FACULDADE IMPACTA DE TECNOLOGIA

Pós-Graduação Lato Sensu em Business Intelligence com Big Data

ESLI MARQUES FLORÊNCIO ARAÚJO

LEO HENRIQUE GRAVINA MARTINS

LUCAS JOSÉ DOS SANTOS

MARCOS FREITAS ALVES

MARCOS XAVIER CAVALCANTE

MICHELY LIMA CATARDO

ANÁLISE DAS RECLAMAÇÕES DE CLIENTES DURANTE O PROCESSO DE COMPRA NO E-COMMERCE DAS EMPRESAS B2W, CNOVA E MAGAZINE LUIZA.

São Paulo

2017

ANÁLISE DAS RECLAMAÇÕES DE CLIENTES DURANTE O PROCESSO DE COMPRA NO E-COMMERCE DAS EMPRESAS B2W, CNOVA E MAGAZINE LUIZA

São Paulo
2017

ESLI MARQUES FLORÊNCIO ARAÚJO
LEO HENRIQUE GRAVINA MARTINS
LUCAS JOSÉ DOS SANTOS
MARCOS FREITAS ALVES
MARCOS XAVIER CAVALCANTE
MICHELY LIMA CATARDO

ANÁLISE DAS RECLAMAÇÕES DE CLIENTES DURANTE O PROCESSO DE COMPRA NO E-COMMERCE DAS EMPRESAS B2W, CNOVA E MAGAZINE LUIZA.

Trabalho de Conclusão de Curso apresentado à Faculdade Impacta de Tecnologia como requisito para obtenção do grau de Pós-Graduação Lato Sensu em *Business Intelligence* com *Big Data.*

Orientador: Prof. Dr. Fabio Teixeira

São Paulo
2017

Araújo, Esli M.F.; Santos, Lucas J.S.: Alves, Marcos F.; Cavalcante, Marcos X.; Catardo, Michely L.; Martins, Leo H. G.

Análise das reclamações de clientes durante o processo de compra no e-commerce das empresas B2W, CNOVA e Magazine Luiza. - São Paulo, 2017 96p.

Trabalho de conclusão de curso (TCC), para obtenção de grau de Especialista em Business Intelligence. – Faculdade Impacta Tecnologia. Curso de Especialização em Business Intelligence com *Big Data*, 2017

Título em inglês: *Analysis of customer complaints during the e-commerce purchase process of B2W, CNova and Magazine Luiza.*

1. *E-commerce*, 2. *Twitter*, 3. Análise de sentimentos, 4. Classificação.

FACULDADE IMPACTA DE TECNOLOGIA

PÓS-GRADUAÇÃO EM BUSINESS INTELLIGENCE COM BIG DATA

Diretora Acadêmica: Profa. Dra. Ana Cristina dos Santos

Coordenadora do Curso de Pós-graduação: Profa. Dra. Ana Cristina dos Santos

ESLI MARQUES FLORÊNCIO ARAÚJO
LEO HENRIQUE GRAVINA MARTINS
LUCAS JOSÉ DOS SANTOS
MARCOS FREITAS ALVES
MARCOS XAVIER CAVALCANTE
MICHELY LIMA CATARDO

ANÁLISE DAS RECLAMAÇÕES DE CLIENTES DURANTE O PROCESSO DE COMPRA NO E-COMMERCE DAS EMPRESAS B2W, CNOVA E MAGAZINE LUIZA.

Trabalho de Conclusão de Curso apresentado à Faculdade Impacta de Tecnologia como requisito para obtenção do grau de Pós-Graduação Lato Sensu em *Business Intelligence* com *Big Data*.

Aprovada em setembro de 2017.

BANCA EXAMINADORA

_____

Prof.
Dr. Fábio Teixeira

_____

Prof.
Rodrigo Carvalho

_____

Prof.
Me. Fernando Sequeira Souza

Dedicatória

Com muito carinho dedicamos este trabalho à nossa família, pelo apoio e força que sempre nos deram para que conseguíssemos conquistar mais esta vitória, são eles a nossa razão para continuarmos sempre nos aprimorando e vencendo os obstáculos.

Agradecimentos

Ao professor Fábio Teixeira da Faculdade Impacta Tecnologia que nos auxiliou na elaboração deste trabalho, com sua devida orientação estamos finalizando mais uma etapa importante na nossa vida. Também agradecemos aos professores que contribuíram com seus conhecimentos e experiências durante o curso, nos tornando um profissional mais crítico, e com isso colaboraram na conclusão deste trabalho.

“Cada sonho que você deixa para trás é um
Pedaço do seu futuro que deixa de existir”.
Charles Chaplin

Resumo

Objetivo: Este trabalho tem por finalidade identificar quais são as principais reclamações postadas no *Twitter* por clientes durante o processo de aquisição de um produto ou serviço através de um sistema de *e-commerce* desenvolvido e/ou administrado pelas empresas B2W (Americanas, Submarino e Shoptime), CNOVA (Ponto Frio, Casas Bahia e Extra) e Magazine Luiza. Então realizar avaliação de qual ponto do processo é mais crítico, em cada empresa, e através desses pontos identificar a loja com maior índice de insatisfação por parte dos clientes, uma vez que estes fatores podem influenciar diretamente na opinião e na compra de outros clientes, também prejudicando a imagem da empresa. **Método:** Para isso foram desenvolvidas duas aplicações, sendo uma na linguagem R e outra na linguagem *Python* para a captura de texto de uma das redes sociais mais populares atualmente, o *Twitter.* Após a coleta, os dados foram armazenados em *cloud,* em um banco de dados *SQL Server* hospedado no serviço da *Microsoft Azure*. Em seguida separado uma parte desses textos e classificados manualmente entre positivo, neutro e negativo. Classificação essa realizada por voluntários através de um site criado especificamente com esse propósito. A próxima etapa do processo foi utilizar material classificado manualmente para o treinamento de um algoritmo classificador, em seguida, o processo automaticamente no restante da base de dados. Com os textos classificados, focamos naqueles que se enquadraram na categoria negativo, pois são neles que se encontram o foco principal deste trabalho as reclamações. Uma vez encontrado os textos relevantes para o desenvolvimento da análise principal, tratou-se de realizar uma divisão das reclamações por etapa do processo de compra através de um algoritmo de cluster. **Resultados:** Na comparação dos resultados obtidos pela classificação humana e pelo modelo treinado no Python para o mesmo conjunto de dados, ambas as classificações indicaram como classes majoritárias a negativa e a neutra concordando na classificação final. Com relação ao desempenho alcançado pelo classificador, os valores de precisão e revocação para classe negativa resultaram bem próximos, 0,69 e 0,64 respectivamente. Quando a medida f-measure considerou, para fins de desempenho do classificador, maior importância à revocação (f2-measure) o melhor resultado foi alcançado (73%). Além disso, resultados das análi-

ses de cluster de reclamação relacionados aos tweets negativos foram apresentados em forma de nuvens de palavras e dendogramas, seguidos da explicação de cada cluster. **Conclusão:** Os resultados obtidos na etapa de avaliação do classificador mostraram um bom desempenho na tarefa de classificar mensagens do Twitter sobre e-commerce comparada a classificação realizada por humanos. Entretanto, o tema escolhido retornou mensagens difíceis de serem rotuladas, até mesmo pelos humanos foram notáveis as discordâncias em algumas classificações. A contribuição desse trabalho visa detectar os problemas mais apresentados pelos clientes, os pontos do processo com maiores níveis de reclamação, permitindo ser criadas estratégias para resolver estes fatores que prejudicam a imagem da empresa e a experiência de compra do cliente. Além de mostrar como é possível realizar uma análise de dados em cima de extrações de informações, possibilitando a criação de estratégias para uma determinada empresa.

Palavras-chave: 1. *E-commerce*. 2. *Twitter*. 3. *Facebook*. 4. Análise de Sentimentos. 5. Classificação. 6. Clusters

Abstract

Objective: This work aims to identify the main complaints posted on Twitter by clients during the process of acquiring a product or service through an e-commerce system developed and / or administered by B2W companies (Submarino and Shoptime), CNOVA (Ponto Frio, Casas Bahia and Extra) and Magazine Luiza. Then carry out an assessment of which point of the process is most critical in each company, and through these points, identifying the store with the highest level of customer dissatisfaction, since these factors can directly influence the opinion and purchase of others customers, also damaging the image of the company. **Method:** Two applications were developed, one in the R language and the other in the Python language to capture Tweets from one of the most popular social networks currently Twitter. After the collection, the data is stored in the cloud, in a SQL Azure database provided by the Microsoft service. Subsequently separate a portion of these texts and manually classified between positive, neutral and negative, classification performed by volunteers through a site created specifically for this purpose. The next step in the process was to use manually sorted material training a classifier algorithm, in sequence, the process automatically in the rest of the database. With the classified texts, we focus on those who fit into the negative category, because they are the ones that are the main focus of this work the complaints. Once the relevant texts were found for the development of the main analysis, it was a question of dividing the complaints by stage of the purchase process through a cluster algorithm. **Results:** In the comparison of the results obtained by the human classification and by the train model in Python for the same data set, both classifications indicated as negative and neutral classes according to the final classification. Regarding the performance achieved by the classifier, the accuracy and recall values for negative class were very close, 0.69 and 0.64 respectively. When the measure f-measure considered, for the purpose of classifier performance, greater importance to the recall (f2-measure) the best result was achieved (73%). In addition, the results of the negative cluster analysis of complaints were presented in the form of word clouds and dendrograms, followed by the explanation of each cluster. **Conclusion:** The results obtained in the evaluation stage of the classifier showed a good performance in the task of classifying Twitter messages on e-commerce compared to the classification performed by humans.

However, the chosen theme returned messages difficult to be labeled, even by humans was remarkable the disagreements in some classifications. The contribution of this work is aimed at detecting the problems most presented by clients, the points of the process with the highest levels of complaints, allowing strategies to be created to solve these factors that harm the image of the company and the customer's buying experience. In addition to showing how it is possible to perform a data analysis over extractions of information, enabling the creation of strategies for a particular company.

Keywords: 1. E-commerce. 2. Twitter. 3. Facebook. 4. Analysis of Feelings. 5. Classification. 6. Clusters.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Símbolos

API	Application Programming Interface
BI	Business Intelligence
EDW	Enterprise Data Warehouse
HTML	Hyper Text Markup Language
PHP	Personal Home Page – Hypertext Preprocessor
SQL	Structured Query Language

Sumário

1 Introdução

Conforme dito por Martins (2016) o *e-commerce* começou a se estruturar na década de 1995 nos Estados Unidos, com o surgimento de empresas como a Amazon. Apenas cinco anos mais tarde o e-commerce ganharia espaço no Brasil. Com base no que foi descrito na história da Magazine Luiza, que desde 1992 são pioneiros no modelo de negócio eletrônico, após implantar um modelo de vendas que poderia ser considerado o embrião do e-commerce no Brasil, através de terminais eletrônicos os clientes poderiam realizar suas compras de produtos sem que existisse mostruário físico em suas lojas, e por se tratar de um período em que não se trabalhava o conceito de *e-commerce* e a falta de experiência dos clientes com esse tipo de equipamento, esses clientes eram auxiliados por funcionários da loja.

Segundo informações encontradas no artigo do Sebrae (2016), devemos nos beneficiar do crescente aumento e popularização da internet, e traçar uma estratégia de venda que englobe esse mercado, estendendo assim suas vendas além do balcão ou área de atuação de seus vendedores.

Com os mais variados perfis dos consumidores, e um crescente número de concorrentes, muitas empresas perceberam a oportunidade de expandirem seus negócios no mundo virtual, com base nesse conhecimento este trabalho procura estudar o grau de satisfação dos clientes com as principais lojas do e-*commerce*. Conforme Karina Takazono Borgato Ribeiro (2007, p.16) com o advento da Internet, vivemos em uma era onde não existem fronteiras e as empresas continuam buscando novos mercados, o comércio eletrônico motiva estas empresas a expandirem-se com o objetivo de obterem vantagens competitivas sobre as concorrentes. Para os clientes, a oferta de mais opções ajuda a atender as suas exigências, como rápido atendimento, qualidade, preço e comodidade, além de outros.

Quando o cliente tem um problema com a aquisição de um produto ou serviço, tudo que ele quer é ter o seu problema solucionado de forma simples e rápida, porém muitas vezes as empresas não estão preparadas para atender a essa expectativa, então este passa a reclamar nas redes sociais, conforme dito por (Deweik, 2016) “*hoje em dia é muito fácil e, cada vez mais comum, os clientes exporem suas insatisfações na internet. E, quando isso acontece, rapidamente centenas de pessoas recebem uma informação ruim sobre sua marca, prejudicando os negócios.*”.

Um cliente satisfeito gera lucro e um insatisfeito gera perda. Com este paradigma as empresas precisam ser mais atentas aos seus negócios e criar uma estratégia para manter seus clientes, ao mesmo tempo em que conquistam novos clientes, por isso mostraremos neste trabalho quais as grandes insatisfações dos consumidores ao realizarem uma compra pela internet e os principais problemas apresentados por esses clientes através de suas reclamações no *Twitter,* gerando perdas para as empresas, inclusive citados por Almeida Lima, Bertarelli e Pereira Alves (2014, p.03) o sucesso organizacional e a capacidade de manter seu negócio vem do lucro gerado pelos clientes satisfeitos. Uma boa organização não deve focar apenas no produto e na marca, mas sim atender as necessidades de seus clientes, tornando este seu grande diferencial competitivo. Pois na chamada "Era dos Serviços", que é uma decorrência irreversível do desdobramento ocorrido na economia mundial dá ao cliente um status de "Rei".

Com o crescimento das redes sociais, *microblogs*, dentre outros, atrelados à facilidade de acesso à *Web*, conforme matéria do site (IT FORUM 365, 2016) onde apresentou dados fornecidos pelo IBGE que apontam o aumento crescente do uso de smartphones no acesso à internet superando o uso de computadores tornou muito mais fácil à interação entre as pessoas no mundo virtual, fazendo com que essas estejam constantemente apresentando críticas, opiniões e conhecimentos e assim produzindo cada vez mais conteúdo, contribuindo para a *Web* se tornar um grande repositório de dados. O mundo *online* se tornou onipresente na vida das pessoas atualmente e consequentemente dos serviços oferecidos por ela.

As redes sociais tornaram-se quase uma extensão de seus usuários, a todo o momento, em qualquer lugar, milhares de pessoas postam algo que reflete o seu sentimento, seja ele positivo ou negativo, de uma boa experiência ou quando há uma experiência mal sucedida.

De forma igual quando uma pessoa tem interesse em algum tipo de produto ou serviço, é normal que procure em *web sites*, *blogs*, *sites* especializados e redes sociais por informações sobre o produto/serviço e passa a analisar as opiniões de pessoas que já tiveram alguma experiência com o mesmo no qual está interessado, e se baseando nas opiniões encontradas, toma a sua própria decisão de adquirir ou não o produto que estava avaliando ou ainda se é viável comprar em uma determinada loja ou em outra.

Esse tipo de pesquisa pode se tornar um tanto quanto perigosa e/ou tenden-

ciosa se não realizada de forma cautelosa, pois um fato que ocorre comumente é a aquisição bem-sucedida e o consumidor não realizar nenhum tipo manifestação positiva, mas quando essa mesma aquisição sofre algum distúrbio em seu processo normal, as pessoas tendem com maior facilidade a realizar postagens negativas sobre o produto adquirido ou sobre a loja em que realizou a compra. Conforme descrito por Alex em seu artigo ao The New York Times (Wright, 2009) a alta no mercado de opiniões pessoais tais como resenhas, classificações, recomendações e outras formas de expressão online originaram-se da ascensão dos blogs e redes sociais. Para os cientistas de dados, essa montanha crescente de informação oferece um vislumbre fascinante da consciência coletiva dos usuários da internet.

1.1 Problema

A Figura 1 mostra um diagrama que relaciona os problemas enfrentados pelos clientes durante uma compra online, e para cada etapa podem ocorrer diversos problemas que causam a insatisfação do cliente, quando este não é atendido parte para expor suas reclamações na internet.

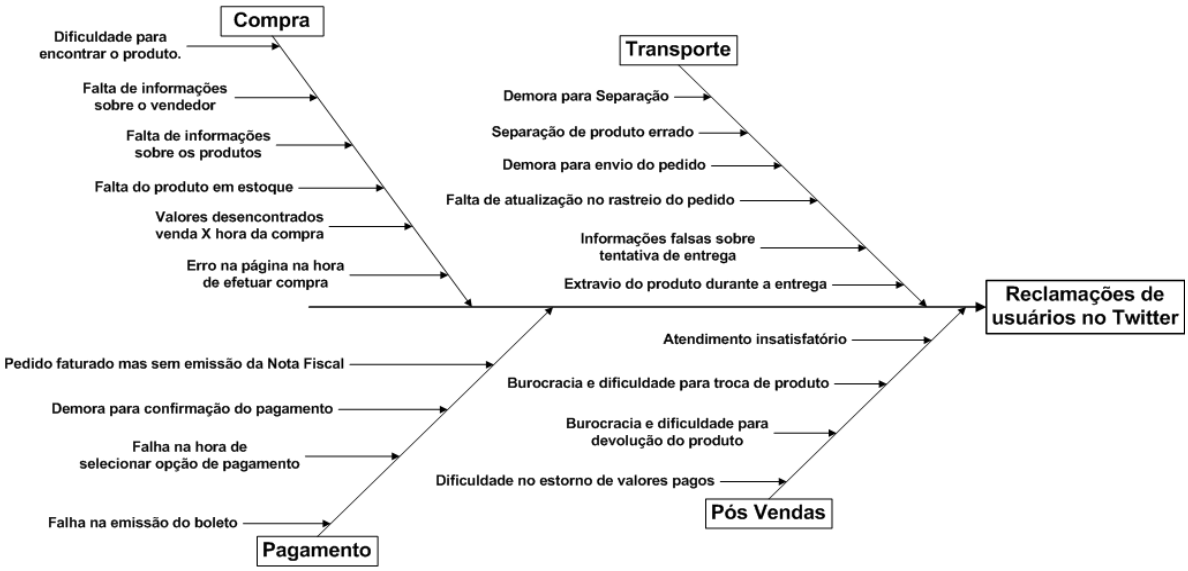

Figura 1 - Diagrama de Ishikawa

1.2 Organização do Documento

O presente documento está organizado nos seguintes capítulos:

- Capítulo 1: o capítulo corrente, contendo a introdução do trabalho;
- Capítulo 2: objetivos gerais e específicos do trabalho;
- Capítulo 3: referencial teórico que expõe trabalhos similares a presente pesquisa;
- Capítulo 4: softwares e ferramentas utilizados para desenvolver o trabalho e atingir os objetivos;
- Capítulo 5: metodologia utilizada para desenvolvimento do projeto;
- Capítulo 6: resultado;
- Capítulo 7: discursões acerca dos resultados;
- Capítulo 8: conclusões do trabalho realizado;
- Capítulo 9: Bibliografia com as referências de citações e outros materiais utilizados para elaboração desse trabalho;

2 Hipótese

Este trabalho possui a seguinte questão norteadora: Quais as principais queixas postadas no *twitter* pelos clientes do *E-Commerce* das empresas CNova, B2w e Magazine Luiza? Procurando responder à pergunta problema foram definidos os seguintes objetivos conforme abaixo.

2.1 Objetivo Geral

Analisar os tipos de reclamações que são mais frequentes dentre os clientes que utilizam *E-commerce* das empresas CNova, B2W e Magazine Luiza no momento ou após a compra de um determinado produto, analisando o que as pessoas postaram no *Twitter* mencionando as lojas avaliadas. Para que assim possam ser criados estratégias de melhoria para estes fatores que prejudicam a imagem da empresa e a experiência de compra do cliente.

2.2 Objetivo Específico

- Detectar os problemas mais apresentados pelos clientes através do *twitter*;
- Apresentar os pontos do processo com maiores níveis de reclamação;

3 Referencial Teórico

Esse espaço apresenta a base teórica necessária para a representação deste estudo e o entendimento do mesmo. Apresenta uma revisão de trabalhos pertinente aos temas relacionados ao estudo proposto. Nas próximas páginas serão apresentados os conceitos de *Business Intelligence,* mineração de dados, algoritmos de classificação, análise de sentimentos, ferramentas de processamento da linguagem natural e avaliação de resultados.

3.1 Business Intelligence

Apoiando a Inteligência Estratégica das organizações, o *Business Intelligence* surge como uma tecnologia que permite transformar dados em informações significativas. Dados são fontes de grande riqueza, podemos extrair informações para realizar estudos dos diferentes comportamentos a partir de diferentes cenários. O termo *Business Intelligence* (Inteligência de Negócios), retrata ao processo de coleta, organização, análise, compartilhamento de informações, suporte a gestão de negócios, que é o conjunto de teorias, processos, metodologias, estruturas e tecnologias mudando uma grande quantidade de dados brutos em informações úteis para tomadas de decisões estratégicas.

Visando o acesso rápido as informações, os recursos de *Business Intelligence* proporcionam a disseminação de conhecimento fazendo com que os usuários estejam alinhados na estratégia da organização.

O *Business Intelligence* define um conjunto de regras e técnicas objetivando organizar adequadamente um grande volume de dados, visando transformá-los em depósitos estruturados de informações, conforme citado por Trindade de Lima e Rouberte de Freitas (2014, p.6) o papel do *Business Intelligence* é coletar dados de diferentes fontes, transformá-los em informação e disponibiliza-la para seus respectivos usuários. Está tarefa é realizada através de vários processos que compõem o *ETL (Extract, Transform, Load),* onde é construído um modelo para armazenamento e entrega destas informações através de um hardware capaz de melhorar diversas consultas às bases analíticas.

3.2 Análise de sentimentos e/ou opiniões

O tipo de pesquisa que falamos até agora se trata de um usuário que deseja realizar a aquisição de um produto, análise essa que geralmente é feita de forma manual, pesquisando, lendo e analisando caso a caso em *sites* específicos de reclamações, *sites* especializados ou ainda em redes sociais. Dessa forma quando uma pessoa lê o *post* realizado por outra pessoa consegue identificar quase de imediato qual o sentimento que ela estava sentindo no momento em que escreveu, mesmo que de forma sarcástica, mas e se uma loja desejar realizar essa mesma avaliação para todos os seus produtos e serviços? Esse seria um processo que demandaria extremo esforço e tempo. Para auxiliar nessa questão surgiu um processo denominado análise de sentimentos e/ou opiniões. Um processo de análise de sentimento é composto pelas etapas: escolha de uma fonte de dados de um determinado assunto de interesse para análise, em seguida a classificação, seleção de palavras chaves, análise sintática e semântica e então a sumarização dos resultados, conforme Ribeiro (2015), de maneira geral, a análise de sentimentos, tem por objetivo analisar o sentimento expresso pelo autor de um texto, quando a intenção é identificar o sentimento em uma classe específica, como por exemplo positivo ou negativo, é tida como uma tarefa de classificação. Para tal são envolvidas áreas de estudo como ciência da computação, linguística, estatística e até psicologia.

3.2.1 Fatores Complexos

Há diversas complexidades na análise de sentimento em suas fases, serão colocados alguns itens frequentes:

- O uso de gírias, como a análise é feita em cima de textos livres, buscando encontrar os padrões com base nos padrões linguísticos. Com a utilização mais constante da *internet* para comunicação informal, acabou-se criando o hábito de escrever da mesma forma que se fala, ou ainda para poupar palavras, utiliza-se das abreviações, além do uso assíduo de gírias, fatores facilmente encontrados em redes sociais como *Twitter* que limitam a quantidade de caracteres dos *posts*, neste caso fonte de coleta dos textos para análise. Com o uso dessas abreviações e gírias a dificuldade de identificar padrões

aumenta, pois não seguem e nem obedecem nenhuma regra ou padrão linguístico.

- Utilização de sátiras e ironias nos comentários, o uso dessas técnicas ou figuração de linguagem tornam as análises extremamente complexas, sendo que uma vez utilizadas palavras positivas em um comentário irônico, acaba tendo seu real contexto invertido.

3.2.2 Etapas da Análise de Sentimentos

Devido à complexidade para treinamento de um sistema de classificação de sentimento, se faz necessário dividi-lo em etapas para aperfeiçoamento das informações, treino efetivo e verificação dos resultados conforme dito por Oliveira (2013).

Sendo assim as etapas são:

- **Seleção dos dados:** nesta etapa é definida a fonte de dados que será utilizada para realização da análise, neste trabalho a fonte selecionada foi o *Twitter*. Em um primeiro levantamento foi utilizado alguns posts dos *facebook* através da coleta de comentários nas páginas das empresas citadas, porém não gerou uma massa de dados que justifica-se a utilização dessa origem de dados.
- **Extração dos Dados:** após definido a origem dos dados, vem a extração, nesta etapa são utilizados ou desenvolvidos processos e aplicativos para realização da captura dos textos e armazenagem em um repositório para posteriormente serem pré-processados.
- **Organização dos Dados:** esta parte é conhecida como pré-processamento, pois os dados passam por um tratamento para tentar melhorar a qualidade do texto realizando o tratamento de gírias, abreviaturas, idiomas, remoção de *stopwords*, remoção de acentuações e números, dentre outras técnicas utilizadas em *text mining*, as técnicas de pré-processamento serão explicadas em detalhes no capítulo 5.8 deste trabalho.
- **Categoria dos sentimentos:** é a fase mais importante, pois nela que as técnicas de análise serão aplicadas, agrupando as informações e definindo-as como positivo, neutro ou negativo. Esta classificação pode ser realizada com algoritmos de máquina ou técnicas com recursos baseados em léxicos conforme representado na Figura 2.

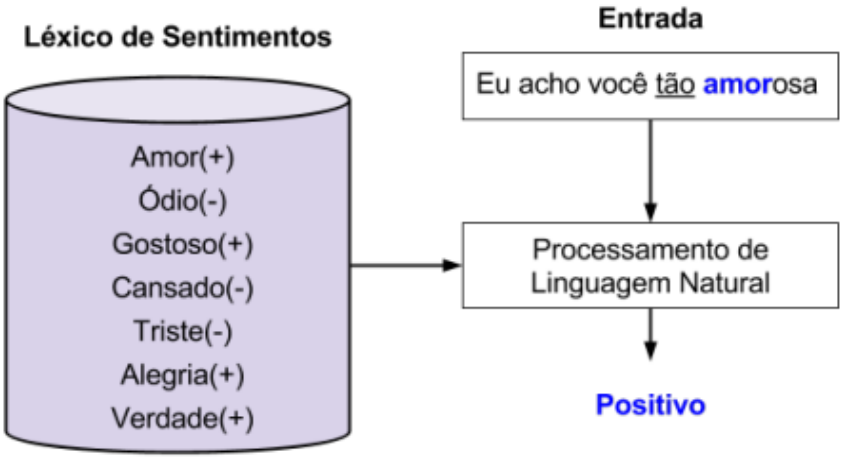

Figura 2 - Processo de análise de sentimentos

- **Sumarização dos Resultados:** tão importante quanto gerar informações, é a disseminação de forma que qualquer pessoa consiga compreender essas informações, para tal feito são gerados relatórios, análises descritivas e gráficas, que facilitam o entendimento dessas informações.

3.3 Data Mining

Conforme dito por Lima (2006) O *Data Mining* é utilizado em processos com enormes repositórios de dados como *Data Warehouse* procurando achar padrões e alguma relação de utilização que não são conhecidas pela empresa e que podem ser usados na tomada de decisão. Existem dificuldades com o processo de transformar dados que as empresas concentram em suas transações diárias em informações que serão úteis no desempenho dos negócios. Com o *Data Mining* este problema é resolvido, pois o processo encontra informações importantes, como os padrões, as associações entre as informações da base de dados, anomalias, mudanças, em uma quantidade muito grande que são colocadas em um repositório de dados.

Raramente encontramos uma empresa que detenha apenas um sistema de gestão, sendo essas empresas de grande porte, ou até mesmo as empresas de médio e/ou pequeno porte. Todas essas empresas acabam acumulando muitos e muitos dados e com o passar do tempo esses dados armazenados acabam sendo deixados de lado ou não sendo mais necessários, pois as empresas não aproveitar todo esse conteúdo armazenado em seus bancos de dados. Com uso do *Data Mining*

é possível encontrar nesses dados até então inúteis, nova utilidade, encontrar padrões, novo insight, transformando tudo isso em novas informações, podendo causar alterações nos processos da empresa, fazendo com que os dados sejam vistos como um todo, e após transformado em informação podendo ser disponibilizada de forma sólida e segura com os conceitos de *Business Intelligence*.

3.3.1 Técnicas de Data Mining

Com o objetivo de extrair o máximo de conhecimento de um determinado grupo de dados, foram criadas diversas técnicas de *Data Mining*. Serão citadas algumas técnicas utilizadas como:

- **Redes Neurais -** As redes neurais artificiais são procedimentos inteligentes parecidos com o funcionamento de um cérebro de um ser humano, isto mostra que uma rede neural é o conjunto interligado com os neurônios. Cada neurônio é interligado para o envio de informações, de acordo com pesos e conexões predefinidas. Tem a capacidade de tratar dados incompletos e distorcidos, gerando resultados de generalizações. Conforme citado por Lima (2016, p.34) uma rede neural artificial, através de um treinamento inicial adequado, tem a capacidade de aprender sozinha. Para isso, para cada etapa do treinamento é feita uma comparação nos dados com um resultado conhecido, caso não atenda a expectativas, uma correção é calculada e processada automaticamente nos nós da rede. Este procedimento é reproduzido até que a condição de parada seja alcançada.
- **Indução de Regras -** É usado para localizar tendências nos grupos de dados.
- **Métodos Estatísticos -** É a forma mais antiga utilizada, sendo a parte fundamental de todas as tecnologias. É utilizada a participação de um usuário, que criará modelos que escreverão o comportamento de um dado através de uma aplicação matemática.
- **Árvore de Decisão -** A mais conhecida das técnicas, divide os dados em subgrupos com base nos valores das variáveis que são utilizados para classificar os dados.

3.3.2 Funções do Data Mining

Segundo dito por Lima (2006) Cada aplicação de *Data Mining* tem como base um conjunto de algoritmos que serão utilizados no processo de extração de informações importante estes algoritmos podem ser definidos como:

- **Classificação –** É a pré-classificação de um conjunto de classes.
- **Estimativa -** É a variação do problema de classificação, onde são gerados valores no decorrer das dimensões dos dados, trabalham com resultados contínuos.
- **Agrupamento por afinidade –** Tem a funcionalidade de encontrar tendências no meio de uma grande quantidade de dados**.**
- **Previsão –** Os registros são classificados através de uma previsão futura.
- **Segmentação -** É obtida automaticamente por algoritmos que localizam características semelhantes e dividem o espaço em dimensões que são definidas por atributos.

3.4 Web Mining

Nos últimos anos houve uma gigantesca evolução da grande rede mundial, consequência essa da evolução exponencial de *hardwares* e *softwares* no mundo tecnológico, o que permitiu que mais pessoas nos mais variados cantos do mundo pudessem acessar a internet através de seus computadores pessoais, *smartphones*, *tablets*, *smartwatches* entre outros dispositivos, e através deles acessando seus sites prediletos, perfis em redes sociais e outros serviços online, e consequente ajudando a produzir os milhões de dados que são gerados todos os dias.

Com esse grande volume de dados que são gerados diariamente, surgiu então a necessidade de encontrar técnicas que conseguisse extrair informações de forma mais customizável, para poder atender a necessidades específicas de alguns usuários e/ou organizações, dando abertura para a criação de uma nova área no campo de descoberta de dados, conhecida como *Web Mining.* Conforme Scoz (2014, p.01) *Web Mining* surge da aplicação de técnicas de mineração de dados na Internet, as aplicações dessas técnicas são motivadas pela grande quantidade e variedade de dados disponíveis. Quando uma pessoa busca informações na web,

normalmente utilizam algum motor de busca, uma das maiores dificuldades é encontrar dados relevantes dentre todos os resultados retornados de modo a conseguir extrair conhecimento útil. Para tal, são necessárias técnicas que aumentem a eficiência das respostas visto que os usuários possuem diferentes padrões de preferências e atividades realizadas online.

Aplicando corretamente técnicas de *Web Mining* sobre os dados providos de forma voluntária ou involuntária, mas muitas vezes sem que o usuário tenha essa percepção, é possível identificar perfis de possíveis clientes, identificando quais tipos de produtos, serviços, locais e, até mesmo, horários que são mais propícios para que esses clientes efetuem uma compra de um determinado produto, e pensando dessa forma muitas empresas acabam investindo de formas substanciais seus recursos para conseguirem essa vantagem que o *Web Mining* ajuda a prover, conforme Srivastava (2010) com o advento do *Web Mining* as empresas podem ir ao encontro dos clientes, oferecendo o que eles realmente precisam, pois baseiam-se na localização, preferência e perfil de seu público-alvo, obtidos a partir do comportamento online dos usuários. Além destes dados, o *Web Mining* pode operar sobre diferentes outras fontes como banco de dados organizacionais, servidores web e servidores *Proxy*.

3.5 Data Mart

Um *Data Warehouse* une banco de dados de toda uma empresa, ou seja, trabalha com um grande volume de dados, já o *Data Mart* é menor e foca em um assunto específico ou um determinado departamento. O *Data Mart* pode ser dependente ou independente, pois no caso de dependente é um subconjunto criado diretamente a partir de um *Data Warehouse*, dando a vantagem de usar um modelo de dados consistente e apresentar dados de qualidade. Os *Data Marts* dependentes suportam apenas um conceito de um único modelo de dados de toda a empresa, mas um *Data Warehouse* tem que ser construído antes. Um *Data Mart* independente é um *Data Warehouse* reduzido, projetado para uma unidade estratégica de negócios ou um departamento, mas cuja a fonte de dados não é um *Enterprise Data Warehouse (EDW)* conforme Turban (2009).

4 Softwares e Ferramentas

Nessa sessão de nosso trabalho, realizaremos uma descrição das ferramentas utilizadas em nosso trabalho, também falando como estas foram aplicadas em nosso desenvolvimento.

4.1 Coleta e Processamento de dados

Para realizar a coleta e processamento dos dados faremos fazer uso das seguintes plataformas:

- **Python**: Uma linguagem desenvolvida pelo holandês Guido Van Rossum, entre o final da década de 1980 e início da década de 1990 em quanto trabalhava em um projeto de um sistema operacional baseado em *microkernel* chamado de Amoeba, com o objetivo de suprir uma lacuna até então existente entre as linguagens C e *Shell Script*.

> Percebi que o desenvolvimento de utilitários para administração de sistema em C (do Amoeba) estava tomando muito tempo. Além disso, fazê-los em *shell Bourne* não funcionaria por diversas razões. O motivo mais importante foi que, sendo um sistema distribuído de microkernel com um design novo e radical, as operações primitivas do Amoeba diferiam muito (além de serem mais refinadas) das operações primitivas disponíveis no *shell Bourne*. Portanto, havia necessidade de uma linguagem que "preencheria o vazio entre C e o *shell*". Por um tempo longo, esse foi o principal objetivo do *Python*.Rossum (2014)

Em nosso trabalho utilizaremos o *Python* junto a uma *API* chamada de *Tweepy*, essa *API* será responsável por realizar a conexão com o *Twitter*, e através dessa conexão acompanhado dos parâmetros apropriados, os dados serão extraídos e passados através de aplicação desenvolvida no *Python* que armazenará esses dados em um Banco de Dados para futura utilização, esta linguagem também será utilizada no treinamento algoritmo classificador *Naive Bayes*.

- **R:** Essa é uma linguagem de programação que teve sua escrita baseada na

antiga linguagem **S** na década de 1990, inicialmente foi concebida por Robert Gentleman e Ross Ihaka, que consequentemente acabaram utilizando suas iniciais no nome da linguagem, e posteriormente recebendo contribuições de desenvolvedores e empresas de todo o mundo. Um pouco diferente do *Python* que possui características de aplicações mais amplas, o R é muito utilizado por pessoas que desejam desenvolver projetos estatísticos.

No desenvolvimento de nosso trabalho, utilizamos a linguagem R acompanhado do pacote o *TwitteR*, similar a *API* do *Python*, que nos permite realizar a conexão de nossa aplicação com a plataforma do *Twitter,* nos permitindo realizar a extração dos dados desejados e sequencialmente armazená-los em um banco de dados.

4.2 SQL Server

O *SQL Server* trata-se de um sistema de Banco de Dados Relacionado, em nosso trabalho foi construído para trabalhar em nuvem, pensando em aplicações em *cloud* ou que necessitam de fácil acesso com garantia de disponibilidade, segurança e capacidade de adequação rápida a variações de demanda.

Desenvolvido e distribuído pela empresa *Microsoft*, o SQL Server hospedado *Azure* traz os mesmos recursos encontrados no *SQL Server* em um servidor físico agregado aos benefícios encontrados nas plataformas *cloud*, dentre elas as listadas a seguir:

- **Autogerenciamento:** oferece a proporção e a funcionalidade de um *Data Center* corporativo, sem a sobrecarga administrativa.
- **Escalabilidade:** proporciona a manipulação do serviço à medida que seus dados crescem, ou até mesmo a demanda de requisições e cargas em horários de picos.

4.3 MySQL

Segundo o Portal Educação (2015) o MySQL teve sua primeira versão lançada em 1995, um ano após Michael Windenius desenvolvedor do banco de dados UNIREG convencer David Hughes criador do banco de dados mSQL a unir esforços e os pontos fortes de seus sistemas, dando origem assim a um novo banco de dados que pudesse ser utilizado na criação de páginas web dinâmicas. Porém os usuários que já trabalhassem com os sistemas antigos necessitariam realizar apenas algumas pequenas modificações em seus códigos para a utilização desse novo servidor. Em 1995 um dos parceiros da empresa de Michaels sugeriu a distribuição desse servidor na Web, o que deu origem ao sucesso do MySQL.

4.4 Pentaho

O *Pentaho* é uma plataforma de código aberto desenvolvido em *java* pela empresa de mesmo nome que pertence ao grupo Hitachi para trabalhos de inteligência de negócio.

Com o *Pentaho* você possui uma gama de ferramentas para trabalhos de *B.I*. Através de seus componentes há o *Data Integration* para realização de *ETL*, *Analysis Services*, Cubos *OLAP*, *Reporting* para criação de relatórios, Data Mining (*Weka*) para realizar a mineração de dados, *CDE (Community Dashboard Editor)* para criação de gráficos e visões em *DashBoard* e por fim integração com *Hadoop* para processamento de grandes massas de dados (*Big Data*).

4.5 HTML

O HTML é a sigla em Inglês para *Hyper Text Markup Language* (Linguagem de Marcação de Hiper Texto), é uma linguagem utilizada na construção de páginas web, e que consegue ser interpretada por qualquer navegador web, sua versão mais recente é conhecida como HTML5.

Essa linguagem foi criada pelo físico britânico Tim Berners Lee no final da década de 1980 e início de 1990, com o propósito de facilitar o uso da web, e assim possibilitando compartilhar suas pesquisas com mais pessoas e com maior facilidade.

4.6 PHP

PHP é um acrônimo para Hypertext Preprocessor (Processador de Hipertexto), trata-se de uma linguagem muito utilizada no desenvolvimento web, que comumente é utilizada em conjunto com a linguagem HTML.

Mas diferente do HTML que é executado e interpretado pelo navegador, o PHP seja executado no servidor, onde este é processado gerando as informações de acordo com o que foi programado e envia para o cliente apenas um código HTML com as informações já processadas.

5 Metodologia

Neste capítulo é descrita a metodologia utilizada para desenvolver o trabalho, detalhando onde cada ferramenta foi utilizada no processo.

Em nosso trabalho faremos uso de *softwares* livres, primeiramente por não deterem custos com licenciamento, e pelo fato de plataformas livres possuírem uma maior gama de informação tanto em seu uso quanto em questões de suporte, *plugins* e *apis*. Também faremos uso de algumas plataformas proprietárias, mas que disponibilizam uso gratuito com algumas restrições ou livres por um período de avaliação.

5.1 Fluxo de Desenvolvimento do Trabalho

Para o desenvolvimento da nossa pesquisa utilizamos um processo que foi divido em duas fases, sendo a primeira fase (Figura 3) subdividida em cinco etapas, nessa fase estão as etapas seguidas no processo de desenvolvimento.

Inicialmente na primeira etapa da fase 01 realizamos a construção de duas aplicações para coletar e armazenar os *tweets* através das seguintes linguagens R e *Python*, ainda nessa etapa além de coleta dos dados as aplicações realizavam a armazenagem destes textos no banco de dados SQL *Server*, que serviu como centralizador dos dados para esse trabalho. No decorrer do desenvolvimento optamos por utilizar somente a aplicação em *Python* por sua versatilidade.

Em uma segunda etapa da fase 01 utilizamos uma amostra de aproximadamente vinte mil *tweets*, realizando uma cópia destes para um banco *MySQL* no qual estava sendo acessado por um site que desenvolvemos, esse site foi utilizado para que voluntários pudessem classificar manualmente os *tweets* coletados em positivo, negativo ou neutro, dentro do escopo do trabalho. Estas classificações foram utilizadas para treinar o algoritmo classificador.

Na terceira etapa da fase 01 utilizamos o *Pentaho Data Integration* para criar um processo de ETL que consolidou as classificações manuais dos *tweets*, uma vez que cada *tweet* poderia ser classificado até três vezes, sendo cada classificação realizada apenas uma única vez pelos voluntários que acessaram o site. Nessa consolidação foi utilizada a seguinte lógica, se um texto foi classificado pelo menos duas

vezes como positivo, este foi consolidado como positivo, caso tenha sido classificado ao menos duas vezes como negativo, então este seria consolidado como negativo, os textos que foram marcados ao menos duas vezes como Neutro ou que tenha havido um empate nas classificações este foi consolidado como Neutro, pois consideramos que sua identificação é complexa até mesmo para humanos que realizaram a sua classificação.

Na etapa quatro da fase 01 realizamos a divisão da amostra em dois grupos, sendo um terço para teste e dois terços para treinamento do algoritmo, em seguida realizamos a remoção de *stopwords*, processo de *tokenização*, *stemmen*, extração de termos únicos e geração do vetor de características, e inserimos as palavras do *Anew-Br, a lista de palavras do Anew-Br é explicada no capítulo 5.5 deste trabalho.* Após esse processo realizamos o treinamento e testes com o algoritmo *Naive Bayes*.

A quinta e última etapa da primeira fase consistiu em utilizar o modelo criado para classificar todos os *tweets* que estavam na base de dados do SQL Server.

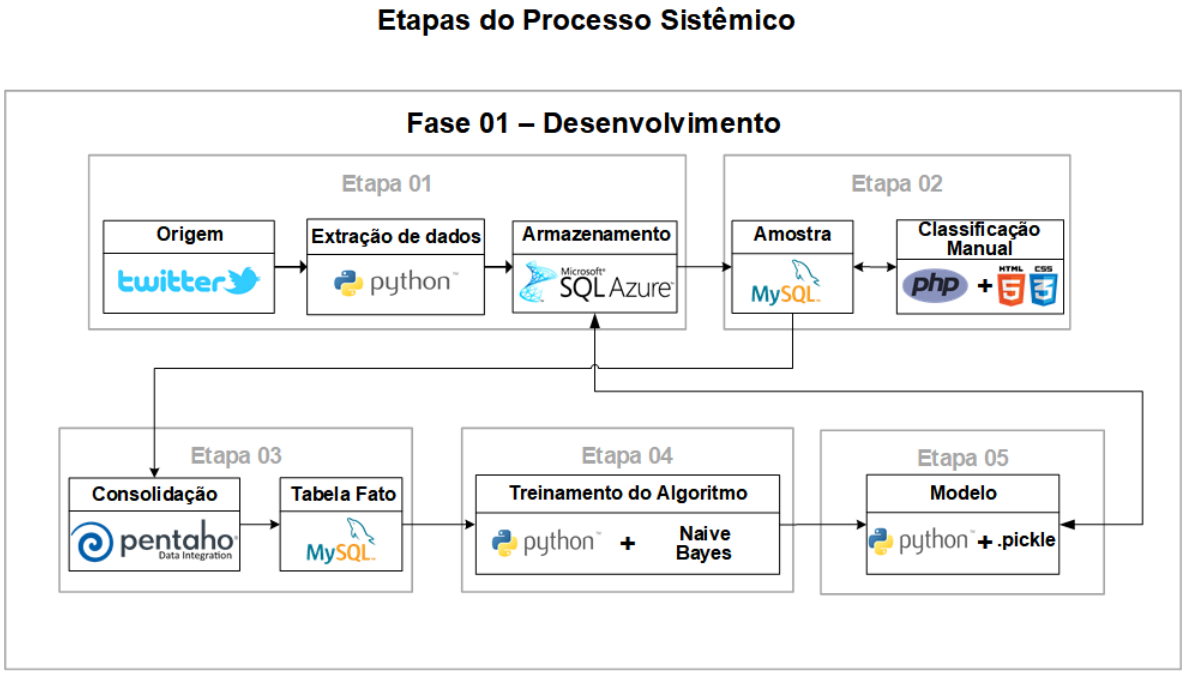

Figura 3 - Etapas do processo Sistêmico – Fase 01 - Desenvolvimento

Na Figura 4, temos as etapas que aplicadas após a conclusão da criação e treinamento do modelo, aplicando a classificação sobre os textos logo após a extração dos mesmos.

Na etapa 01 da Fase 02, os dados são extraídos do *Twitter* utilizando uma aplicação desenvolvida em *Python* e utilizando a API *Tweepy*. Esses *tweets* são encaminhados então para a segunda etapa da fase 02, onde os textos passam pelo Modelo que os classificam e os gravam no banco de dados do *MySQL* já classificados. Foi necessário a troca do banco de dados utilizados por questões de melhor compatibilidade entre o *Pentaho*.

Na terceira etapa da Fase 02 é realizada a clusterização, utilizando somente os textos que foram classificados como sendo negativos, e os gravam em uma nova tabela.

A quarta etapa da fase 02 utiliza o Pentaho junto ao pacote Saiku e CDF (Community Dashboard Framework) para acessar os dados processados e gerar as visões com esses dados.

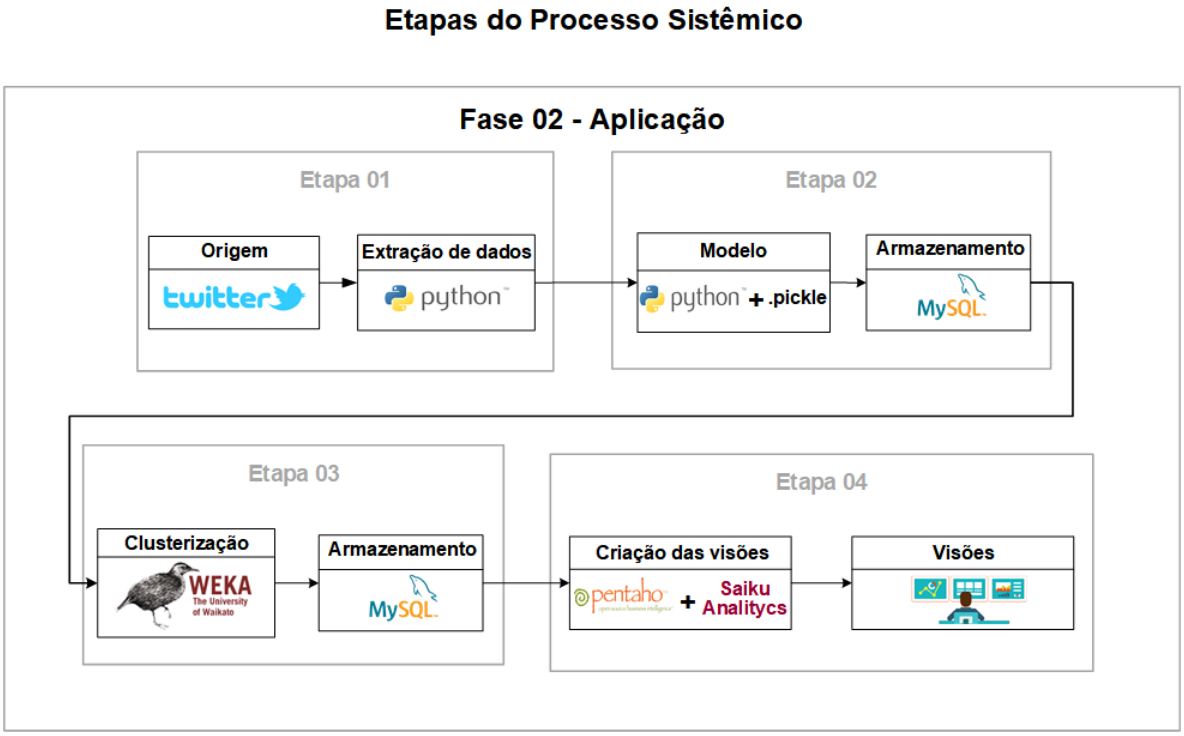

Figura 4 - Etapas do processo Sistêmico – Fase 02 - Aplicação

Durante o processo de desenvolvimento foi necessário realizar algumas escolhas que seriam determinantes para o sucesso deste trabalho, essas escolhas determinaram desde a origem dos dados que seriam analisados, até as ferramentas e tecnologias a serem utilizadas para realizar os processos de coleta, classificação, clusterização e análise dos dados.

Para auxiliar a tomada de decisão realizamos em alguns casos testes com as múltiplas alternativas que detínhamos sobre um determinado assunto, na Figura 5 é possível visualizar algumas das opções que foram testadas e que não obtiveram êxito pelos seguintes motivos:

- Coleta de Dados > Twitter > R + TwitterR: Optou-se por não utilizar essa tecnologia por questões de conhecimento técnico sobre a ferramenta e limitações na utilização da API
- Coleta de Dados > Facebook: inicialmente cogitou-se em utilizar postagens realizadas no *Twitter* e no *Facebook*. Porém durante a fase inicial do projeto, verificou-se que a massa de dados coletadas no *Twitter* durante duas semanas superava consideravelmente a quantidade de textos coletados no *Facebook* durante três meses. Desta forma não justifica os esforços empregados para coleta e tratamento dos dados provenientes do *Facebook*.
- Classificação automatizada > API da Microsoft + R: não houve continuidade na utilização dessa API pois o objetivo do trabalho consistia em aplicar as técnicas de text mining, data mining e machine learning de forma que houvese um poder maior de controle sobre o processo, assim optou-se por utilizar técnicas aplicadas com o Python.
- Classificação automatizada > API Repustate + Python: inicialmente foram realizados testes com essa API proprietária de uma empresa de mesmo nome, porém essa não conseguiu um bom desempenho, porque foi desenvolvida e reconhece somente palavras em inglês e português de Portugal, além dos fatores citados também para a não utilização da API da Microsoft.

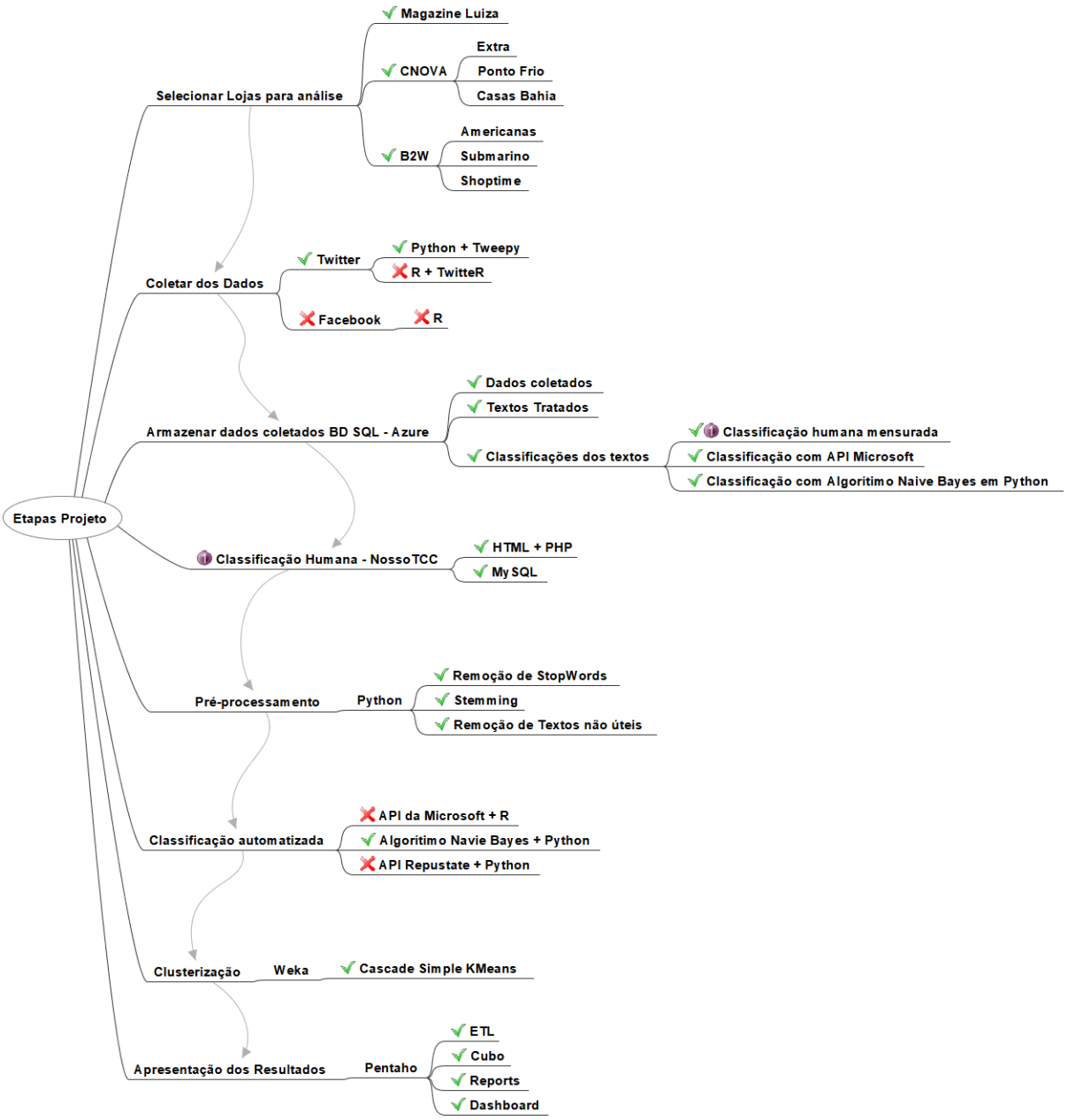

Figura 5 - Mapa mental da etapa de Desenvolvimento

5.2 Seleção das lojas avaliadas

Para seleção das lojas avaliadas, foi realizado um debate entre os membros do grupo, a princípio foram selecionadas apenas as lojas Americanas, Submarino e Extra, como era de conhecimento dos integrantes, esses sites de *e-commerce* são pertencentes a dois grupos distintos, sendo Americanas, Submarino e Shoptime pertencentes ao grupo B2W (https://www.b2wmarketplace.com.br) e as lojas Extra, Ponto Frio e CasasBahia são lojas pertencentes ao grupo CNOVA (http://marketplace.br.cnova.com/cnova/).

Por se tratarem de lojas conceituadas e muito conhecidas, optamos por trabalhar com essas seis lojas, sendo três lojas pertencentes a cada grupo. Além destas selecionamos também a loja Magazine Luiza que pertence à rede de lojas físicas de mesmo nome que foi a primeira a trabalhar com um modelo de negócio que poderia ser considerado o embrião do *e-commerce* no Brasil.

5.3 Etapas de compra em um E-Commerce

Como o foco deste trabalho será analisar em qual etapa do processo de compra em um e-commerce os clientes apresentam um maior volume de reclamações, a Figura 6 auxilia no entendimento de cada etapa do processo:

- Etapa 01: consiste basicamente da pesquisa sobre o produto, procura pela oferta do mesmo e seleção da melhor loja para aquisição, e iniciar o processo de compra do produto desejado.
- Etapa 02: após a compra efetuada, inicia a etapa de pagamento, que pode variar de acordo com a loja e opção do cliente, em seguida se faz necessário a confirmação desse pagamento, seja pelo banco para pagamentos de boleto, operadora de cartão para compras no cartão de crédito ou outra forma que tenha sido selecionada durante a compra. Para finalizar essa etapa realiza-se o faturamento e emissão da nota fiscal do produto.
- Etapa 03: realizado a emissão da nota fiscal, inicia-se o processo de separação e embalagem dos produtos que serão enviados para o correio ou transportadora, que realizarão a entrega do produto ao cliente final.

- Pós-Venda: A etapa de pós-venda geralmente é acionada quando existe uma anomalia no processo habitual da compra. Geralmente quando existe uma demora para confirmação de um pagamento, dúvida sobre um produto, demora para entrega. Nesses casos o cliente entra em contato com a central de atendimento (via telefone, chat, e-mail) para verificação do problema ocorrido e verificar a solução que será tomada para resolução do mesmo.

  Ainda nessa etapa o cliente pode desejar realizar a troca de um produto que veio com problema, entregue diferente do que havia sido comprado, cancelamento de uma compra, devolução de um produto e solicitação de estorno do valor pago ou outras situações que saem do processo principal.

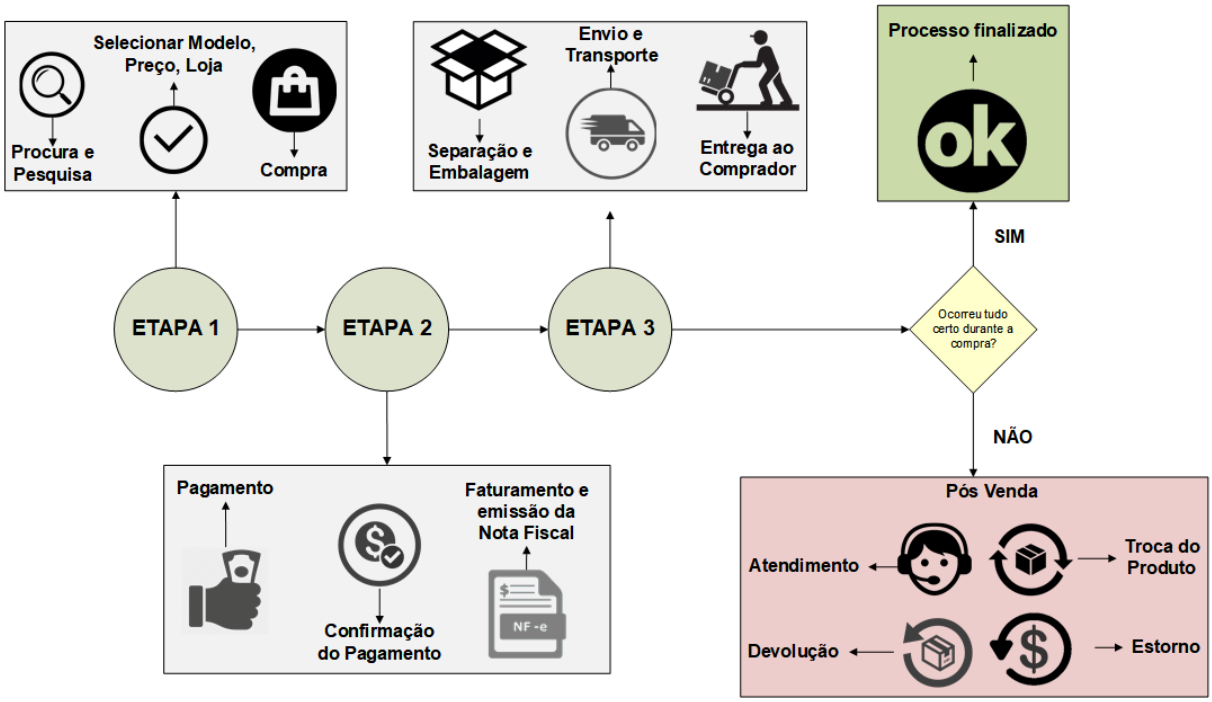

Figura 6 - Etapas da compra em E-Commerce

5.4 Origem dos dados

5.4.1 Twitter

Optou-se por utilizar como origem dos dados o *Twitter,* por ser uma das redes sociais mais utilizadas na atualidade pelas empresas para comunicar-se com seus clientes, além de disponibiliza bibliotecas e *APIs* (https://dev.twitter.com/docs), que possibilitam a conexão através de sistemas terceiros, permitindo uma interação com sua plataforma, inclusive a extração e coleta dos textos postados por seus usuários.

5.4.2 Facebook

Inicialmente também realizamos coleta de textos providos das páginas oficiais das lojas no *Facebook*, porém a quantidade de dados extraídos não justificou os esforços empregados. Enquanto coletaram-se do *twitter* no período de um mês 27 mil textos provenientes do *Twitter*, no *Facebook* foi possível a coleta de apenas 1000 postagens em um período de seis meses.

Além da pequena massa de dados, os textos provenientes dessa rede demandam um maior esforço para tratamento, uma vez que não há limites para o tamanho dos textos escritos pelos usuários, e em sua maior parte são postagem não relacionadas ao foco deste trabalho.

A primeira etapa considerou a construção das aplicações em *Python* e R, que foram criadas com o objetivo de coletar os *Tweets* em tempo real do *Twitter*, armazenando-os dentro de uma tabela no banco de dados *SQL Server*.

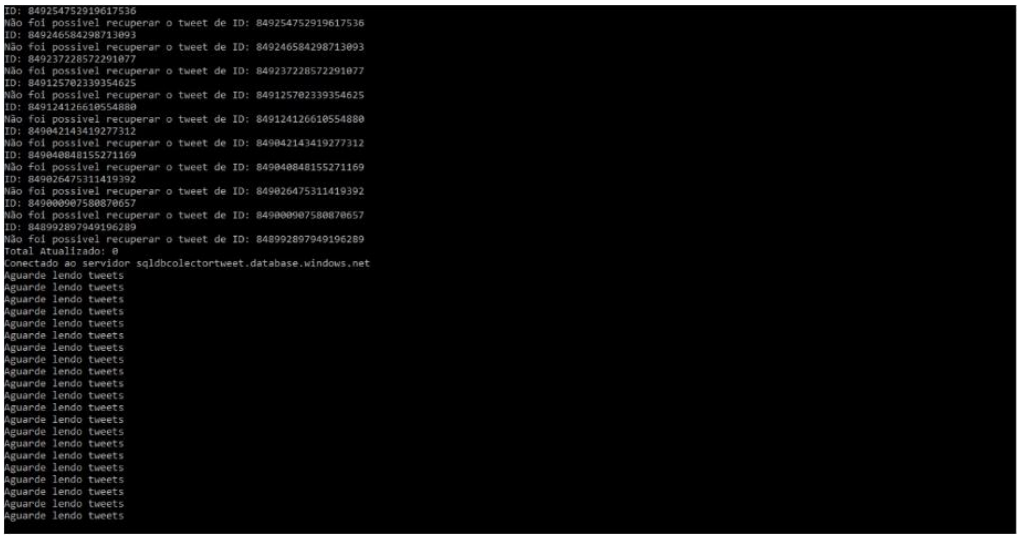

Figura 7 - Mostra a execução da conexão do Servidor do banco de dados para a coleta das informações dos *tweets* com *Python*.

Na Figura 7 podemos visualizar a execução da aplicação extraindo os dados e realizando a inserção no banco de dados competente a etapa 01 da primeira fase do trabalho.

Não é necessário interagir com as aplicações, bastando apenas digitar o comando para executá-las e a partir deste ponto todos os dados coletados serão exibidos na tela à medida que são coletados e armazenados, tudo em tempo real, realizando várias inserções dentro do nosso banco de dados. A exibição no console é apenas um log predefinido durante o desenvolvimento do script de coleta. Neste processo de análise não consideramos os famosos *emotions*, devido a difícil identificação no meio do texto dos *tweets ou posts*. As ilustrações (Figura 8) que serão apresentadas têm uma amostragem de 10 registros que foram armazenados no banco de dados.

	tweet_id	tweet_json	tweet_hashtag	tweet_source
▶	836895919287201793	Se existisse um hanking de pior atendimento e atenção ao cliente, certamente o primeir	[]	Twitter
	836896114922176512	Nunca vi tanto descaso com o cliente e ineficiência no atendimento! Seis tão de parabén	[]	Twitter
	836901418049110018	a @americanascom ligou agr pouco pro meu pai dizendo q "não tem" as rodinhas (skate	[]	Twitter
	836902010330951680	@americanascom ele fez o pagamento no boleto pra não ter nenhum problema e agr o s	[]	Twitter
	836904033180921860	@americanascom comprei na pré venda um produto que está pronto pra entrega há dia	[]	Twitter
	836930836100415488	@americanascom @lojasamericanas tenho um pedido com pagamento aprovado desde	[]	Twitter
	836931499089870848	@americanascom Quando terá ovos de páscoa no site?	[]	Twitter
	836932434033786882	Não está passando nenhum cartão na @americanascom. Sem internet ?. Sinceramente,	[]	Twitter
	836936849901912064	@americanascom americanas sos! Ganhei um notebook do meu pai mas nao fui com a c	[]	Twitter
	836956140986646528	Pô, @americanascom, e essa zuera ai? De 19,90 para 1,99+17,26 de frete? ¬¬ https://t	[]	Twitter
*				

Figura 8 - Amostra dos dados dos Twitter que foram coletados.

Podemos notar através da amostra, que existem muitos tipos de reclamações dos usuários sobre as lojas virtuais de *E-commerce*. Nos campos das figuras chamados *tweet_json*, apresentam os comentários e opiniões dos clientes das lojas utilizadas na nossa pesquisa.

Com as informações que obtemos, criamos uma nuvem de palavras (Figura 9) da qual percebemos que as mais citas são pedidos, dia e produto pelo tamanho da fonte da letra que se encontra em destaque.

Figura 9 - Mostra a nuvem de palavras feita com base nos dados coletados do *Twitter*.

5.5 ANEW-BR

O ANEW-BR é um conjunto de palavras criado por pesquisadores do Center for the Study of Emotion and Attencion do National Institute of Mental Health da University of Florida, utilizados para na realização de estudos de estímulos emocionais. Originalmente esse conjunto foi desenvolvido para aplicações em estudos nos Estados Unidos, possuindo um total de 1.034 palavras, com sua ampla utilização e disseminação para outros países, e não encontrando uma versão em português do Brasil para esse conjunto, um grupo de pesquisadores do Rio Grande Do Sul decidiu realizar a tradução e adaptação dessas palavras para o português do Brasil, criando um conjunto de 1.046 palavras.

Cada palavra desse conjunto tem sua avaliação realizada em três dimensões, na primeira dimensão é realizada a medição de quanto à palavra é agradável ou desagradável que é chamada de valência. Na segunda dimensão é realizada a medição de alerta que verifica o quanto a pessoa fica relaxada ou estimulada. E na terceira dimensão é avaliado qual o controle ou dominação é percebida pelo indivíduo chamada de Dominância.

Durante o processo de tradução e adaptação, gerou-se uma medida chamada média que consiste no cálculo da média e desvio padrão dos julgamentos de valência e alerta, uma vez que esses por si só, já podem ser utilizadas para identificar o sentimento atrelado àquela palavra conforme descrito por Kristensen, Gomes, Justo, & Vieira (2011).

Em nosso trabalho realizamos o treinamento do modelo com base nas classificações realizadas de forma manual por voluntários e associamos as palavras e média que compõem o ANEW-Br, porém notamos uma queda no desempenho e assertividade do modelo após a inclusão desse conjunto de palavras. A lista de palavras e os valores de sentimento média encontram-se no apêndice A deste trabalho.

5.6 Classificações Manuais de Tweets

5.6.1 Site para classificação de textos por voluntários.

Para realização do treinamento do algoritmo *Naive Bayes* conforme descrito no capítulo 5.9 deste trabalho, necessitamos ter previamente uma base com textos já classificados, para assim realizar um treinamento supervisionado do algoritmo, e somente depois de treinado e testado, é que aplicamos o modelo aos demais textos que estão tanto em base de dados quanto nos novos *tweets* que forem coletados.

Para conseguir gerar uma base de dados com textos já classificados e com uma massa de tamanho significativo, foi desenvolvido um site (www.nossotcc.com.br) onde pessoas voluntárias nos auxiliaram com essa tarefa.

O site foi construído com o princípio de ser prático e simples, partindo dessa premissa o site foi construído com basicamente apenas duas páginas, sendo a primeira tela a identificação do voluntário conforme pode ser visto na

Figura 10, onde o mesmo irá digitar seu nome, sobrenome e endereço de e-mail.

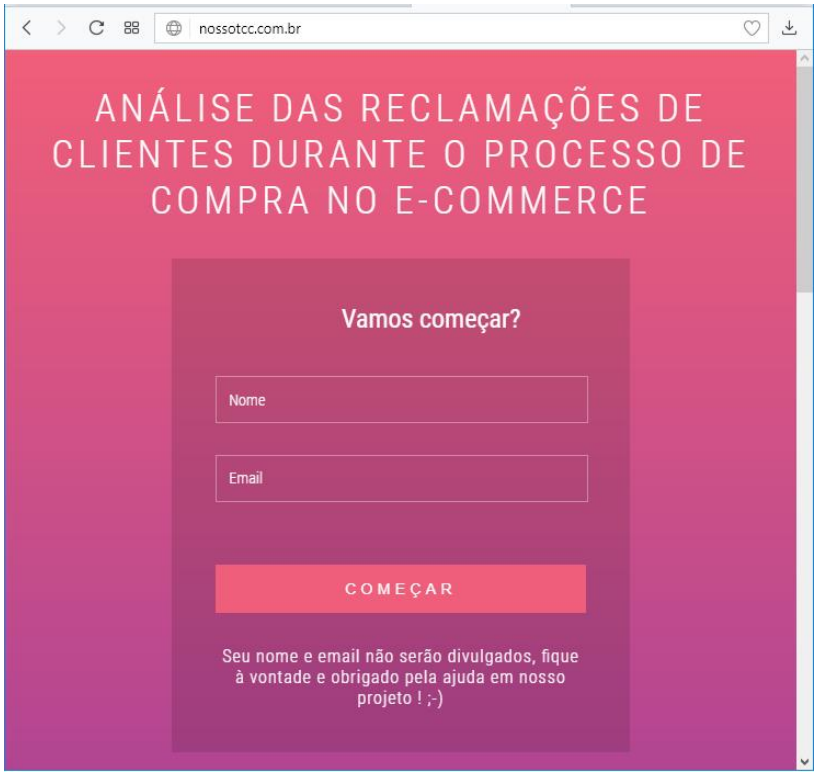

Figura 10 - Tela de Login NossoTCC

Na mesma tela foi disposto um texto com os tópicos **quem Somos, objetivo e como funciona,** onde é explicado quem são os envolvidos no projeto, e por que desenvolvemos o site, e como utilizá-lo Figura 11.

Figura 11 - Topicos NossoTCC

A segunda (Figura 12) tela funciona da seguinte forma, um texto coletado do *twitter* é exibido para que o voluntário possa ler e decidir se o mesmo é um texto positivo, neutro ou negativo. Abaixo do texto existem três botões no qual o usuário clicará na opção correspondente a sua classificação de acordo com o seu entendimento do texto.

Para evitar que a disposição dos botões influenciasse na opinião ou classificação do usuário, os botões correspondentes a negativo e positivo foram alterados a cada quatro semanas.

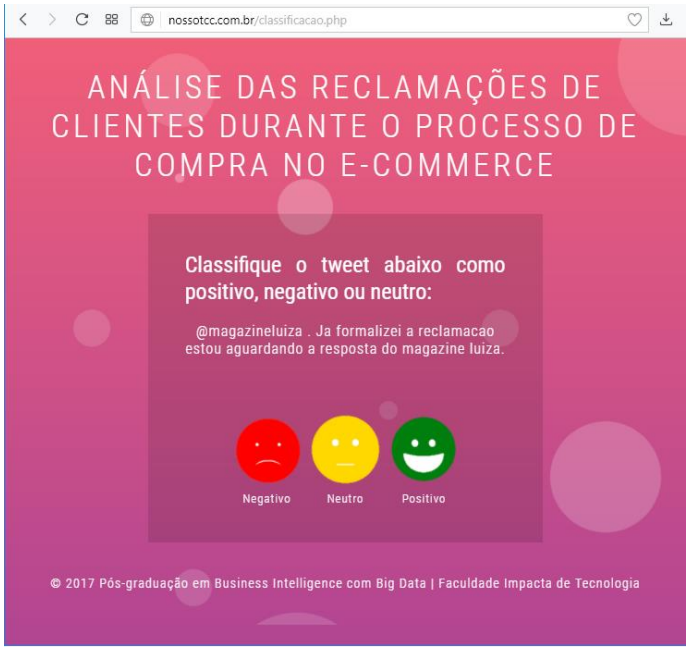

Figura 12 - Classificação de Twitters Nosso TCC

5.6.2 Desenvolvimento do Site.

O desenvolvimento do site se deu através de tecnologias como HTML5 e CSS3 para construção de sua interface gráfica, e um *layout* responsivo que facilita a utilização do site em dispositivos móveis assim permitindo que os voluntários realizassem a classificação em momentos em que estivessem mais ociosos, como por exemplo, o transporte coletivo no deslocamento de suas residências para o local de trabalho ou caminho inverso.

A linguagem PHP foi utilizada para realizar a conexão com o banco de dados, questões de segurança como verificar se o usuário realizou o preenchimento da tela de login, coleta e armazenamento dos dados como respostas dos voluntários e data hora da classificação, exibir textos armazenados no banco para que o voluntário pode-se realizar a classificação, utilizando a seguinte lógica, permitir a classificação de cada texto até três vezes, sendo que cada voluntário somente poderia realizar a classificação de cada texto apenas uma única vez.

Para o armazenamento dos dados utilizou-se um banco de dados MySQL, onde inicialmente foram inseridos uma amostra de aproximadamente vinte e sete mil textos.

5.6.3 Tweets Classificados Por Voluntários.

O site contava com uma base de dados de aproximadamente vinte e sete mil *tweets* sendo que cada *tweet* poderia ser classificado até três vezes, porém não mais do que uma única vez por voluntário e cada voluntário poderia classificar quantos textos deseja-se. No período de dia entre as datas 22/06/2017 até 20/08/2017 conseguimos obter a classificação de pouco mais de onze mil textos, nos quais aproximadamente oito mil e quatrocentos foram classificados de forma unânime que foram classificados por três voluntários na mesma categoria de sentimento e pouco mais de dois mil e setecentos foram classificados de formas distintas pelos voluntários. O que nos leva a hipótese de que são textos complexos até mesmo para compreensão humana e por este motivo foi realizada a ponderação das classificações com o intuito de não afetar negativamente o treinamento do modelo.

5.7 Entendimento dos Dados

Nesse capítulo serão realizadas algumas explanações referentes aos dados que conseguimos capturar para darmos início aos trabalhos de limpeza, agrupamento e classificação.

Atualmente nossa base de dados conta com 49.381 (quarenta e nove mil trezentos e oitenta e um) registros capturados do *twitter*, sendo essa captura realizada desde o dia primeiro de março de 2017 até setembro do mesmo ano. Também fizemos a captura de aproximadamente três mil e trezentos posts do *Facebook* iniciando na mesma data, porém buscando posts mais antigos, onde conseguimos obter postagens a partir de primeiro de outubro de 2016, dentre estes foram classificados voluntariamente através do site cerca de onze mil em positivo, negativo ou neutro.

5.7.1 Amostragem dos dados

Uma amostragem dos dados coletados, e classificados por humanos voluntariamente, podem ser visualizados Figura 13.

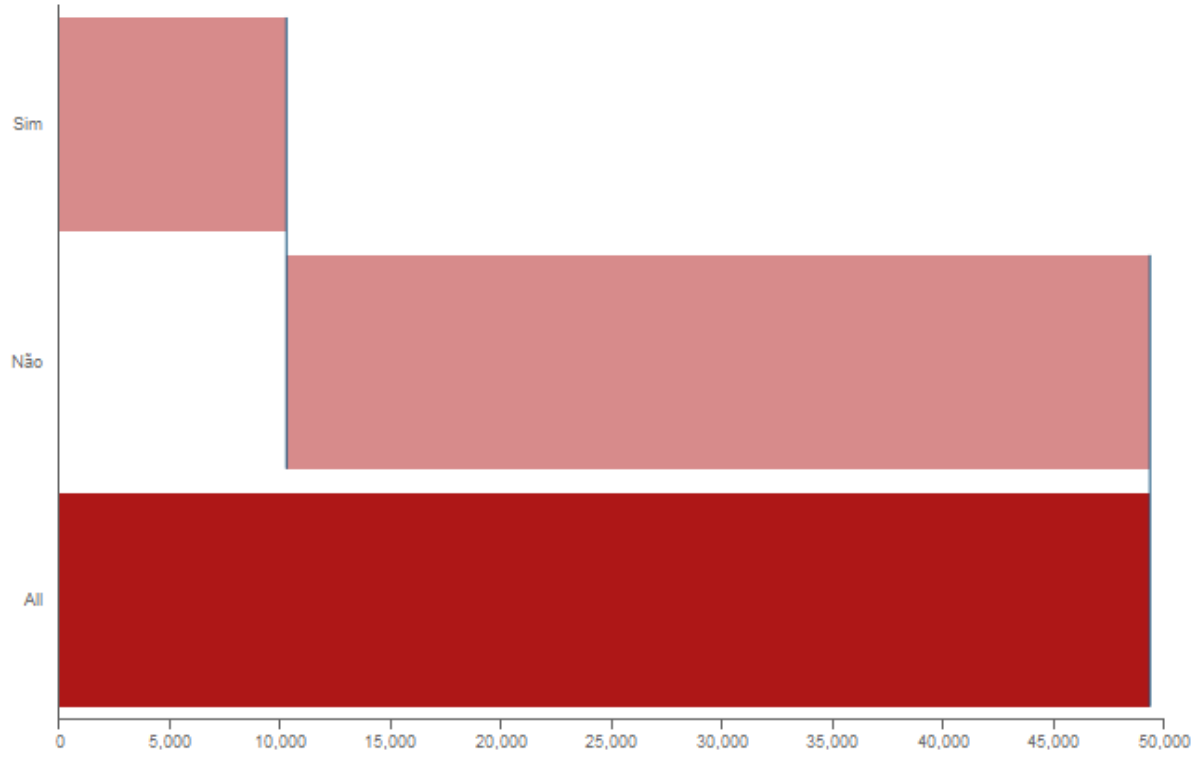

Figura 13 - Visão Geral dos Tweets

Durante o processo de classificação manual através do site (Figura 14), foram coletadas as classificações dos voluntários, estas classificações serão utilizadas no treinamento do algoritmo classificador.

Figura 14 - Entendimento dos dados Análise de sentimentos

5.7.2 Sistemas origem dos dados

Para iniciarmos nossos estudos, procuramos entender também quais são os meios utilizados pelos usuários para realizar as suas postagens, onde se mostra claro na Figura 15 que a maioria dos usuários utiliza seus aparelhos *smartphone*. Realizando o cruzamento dessa informação com o gráfico visto na Figura 16, conseguimos definir que os usuários acabam expressando suas opiniões quando tem mais tempo ocioso, sendo estes na hora em que estão indo ou voltando do trabalho, escola ou compromissos em horário comercial.

Esta informação pode se tornar um ponto de atenção, pois acontecimentos imprevistos durante o trajeto podem interferir nos sentimentos usuário que por sua vez transmite de forma errônea ou focalizando no alvo errado o seu descontentamento ou felicidade.

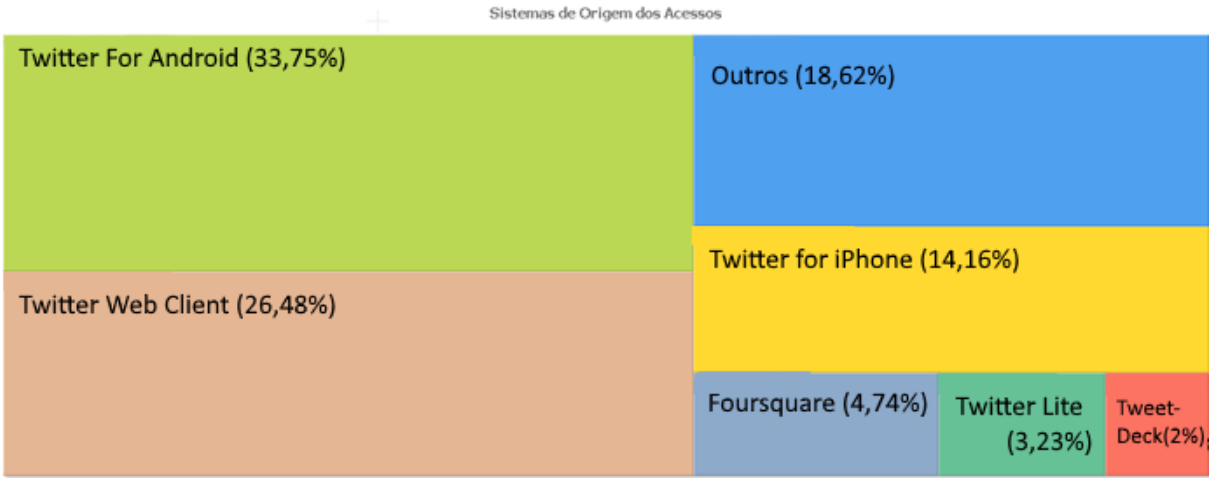

Figura 15 - Sistemas origem dos dados

Visualizamos no gráfico exibido na Figura 16 que os horários com maior quantidade postagem ocorrem nos chamados horários de pico, onde a maior parte da população está em deslocamento entre suas residências e seu local de trabalho, ou durante o caminho inverso.

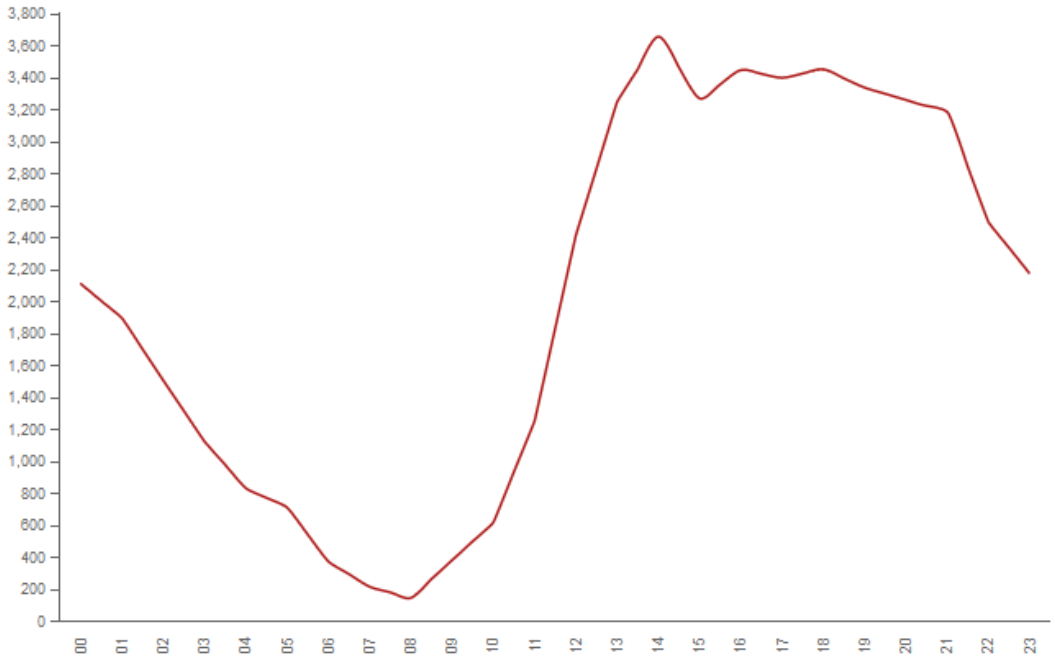

Figura 16 - Horários de postagem

5.8 Pré-Processamento dos dados

O processo de classificação e indexação dos tweets foi composto, além da construção da base de dados a ser avaliada, pela transformação dos documentos textuais em vetores numéricos, capazes de representá-los de maneira unívoca (Zhang, Yoshida, Tang, 2011).

O modelo de espaço vetorial é um dos métodos amplamente utilizados pela comunidade científica para tal representação (Salton, Wong, Yang, 1975). Nestes modelos, um documento é representado como um vetor, formado por um conjunto de palavras representados pela expressão $d_j = (w_{1,k}, \dots, w_{j,k})$, na qual k é o tamanho do conjunto de palavras únicas da base de dados e $w_{j,k}$ é a importância da palavra em relação ao documento.

Este estudo utilizou 9679 (Nove mil seiscentos e setenta e nove) termos para compor a dimensão dos vetores numéricos que compunham os tweets. A origem dos termos se deu a partir das palavras únicas presentes nos textos dos tweets coletados.

Para alcançar o número total de termos utilizados foram executados processamentos preliminares, como:

1. Remoção de palavras presentes em uma lista de *stopwords* (Baeza-Yates RA, Ribeiro-Neto B, 1999).
2. Aplicação de *stemming* (Porter M, 1980) para cada palavra.

O processo de remoção de *stopwords* conta com a identificação, nos textos da base de dados, de pronomes, conjunções, preposições e artigos que são irrelevantes para a tarefa de classificação ou indexação. Este trabalho utilizou a lista de *stopwords* disponível na biblioteca NLTK do Python, utilizada para a mineração dos textos. O *stemming* das palavras refere-se à redução das mesmas a sua raiz morfológica, por meio da eliminação de prefixos e sufixos, para tal foi utilizada a biblioteca *RSLPStemmer* também presente no NLTK.

Para a construção do vetor de características foi adotada a técnica de ocorrência binária (*BO*), pois devido ao tamanho dos textos a serem processados, não haverá muita repetição de palavras para ponderação, inviabilizando a utilização de técnicas como *TF* (*Term Frequence), TF-IDF (Term frequence – inverse document frequence)* e *TO (Term Ocorrence).*

A descrição de cada técnica, bem como suas restrições, é mostrada no Tabela 1:

Técnica	Descrição
tf	Dado um documento, a técnica calcula a divisão entre o n° de ocorrências de um determinado termo e a quantidade de termos existentes no mesmo documento.
bo	Dado o conjunto de termos de um documento, a técnica indica a presença ou não de um determinado termo no mesmo.
to	Esta técnica utiliza a quantidade de vezes que um termo ocorreu em documento para compor o vetor de características do mesmo. Não ocorre a divisão pela quantidade de termos do documento avaliado, como na técnica tf.
tf.idf	Esta técnica explora a relação entre a quantidade de vezes que um termo ocorre em um documento e a ocorrência do mesmo em todos os documentos avaliados.

Tabela 1 - Descrição das técnicas não-supervisionadas de extração de características

5.8.1 Ponderação da classificação humana dos tweets

Foram disponibilizados 27.775 (vinte e sete mil setecentos e setenta e cinco) tweets em uma página web que fossem classificados por indivíduos humanos. Estes tweets foram carregados aleatoriamente após cada classificação, onde o mesmo não poderia ser classificado mais de uma vez pela mesma pessoa, a posição dos ícones de resposta também foi alterada com passar do tempo durante a disponibilização da página para a classificação, onde nas 2 primeiras semanas ficaram disponíveis na ordem positivo, neutro, negativo. Nas seguintes negativo, neutro, positivo. Este modelo foi proposto para evitar uma classificação tendenciosa do usuário.

Foram definidas algumas regras para as classificações onde número máximo de classificações por tweet fosse de 3 classificações. Os pesos das classificações foram definidos tendo o valor -1 para as negativas, 0 para neutras e 1 para os sentimentos positivos. Desta forma ao realizar a coleta, a classe do tweet será definida pela expressão $C = \sum c$ onde C corresponde à classificação humana resultante da soma das classificações c, onde se o resultado da soma das classificações for um

número negativo, a classificação para este tweet será negativo, se for zero o mesmo será considerado neutro e se positiva será considerado um tweet positivo. Esta medida foi tomada visto que um mesmo tweet pode ser interpretado de diferentes formas por pessoas diferentes.

5.9 Treinamento do Algoritmo Naive Bayes

Para realizar o treinamento do modelo, adotamos o algoritmo classificador Naive Bayes. A escolha foi baseada na simplicidade do método e em sua eficiência na tarefa de classificação supervisionada de textos comprovada ao logo dos anos por estudos científicos (Sohn S, Kim W, Comeau DC, Wilbur WJ., 2008).

Para o treinamento do algoritmo foram utilizadas as informações coletadas no site onde foi utilizado $^{2}/_{3}$ para treinamento e $^{1}/_{3}$ dos dados para avaliação e teste do modelo.

Após criada as matrizes de características de ocorrência binária utilizando os exemplos coletados, as mesmas foram utilizadas como valores de entrada para o classificador, a fim de realizar a tarefa de validação da classificação dos tweets que compunham a base de dados de validação deste projeto.

O classificador Naive Bayes assume que os termos que fazem a composição da base de dados são únicos, por isso possui o adjetivo *naive*, ingênuo em português, que é utilizado para indicar que a premissa de independência ocorre, ao avaliarmos, por exemplo, a semântica de um texto (Teixeira, 2011). Desta forma é possível dizer que a ordem que as palavras não interferem no resultado da classificação. Assim as frases “A festa foi legal”, “Legal a festa foi” e “Foi legal a festa” são analisadas da mesma forma pois o vetor de características gerados serão os mesmos e as palavras serão tratadas de forma independente.

O modelo bayesiano utilizado neste trabalho é fundamentado na teoria das probabilidades regida pela Equação 1:

$$P(C = C_k|X = x) = P(C = C_k) \times \frac{P(X = x|C = C_k)}{P(x)}$$

onde,

$$P(X = x) = \sum_{k=1}^{e_C} P(X = x|C = C_k) \times P(C = C_k)$$

Equação 1 – **Teorema da probabilidade de Bayes** (a)

O denominador $P(x)$ representa o somatório da probabilidade de todos os eventos possíveis. Neste estudo, significa documentos pertencerem a uma determinada classe $(C_1, C_2, \dots, C_n)$. O parâmetro X reúne o conjunto de características de um documento,$x = (x_1, \dots, x_j, \dots, x_d)$.

O numerador $P(X = x|C = C_k)$ é obtido considerando a premissa da independência das características dos documentos, na qual os elementos contidos no vetor x_j são estatisticamente independentes. A Equação 2 apresenta este cálculo.

$$P(X = x|C = C_k) = \prod_{1}^{d} P(x_j|C_k)$$

Equação 2 – **Teorema de probabilidade de Bayes (b)**

Portanto, $P(X = x|C = C_k)$ é a probabilidade condicional de um determinado documento em pertencer a uma classe, uma vez que o vetor de características x é conhecido.

Os valores presentes nos conjuntos de características dos documentos, variável X, foram obtidos pela técnica de extração ocorrência binária e submetidos ao classificador de padrões probabilístico definido neste estudo.

5.10 Avaliação de desempenho do modelo

Para avaliar o desempenho do resultado obtido pelo modelo, medidas de convencionais de desempenho foram aplicadas, são elas precisão, revocação e *f-score*, desta forma foi possível avaliar as variações nos resultados de acordo com as alterações nos parâmetros de entrada.

Precisão consiste de uma medida de fidelidade, enquanto revocação trata-se de uma medida de completude. Ambas são medidas padrões da Recuperação de Informação (RI) segundo Cleverdon (1966 apud SILVA, 2006) estas medidas são amplamente utilizadas para avaliar sistemas de recuperação de informação a partir da entrada de um usuário, mas também são bastante utilizadas em sistemas de aprendizado de máquina e processamento de linguagem natural para avaliação de desempenho.

A seguir são explicadas as medidas do ponto de vista do processamento de linguagem natural.

5.10.1 Precisão e Revocação

Uma matriz de confusão oferece uma medida concreta do modelo de classificação, pois esta demonstra o número de classificações corretas e as classificações preditas para cada classe num conjunto de exemplos.

Na tabela 2 é demonstrada a matriz de confusão utilizada para avaliar as 3 classes utilizadas para classificar os tweets.

Resultado / Modelo / Classif. Manual	Negativo	Neutro	Positivo
Negativo	VP (Verdadeiro Positivo)	FN (Falso Negativo)	FN (Falso Negativo)
Neutro	FP (Falso Positivo)	VN (Verdadeiro Negativo)	FN (Falso Negativo)
Positivo	FP (Falso Positivo)	FN (Falso Negativo)	VN (Verdadeiro Negativo)

Tabela 2 - Matriz de confusão do modelo gerado

Baseado no resultado extraído da matriz de confusão é possível calcular as medidas de precisão e revocação. A primeira é calculada com base no percentual de acertos do modelo, utilizando os valores verdadeiros positivos (VP) em relação ao total de falsos positivos (FP) que estão acima da diagonal e é definida pela Equação 3:

$$\boldsymbol{P} = \frac{\boldsymbol{VP}}{(\boldsymbol{FP} + \boldsymbol{VP})}$$

Equação 3 – Cálculo da precisão

A revocação é calculada por meio de todos os exemplos classificados corretamente VP presente no numerador da Equação 4 em relação a todos os exemplos que deveriam ter sido associados a uma determinada classe pelo mecanismo avali-

ado, representado pelo denominador da mesma equação, que calcula a soma entre VP e FN (verdadeiros positivos e falsos negativos) onde FN são os valores que o modelo classificou incorretamente.

$$R = \frac{VP}{VP + FN}$$

Equação 4 – Cálculo de Revocação

O f-score, por sua vez, é uma medida harmônica entre precisão e revocação, regida pela Equação 5, na qual P e R representam os valores de precisão e revocação, respectivamente, e β é um parâmetro de ponderação da revocação em relação a precisão, determinando a importância da mesma para o sistema de recuperação de informação avaliado. Neste estudo, os experimentos utilizaram três variações para o parâmetro β, sendo 0.5, 2 e 1, que determinaram maior importância a precisão, revocação e pesos iguais, respectivamente, aos parâmetros utilizados na fórmula.

$$F - score_{\beta} = \frac{(1 + \beta^2) * (P + R)}{\beta^2 * P + R}$$

Equação 5 – **Medida f-score**

5.11 Clusterização com Cascade Simple K-Means

Após a classificação dos *tweets* em suas respectivas classes utilizando o algoritmo *Naive Bayes*, foi realizada uma nova classificação para que fosse possível identificar a etapa do processo em que as reclamações estão focadas. Para tal a ferramenta *Weka* foi utilizada.

O primeiro passo utilizado para realizar este procedimento foi filtrar os tweets classificados como negativo pelo classificador, pois este é o foco do trabalho. Uma vez importados esta coleção de documentos no *Weka*, foi utilizado o filtro *StringToWordVector* que possibilita executar uma série de pré-processamentos, tais como *tokenização, stemming* e *TF* (estes procedimentos foram descritos no capítulo 5.8 deste trabalho), este filtro foi utilizado para criar o vetor de características a ser utilizado para aplicação do algoritmo de cluster, ainda no filtro foi determinado que só comporiam o vetor de características as palavras que estivessem contidas 100 vezes ou mais em todo o corpus analisado, desta forma os ruídos, ou seja, palavras menos importantes não atrapalharão na clusterização.

5.11.1 Cascade Simple K-Means

Antes de descrever o Cascade Simple K-Means, será descrito primeiro o K-Means, pois o primeiro consiste em uma derivação do segundo. O algoritmo K-means, também chamado como K-media. É baseado em análise e comparações entre valores numéricos dos dados utilizados. O algoritmo analisa todos os dados do vetor de características para criar os clusters, e classificar os exemplos disponibilizados. A quantidade de clusters no algoritmo é chamado de K, podendo assim dar a nomenclatura da primeira letra do algoritmo: K-Means.

O algoritmo calcula a distância entre cada ponto no grafo. Enquanto calcula a distância entre os vetores e os centroides para cada um dos clusters, a cada repetição, o valor de cada centroide é refinado pela média dos valores de cada atributo de cada ocorrência que pertence a este.

O *Cascade Simple K-Means* foi escolhido para a clusterização dos tweets negativos, pois em comparação os demais algoritmos utilizados, apresentou um melhor desempenho ao criar clusters distintos. Este algoritmo seleciona os melhores clusters de acordo com o critério de *calinski-harabasz* (T. Calinski, 1974) que consiste em subdividir os documentos em um grafo de arvores conectadas umas às outras, buscando seus vizinhos mais próximos e criando também sub-arvores, segundo Calinski e Harabasz este procedimento é realizado de forma a necessitar menos processamento em relação a um K-means convencional. Para calcular a distância entre os vetores foi utilizada a distância Euclidiana que consiste basicamente em calcular a distância da linha formada entre os pontos em um grafo.

6 Resultados

Será apresentado nesta seção os resultados obtidos na avaliação do classificador *Naive Bayes* nos tweets coletados, e também os grupos gerados com *Cascade Simple K-means*. Primeiramente serão exibidos os resultados obtidos pela classificação humana e em seguida os resultados obtidos pelo classificador no mesmo conjunto de dados. Posteriormente é realizada uma comparação entre as duas classificações e uma análise de desempenho do classificador *Naive Bayes*. Na sequência são exibidas as proporções de todos tweets classificados pelo classificador e, por fim, são exibidas nuvens de palavras e dendogramas para explicação dos grupos gerados a partir dos tweets negativos.

6.1 Análise de Sentimentos

Para a identificação do sentimento dos tweets coletados, voluntários classificaram os tweets, através do site criado para este fim, em positivo, neutro ou se a mensagem era negativa. No período entre 19 de junho de 2017 à 31 de agosto do mesmo ano, 159 voluntários classificaram 14.707 tweets, sendo 2.315 classificados como positivo, 5.816 neutros e 6.576 identificados como negativo. Na Figura 17**Erro! Fonte de referência não encontrada.** podemos observar a distribuição das classificações feitas por voluntários.

Com a classificação feita por voluntários foi criado um padrão ouro, que é a base de dados já classificada, para o treinamento e teste do modelo de classificação.

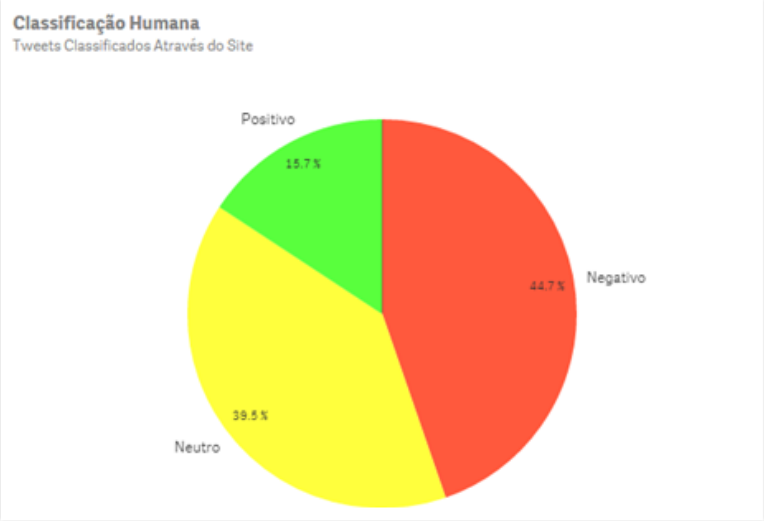

Figura 17 - Classificação Humana

6.1.1 Validação do Classificador

Após o treinamento e teste do classificador foi gerada uma matriz de confusão para a validação do desempenho do modelo. Podemos observar o resultado dessa matriz na Tabela 3.

		Classe Predita (Classificação Naive Bayes)		
		Negativo	Neutro	Positivo
Classe Ouro	Negativo	1562	406	153
	Neutro	511	1090	336
	Positivo	205	328	251

Tabela 3 - Matriz de Confusão

Com base na matriz de confusão gerada podemos aferir uma acurácia de 0,60 do modelo. Porém ao avaliarmos a classe objeto de nosso estudo, a negativa, os resultados são ainda melhores. A precisão aferida para a classe negativo foi de 0,69 enquanto a revocação resultou em 0,74. Já na medida f-measure o resultado atribuindo pesos iguais à precisão e revocação foi de 0,71, quando consideramos

maior importância à precisão para de avaliar o desempenho do classificador, o f-measure resultou em 0,70, um resultado um pouco abaixo do f-measure que atribuiu uma relevância maior à revocação que foi de 0,73. Na Tabela 4 podemos observar as medidas de desempenho para cada classe.

	Precisão	Revocação	$f_{1\text{-measure}}$	$f_{0,5\text{-measure}}$	$f_{2\text{-measure}}$
Negativo	0,69	0,74	0,71	0,70	0,73
Neutro	0,28	0,56	0,37	0,31	0,47
Positivo	0,28	0,32	0,30	0,28	0,31

Tabela 4 - Medidas de Desempenho

6.1.2 Resultados do Classificador

Uma vez realizada a validação do classificador, todos os tweets que haviam sido coletados foram submetidos ao modelo de classificação, além disso o sistema desenvolvido passou a coletar, classificar e armazenar os tweets já classificados na base de dados. No total 29.246 tweets foram classificados pelo modelo sendo 4.353 positivos, 11.286 classificados como neutro e 13.607 como negativo. Na Figura 18 podemos verificar a distribuição dos tweets classificados.

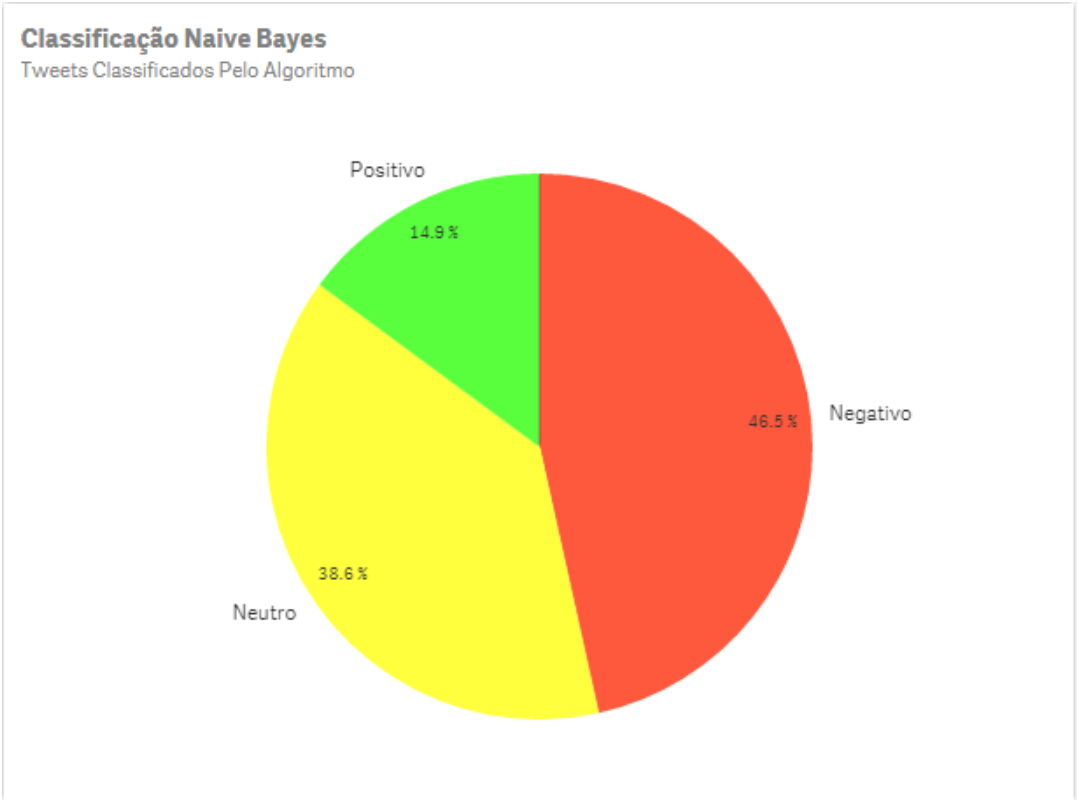

Figura 18 - Classificação Naive Bayes

Observando a distribuição dos tweets classificados pelo modelo, podemos notar que seus resultados ficaram muito parecidos com a classificação humana, o que podemos avaliar como algo muito positivo para o nosso estudo.

6.2 Clusters

Após realizada a classificação dos tweets coletados, foi realizada uma classificação de *Clusters* dos tweets negativos com o intuito de identificar nos tweets os principais problemas relatados pelos usuários de *e-commerce.* Após a classificação destes clusters, foram encontrados 6 grupos onde foram geradas nuvens de palavras e dendogramas para facilitar a explicação dos mesmos. A seguir estas nuvens serão demonstradas e explicadas. Na Figura 19 pode ser visto como foram distribuídos os tweets entre os clusters.

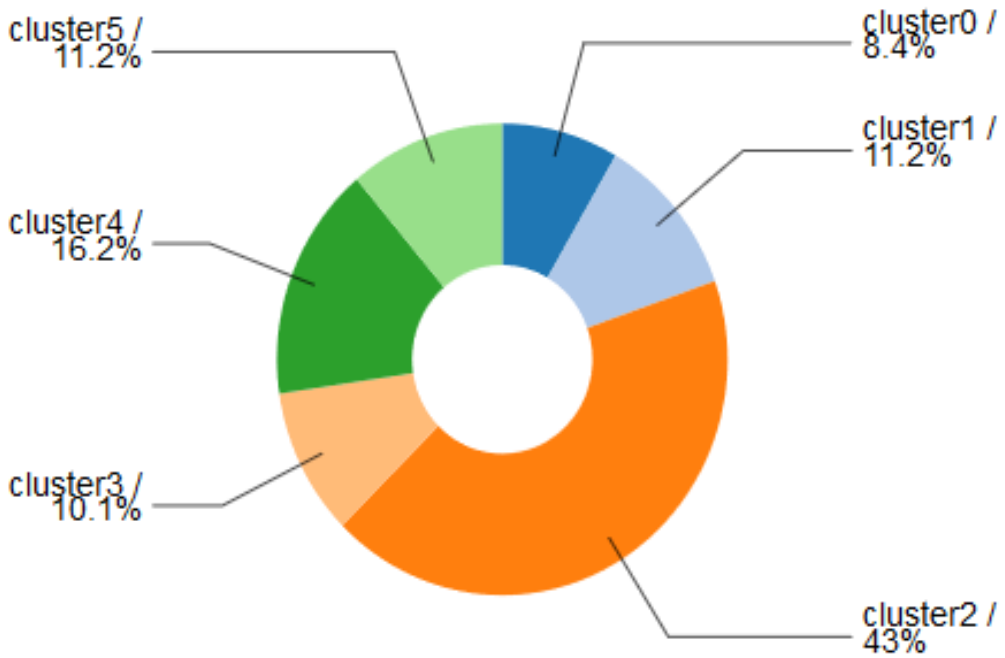

Figura 19 - Divisão dos Clusters

6.2.1 Descrição dos Clusters

Na Figura 20 é demonstrada uma nuvem de palavras gerada a partir de todos os tweets classificados como negativo pelo algoritmo Naive Bayes. Nela as palavras que se destacam são produto, pedido, comprou, entrega e dias. Podemos dizer que em geral, as queixas são sobre problemas pedidos e compras de produtos e sua entrega.

Figura 20 - Nuvem Tweets Negativos

A Figura 21 representa o dendograma criado a partir destes tweets, nele é possível visualizar com maior clareza o relacionamento entre as palavras.

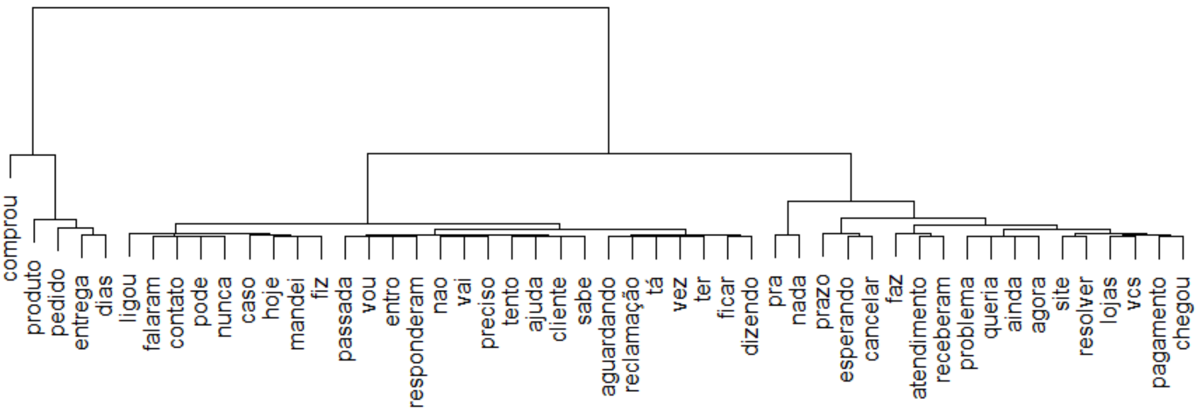

Figura 21 - Dendograma Tweets Negativos

A seguir serão exibidas nuvens de palavras criadas para cada cluster gerado no Weka.

A Figura 22 representa o *Cluster 0* com 8,4% dos tweets negativos, este cluster refere-se aos clientes que informam que “Nunca comprarão determinado produto devido a problemas com recebimento”.

Figura 22 - Cluster 0

A Figura 23 demonstra o dendograma dos tweets classificados no *Cluster 0* e o relacionamento entre as principais palavras do mesmo.

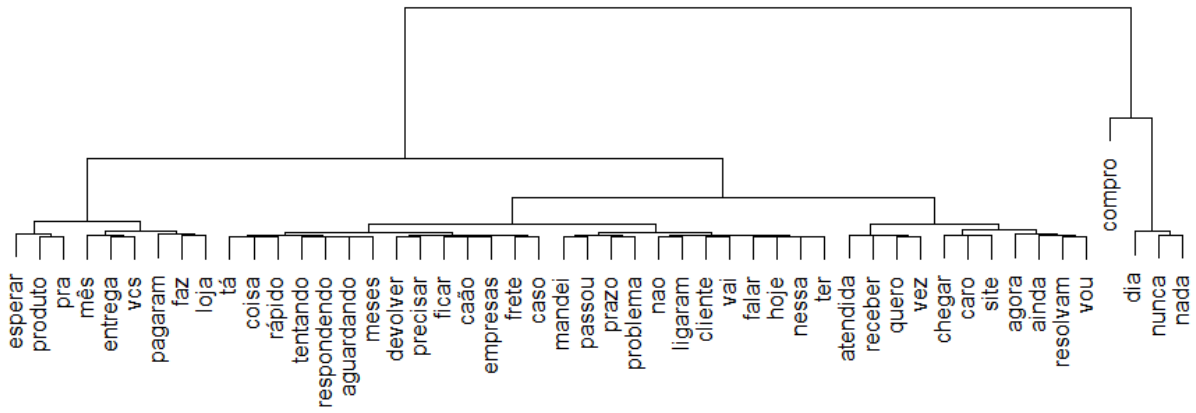

Figura 23 - Dendograma Cluster 0

A Figura 24 representa a nuvem criada a partir dos tweets negativos do *Cluster 1* que abrange 11,2% dos tweets classificados como negativo. Analisando as palavras da nuvem é possível identificar que os usuários reclamam do prazo de entrega dos produtos e pedidos.

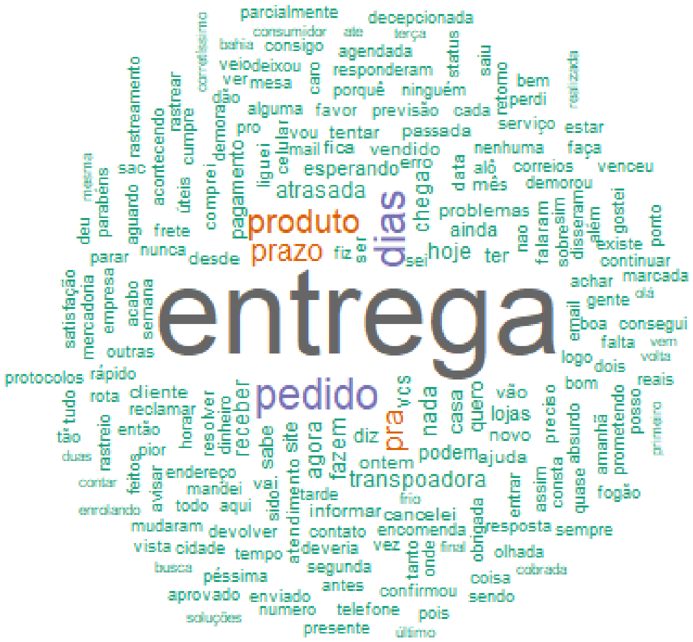

Figura 24 - Cluster 1

Na sequência a Figura 25 representa o relacionamento das palavras do *Cluster 1*.

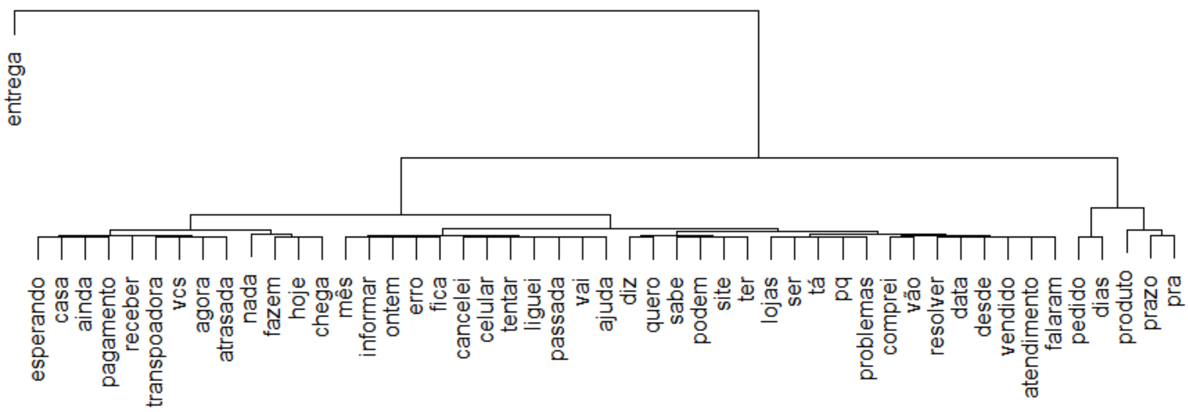

Figura 25 - Dendograma Cluster 1

A Figura 26 corresponde a nuvem gerada a partir dos tweets do *Cluster 2*, este cluster foi o que abrangeu a maior parte dos tweets, com 43% dos tweets classificados como negativo, é possível dizer que este é o principal ponto de reclamação dos usuários. Analisando as palavras em destaque da nuvem pode-se observar que a principal queixa é sobre o atendimento ao cliente onde também são realizadas as reclamações em diversos canais.

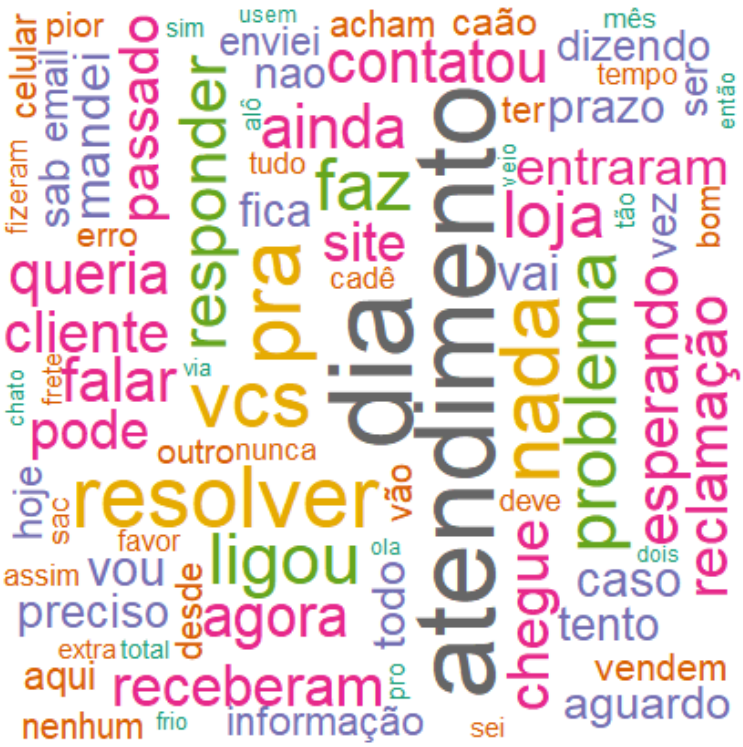

Figura 26 - Cluster 2

A Figura 27 representa o relacionamento entre as palavras do *Cluster 2*, reforçando o entendimento do mesmo.

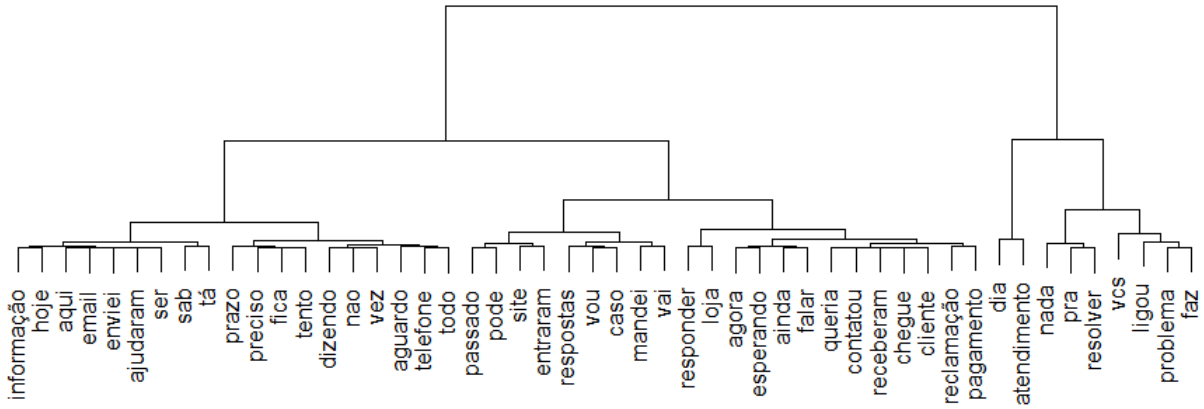

Figura 27 - Dendograma Cluster 2

A Figura 28, corresponde a nuvem gerada a partir dos tweets classificados no *Cluster 3*, com 10,1% dos tweets coletados, onde podemos observar que as palavras mais frequentes são compra e produto. Porém, na nuvem, as demais palavras possuíram frequência e distância semelhantes as estas duas, utilizando o dendograma é possível identificar alguns relacionamentos, desta forma este cluster pode ser entendido como “Problemas na compra de produto, estorno e recebimento”.

Figura 28 - Cluster 3

A Figura 29 demonstra o dendograma do *Cluster 3* mostrando o relacionamento entre as principais palavras dos tweets deste cluster.

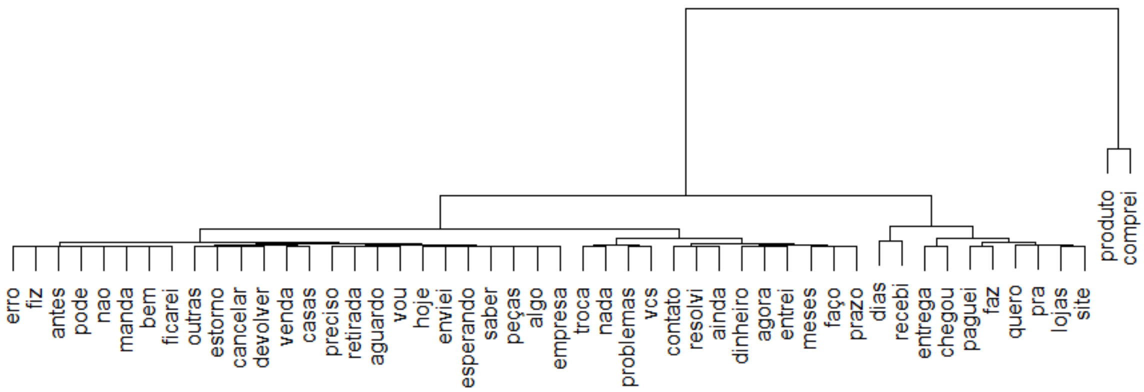

Figura 29 - Dendograma Cluster 3

A Figura 30 corresponde a nuvem criada a partir dos tweets classificados no *Cluster 4*, com 16,2% dos tweets. Analisando a nuvem podemos observar através das palavras destacas que a principal queixa dos usuários é de “Compra no site e problemas relacionados ao prazo de entrega” ainda neste cluster pode ser verificado que um produto específico foi destacado, os aparelhos celulares.

Figura 30 - Cluster 4

A Figura 31 demonstra o relacionamento entre as palavras do *Cluster 4*.

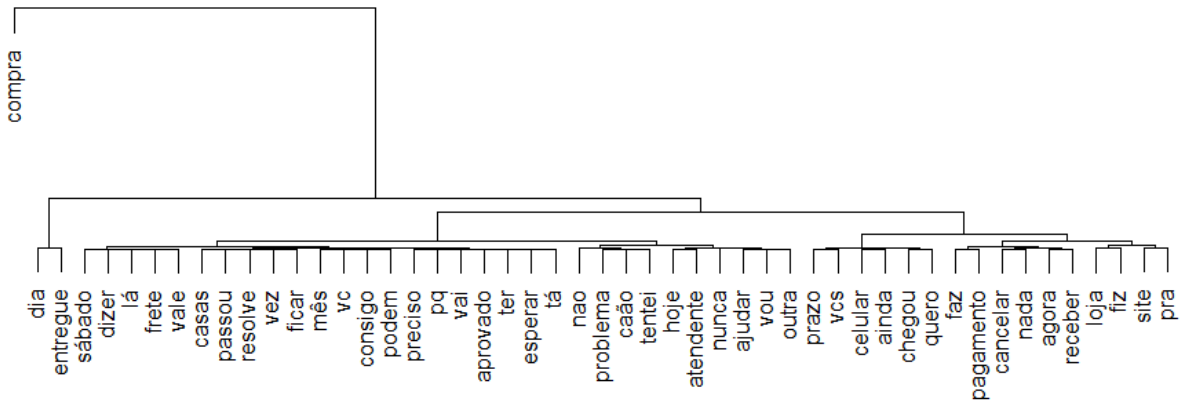

Figura 31 - Dendograma Cluster 4

A Figura 32 representa o *Cluster 5*, com 11,2% dos tweets classificados como negativo é possível observar que as palavras em destaque para esta nuvem são: pedido, dia, compra e cancelar. Analisando a nuvem e o dendograma pode-se dizer que os usuários estão reclamando de “Não conseguir cancelar pedido ou compra, nos canais de atendimento”.

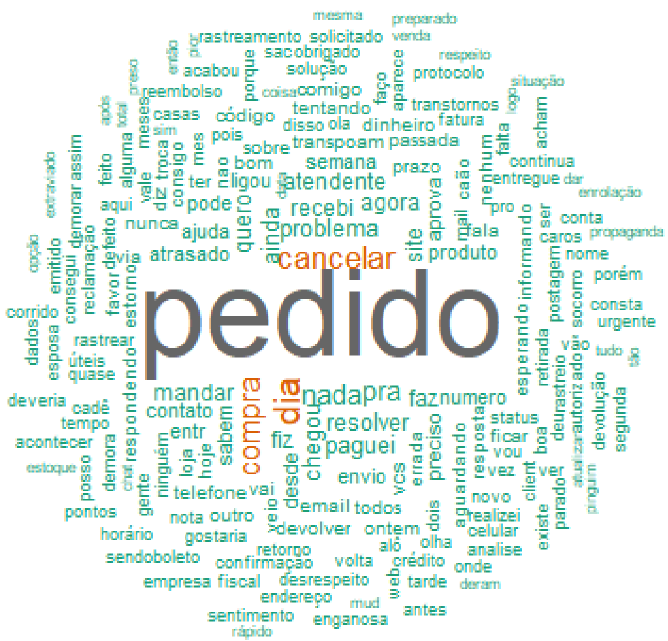

Figura 32 - Cluster 5

Por fim a Figura 33 demonstra o dendograma do *Cluster 5* facilitando a visualização do relacionamento das palavras do mesmo.

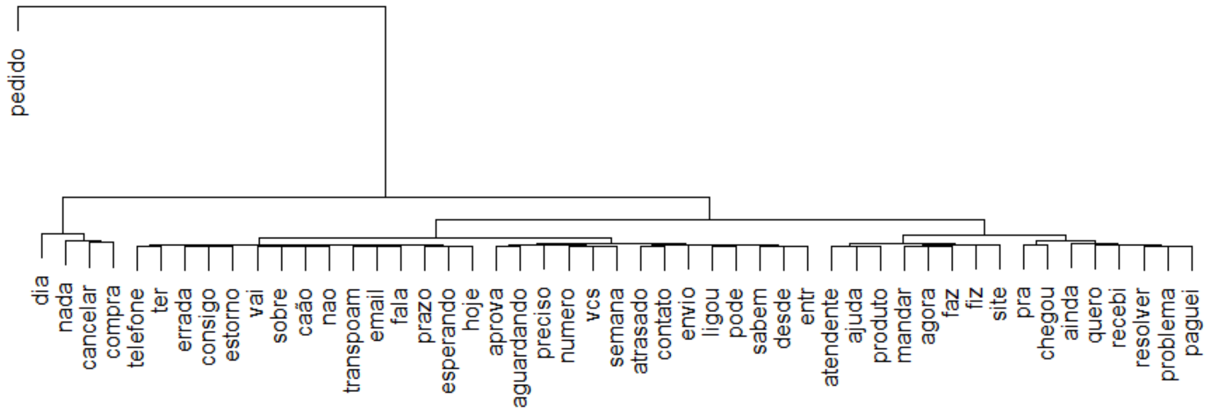

Figura 33 - Dendograma Cluster 5

6.3 ETL

Para criação das visões no Pentaho foi criado um ETL utilizando a ferramenta *Spoon* do Kettle. No primeiro momento é criada a tabela fato, na sequência é realizada a coleta dos dados nas diversas fontes utilizadas. Após a coleta e consolidação dos dados são realizadas as categorizações de algumas dimensões utilizando scripts java, em seguida é realizada a identificação da loja a que se refere o tweet e por último o input dos dados. A Figura 34 demonstra o ETL na interface do *Spoon*.

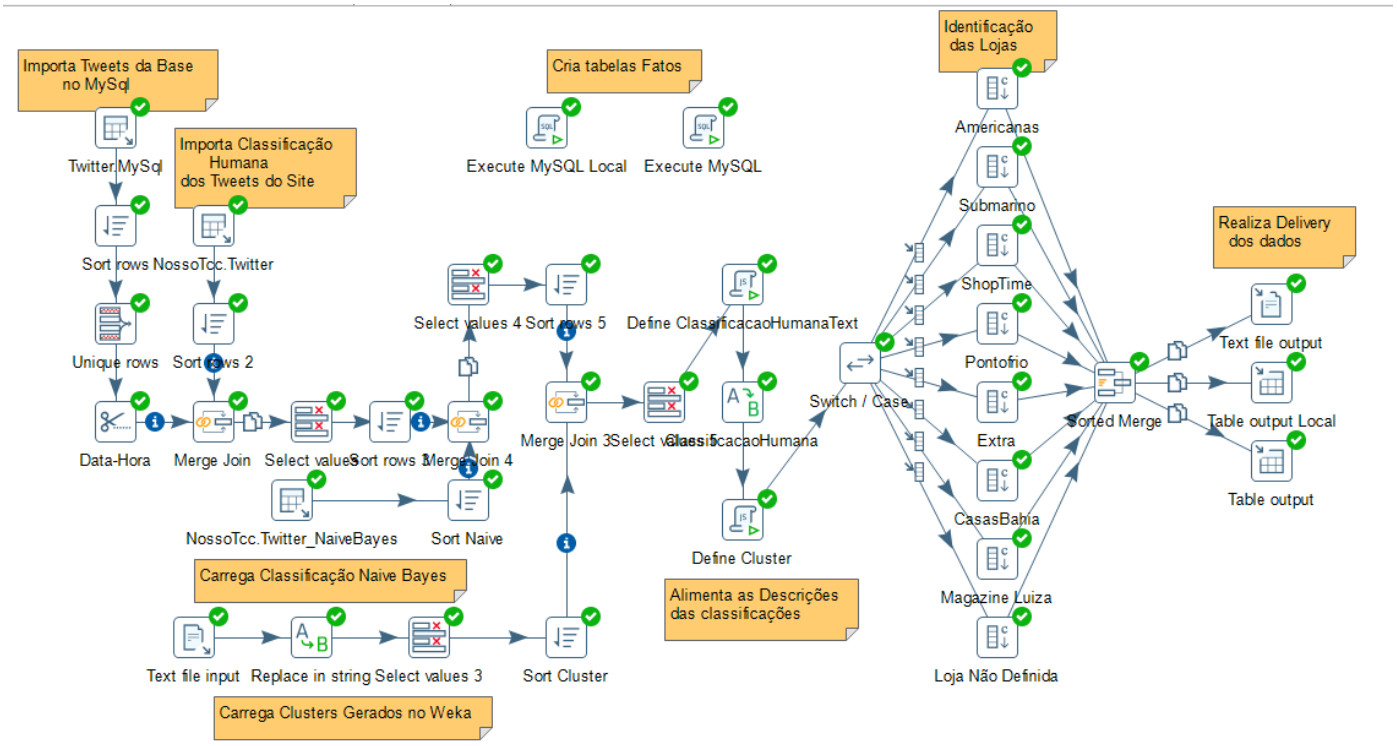

Figura 34 – ETL

6.4 Cubo e Dashboard

Com a tabela fato criada, foram criadas as conexões no Pentaho, a tabela fato completa foi utilizada como Cubo no *Saiku Analitycs* permitindo análises e exploração dos dados de acordo com a necessidade do usuário. A Figura 35 representa a interface do *Saiku Analitycs*.

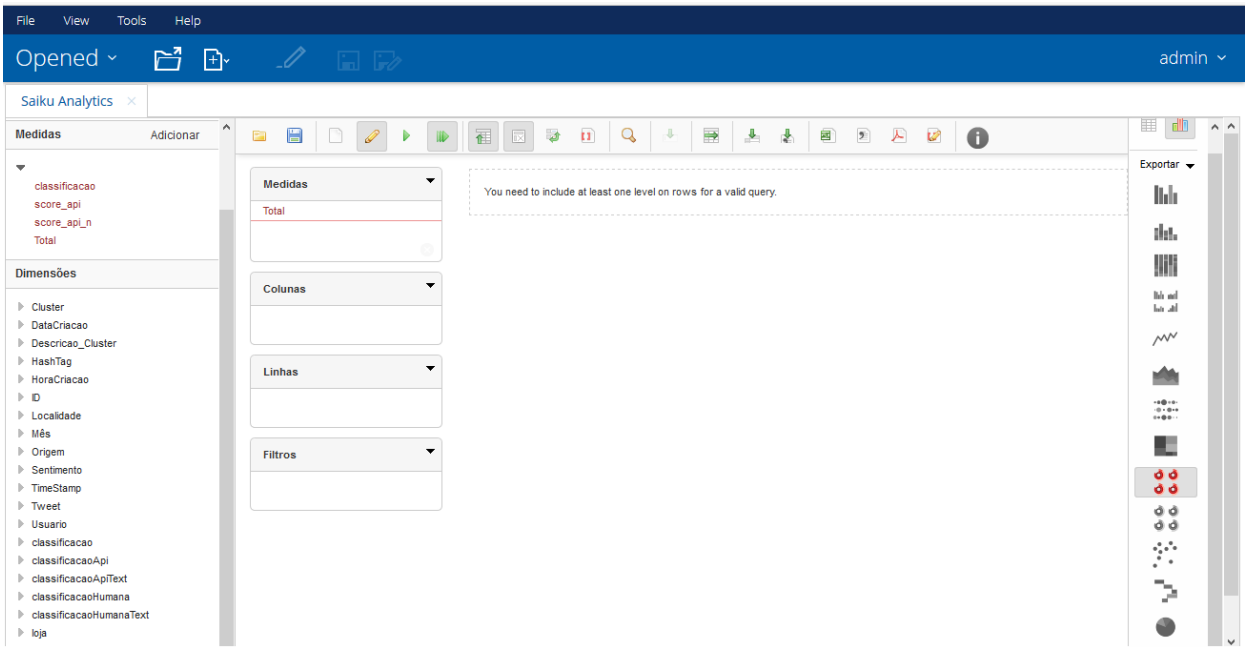

Figura 35 - Saiku Analytics

A partir da mesma tabela fato, foi criado um *dashboard* onde são exibidas as classificações de sentimentos de todas as empresas e logo abaixo os clusters e a distribuição dos tweets no decorrer do tempo, na parte inferior é possível também utilizar um filtro por loja. A Figura 36 representa o *dashboard* criado.

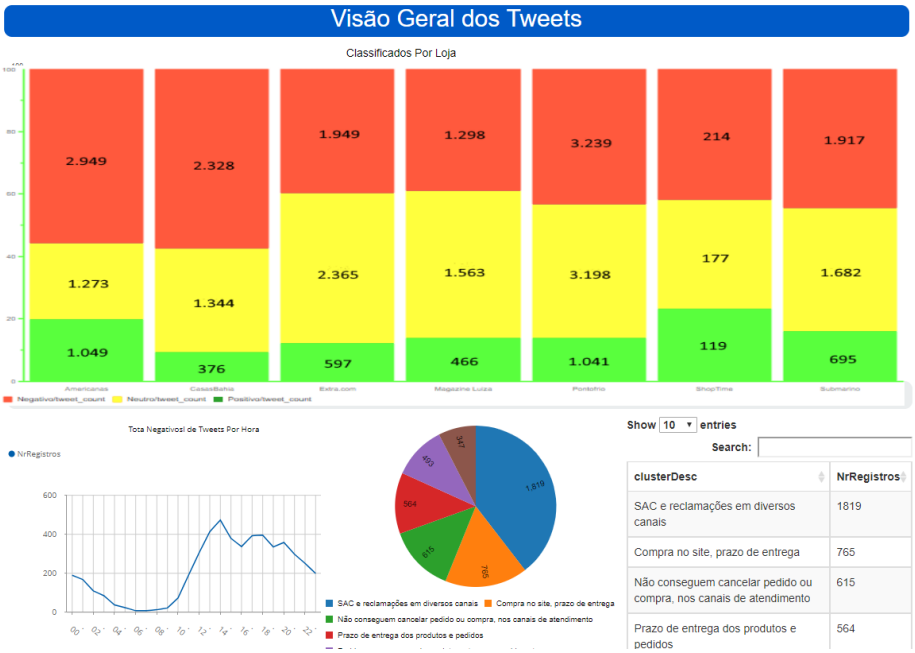

Figura 36 - Dashboard

7 Discussão

Este capítulo aborda a discursão sobre os resultados obtidos no capítulo anterior, será divido em duas partes, primeiramente sobre a avaliação da classificação do algoritmo *Naive Bayes*, e na sequência sobre os clusters de tweets negativos gerados com *Cascade Simple K-means*.

7.1 Avaliação do Algoritmo *Naive Bayes*

Nesta parte do trabalho serão discutidas a classificação humana realizada pelos voluntários e também a classificação realizada pelo algoritmo, na segunda veremos tanto o comparativo de ambas as classificações como também o resultado total da classificação do algoritmo.

Classificação Humana

Apenas com o apoio dos voluntários através do site para classificar os tweets foi possível realizar o treinamento do algoritmo, porém, como alguns textos eram complexos, com ironias, gírias ou textos fora de contexto, algumas classificações realizadas por humanos, mesmo após a ponderação, não poderiam ser aceitos como um valor correto para a classe em um tweet especifico. Este tipo de classificação errônea pode ter um efeito negativo no desempenho do treinamento do classificador. Na Tabela 5 temos alguns exemplos destes tweets que poderiam causar dúvidas nos voluntários.

Tweet	Sentimento
Que absurdo esse preco.... Praticamente um roubo...	Neutro
@americanascom Bom dia! Comprei uma mercadoria pelo site e ate agora nao chegou em minha residencia. Preciso de uma posicao, por favor!	Neutro
N compro mais na @pontofrio e aconselho a n comprarem.Minha mercadoria foi extraviada, abri 4 protocolos e meu problema n foi resolvido	Neutro
@CasasBahia obrigada, nao preciso mais..alias procurei bastante por quem pudesse solucionar meu problema..todo o caso foi contado no reclame	Positivo
Me ligaram sabado das @CasasBahia dizendo que eu poderia vir na loja pegar um novo ar condicionado. Chego aqui e...???? Nao posso!	Positivo
@magazineluiza ...extendida. Tenho a nota fiscal e estou tentando falar no 0800 e nao consigo, fiquei duas vezes mais de 15 min esperando...	Neutro

Tabela 5 - Classificação Humana Errônea

Algoritmo *Naive Bayes*

Com os tweets classificados por humanos o algoritmo *Naive Bayes* foi capaz de classificar os demais tweets coletados durante o desenvolvimento do trabalho, no geral os resultados dos tweets classificados no algoritmo mantiveram proporção relativa aos da classificação humana, com uma variação bastante sucinta entre os percentuais de cada classe, considerando que o algoritmo classificou um pouco mais que o dobro de tweets que os humanos, pode-se dizer que o resultado superou as expectativas.

Comparação dos resultados

Realizando a validação dos tweets classificados por humanos e pelo algoritmo no mesmo conjunto de dados, foi possível identificar que em alguns casos, o desempenho do classificador foi superior ao dos humanos para identificar tweets negativos, utilizando o mesmo exemplo de tweets citado anteriormente na Tabela 6 é possível ver que o algoritmo realizou a classificação correta destes.

Tweet	Classif. Hum.	Classif. Alg.
Que absurdo esse preco.... Praticamente um roubo...	Neutro	Negativo
@americanascom Bom dia! Comprei uma mercadoria pelo site e ate agora nao chegou em minha residencia. Preciso de uma posicao, por favor!	Neutro	Negativo
N compro mais na @pontofrio e aconselho a n comprarem.Minha mercadoria foi extraviada, abri 4 protocolos e meu problema n foi resolvido	Neutro	Negativo
@CasasBahia obrigada, nao preciso mais..alias procurei bastante por quem pudesse solucionar meu problema..todo o caso foi contado no reclame	Positivo	Negativo
Me ligaram sabado das @CasasBahia dizendo que eu poderia vir na loja pegar um novo ar condicionado. Chego aqui e...???? Nao posso!	Positivo	Negativo
@magazineluiza ...extendida. Tenho a nota fiscal e estou tentando falar no 0800 e nao consigo, fiquei duas vezes mais de 15 min esperando...	Neutro	Negativo

Tabela 6 - Classificação Humana X Classificação Naive Bayes

Da mesma maneira, houve alguns casos onde o algoritmo classificou erroneamente alguns tweets, esta classificação pode ser explicada devido as palavras chaves dos tweets serem mais recorrentes em tweets negativos em relação tweets positivos ou neutros. Alguns exemplos destes tweets podem ser vistos na Tabela 7.

Tweet	Classif. Hum.	Classif. Alg.
@HamamLuiz_ Da uma olhada nesse, muita gente recomenda	Positivo	Negativo
Eu so acho que a @submarino fica me tentando com promocoes maravilhosas pelo App, so essa semana fiz duas compras. Socorro!!!	Positivo	Negativo
Ultimo dia pra entregarem minha encomenda, sera que chega @magazineluiza ?	Neutro	Negativo
A @americanascom e a melhor loja da internet sempre entrega os produtos na data marcada	Neutro	Negativo
@americanascom bom demais, recebi a uma hora atras! :) Valeu!!!!!!!!	Positivo	Negativo
A @americanascom e tao foda, eles me deram 5 dias uteis pra entrega, passou 2 dias e ja chegou <3	Positivo	Negativo

Tabela 7 - Classificação Errônea Naive Bayes

Considerando as ocorrências citadas anteriormente, pode ser levantada a hipótese de que a precisão do modelo para detectar os tweets negativos é maior do que a precisão retornada pela matriz de confusão.

7.2 Clusters Tweets Negativos

Com os tweets classificados como negativos pelo *Naive Bayes* foi possível aplicar o algoritmo de Cluster (*Cascade Simple K-means*) para que desta forma fosse possível identificar a quais etapas da compra no e-commerce estavam ocorrendo o maior número de reclamações.

Com base nos resultados obtidos, perceptível que não há distinção na distribuição relativa das reclamações, onde os grupos observados ocorrem de forma semelhante entre os grupos de empresas estudados. A Figura 37 demonstra a distribuição dos tweets nos clusters em cada empresa estudada.

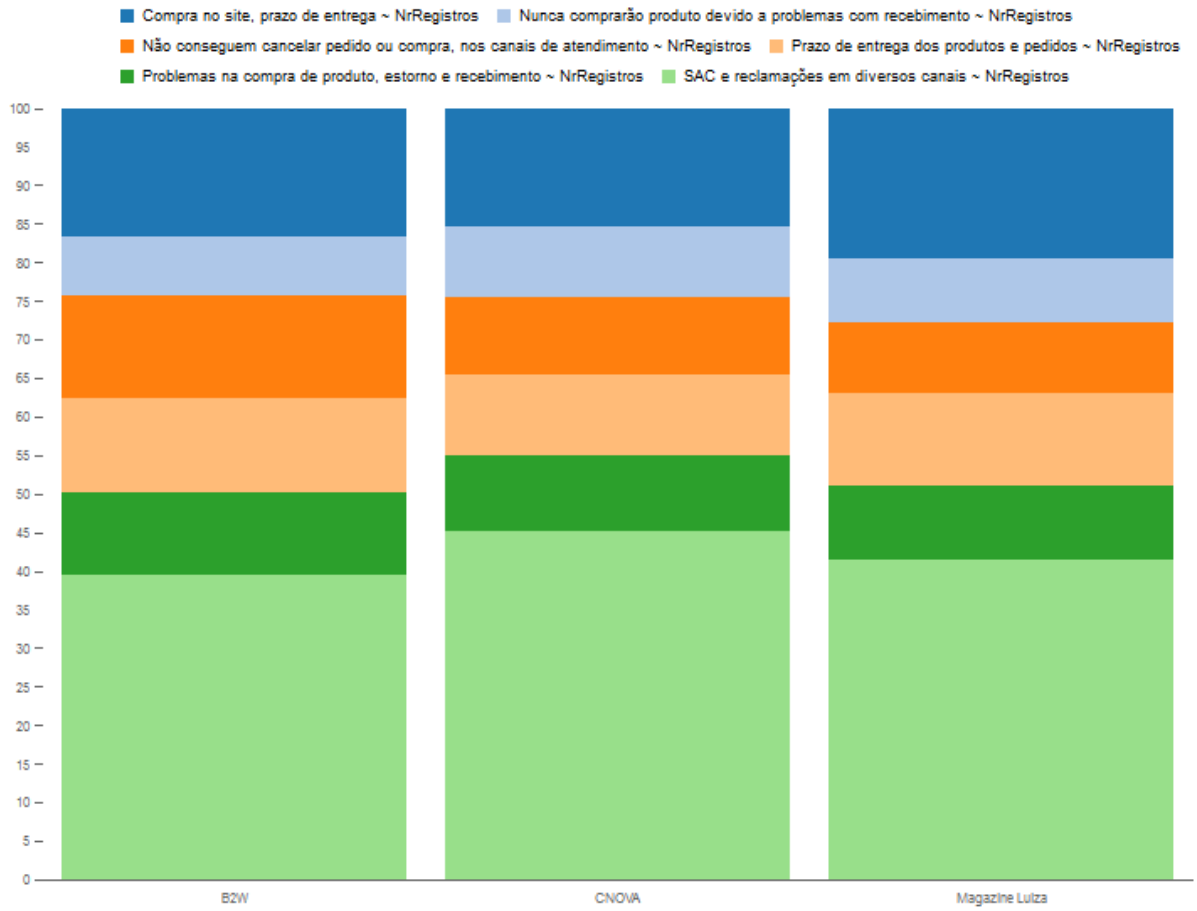

Figura 37 - Cluster por loja

Após a criação dos clusters foi possível observar que 43% dos tweets negativos estavam relacionados ao pós-venda, ou seja, problemas que ocorriam após realizada a compra, nos serviços de atendimento ao consumidor, enquanto os demais foram divididos em 5 grupos, cada um contendo entre 8% e 11% dos tweets negati-

vos, nestes 5 grupos os problemas relatados são basicamente sobre prazo de entrega ou recebimento e dificuldades com cancelamento e estorno das compras.

As nuvens de palavras e principalmente os dendogramas foram determinantes na identificação destes padrões nos clusters, de outra forma seria quase impossível categorizar cada cluster e obter o significado dos mesmos.

8 Conclusão

Apresentamos neste trabalho a construção de um método de classificação de sentimento aplicada em cima dos posts do Twitter em português brasileiro relacionados aos tipos de reclamações que são mais frequentes dentre os clientes que utilizam E-commerce das empresas CNova, B2W e Magazine Luiza no momento de realizar a compra de um determinado produto.

Inicialmente foi realizado a construção de duas aplicações para coletar e armazenar os *tweets* através das linguagens R e *Python*, além de coleta dos dados as aplicações realizavam a armazenagem destes textos no banco de dados SQL *Azure*, que foi utilizada como centralizador dos dados para esse trabalho. Foi utilizada uma amostra de aproximadamente vinte mil *tweets*, realizando uma cópia destes para um banco *MySQL* utilizado em um site criado para que voluntários pudessem classificar manualmente os *tweets* coletados em positivo, negativo ou neutro, dentro do escopo do trabalho. Estas classificações foram utilizadas para treinar o algoritmo classificador. O processo de ETL que consolidou as classificações humanas dos *tweets*, pois consideramos que sua identificação é complexa até mesmo para humanos que realizaram a sua classificação. A amostra foi dividida em dois grupos, sendo um terço para teste e dois terços para treinamento do algoritmo, em seguida foi feita a remoção de *stopwords*, processo de *tokenização*, *stemmen*, extração de termos únicos e geração do vetor de características. Após esse processo foi realizado o treinamento e testes com o algoritmo *Naive Bayes*.

Com os tweets classificados como negativo, foi possível aplicar um algoritmo de cluster para encontrar padrões nos tweets, após gerados os clusters foram criadas nuvens de palavras e dendogramas que permitiram observar que de todos os tweets classificados como negativo pelo algoritmo Naive Bayes os principais problemas no processo de compra em e-commerce conforme a lista abaixo:

- Nunca comprarão determinado produto devido a problemas com recebimento;
- Prazo de entrega dos produtos e pedidos;
- Atendimento ao cliente onde também são realizadas as reclamações em diversos canais;
- Problemas na compra de produto, estorno e recebimento;
- Compra no site e problemas relacionados ao prazo de entrega;

- Não conseguir cancelar pedido ou compra, nos canais de atendimento;

A contribuição desse trabalho é identificar os problemas mais apresentados pelos clientes, os pontos do processo com maiores níveis de reclamação, permitindo ser criadas estratégias para resolver estes fatores que prejudicam a imagem da empresa e a experiência de compra do cliente. Além de mostrar como é possível realizar uma análise de dados em cima de extrações de informações, possibilitando a criação de estratégias para uma determinada empresa.

Bibliografia

TAKAZONO BORGATO RIBEIRO, Karina. ***E- commerce* – Atraindo e conquistando clientes para o varejo virtual.** Mato Grosso: SINOP, 2007.

ALMEIDA LIMA, Márcia C.; BERTARELLI, Rosana; PEREIRA ALVES, Rosilene. **Fidelização de Clientes: Uma ferramenta estratégica de marketing.**

TRINDADE DE LIMA, Felipe; ROUBERTE DE FREITAS, Jean. **Uma comparação entre plataformas de *Business Intelligence* usando bases de dados governamentais.** Rio de Janeiro: Universidade Federal do Estado do Rio de Janeiro, 2014.

BRAGA RIBEIRO, Lucas. **Análise de sentimento em comentários sobre aplicativos para dispositivos móveis: Estudo do impacto do pré-processamento.** Brasília: Universidade de Brasília, 2015.

COSTA E LIMA, Helena C.. **Utilização de Data Warehouse e Data Mining no acompanhamento das atividades de pesquisa do CEULP/ULBRA.** Palmas/ TO: Centro Universitário Luterano de Palmas Ulbra ,2006.

Business Inteligence: Um enfoque gerencial para a inteligência do negócio /Efraim Turban...[et al.]; tradução Fabiano Bruno Gonçalvez. – Porto Alegre: Brookman, 2009.

SMAAL, Beatriz. **A História do Twitter.** O Estado do Paraná, 19 de fev. de 2010. Disponível em: **<**https://www.tecmundo.com.br/rede-social/3667-a-historia-do-twitter.htm**>.** Acesso em: 15 abr. 2017.

TEIXEIRA, Carlos A. **A origem do Facebook: Saiba sobre a história da rede social mais popular do mundo que abre a capital nesta sexta – feira.** 18 mai. 2012. Disponível em:< https://oglobo.globo.com/sociedade/tecnologia/a-origem-do-facebook-4934191 > Acesso em: 15 abr. 2017.

MARQUES, José R**.. Conheça a história de sucesso de Luiza Helena Trajano, da**

rede Magazine Luiza. 11 abr. 2014. Disponível em: < http://www.ibccoaching.com.br/portal/exemplo-de-lideranca/historia-sucesso-luiza-helena-trajano-magazine-luiza/> Acesso em: 01 mai. 2017

LOPES DOS SANTOS, Carolina L**.. Melhoria da Atratividade Internacional das Instituições de Ensino Superior através de Análise de Sentimentos**. Lisboa: Instituto Universitário de Lisboa,2016.

Deweik, Albert. 2016. Como reduzir reclamações de clientes nas redes sociais. [Online] 25 de 10 de 2016. [Citado em: 15 de 07 de 2017.] Albert Deweik é CEO da NeoAssist. https://www.ecommercebrasil.com.br/artigos/reduzir-reclamacoes-redes-sociais/.

Figueiredo, Pedro. 2017. Solução híbrida da VASP aposta em Microsoft Azuere para a componete na nuvem. *Microsoft.* [Online] Microsoft, 27 de 02 de 2017. [Citado em: 30 de 04 de 2017.] Página sobre Casos de Sucesso da Mircrosoft - Citação de Pedro Figueiredo: Information Architect. https://customers.microsoft.com/pt-PT/story/vasp-retail-and-consumer-goods-azure-storage-app-service-sql-database-pt.

IT FORUM 365. 2016. Brasil tem mais de 100 milhões de internautas, aponta IBGE. *IT FORUM 365.* [Online] 25 de 11 de 2016. [Citado em: 16 de 07 de 2017.] http://www.itforum365.com.br/conectividade/dispositivos/brasil-tem-mais-de-100-milhoes-de-internautas-aponta-ibge.

—. 2017. Internet móvel ganha mais de um milhão de novos usuários por dia. *IT FORUM 365.* [Online] 17 de 06 de 2017. [Citado em: 16 de 07 de 2017.] http://www.itforum365.com.br/conectividade/estrategia-movel/internet-movel-ganha-mais-de-um-milhao-de-novos-usuarios-por-dia.

Kristensen, Chirstian Haag, et al. 2011. Normas brasileiras para o Affective Norms for English Words. *Normas brasileiras para o ANEW-Br.* 2011, pp. 135-146. Traduzido pelo autor.

Lima, Helena Costa e. 2006. Utilização de Data Warehouse e Data Mining no acompanhamento das atividades de pesquisa do CEULP/ULBRA. Palmas, Tocantins, Brasil : Centro Universitário Luterano de Palmas Ulbra, 2006.

Magazine Luiza. Nossa História. [Online] [Citado em: 09 de 07 de 207.] http://www.magazineluiza.com.br/quem-somos/historia-magazine-luiza/.

Martins, Sergio. 2016. Evolução do e-commerce significou o desenvolvimento de um novo mercado com uma lógica própria, que as empresas precisam compreender. *Computer World.* [Online] 08 de 09 de 2016. [Citado em: 09 de 07 de 2017.] Sergio Martins é diretor de mídias e e-commerce da Exceda.. http://computerworld.com.br/uma-breve-historia-do-e-commerce.

Oliveira, Frank Willian Cardoso de. 2013. Análise de sentimentos de comentários em português utilizando SENTIWORDNET. *Análise de sentimentos de comentários em português utilizando SENTIWORDNET.* Maringá, Paraná, Brasil : Universidade Estadual de Maringá, 2013. p. 45.

Portal Educação. 2015. História do MySQL. [Online] 21 de 08 de 2015. [Citado em: 01 de 08 de 2017.] https://www.portaleducacao.com.br/conteudo/artigos/informatica/historia-do-mysql/66679#.

Ribeiro, Lucas Braga. 2015. Análise de sentimentos em comentários sobre aplicativos para dispositivo móveis: Estudo do impacto do pré-processamento. *Análise de sentimentos em comentários sobre aplicativos para dispositivo móveis: Estudo do impacto do pré-processamento.* Brasília, Brasília, Brasil : Universidade de Brasília, 2015. p. 82.

Rossum, Guido Van. 2014. A História do Python. *Mind Bending.* [Online] 08 de 10 de 2014. [Citado em: 23 de 04 de 2017.] http://mindbending.org/pt/a-historia-do-python.

Sebrae. 2016. COMÉRCIO ELETRÔNICO A tendência do comércio eletrônico no varejo de autopeças. *Sebrae.* [Online] Sebrae, 21 de 12 de 2016. [Citado em: 15 de 07 de 2017.] https://www.sebrae.com.br/sites/PortalSebrae/artigos/a-tendencia-do-comercio-eletronico-no-varejo-de-autopecas,a97872eefd4e8510VgnVCM1000004c00210aRCRD.

Srivastava, Jaideep. Web Mining: Accomplishments & Future Directions. *IEEE Sección Argentina.* [Online] [Citado em: 07 de 05 de 2017.] Traduzido por Scoz, Diogo Pietro. http://www.ieee.org.ar/downloads/Srivastava-tut-pres.pdf.

T. Calinski, J. Harabasz. 1974. *A dendrite method for cluster analysis.* 1974.

Teixeira, Fabio Oliveira. 2011. Tese (Mestrado) - Universidade Federal de São Paulo. Escola Paulista de Medicina. Programa de Pós Graduação em Gestão de Informática em Saúde. *Classificação e indexação de artigos científicos internacionais de infromática em saúde.* São Paulo : s.n., 2011.

Turban, Efrain. 2009. Business Inteligence: Um enfoque gerencial para a inteligência de negócio. [trad.] Fabiano Bruno Gonçalvez. *Business Inteligence: Um enfoque gerencial para a inteligência de negócio.* [Trabalho de Conclusão de Curso - TCC]. Porto Alegre, Rio Grande do Sul, Brasil : s.n., 2009.

Web Mining: Conceitos e Aplicações. **Scoz, Diogo Pietro. 2014.** 2014, Revista Científica do Alto VAle do Itajaí, p. 5.

Wright, Alex. 2009. Mining the Web for Feelings, Not Facts. *The New York Times.* [Online] 23 de 08 de 2009. [Citado em: 23 de 04 de 2017.] Tradução: Paulo Migliacci ME - Terra Tecnologia. http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html.

Apêndice A – Lista de palavras e Média que compõem o ANEW-Br

Palavra	Sentimento_Media
ABALADO	2,58
ABANDONADO	1,85
ABDUÇÃO	5
ABELHAS	4,58
ABENÇOADO	8,03
ABORTO	1,67
ABRAÇAR	8,63
ABRAÇO	8,77
ABRASADOR	5,43
ABRIGADO	7,61
ABSURDO	2,79
ABUNDÂNCIA	6,94
ABUSO	2,76
ACALMAR	7,23
ACANHADO	3,89
ACASO	6,17
ACEITAÇÃO	7,21
ACIDENTE	1,67
ACONCHEGANTE	8,29
ACONCHEGO	8,4
ACORDO	6,89
AÇÚCAR	7,04
ADAGA	4,32
ADMIRADO	7,67
ADORÁVEL	8,24
ADULTO	6,49
AFEIÇÃO	7,82
AFINAR	5,8
AFOGAR	2,53
AGILIDADE	7,5
AGONIA	2,05
AGRADÁVEL	8,15
AGRADECIDO	8,42
AGRESSIVO	1,89
ÁGUA	8,35
AGULHA	4,04
ALCOÓLICO	2,76
ALEGRE	8,44
ALEGRIA	8,61
ALERGIA	4,02
ALERTA	6,08
ALERTO	5,67
ALIMENTO	8,26
ALVORECER	6,24
AMABILIDADE	7,68
AMADO	8,31
AMARELO	6,76
AMÁVEL	8,09
AMBIÇÃO	5
AMBULÂNCIA	2,47
AMEAÇA	1,87
AMIGÁVEL	8,25
AMIGO	8,74
AMOR	8,75
ANGUSTIADO	2,22
ANIMAÇÃO	8,33
ANIVERSÁRIO	8,39
ANJO	8,06
ANSEIO	5,11
ANSIOSO	3,5
APARELHO	6
APÁTICO	3,21
APLAUSO	8,21
APRENDER	7,86
AR	8,39
ARANHA	3,49
ARMA	2,24
ARMAMENTO	2,17
ARMÁRIO	6,11
ARREPENDIDO	3,46
ARROGANTE	2,05
ARRUMADO	7,78
ARTE	8,08
ÁRVORE	7,98
ÁS	5,71
ÁSPERO	3,38
ASSALTANTE	2,03
ASSALTO	1,47
ASSAR	6,24
ASSASSINO	1,16
ASSENTO	5,87

ASSOVIO	6,15	BONECA	5,57
ASSUSTADO	2,89	BONITO	8,42
ASTRONAUTA	5,81	BORBOLETA	7,38
ATADURA	2,87	BOXEADOR	4,13
ATERRORIZADO	1,79	BRABEZA	2,83
ATIVAR	7,19	BRABO	2,26
ATLETISMO	7,31	BRAÇO	6,35
ATRAÇÃO	8,22	BRANCO	7,18
AURORA	6,96	BRAVO	2,56
AUTONOMIA	7,71	BRILHANTE	7,45
AVALANCHE	2,32	BRINQUEDO	7,54
AVENIDA	5,84	BRISA	7,39
AVENTURA	8,06	BRUTAL	1,98
AVÔ	7,18	BUQUÊ	7,42
AZEDO	2,87	BURRO	2,96
AZUL	7,42	CABANA	7,21
BACIA	4,68	CABELO	6,67
BAGUNÇADO	3,21	CACHOEIRA	8,14
BALA	6,98	CACHORRO	7,98
BANDEIRA	6,32	CADÁVER	1,7
BANHEIRA	8	CADEIA	2,02
BANHEIRO	6,82	CADEIRA	6,3
BANHO	8,29	CADERNO	5,76
BANQUETA	5,84	CAIXÃO	1,92
BARATA	2,83	CALOR	6,14
BARRA	4,74	CAMA	8,08
BARRIL	5,85	CAMINHÃO	4,96
BASTARDO	3,42	CAMPEÃO	8,43
BEBÊ	8,21	CAMPO	7,61
BEBIDA	6,03	CANÇÃO	8,35
BECO	3,02	CÂNCER	1,49
BEIJO	8,76	CANHÃO	2,38
BELEZA	8,09	CANSADO	2,39
BELISCAR	3,71	CAOS	2,05
BELO	7,73	CAPAZ	7,63
BENZER	6,5	CARCAÇA	3,75
BERÇÁRIO	7,6	CÁRCERE	1,89
BERRAR	3,63	CARÍCIA	8,78
BESTA	2,45	CÁRIE	2,27
BISPO	5,68	CARINHOSO	8,73
BLASFÊMIA	2,63	CAROÇO	3,81
BOBAGEM	4,84	CARRO	7,95
BOLINHO	7,5	CARTA	7,02
BOLO	7,37	CASA	8,14
BOM	8,19	CASAL	7,84
BOMBA	1,53	CASAMENTO	7,59

CASSINO	5,21
CAVALO	6,89
CAVEIRA	2,61
CÉDULA	7,43
CEFALÉIA	2,92
CEGO	2,35
CÉLULA	5,06
CEMITÉRIO	2,22
CESTA	6,06
CÉTICO	4,89
CÉU	8,1
CHACINA	1,43
CHALEIRA	5,19
CHAMUSCAR	4,46
CHANTAGEM	1,61
CHAPÉU	6,16
CHARME	7,63
CHATEAÇÃO	2,3
CHAVE	5,46
CHOCALHO	5,73
CHOCOLATE	8,08
CHUTE	4,4
CHUVA	5,54
CICATRIZ	3,23
CICLONE	2,67
CIDADE	6,41
CINEMA	8,25
CIRCO	6,77
CÍRCULO	5,29
CIRURGIA	2,85
CIÚME	2,86
COBERTURA	7,33
COBRA	2,68
COELHINHO	7,31
COELHO	6,96
COFRE	6,13
COGUMELO	5,35
COLETE	4,63
COLISÃO	2,4
COLUNA	4,65
COMÉDIA	8,39
COMER	7,42
COMPLACENTE	4,71
COMPROMETIDO	6,33
COMPUTADOR	6,62
CONCENTRADO	6,71
CONFIANÇA	7,97
CONFIANTE	7,42
CONFORTO	8,03
CONFUSO	2,54
CONHECIMENTO	8,29
CÔNJUGE	6,02
CONSOLADO	5,36
CONSTRANGIDO	2,96
CONTENTAMENTO	6,72
CONTEÚDO	7,04
CONTEXTO	5,36
CONTINÊNCIA	4,4
CONTROLE	5,75
COR	7,65
CORAÇÃO	7,73
CORAGEM	7,96
CORDA	4,53
CORDEIRO	5,56
COROA	5,58
CORPO	6,94
CORREDOR	5,3
CORRUPTO	1,63
CORTE	3,3
CORTESIA	8,02
CORTIÇA	5
CORTINAS	5,88
CORUJA	5,68
COSTA	5,54
COSTUME	5,42
COTOVELO	4,88
COVARDE	2,51
COZINHEIRO	6,79
CREPÚSCULO	5,34
CRIANÇA	7,92
CRIME	1,72
CRIMINOSO	1,75
CRISE	2,31
CRU	3,58
CRUCIFICAR	2,36
CRUEL	1,67
CULINÁRIA	7,03
CULPADO	2,27
CUPIM	2,99
CURAR	7,83
CURIOSO	6,77
DÁDIVA	7,73

DANÇARINO	6,7
DANO	2,32
DÉBIL	3,13
DÉBITO	2,47
DECEPCIONAR	1,76
DECOMPOR	3,94
DECORAR	6,22
DEDO	5,75
DEFEITO	3,15
DEFICIENTE	3,13
DEFORMADO	2,43
DELEITE	5,53
DELICADO	6,3
DEMÔNIO	2,1
DEMORADO	2,94
DENTISTA	5,18
DEPRESSÃO	1,96
DEPRIMENTE	1,82
DEPRIMIDO	2,2
DERROTADO	1,69
DESAFIANTE	4,37
DESAFIO	7,2
DESAGRADADO	4,83
DESAJEITADO	3,67
DESAMPARADO	2,48
DESANIMADO	1,84
DESASTRE	1,84
DESAVENÇA	1,98
DESCONFORTO	2,31
DESCULPA	5,9
DESDENHOSO	3,01
DESEJO	7,78
DESERTOR	3,72
DESESPERADOR	1,92
DESINTERESSADO	3,46
DESLEAL	1,63
DESLIGADO	3,31
DESPEJAR	4,08
DESPERDÍCIO	3,45
DESPREOCUPADO	5,33
DESPREZAR	2,08
DESPREZO	1,71
DESTACADO	6,82
DESTROÇAR	2,53
DESTRUIÇÃO	2,13
DESTRUIR	2
DETALHE	6,36
DETESTAR	2,29
DEUS	8,11
DEVOTADO	5,8
DIABO	2,67
DIAMANTE	7,22
DIGNO	7,69
DINHEIRO	7,2
DIPLOMA	8,33
DIVERSÃO	8,31
DIVERTIDO	8,57
DIVERTIMENTO	8,27
DIVÓRCIO	2,97
DOCE	7,88
DOENÇA	1,77
DOENTE	1,69
DÓLAR	5,45
DOMINADOR	3,96
DOR	1,91
DOUTOR	4,84
DURO	4,41
EDIFÍCIO	5,76
EDUCAÇÃO	8,04
EGOÍSTA	2,04
ELEGANTE	7,38
ELEVADOR	5,37
EMPREGO	7,31
ENCARDIDO	3,08
ENCHARPE	5,59
ENCONTRO	7,67
ENFERMEIRA	5,33
ENFERRUJADO	3,41
ENFURECIDO	2,67
ENGANAÇÃO	1,76
ENJOATIVO	2,63
ENLAMEADO	3,57
ENTEDIADO	2,44
ENTERRO	1,65
ENTULHO	3,61
ENTUSIASMO	8,02
ENVERGONHADO	3,29
ERÓTICO	7,31
ERRO	2,43
ERUDITO	5,47
ESBANJAMENTO	3,47
ESCALDANTE	4,2

ESCÂNDALO	3,31
ESCONDER	4,08
ESCORBUTO	3,77
ESCORPIÃO	3,53
ESCRAVO	2,36
ESCRITOR	7,02
ESCRITÓRIO	5,7
ESCURO	4,49
ESFERA	5,43
ESFOMEADO	3,61
ESMAGADO	2,54
ESNOBE	1,96
ESPAÇO	6,88
ESPANTADO	4,47
ESPERANÇA	8,29
ESPERANÇOSO	7,55
ESPINGARDA	2,82
ESPINHO	2,98
ESPÍRITO	6,87
ESPOSA	5,8
ESPUMA	6,77
ESQUINA	4,67
ESTAGNADO	3,54
ESTÁTUA	4,88
ESTERCO	3,26
ESTÔMAGO	5,14
ESTRANGEIRO	6
ESTRANHO	4,58
ESTRELA	7,75
ESTRESSE	2,22
ESTUPENDO	3,88
ESTÚPIDO	2,21
ESTUPRO	4,94
EVENTO	7,03
EXAME	4,22
EXCELÊNCIA	6,24
EXCITAÇÃO	7,73
EXCURSÃO	7,82
EXECUÇÃO	3,92
EXERCÍCIO	6,68
EXÉRCITO	4,67
ÊXTASE	7,34
EXULTANTE	5,97
FACA	3,9
FÁCIL	7,31
FAIXA	4,67
FALCÃO	5,82
FALHA	2,03
FALIDO	2
FALSO	5,45
FAMA	6,73
FAMÍLIA	7,07
FAMINTO	2,94
FAMOSO	5,76
FANTASIA	7,06
FAROL	6,75
FAROLETE	5,2
FASCINAR	5,37
FASE	5,44
FATIGADO	5,1
FAVELA	1,64
FAVOR	7,62
FAVORITO	7,27
FAZENDA	4,32
FEBRE	2,2
FEDOR	5,58
FEITO	3,17
FELIZ	8,69
FENO	4,65
FERIADO	4,99
FÉRIAS	8,62
FERIDAS	4,29
FERIMENTO	2,54
FERRAMENTA	7,07
FERRO	4,91
FESTA	8,22
FESTIVO	8,09
FILHOTE	6,88
FIRMAMENTO	6,66
FIRME	6,97
FLÁCIDO	2,43
FLEXÍVEL	7,25
FLOR	8,11
FLORESCER	6,38
FOFOCA	2,44
FOGÃO	4,71
FOGO	4,55
FORÇA	4,62
FORTE	7,66
FORTUITO	6,61
FOTOGRAFIA	7,71
FRAGRÂNCIA	5,98

FRAUDE	1,6
FREIRA	4,89
FRÍGIDA	2,43
FRIO	4,45
FRUSTRADO	1,84
FUGA	3,63
FUNERAL	1,52
FUNGO	3,84
FURACÃO	2,72
GABINETE	5,35
GANGRENA	2,64
GARFO	6,12
GAROTOS	6,94
GARRAFA	6,49
GATINHO	7,06
GATO	5,75
GATO	3,47
GELADEIRA	6,86
GELÉIA	5,84
GELEIRA	4,33
GÊNERO	4,33
GENTIL	8,23
GERMES	4,69
GINASTA	6,4
GLAMOUR	5,26
GLÓRIA	8,16
GOLFISTA	4
GOLPE	2,2
GORDO	4,78
GOSTO	7,57
GRACINHA	7,15
GRADUADO	8,17
GRAMA	5,75
GRAMADO	7,75
GRAMPOS	6,54
GRANADA	2
GRITO	4,42
GROSSO	2,71
GUERRA	1,61
GUILHOTINA	5,56
GULA	3,88
HABILIDADE	6,91
HABITANTE	5,55
HÁBITO	4,13
HEMODIÁLISE	3,78
HEROÍNA	5,35
HIDRANTE	6,45
HIDROFOBIA	3,48
HISTÓRIA	3,54
HOMEM	6,85
HOMICIDA	5,13
HONESTO	8,6
HONRA	4,75
HORRÍVEL	1,98
HORROR	2,54
HOSPITAL	2,75
HOSTIL	5,6
HOTEL	7,11
HUMANITÁRIO	4,58
HUMBÚRGUER	6,91
HUMILDE	7,66
HUMILHAR	4,36
HUMOR	7,97
IATE	6,73
IDÉIA	8,14
IDENTIDADE	7,35
IDIOTA	2,28
ÍDOLO	5,99
IGNORÂNCIA	2,11
IGREJA	4,22
IMAGINAR	8,02
IMATURO	2,85
IMORAL	2,03
IMPLICAR	4,69
IMPOTENTE	2,08
IMPRESSIONADO	6,97
IMUNDÍCIE	1,86
INCENTIVO	4,76
INCOMODADO	2,45
INCOMODAR	2,97
INCUMBÊNCIA	5,02
INDIFERENTE	4,37
INDÚSTRIA	5,63
INFANTE	2,99
INFECÇÃO	1,77
INFELIZ	1,54
INFERIOR	2,49
INFERNO	3,81
INFIEL	2,08
INOCENTE	3,74
INSANO	3,11
INSEGURO	2,26

INSETO	3,54
INSOLENTE	5,67
INSOSSO	3,55
INSPIRADO	4,72
INSPIRAR	7,31
INSULTO	5,28
INTELECTO	6,8
INTELIGENTE	7,81
INTERCURSO	5,02
INTERESSE	4,74
ÍNTIMO	7,27
INTROMETER	2,4
INTROMETIDO	2,52
INTRUSO	2,29
INUNDAÇÃO	1,88
INÚTIL	4,81
INVASOR	2,2
INVESTIR	4,35
IRMÃO	7,78
IRRITAR	4,45
ITEM	5,08
JANELA	7,29
JANTA	7,47
JARDIM	5,85
JARRA	5,72
JIBÓIA	4,96
JOGO	6,34
JÓIA	7,76
JUSTIÇA	6,52
JUVENTUDE	4,51
KETCHUP	5,31
LADRÃO	2,9
LAGO	7,38
LAMA	2,58
LÂMPADA	6,51
LANTERNA	5,95
LÁPIS	6,24
LAR	8,23
LARVA	3,07
LEAL	8,27
LEÃO	5,89
LEITE	6,68
LENDA	6,36
LENTO	3,77
LEPRA	1,87
LÉSBICA	4,01
LETÁRGICO	3,93
LETRA	6,5
LIBERDADE	8,8
LÍDER	6,92
LIGA	5,88
LINDO	8,17
LIVRAR	6,26
LIVRE	8,37
LIVRO	7,18
LIXO	2,48
LODO	2,91
LOIRO	5,97
LOTERIA	7,03
LOUCO	3,87
LUCRO	7,64
LUSTRE	5,65
LUTA	5,66
LUTO	1,7
LUXO	6,82
LUXÚRIA	4,15
LUZ	7,96
MACHUCADO	2,38
MACIO	7,74
MÃE	8,75
MÁGICO	7,38
MAGOAR	1,88
MAL	1,58
MALÁRIA	2,12
MALCHEIROSO	1,8
MALÍCIA	4,47
MALUCO	4,31
MALVADO	2,09
MAMILO	5,51
MANEIRA	5,85
MANÍACO	1,95
MANSO	6,73
MANTEIGA	5,68
MÃO	6,88
MÁQUINA	5,61
MARAVILHA	6,8
MAREADO	4,78
MARICAS	4,23
MARTELO	5,04
MASSA	7,13
MASSACRE	1,31
MASTIGAR	6,62

MASTURBAR	5,6
MATADOR	1,49
MATERIAL	6,07
MEDO	2,13
MEL	7,28
MELHORAR	8,21
MELODIA	7,73
MEMÓRIA	6,36
MEMÓRIAS	6,96
MENINA	7,1
MENINO	7,36
MENOSPREZADO	2,03
MENSAGEIRO	6,05
MENTE	7,03
MENTIRA	1,35
MERCADO	6,59
MERETRIZ	3,82
MERGULHADOR	5,71
MÊS	6,11
MESA	6,18
METAL	5,15
MÉTODO	6,06
MILAGRE	7,95
MILIONÁRIO	6,55
MISÉRIA	1,21
MÍSTICO	6,6
MOBILIDADE	6,17
MODESTO	6,79
MOEDA	6,76
MOFO	2,48
MOINHO	5,45
MOLDE	5,21
MOMENTO	7,09
MONTANHA	7,13
MORAL	6,72
MÓRBIDO	2,07
MORTE	1,4
MORTO	2,01
MOSQUITO	2,09
MOTIM	3,77
MOTOR	5,09
MUCO	5,43
MULETA	2,81
MULHER	6,57
MUNDO	6,17
MUSCULAR	6,99
MUSEU	6,01
MÚSICA	5,18
MUTAÇÃO	4,53
MUTILAR	4,66
NADADOR	6,38
NAMORADA	7,59
NARCÓTICO	2,77
NASCIMENTO	8,31
NATAL	8,06
NATURAL	4,97
NATUREZA	8,58
NÁUSEA	3,95
NAVALHA	2,64
NAVIO	6,55
NECROTÉRIO	2,1
NÉCTAR	4,66
NEGLIGÊNCIA	2,4
NERVOSO	4,21
NEURÓTICO	2,49
NEVE	6,93
NÓ	3,9
NOIVA	6,84
NOME	7,5
NOTÍCIA	6,49
NOVO	7,96
NU	7,14
NUBLADO	3,91
NUTRIR	6,42
NUVEM	6,67
OBEDECER	4,55
OBESIDADE	2,14
OBSCENO	3,89
OBSESSÃO	3
OBSTRUIR	2,64
OCEANO	7,77
ÓDIO	2,86
OFENDER	1,78
OFUSCAR	4,51
ÔNIBUS	4,18
ONIPOTENTE	5,8
OPÇÃO	6,27
OPINIÃO	6,78
ORGASMO	7,92
ORGULHO	6,38
ORGULHOSO	5,63
ORQUESTRA	7,13

OTIMISMO	8,41
OURO	6,99
OUSADO	6,57
OUTONO	5,95
OVO	6,12
PACIENTE	5,96
PADRE	5,29
PAI	8,08
PAÍS	5,8
PAIXÃO	6,34
PALÁCIO	6,28
PANFLETO	6,28
PÂNICO	2,14
PANQUECA	6,88
PÂNTANO	4,15
PAPEL	7,29
PAQUERAR	7,99
PARAÍSO	7,07
PARALISIA	1,86
PARTE	6,43
PASSAGEM	5
PÁSSARO	7,11
PATENTE	4,77
PATRIOTA	6,21
PAZ	8,64
PAZINHA	4,41
PÉ	5,99
PECADO	3,29
PECAMINOSO	3,43
PEÇONHA	4,67
PEDINTE	3,23
PEITO	6,22
PEIXE	6,45
PELADO	4,72
PÊLO	4,29
PENALIDADE	5,43
PENHASCO	3,43
PÊNIS	6,85
PENITENTE	4,06
PENSAMENTO	4,88
PENSATIVO	5,99
PERDEDOR	4,22
PERDIDO	2,93
PERFEIÇÃO	6,85
PERFUME	8,2
PERIGO	2,56
PERSEGUIR	3,46
PERTURBADO	2,49
PERTURBAR	2,32
PERVERTIDO	3,35
PESADELO	2,11
PESAR	3,41
PESSOA	7,85
PESTE	2,17
PIADA	8,14
PICADA	2,58
PIEDADE	5,82
PINTAR	7,26
PIOLHO	2,08
PISCAR	6,38
PISTOLA	2,18
PIZZA	8,35
PLANÍCIE	6,24
PLANO	7,11
PLANTA	7,83
POBREZA	1,41
PODER	5,53
PODEROSO	6,09
PODRE	1,81
POENTE	6,75
POESIA	7,66
POLUIR	1,51
POMBA	6,04
PORÃO	3,86
PORCO	4,29
PORTA	5,92
PÔSTER	6,05
POVO	6,22
PRADARIA	5,77
PRAIA	8,29
PRANTO	2,52
PRAZER	8,65
PREGUIÇOSO	3,66
PREJUDICADO	2,13
PREJUDICAR	1,68
PRESENTE	8,33
PRESSÃO	3,16
PRESTÍGIO	7,75
PRETO	5,26
PRIMAVERA	8,33
PRIMO	6,63
PRISÃO	1,68

PRIVAÇÃO	2,86
PRIVACIDADE	7,99
PROBLEMA	2,17
PROCESSO	4,31
PROEMINENTE	5,43
PROFESSOR	6,74
PROGRESSO	7,9
PROMOÇÃO	7,89
PRÓSPERO	7,78
PROSTITUTA	2,58
PROTEGIDO	7,26
PROVA	4,21
PULGA	2,68
PULVERIZADOR	4,72
PUNIÇÃO	2,89
PUS	2,74
PÚTRIDO	3,25
QUADRADO	5,18
QUADRO	5,87
QUALIDADE	8,29
QUEBRADO	2,57
QUEDA	2,6
QUEIMADURA	1,94
QUEIXO	5,79
QUERIDO	8,36
QUEROSENE	4,04
QUIETO	5,23
QUIMIOTERAPIA	1,97
RÃ	4,01
RADIADOR	4,75
RADIANTE	8,08
RÁDIO	7,84
RAINHA	6,27
RAIVA	1,97
RALÉ	2,94
RANÇOSO	2,49
RÁPIDO	6,32
RATO	3,19
RAZÃO	6,49
REALIZAÇÃO	8,51
RECEOSO	3,89
RECOMPENSA	7,59
RECREIO	7,65
REFÉM	1,84
REFRESCO	7,86
REI	5,73
REJEITADO	1,61
RELÂMPAGO	3,77
RELAXADO	6,36
RELÓGIO	5,73
REMÉDIO	3,93
RÉPTIL	4,47
REPUGNADO	2,67
REPULSIVO	3,1
RESERVADO	5,75
RESPEITO	8,43
RESPEITOSO	7,52
RESPOSTA	6,96
RESSENTIDO	2,83
RESTAURANTE	7,7
REUNIÃO	5,06
REVERENTE	5,41
REVISTA	6,9
REVOLTA	2,44
REVÓLVER	1,94
RICO	6,49
RIDÍCULO	2,61
RIFLE	2,36
RÍGIDO	4,56
RIO	6,94
RIQUEZAS	6,75
RISADA	8,59
ROCHA	5,49
RODOVIA	5,46
ROMÂNTICO	8,34
ROSTO	7,53
ROUPA	7,41
RUA	6,19
RUDE	2,96
RUIDOSO	3,97
SÁBIO	7,92
SABOROSO	8,24
SAFIRA	6
SALADA	6,88
SALVADOR	7,67
SALVAR	8,1
SANGRENTO	2,29
SANTO	6,65
SARAMPO	2,49
SATISFEITO	8,09
SAUDAR	8
SAÚDE	8,39

SECADOR	5,61	TEMPERAMENTAL	4,37
SEDA	6,38	TEMPESTADE	3,49
SEGURO	7,06	TEMPO	5,66
SENTIMENTO	7,41	TÊNIS	6,49
SÉRIO	5,39	TENSO	2,52
SEVERO	3,29	TEORIA	5,65
SEXO	7,93	TERMÔMETRO	4,49
SEXY	7,66	TERRA	7,76
SÍFILIS	1,93	TERRÍVEL	2,12
SOBRECARREGADO	2,58	TERRORISTA	1,45
SOBRESSALTO	4,76	TESOURA	4,62
SOCIAL	6,74	TESOURO	7,6
SOCIEDADE	5,53	TÍMIDO	4,12
SOFRIMENTO	1,68	TINTA	6,18
SOL	8,28	TIO	6,9
SOLENE	5,29	TOBOGÃ	4,88
SOLIDÃO	1,58	TOLO	2,71
SOLITÁRIO	2,47	TORNADO	2,41
SOMBRA	6,19	TORNOZELO	5,49
SOMBRINHA	5,09	TORRE	5,51
SONHO	7,59	TORTA	7,71
SONO	6,1	TORTURA	1,32
SÓRDIDO	2,9	TÓXICO	1,8
SORRISO	8,5	TRAGÉDIA	1,38
SORTUDO	7,46	TRAIDOR	1,29
SOZINHO	2,47	TRAIR	1,4
SUAVE	7,79	TRANQUILAMENTE	7,89
SUBJUGADO	2,95	TRANQÜILO	8,07
SUCESSO	8,25	TRATAR	6,55
SUFOCAR	2,05	TRAUMA	1,86
SUICÍDIO	1,35	TRAVESSEIRO	8,19
SUJEIRA	1,9	TRAVESSURA	6,56
SUJO	2,43	TREVAS	1,74
SURPRESO	6,6	TRISTE	1,73
SURRA	1,94	TRIUNFANTE	8,14
SUSPEITO	3,14	TRIUNFO	7,93
TABACO	2,04	TROFÉU	8,04
TALENTO	8,31	TROMPETE	5,84
TANQUE	4,64	TRONCO	5,44
TAPA	2,22	TUBARÃO	3,4
TÁXI	5,1	TUMOR	1,44
TECIDO	5,97	TÚMULO	2,06
TÉDIO	2,69	ÚLCERA	1,54
TELEVISÃO	6,69	ULTRAJE	3,71
TEMIDO	4,23	UNIDADE	5,76
TEMÍVEL	2,95	UNTENSÍLIO	6,63

URINA	5,38
ÚTIL	8,02
VACA	6,29
VAGÃO	4,97
VAGINA	6,8
VAIDADE	6,18
VAMPIRO	3,55
VÂNDALO	1,76
VANTAGEM	5,71
VARA	3,24
VARÍOLA	1,95
VEÍCULO	7,41
VELEIRO	6,57
VELOZ	6,23
VENCER	8,55
VENENO	1,54
VENTILADOR	6,73
VERDADE	8,56
VERDE	7,39
VERMELHO	5,84
VESPA	3,63
VESTIBULAR	4,93
VESTIDO	6,69
VESTUÁRIO	6,8
VIAGEM	8,55
VIBRAÇÃO	6,96
VICIADO	1,46
VÍCIO	2,16
VIDA	8,63
VIDRO	4,96
VIGIAR	4,13
VIGOROSO	6,85
VILA	4,94
VINHO	7,19
VIOLENTO	1,46
VIOLINO	6,42
VIRGEM	5,84
VIRTUDE	8,02
VISÃO	7,33
VÍTIMA	2,26
VITÓRIA	8,74
VÍVIDO	7,02
VIVO	8,34
VÔMITO	2,31
VULCÃO	3,74