRESA** doravante denominadas em conjunto como **"PARTES"** e isoladamente como **"PARTE"**;

CONSIDERANDO que o COMPROMITENTE está mantendo tratativas com a EMPRESA, de acordo com os termos firmados no presente instrumento, o COMPROMITENTE terá acesso a informações sobre a EMPRESA, em seu âmbito financeiro, operacional, relativas ao seu conceito, bem como sobre as estratégias a ela relacionadas ("INFORMAÇÕES CONFIDENCIAIS").

As PARTES, de mútuo e comum acordo, decidem celebrar o presente Acordo de Confidencialidade com o intuito de evitar a divulgação e utilização das INFORMAÇÕES CONFIDENCIAIS, nos seguintes termos e condições:

1. - Todas as INFORMAÇÕES CONFIDENCIAIS que sejam fornecidas à COMPROMITENTE e respectivos consultores, advogados ou representantes, de forma escrita (incluindo registros eletrônicos) ou verbal, serão tratadas sob a mais estrita confidencialidade.

2. - O COMPROMITENTE obriga-se a manter as INFORMAÇÕES CONFIDENCIAIS em sigilo, utilizando o mesmo nível de cuidado e discrição para evitar a divulgação, publicação ou disseminação de tais INFORMAÇÕES CONFIDENCIAIS a qualquer terceiro que aquele dispensado a suas próprias informações similares que não deseja sejam divulgadas, publicadas ou disseminadas.

3. - As INFORMAÇÕES CONFIDENCIAIS não deverão ser copiadas, reproduzidas sob nenhuma forma, ou armazenadas sob qualquer forma, pelo COMPROMITENTE.

4. - Caso o COMPROMITENTE torne-se legalmente obrigada a revelar qualquer das Informações, ela prontamente notificará a EMPRESA sobre tal obrigação. Adicionalmente, o COMPROMITENTE somente revelará a parte das INFORMAÇÕES CONFIDENCIAIS a que for legalmente requisitada, e fará seus melhores esforços para utilizar todos os procedimentos disponíveis para assegurar que as Informações assim reveladas permaneçam em sigilo.

5. - As obrigações acima não serão aplicáveis a quaisquer INFORMAÇÕES CONFIDENCIAIS que, (1) anteriormente ao seu recebimento pelo COMPROMITENTE tenham tornado-se públicas ou chegado ao poder do COMPROMITENTE por uma fonte que não a EMPRESA, ou (2) após o recebimento pelo COMPROMITENTE, tenham tornado-se públicas por qualquer meio que não como conseqüência de uma violação de sua obrigação aqui prevista, ou (3) tenham sido legalmente adquiridas pelo COMPROMITENTE sem uma obrigação de sigilo, de um terceiro que não estivesse sob obrigação de manter sigilo das Informações, ou (4) tenham sido independentemente desenvolvidas pelo COMPROMITENTE.

6. - Este Acordo vincula as Partes e seus respectivos sucessores.

7. - Este Acordo será regido e interpretado pelas leis do Brasil, pelo prazo de 12 (doze) meses, sendo o foro da Cidade de Taubaté, Estado de São Paulo, Brasil, eleito para dirimir quaisquer dúvidas ou controvérsias oriundas do presente.

EM TESTEMUNHO DO QUE, as partes assinam o presente instrumento em 2 (duas) vias de idêntico conteúdo e forma, no dia e ano abaixo apostos.

Taubaté, 12 de Abril de 2019.

Debora Vanessa Gobbo

CPF: [redacted]

[redacted]

CPF: [redacted]