UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

BACHARELADO EM ENGENHARIA DE SOFTWARE

CARLOS MATHEUS DA SILVA QUIXADÁ

CLASSIFICAÇÃO DE TEXTO: MULTINOMIAL X BERNOULLI UTILIZANDO REVIEWS DE E-COMMERCES

QUIXADÁ

2019

CARLOS MATHEUS DA SILVA QUIXADÁ

CLASSIFICAÇÃO DE TEXTO: MULTINOMIAL X BERNOULLI UTILIZANDO REVIEWS DE E-COMMERCES

Monografia apresentada no curso de Engenharia de Software da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Engenharia de Software. Área de concentração: Computação.

Orientador: Dr. Regis Pires Magalhães

QUIXADÁ

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

Q57c Quixadá, Carlos Matheus da Silva.
Classificação de texto : multinomial x bernoulli utilizando reviews de E-Commerces / Carlos Matheus da Silva Quixadá. – 2019.
36 f.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2019.
Orientação: Prof. Dr. Regis Pires Magalhães.

1. Classificação-textos. 2. Teoria bayesiana de decisão estatística. 3. Comércio eletrônico. I. Título.
CDD 005.1

CARLOS MATHEUS DA SILVA QUIXADÁ

CLASSIFICAÇÃO DE TEXTO: MULTINOMIAL X BERNOULLI UTILIZANDO REVIEWS DE E-COMMERCES

Monografia apresentada no curso de Engenharia de Software da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Engenharia de Software. Área de concentração: Computação.

Aprovado em: _____/_____/_____.

BANCA EXAMINADORA

Dr. Regis Pires Magalhães (Orientador)
Universidade Federal do Ceará – UFC

Dra. Ticiana Linhares Coelho da Silva
Universidade Federal do Ceará - UFC

Dr. Marcos Antonio de Oliveira
Universidade Federal do Ceará - UFC

À Deus.

Aos meus familiares e amigos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por sempre ter me dando forças e coragem para continuar nos momentos difíceis. À todos os meus familiares em especial a minha mãe, Maria Arineuda da Silva Quixadá, que desde de sempre me incentivou a seguir o caminho acadêmico e sempre foi um exemplo de força. Ao meu pai, Luiz Uramar Quixadá Joca, que foi quem me fez ter os primeiros contatos com tecnologia ainda na minha infância e aos meus irmãos, que sempre estiveram do meu lado.

Aos meus amigos Bruno Barreto, Jacques Nier, Amarildo Barros, que dividiram o icônico "AP 301"comigo durante todo o período de graduação e hoje os considero meus irmãos. Aos amigos de turma Lucas Sales, Leonardo Gomes e tantos outros que me auxiliaram nas disciplinas, ao Zarathon Maia, que durante o estágio foi um mentor fantástico. Agradeço a todos meus amigos em geral que não pude colocar os nomes.

Ao Dr. Régis Magalhães, por toda atenção e comprometimento com este trabalho e orientações. À Dra. Ticiana Linhares que me orientou durante TCC1 e boa parte do TCC2, pelas conversas, incentivos, paciência ao longo da nossa convivência e por ser uma fonte de inspiração para minha carreira. Ao Dr. Marcos de Oliveira, que compôs a banca deste trabalho e contribuiu tão significativamente para a melhoria deste trabalho. À todo o Campus Quixadá por ter me dado a oportunidade, aos docentes e os demais funcionários que compõe a Universidade Federal do Ceará - Campus Quixadá que contribuíram direta, ou indiretamente na minha vida acadêmica

“A melhor maneira de prever o futuro é inventá-lo.”

(Alan Kay)

RESUMO

Com a popularização da internet muitos serviços que antes era possível ter acesso apenas fisicamente passou a oferecer sua versão digital, um desses serviços é o de *e-commerce* que é a representação da loja virtualmente. Como essa relação entre cliente e prestador de serviço é apenas virtual, alguns clientes passaram a enviar *reviews* para as empresas de modo a explicitar sua satisfação ou insatisfação com o serviço. Dessa forma algumas empresas como ReclameAqui e Trustvox passaram a auxiliar *e-commerces* e clientes a facilitarem a comunicação e consequentemente aumentar a qualidade dos produtos e serviços oferecidos. O trabalho a seguir tem como objetivo desenvolver um modelo preditivo capaz de classificar, dado um *review*, qual categoria ele se refere, as *labels* utilizadas são: informação, característica, preço e entrega. Para isso é realizada a etapa de pré-processamento e limpeza na base utilizada que foi fornecida pela empresa Trustvox, em seguida é realizado o treinamento e teste dos modelos Multinomial e Bernoulli e por fim apresentado o resultado da comparação de performance dos modelos para cada *label*.

Palavras-chave: Classificação-textos. Teoria bayesiana de decisão estatística. Comércio eletrônico.

ABSTRACT

With the popularization of the internet many services that previously could only be accessed physically came to offer its digital version, one of these services is the e-commerce which is the representation of the store virtually. As this relationship between customer and service provider is only virtual, some customers have sent reviews to companies in order to explain their satisfaction or dissatisfaction with the service. In this way, some companies like ReclameAqui and Trustvox started to assist e-commerces and clients to facilitate communication and consequently to increase the quality of the products and services offered. The following work aims to develop a predictive model capable of classifying, given a review, which category it refers to, the labels used are: information, characteristic, price and delivery. For this, the pre-processing and cleaning step in the used base was performed, which was provided by the company Trustvox, then the training and testing of the Multinomial and Bernoulli models were performed and finally the result of the comparison of the performance of the models for each label .

Keywords: Classification-texts. Bayesian theory of statistical decision. E-commerce.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KDD	Knowledge Discovery in Databases

SUMÁRIO

1 INTRODUÇÃO

Com a popularização da internet, muitos serviços que antes só tínhamos acesso fisicamente, passaram a ter suas versões digitais, visando oferecer maior comodidade para seus clientes. Como é o caso de aplicativos de *delivery* de comida, aplicativos para solicitar táxis, compras de ingressos, e também lojas digitais que são conhecidas como *e-commerces*.

Mesmo o Brasil passando por um momento difícil economicamente, o *e-commerce* conseguiu se destacar do mercado físico. Segundo ABComm (Associação Brasileira de Comércio Eletrônico) os *e-commerces* tiveram um crescimento de 12% em 2017 em relação ao ano anterior e faturamento de 59,9 bilhões de reais. De acordo com o 39º Webshoppers a projeção para 2019 é que aumente em 15% e tenha um faturamento de 61,2 bilhões de reais.[1] Além das projeções para 2019 serem de aumento em 15% e faturamento de mais de 60 bilhões de reais, com esses números positivos muitas empresas já consolidadas disponibilizaram suas versões digitais [2].

Mas só esses números positivos não são suficientes para indicar que os serviços de *e-commerce* estão agradando aos clientes. Com isso algumas empresas resolveram auxiliar os clientes a pesquisarem melhor sobre a empresa onde eles pretendem fazer uma compra, como é o caso do Reclame Aqui[3], que baseada nas reclamações dos usuários e no atendimento das empresas que sofreram reclamação gera uma nota média da reputação da empresa, onde um futuro cliente pode consultar e decidir se irá realizar compras ou utilizar serviços daquela empresa. Um outro exemplo, de empresas que auxilia os *e-commerces* a venderem melhor baseando-se na experiência de compra de seus clientes, é a Trustvox[4] que oferece uma plataforma para *e-commerces* poderem ter acesso aos *reviews* fornecidos por seus clientes podendo ver histórico da sua nota de satisfação fornecida pelos clientes entre outros *insights*.

Este trabalho tem como objetivo principal desenvolver um modelo preditivo capaz de classificar, dado um *review*, qual categoria ele se refere, por exemplo: informação, característica, preço e entrega. Esse modelo preditivo é baseado nos dados fornecido pela empresa Trustvox. Para isso será realizado treinamento e teste dos modelos Multinomial e Bernoulli e em seguida, a comparação entre os resultados dos testes entre os modelos com o objetivo de identificar qual modelo apresenta melhor performance mediante as métricas para as categorias propostas nesse trabalho.

[1] https://www.ecommercebrasil.com.br/noticias/e-commerce-crescer-15-faturar-61-bi-ebit-nielsen/
[2] https://abcomm.org/noticias/e-commerce-brasileiro-espera-faturar-r-599-bilhoes-em-2017/
[3] https://www.reclameaqui.com.br/
[4] https://site.trustvox.com.br/

Este estudo tornará possível que se verifique automaticamente sobre o que um cliente pode estar comentando, assim podendo classificar um maior número de comentários em um menor espaço de tempo. Além de verificar quais categorias que mais aparecem nos comentários, fornecer dados para gerar soluções especificas para cada categoria, e assim aumentar o nível de satisfação dos clientes daquele *e-commerce*.

Este trabalho está organizado da seguinte forma: O Capítulo 2 discute os conceitos chave envolvidos. No Capítulo 3, os principais trabalhos relacionados a este trabalho são discutidos. O Capítulo 4 discute a metodologia a ser aplicada neste trabalho. O Capítulo 5 apresenta os resultados encontrados. O Capítulo 6 apresenta as considerações finais sobre os resultados encontrado, e sugestões para a continuidade desse trabalho.

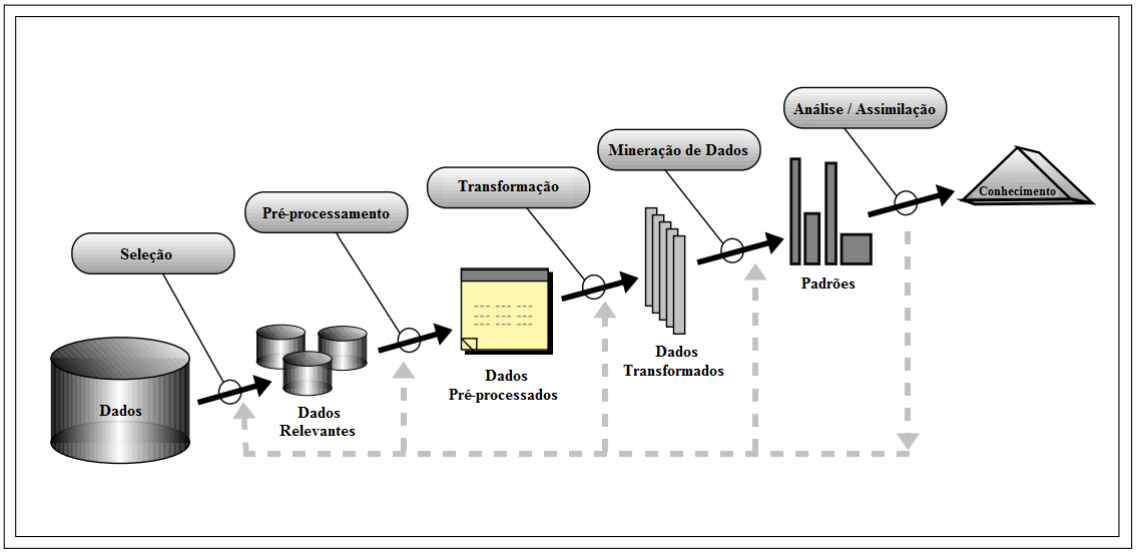
2 FUNDAMENTAÇÃO TEÓRICA

Nesse Capítulo são apresentados os conceitos de mineração de dados e mineração de texto, bem como o teorema de classificação que é utilizado nesse trabalho: *Naive Bayes*. Além disso, é abordado o processo de descoberta de conhecimento em banco de dados (*Knowledge Discovery in Databases* - KDD), que será o processo utilizado nesse trabalho.

2.1 Descoberta de Conhecimento em Banco de Dados

Fayyd, Shapiro e Smyth (1996) definem o processo de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases* – KDD) como um processo de extração de informações em dados armazenados em um banco de dados, onde essas informações são previamente desconhecidas. O KDD é dividido em cinco etapas que serão apresentadas nas seções a seguir. A Figura 1 apresenta uma visão geral das etapas do processo KDD.

Figura 1 – Uma visão geral das etapas do processo KDD



Fonte – Fayyd, Shapiro e Smyth (1996)

2.1.1 Seleção

A primeira etapa do processo KDD é a seleção do conjunto de dados que fazem parte da análise do estudo realizado. Essa etapa possui um peso significativo sobre a qualidade dos resultados finais.

Uma vez que os dados podem vir de várias fontes diferentes e podem apresentar

vários formatos diferentes, a etapa de seleção torna-se bastante complexa. Em muitos casos, é necessário desenvolver *scripts* específicos para realizar um pré-processamento e muitas vezes limpezas dos dados que serão utilizados.

2.1.2 Pré-processamento e Limpeza

Na etapa de pré-processamento e limpeza dos dados, são utilizados métodos de redução ou transformação para diminuir quantidade de variáveis envolvidas no processo de descoberta de conhecimento, com o intuito de melhorar o desempenho do algoritmo de análise. Nessa etapa também são realizadas atividades que eliminam dados que possam ser redundantes, inconsistentes ou dados desnecessários para o estudo. Por ser uma etapa que tem como objetivo filtrar os dados que são utilizados, torna-se uma parte crucial dentro do processo KDD.

Posteriormente à execução do pré-processamento e a limpeza dos dados, esses dados são armazenados adequadamente para que possam ser utilizados da melhor forma para o estudo. Esse processo de armazenamento é realizado na etapa de transformação dos dados que será apresentada a seguir.

2.1.3 Transformação dos Dados

Na etapa de transformação dos dados é realizado o armazenamento dos dados em um repositório, após passarem pelas etapas de seleção, pré-processamento e limpeza. Esse armazenamento é necessário para que os algoritmos que utilizam esses dados como entrada, possam ser aplicados.

Além do armazenamento, nessa etapa, é possível obter dados faltantes através da transformação ou combinação de outros dados. Por exemplo, pode-se obter a idade de um indivíduo a partir da sua data de nascimento. Essa informação é denominada dado derivado, porque essa informação não estava na seleção e foi gerada a partir da execução de um cálculo utilizando um dado existente na seleção.

Após o armazenamento dos dados que serão utilizados e a obtenção de dados faltantes, é possível aplicar a técnica de mineração de dados para realizar a descoberta de conhecimento utilizando essa base de dados já tratada.

2.1.4 Mineração de Dados

Na etapa de mineração de dados são executadas técnicas especificas de acordo com a finalidade da descoberta de conhecimento proposta pelo estudo.

Berry e Linoff (1997) definem mineração de dados como a análise e exploração, que pode ser realizada de forma automática ou semi-automática, de grandes bases de dados com a finalidade de realizar descoberta de padrões e regras.

Segundo Fayyd, Shapiro e Smyth (1996), a mineração de dados pode ser utilizada em várias tarefas como: associação, que tem como finalidade determinar quais fatos ou objetos tendem a ocorrerem juntos em um mesmo evento; classificação, que tem como finalidade a construção de um modelo que possa ser aplicado a dados não classificados visando categorizar os objetos em classes; regressão, que tem como finalidade definir um provável valor para uma ou mais varáveis; segmentação, que tem como finalidade realizar a divisão de uma população em subgrupos o mais heterogêneo possível entre si; sumarização, que tem como finalidade encontrar uma descrição compacta para um subconjunto de dados.

Após a execução da etapa de mineração de dados são geradas informações que deverão ser interpretadas e avaliadas para se ter a validação da relevância das informações geradas.

2.1.5 Interpretação e Avaliação

Na etapa de interpretação e avaliação, as informações obtidas através da mineração de dados serão interpretadas e avaliadas. Caso o resultado encontrado não seja satisfatório, de acordo com os critérios estabelecidos no estudo, o processo pode retornar a qualquer uma das etapas anteriores. Algumas das ações mais comuns, caso o resultado não seja satisfatório, são modificar o conjunto de dados inicial ou trocar o algoritmo de mineração de dados.

O processo KDD é utilizado nesse trabalho por oferecer a divisão das etapas necessárias para se realizar a descoberta de conhecimento, assim auxiliando a construção do processo metodológico do trabalho.

2.2 Mineração de Texto

Segundo Feldman e Sanger (2007), a mineração de texto, ou mineração de dados textuais, é uma área da mineração de dados. Pois, como na mineração de dados, a mineração de

texto tem como objetivo a extração de informações úteis utilizando técnicas de identificação e exploração de padrões.

Hotho, Nürnberger e Paaß (2005) definem mineração de textos como a identificação de informações significativas a partir de uma grande quantidade de textos escritos em linguagem natural. Para realizar essa identificação de informações úteis, é possível utilizar a aplicação de algoritmos e métodos de aprendizagem de máquina, o uso de estatísticas aplicadas aos textos, assim como na mineração de dados convencional. Para Tan, Steinbach e Kumar (2009) a mineração de texto é o processo de extração de padrões ou conhecimento de documentos de texto não estruturados. Por ser feita uma extração em dados não estruturados, o processo de mineração de texto torna-se uma tarefa nada trivial. Essa ideia é reforçada por Fayyd, Shapiro e Smyth (1996) que afirmam que extrair informações de texto é mais difícil, na maioria das vezes, do que se comparados aos dados organizados em banco de dados. Essa dificuldade ocorre, pois em banco de dados estruturados, os dados geralmente estão organizados e estruturados em tabelas, e as mesmas possuem algumas relações, o que nos garante uma maior coesão nas informações que em textos não estruturados.

A mineração de texto é utilizada nesse trabalho para identificar padrões nas *reviews*, e assim poder treinar algoritmos para realizar uma classificação automática. Ela é utilizada na etapa de mineração de dados conforme o processo KDD. Na próxima seção é apresentado a técnica de mineração de dados Naive Bayes que será utilizado nesse trabalho.

2.3 Teorema de Bayes e Classificador Naive Bayes

Naive Bayes é um algoritmo probabilístico de classificação relativamente simples. O algoritmo se baseia no teorema de Bayes, esse teorema descreve a probabilidade de um evento, baseado em um conhecimento anterior que pode estar relacionado a um determinado evento (MORETTIN; BUSSAB, 2017). Uma de suas primeiras descrições é encontrada em (DUDA; HART; STORK, 1973).

Na área de inteligência artificial este teorema é utilizado na técnica de classificação Naive Bayes. Oguri, LUIZ e RENTERIA (2006) afirmam que os classificadores Naive Bayes, utilizados para classificar dados baseados em um modelo computacional, são um dos mais utilizados no mundo para o aprendizado de máquinas. O classificador é tido como Naive (ingênuo) pois assume que a informação de um determinado evento, não serve de informação para outro evento, dentro do mesmo contexto. Segundo Domingos e Pazzani (1996) as principais

razões para o classificador Naive Bayes ser ter utilizado é por sua facilidade e rapidez para implementar e apresentar, relativamente, uma boa eficácia.

A seguir é apresentado um exemplo simples da utilização do teorema de Bayes para classificar se a palavra "produto"é positiva ou negativa. Esse exemplo é baseado no exemplo apresentado por Tan, Steinbach e Kumar (2009):

A Tabela 1 apresenta base de dados utilizada para o exemplo da utilização do teorema de Bayes.

Tabela 1 – Base de Dados Utilizada no Exemplo

Palavra	Label
rasgado	negativo
produto	negativo
ruim	negativo
produto	positivo
produto	positivo
rasgado	negativo
loja	positivo
camisa	positivo
camisa	negativo
camisa	negativo
produto	negativo
loja	positivo
produto	positivo

Fonte – Produzido pelo Autor

1. Primeiro é necessário identificar as frequências das palavras e suas *labels* baseado-se na base de dados:

Tabela 2 – Frequência das Palavras na Base de Dados

Palavra	Positivo	Negativo
rasgado	0	2
produto	3	2
ruim	0	1
camisa	1	2
loja	2	0

Fonte – Produzido pelo Autor

2. Após identificar a frequência de cada palavra, será calculado a probabilidade de

"produto"ser positivo e a probabilidade de ser negativo. O Naive Bayes utiliza a seguinte formula para calcular a probabilidade de "produto"ser positivo:

$$P(positivo|produto) = \frac{P(produto|positivo) * P(positivo)}{P(produto)}$$

Onde os termos da formula são: P(positivo|produto), representa a probabilidade de "produto"ser positivo; P(produto|positivo), representa a probabilidade de uma palavra positiva ser "produto"; P(positivo), representa a probabilidade de alguma palavra ser positivo; e P(produto), representa a probabilidade de uma palavra ser "produto". Logo tem-se que:

$$P(produto|positivo) = \frac{\text{Total de vezes que "produto"é positivo}}{\text{Total de ocorrências positivas}}$$

$$P(produto|positivo) = \frac{3}{6} = 0.5$$

Agora será calculado o valor de P(positivo). Logo tem-se que:

$$P(positivo) = \frac{\text{Total de palavras positivas}}{\text{Total de palavras}}$$

$$P(positivo) = \frac{6}{13} = 0.46$$

Após calcular o valor para P(positivo), será calculado o valor para P(produto). Logo tem-se que:

$$P(produto) = \frac{\text{Total de ocorrências de produto}}{\text{Total de palavras}}$$

$$P(produto) = \frac{5}{13} = 0.38$$

Agora que foi calculado os valores necessários será possível calcular a probabilidade de produto ser positivo. Logo tem-se que:

$$P(positivo|produto) = \frac{0.5 * 0.46}{0.38} = \frac{0.23}{0.38} = 0.60$$

Assim tem-se que a probabilidade de produto ser positivo é de 0.60, mas para ter certeza é necessário saber a probabilidade de produto ser negativo também para que seja comparado os valores. A fórmula para calcular a probabilidade de "produto"ser negativo é:

$$P(negativo|produto) = \frac{P(produto|negativo) * P(negativo)}{P(produto)}$$

Onde os termos da fórmula são: P(negativo|produto), representa a probabilidade de "produto"ser negativo; P(produto|negativo), representa a probabilidade de uma palavra negativa ser "produto"; P(negativo), representa a probabilidade de alguma palavra ser negativo; e P(produto), representa a probabilidade de uma palavra ser "produto". Como os cálculos para se obter os valores de cada termo da formula são semelhantes com o que já foi mostrado para calcular a probabilidade de ser positivo essa etapa será omitida e será feito de imediato o cálculo da probabilidade de "produto"ser negativo. Logo tem-se que:

$$P(negativo|produto) = \frac{0.28 * 0.53}{0.38} = 0.39$$

Após calcular a probabilidade de "produto"ser positivo, com valor de 0.60, e a probabilidade de "produto"ser negativo, com valor de 0.39, é possível inferir que a probabilidade da palavra "produto"ser positiva é maior que a probabilidade dela ser negativa. Logo a palavra "produto"é positivo nessa base de dados e esse resultado foi obtido utilizando o teorema de Bayes que é o mesmo utilizado nos classificadores Naive Bayes.

Nesse trabalho será utilizado os modelos: Multinomial e Bernoulli. Segundo (MCCALLUM; NIGAM et al., 1998) o modelo Multinomial assume que cada documento terá sua representação a partir da frequência que cada termo ocorreu no documento. Já o modelo Bernoulli utiliza recursos de ocorrências de termos. A razão para utilizar os modelos apresentados é por cada um ter uma abordagem diferente. Onde o Multinomial trabalha com a frequência e o Bernoulli com a ocorrência.

O classificador Naive Bayes é utilizado nesse trabalho por ser umas das técnicas mais fáceis e simples de implementar e ser bastante utilizada na classificação de textos como já dito anteriormente.

3 TRABALHOS RELACIONADOS

Nessa seção serão brevemente apresentados alguns trabalhos que serviram de embasamento para a construção deste trabalho.

Evangelista e Padilha (2013) apresenta uma ferramenta *Web* para classificar publicações em redes sociais como positivas, negativas e neutras. Para desenvolver essa ferramenta, foi utilizado o recurso léxico *SentiWordNet* e o modelo Multinomial do algoritmo Naive Bayes. A base utilizada para treinamento da ferramenta foi de *review* de usuários do site Buscapé e testaram com publicações de usuários sobre as empresas nas redes sociais Twitter e Facebook. Eles observaram que as classificações utilizando os *posts* da rede social Twitter tiveram melhor resultado utilizando o algoritmo Naive Bayes. Nos experimentos, a sua taxa de acerto variaram de 71% a 95% e constataram que há uma tendência muito maior das pessoas postarem reclamações ao invés de elogios nas redes.

Martinazzo (2010) apresenta um experimento de um sistema que classifica emoções como: alegria, raiva, tristeza, desgosto, medo; surpresa, de manchetes de notícias em páginas de jornais. O método utilizado foi baseado em *Latent Semantic Analysis*. Após realizado o experimento, os autores observaram que é possível a identificação de emoções em textos de forma automatizada, através de um algoritmo de mineração de textos. Embora a eficácia comprovada do sistema ainda seja inferior a 80%, percebe-se que a média atingida através de outros estudos realizados na área é semelhante a aqui obtida, o que comprova a eficácia do método escolhido.

Santos (2016) apresenta um protótipo de uma aplicação *Web* que classifica comentários em positivo, negativo ou neutro. Para analisar o sentimento de cada comentário foi utilizado o Sentilex-PT. A base de dados foi construída extraindo as palavras do arquivo disponibilizado pelo Sentilex-PT, onde dentre os atributos de cada palavra, apenas a polaridade da palavra foi considerada. O experimento contou com uma base de 328 comentários aleatórios coletados da base existente. Os 328 comentários foram analisados manualmente por três especialistas da área de linguística, graduandos em Letras Português, definindo cada comentário como positivo, negativo ou neutro. Apos a conclusão da analise manual, os comentários foram submetidos ao protótipo da aplicação Web. Observaram que houve aumento na detecção dos comentários positivos e a queda na detecção dos comentários negativos, segundo os autores isso ocorreu devido ao modo com que o protótipo define a polaridade do comentário, somando as polaridades das palavras opinativas referentes a cada caracterıstica encontrada.

A tabela 3 apresenta a comparação entre o trabalho proposto e os trabalhos relacionados apontando diferenças e semelhanças entre eles. As linhas da tabela são as principais características dos trabalhos, como: "Finalidade da Classificação", refere-se ao objetivo que o autor deseja alcançar com a classificação de texto; "Base Utilizada", refere-se ao conteúdo da base utilizada na classificação de texto; "Ferramenta Utilizada", refere-se as ferramentas utilizadas para realizar a classificação de texto; e "Quantidade de Labels", refere-se a quantas *labels* foram utilizadas na classificação de texto no trabalho. Já as colunas são os trabalhos relacionados com esse e o trabalho proposto.

Tabela 3 – Trabalhos Relacionados

	Evangelista e Padilha (2013)	Martinazzo (2010)	Santos (2016)	Trabalho Proposto
Finalidade da Classificação	Análise de Sentimento	Análise de Sentimento	Análise de Sentimento	Categorização
Base Utilizada	Reviews sobre Buscapé das Redes Sociais	Manchetes de Notícia	Comentários do Sentilex-Pt	Reviews de E-Commerce
Ferramenta Utilizada	Léxico(SentiWordNet) /Multinomial	Latent Semantic Analysis	Sentilex-Pt	Multinomial/Bernoulli
Quantidade de Labels	3	6	3	4

Fonte – Produzido pelo Autor

4 PROCEDIMENTOS METODOLÓGICOS

Nessa seção serão apresentadas as etapas dos procedimentos metodológicos que tem como base o processo de KDD.

4.1 Seleção e Coleta de Dados

A primeira etapa deste trabalho consiste na realização da coleta dos *reviews*. Nessa fase, os dados coletados são uma amostra da base de dados fornecida pela empresa Trustvox que forneceu os dados para a realização deste trabalho. Esses dados são fornecidos em um arquivo CSV para que se possa facilitar a manipulação. O arquivo possui 19 atributos, mas para esse trabalho são consideradas apenas as colunas "id"e "*review*"pois os outros atributos são de interesse exclusivo para aplicações da empresa Trustvox e não influenciará nos objetivos desse trabalho. Id e *review* são respectivamente: o id que representa uma *review*; e a descrição do comentário (que nesse trabalho será chamado de *review*) enviado por um cliente. Para esse trabalho são utilizados *reviews* que são referentes as seguites *labels*: Informação, que são *reviews* que falam sobre características que não estão relacionadas ao produto diretamente; Característica, são *reviews* referentes as características diretamente ligadas ao produto; Entrega, são *reviews* referentes ao serviço de entrega do produto; e Preço, são *reviews* referentes ao preço do produto. A Tabela 4 apresenta um exemplo para cada *label* que será utilizada nesse trabalho:

Tabela 4 – Exemplo de *Reviews*

Label	Exemplo
Informação	A cor na foto do produto é diferente do original
Característica	Adorei o tamanho da calça
Entrega	Produto demorou muito para chegar
Preço	Está muito caro

Fonte – Produzido pelo Autor

Antes desses dados serem classificados, é necessário que sejam submetidos a um pré-processamento e limpeza dos dados para assegurar que apenas dados relevantes são utilizados na etapa de classificação.

4.2 Pré-processamento e Limpeza

Na etapa de pré-processamento e limpeza, os dados que foram coletados passam por um pré-processamento e limpeza. É realizada a remoção das *stopwords*, por exemplo: e, mas, para, já, entre outras, pois são palavras que não afetam o significado do *review*; a remoção dos acentos; a remoção das desinências das palavras. Essa remoção é realizada para diminuir a variáveis no momento de treinar o classificador, já que palavras que possuem o mesmo radical terão o mesmo significado na frase; a última fase do pré-processamento é deixar todas a palavras em minúsculo para que não haja diferenciação entre palavras iguais que estão em maiúsculo ou em minúsculo.

A Tabela 5 apresenta exemplos de *reviews* antes e depois de execução de pré-processamento e limpeza do texto da *review*.

Tabela 5 – Antes e Depois do Pré-Processamento e Limpeza do *Review*

Label	Antes	Depois
Informação	A cor na foto do produto é diferente do original	cor fot produt difer orig
Característica	Adorei o tamanho da calça	ador tamanh calc
Entrega	Produto demorou muito para chegar	produt demor cheg
Preço	Está muito caro	est car

Fonte – Produzido pelo Autor

O resultado dessa etapa são os dados tratados e prontos para serem utilizados na etapa de treinamento do classificador.

4.3 Treinamento dos Modelos

Com os dados pré-processados e limpos é realizado a classificação manual da base de dados para que possa ser utilizado no treinamento e teste dos modelos. Após realizada a classificação o próximo passo é realizar separação da base em dois conjuntos, onde um conjunto será para o treino dos modelos e outro para o teste dos modelos, sendo que 80% da base é para o treino e 20% para o teste.

Os modelos Multinomial e Bernoulli são implementados utilizando códigos Python, pois a linguagem apresenta a biblioteca "sklearn"que possui a implementação dos modelos, dessa forma será necessário apenas inserir a base de dados para treino para que seja realizado o treinamento dos modelos

Com os modelos treinados, iremos para fase de avaliação que é utilizado o conjunto de teste.

4.4 Avaliação dos Modelos

Com os dados das avaliações gerados pelos modelos na fase anterior, que expressa tipos de problemas com a compra, iremos utilizar algumas técnicas para medir a eficiência desse modelo gerado. Como por exemplo as métricas: *Precision*, é o número de vezes que uma classe foi predita corretamente dividido pelo número de vezes que a classe foi predita; *Recall*, o número de vezes que uma classe foi predita corretamente dividido pelo número de vezes que a classe aparece no dado de teste; e *F1-Score*, é a média harmônica entre *Precision* e *Recall*.

Utilizando essas métricas poderemos mensurar a evolução da performance de cada modelo que será treinado e comparar entre eles qual apresenta melhores resultados para o contexto desse trabalho.

5 DESENVOLVIMENTO E RESULTADOS

Nesse capitulo, é apresentado a execução das etapas definidas para esse trabalho e os resultados observados.

5.1 Seleção de Dados e Pré-Processamento

Na etapa de seleção de dados foram coletados mais de 60.000 *reviews* de compras cedidos pela empresa Trustvox e armazenados em um arquivo CSV (*Comma Separated Values*). Após a coleta foram removidos os registros que apresentavam *review* vazio, também foram removidos *reviews* que não eram referentes as *labels*: informação, característica, entrega e preço, ou *reviews* que possuíam palavrões.

Após passar por essa etapa de filtragem inicial, os dados passaram por outra etapa de pré-processamento onde foram removidos dos *reviews* as *stopwords*, sinais de pontuação, caracteres especiais, acentuação e a desinências. Para etapa de pré-processamento foi desenvolvido um *script* em Python para automatizar essa tarefa.

Foi identificado na base que alguns *reviews* poderiam pertencer a mais de uma *label*, dessa forma para esse trabalho será utilizado dois contextos diferentes onde: o contexto Homogêneo é considerado quando o *review* apresenta em sua estrutura referência a apenas uma *label*; já o contexto Heterogêneo é considerado quando o *review* apresenta em sua estrutura referência a mais de uma *label*. Como nesse trabalho são utilizados modelos que não apresentam classificação para *reviews* que pertencem a mais de uma *label*, ao mesmo tempo, nesse caso o *review* que for Heterogêneo a *label* que apresentar maior representatividade em sua estrutura será a considerada para treinamento e teste. A tabela 6 apresenta exemplos de *reviews* para o contexto Heterogêneo e Homogêneo onde a coluna "Review"apresenta o texto do *review*, a coluna "Labels Presentes"apresenta as *labels* que o *review* se refere e a coluna "Label Utilizada"apresenta a *label* utilizada para o treinamento e teste dos modelos.

Tabela 6 – Exemplo de Reviews Homogêneos e Heterogêneos

(a) Reviews Heterogêneos

Review	Labels Presentes	Label Utilizada
Produto estava rasgado e é mais e caro	Característica/Preço	Característica
Chegou muito rápido e sem amassado mas não era igual o anuncio	Entrega/Informação	Entrega
Excelente produto: belo acabamento e de ótimo robustez No anúncio não dizia que acompanhava o soquete de porcelana Porém ótima qualidade e preço bom	Característica/Informação/Preço	Característica
Excelente preço principalmente pelo desconto, só demorou para chegar	Preço/Entrega	Preço

(b) Reviews Homogêneos

Review	Labels Presentes	Label Utilizada
A cor na foto do produto é diferente do original	Informação	Informação
Adorei o tamanho da calça, tem um excelente caimento	Característica	Característica
Produto demorou muito para chegar	Entrega	Entrega
Está muito caro	Preço	Preço

Fonte – Produzido pelo Autor

Após essa etapa de seleção e pré-processamento, os dados estão prontos para serem utilizados na etapa de treinamento dos modelos.

5.2 Treinamento dos Classificadores

Na etapa de treinamento são implementados os modelos Bernoulli e Multinomial do algoritmo Naive Bayes. Para que seja possível acompanhar a evolução da performance dos modelos é utilizado versionamento da base de dados, onde é considerado uma nova versão quando a quantidade de *reviews* na base de dados aumenta, em pelo menos 1000 *reviews*, por exemplo: na versão 1 a quantidade de *reviews* era de 2112, já na versão 2 a quantidade de *reviews* passou a ser de 3112.

Em cada versão da base de dados é utilizado um algoritmo para realizar a separação desses dados em base de treino e base de teste. Para garantir que após a separação da base de dados as proporções das *labels* se mantivessem, é utilizado a técnica de estratificação nos dados, onde para cada *label* é utilizado 80% dos *reviews* para treino e 20% para teste.

Os dois modelos foram treinados com cada versão e com cada contexto para que não afetasse os resultados na etapa de avaliação dos modelos. A Tabela 7 apresenta a base de dados

para treino, onde a Tabela 7a apresenta para o contexto Heterogêneo e a Tabela 7b apresenta os dados para o contexto Homogêneo. Já a Tabela 8 apresenta a base de dados para teste, onde a Tabela 8a apresenta dados para o contexto Heterogêneo e a Tabela 8b apresenta dados para o contexto de Homogêneo.

Os campos das tabelas são: a coluna "Versão", lista o nome de cada versão; a coluna "Informação", lista a porcentagem de *reviews* que compõe a base de dados que são referentes a *label* "Informação", conforme o exemplo mostrado na tabela 4 no capitulo 4; a coluna "Característica", lista a porcentagem de *reviews* que compõe a base de dados que são referentes a *label* "Caracteristica", conforme o exemplo mostrado na tabela 4; a coluna "Entrega", lista a porcentagem de *reviews* que compõe a base de dados que são referentes a *label* "Entrega", conforme o exemplo mostrado na tabela 4; a coluna "Preço", lista a porcentagem de *reviews* que compõe a base de dados que são referentes a *label* "Preço", conforme o exemplo mostrado na tabela 4; e a coluna "Total de Reviews", lista a quantidade total de *reviews* utilizada em uma determinada versão. A seguir a Tabela 7 apresenta a versões da base de dados para treino.

Tabela 7 – Distribuição das *Labels* na Base de Treino

(a) Treino-Heterogêneo

Versão	Informação	Característica	Preço	Entrega	Total de Reviews
Versão 1	2.43 %	87.97 %	8.69 %	0.89 %	1231
Versão 2	2.47 %	88.28 %	8.39 %	0.85 %	2466
Versão 3	2.48 %	92.30 %	4.32 %	0.87 %	3705
Versão 4	2.63 %	91.51 %	4.93 %	0.91 %	4698
Versão 5	2.59 %	91.37 %	5.09 %	0.93 %	5981

(b) Treino-Homogêneo

Versão	Informação	Característca	Preço	Entrega	Total de Reviews
Versão 1	2.49 %	90.10 %	6.48 %	0.91 %	1203
Versão 2	2.50 %	92.02 %	4.58 %	0.89 %	2356
Versão 3	2.48 %	92.30 %	4.32 %	0.87 %	3535
Versão 4	2.63 %	91.51 %	4.93 %	0.91 %	4478
Versão 5	2.59 %	91.37 %	5.09 %	0.93 %	5749

Fonte – Produzido pelo Autor

Analisando a Tabela 7a é possível identificar que: no contexto Heterogêneo em todas as versões a *label* "Característica"apresenta a maior parte das bases de treino, a *label* "Informação"mantém um crescimento da versão 1 até a versão 4, já na versão 5 tem uma leva queda, a *label* "Preço"tem uma queda da versão 1 até a versão 4 e aumenta na versão 5, e a *label*

"Entrega"mantém um crescimento gradual em todas as versões, mesmo que sua porcentagem na base mantenha-se bem menor em comparação as outras *labels*.

Já na Tabela 7b é possivel identificar que: o comportamento da *label* "Característica"se mantém, comparado com o contexto Heterogêneo, para todas as versões, a *label* "Informação"e a *label* "Entrega"possui um comportamento parecido onde ambas as *labels* ficam oscilando de uma versão para outra, já a *label* "Preço"apresenta uma uma queda da versão 1 até versão 3 depois apresenta uma crescente na base de dados.

Após analisarmos as bases de treino, a seguir é analisado a base de teste de cada versão e para os contextos Heterogêneo e Homogêneo. A Tabela 8 apresenta as bases de teste.

Tabela 8 – Distribuição das *Labels* na Base de Teste

(a) Teste-Heterogêneo

Versão	Informação	Característica	Preço	Entrega	Total de Reviews
Versão 1	1.66 %	90.00 %	7.85 %	0.47 %	420
Versão 2	1.77 %	89.67 %	7.94 %	0.59 %	843
Versão 3	1.81 %	89.65 %	7.97 %	0.55 %	1266
Versão 4	1.83 %	89.63 %	7.93 %	0.59 %	1688
Versão 5	1.84 %	89.53 %	8.00 %	0.61 %	2112

(b) Teste-Homogêneo

Versão	Informação	Característica	Preço	Entrega	Total de Reviews
Versão 1	1.70 %	91.72 %	6.08 %	0.48 %	411
Versão 2	1.84 %	92.61 %	4.92 %	0.61 %	813
Versão 3	1.80 %	92.92 %	4.68 %	0.57 %	1216
Versão 4	1.84 %	92.45 %	5.09 %	0.61 %	1630
Versão 5	1.85 %	92.08 %	5.42 %	0.63 %	2046

Fonte – Produzido pelo Autor

Analisando a Tabela 8a é possível identificar que: no contexto Heterogêneo, assim como na base treino, em todas as versões a *label* "Característica"representa a maior parte das bases de teste, a *label* "Informação"mantém um crescimento da versão 1 até a versão 5, a *label* "Preço"tem um aumento da versão 1 até a versão 3 e queda na versão 4, mas aumenta na versão 5, e a *label* "Entrega"mantém uma oscilação de uma versão para outra.

Na Tabela 8b é possível identificar que: a *label* "Característica"manteve em todas as versões uma maior representatividade para todas as versões, a *label* "Informação"e a *label* "Entrega"possue um comportamento parecido onde ambas as *labels* ficam oscilando da versão 1 até versão 4 e na versão 5 tem um leve aumento, já a *label* "Preço"apresenta uma uma queda da

versão 1 até versão 3 depois apresenta uma crescente na base de dados, assim como na base de treino.

Após a analise e treinamento das bases de treino e teste para os contextos de Heterogêneo e Homogêneo algumas hipóteses aparecem, essas hipóteses são:

1. H1. O modelo Multinomial é melhor que o Bernoulli por trabalhar com frequência, ideal para texto longos como é o caso da base.
2. H2. Label com maior representatividade tem melhor perfomance, independente do modelo utilizado.
3. H3. O contexto Homogêneo é melhor que o contexto Heterogêneo

Essas questões serão respondidas no próximo capitulo, onde os modelos são submetidos a avaliação das métricas onde são apresentados e comparados os valores das métricas que foram estabelecidas na seção 4.4 para cada modelo, em versões diferente e contextos diferentes. A fase de avaliação é melhor abordada no capitulo seguinte.

5.3 Validação dos Classificadores

Após realizada a etapa de treinamento, os modelos são submetidos a avaliação das métricas que foram estabelecidas na seção 4.4. Essa avaliação é necessária para se ter uma melhor visão de como está a performance do modelo em cada versão, para isso é analisamos os resultados das métricas para cada *label*, pré determinada anteriormente, e em cada contexto, descrito na seção anterior. A tabela 9 apresenta os valores das métricas *label* Informação nos modelos Multinomial e Bernoulli nos contextos de reviews Homogêneos e Heterogêneos.

Tabela 9 – Valores das Métricas para *Label*: Informação

(a) Multinomial-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	100 %	20 %	33 %
Versão 3	67 %	27 %	39 %
Versão 4	70 %	23 %	35 %
Versão 5	100 %	18 %	31 %

(b) Bernoulli-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	0 %	0 %	0 %
Versão 3	0 %	0 %	0 %
Versão 4	33 %	7 %	11 %
Versão 5	25 %	5 %	9 %

(c) Multinomial-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	100 %	20 %	33 %
Versão 3	100 %	13 %	23 %
Versão 4	50 %	10 %	16 %
Versão 5	89 %	21 %	33 %

(d) Bernoulli-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	0 %	0 %	0 %
Versão 3	29 %	17 %	22 %
Versão 4	50 %	19 %	28 %
Versão 5	33 %	13 %	19 %

Fonte – Produzido pelo Autor

Analisando a tabela 9 nota-se que o modelo Multinomial apresentou melhor performance ou teve valores iguais, em comparação ao Bernoulli, em todas as versões tanto para contexto Homogêneo, quanto Heterogêneo.

Outro ponto interessante é que em contexto Homogêneo o Multinomial apresentou melhor performance na maioria das versões do que no Heterogêneo.

Tabela 10 – Valores das Métricas para *Label*: Característica

(a) Multinomial-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	97 %	100 %	99 %
Versão 2	97 %	100 %	98 %
Versão 3	97 %	100 %	98 %
Versão 4	97 %	100 %	98 %
Versão 5	97 %	100 %	98 %

(b) Bernoulli-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	92 %	99 %	96 %
Versão 2	93 %	100 %	96 %
Versão 3	93 %	100 %	96 %
Versão 4	93 %	100 %	96 %
Versão 5	92 %	100 %	96 %

(c) Multinomial-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	96 %	100 %	98 %
Versão 2	96 %	100 %	98 %
Versão 3	96 %	100 %	98 %
Versão 4	96 %	100 %	98 %
Versão 5	96 %	100 %	98 %

(d) Bernoulli-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	95 %	99 %	97 %
Versão 2	94 %	99 %	97 %
Versão 3	96 %	99 %	97 %
Versão 4	95 %	99 %	97 %
Versão 5	96 %	99 %	98 %

Fonte – Produzido pelo Autor

Analisando a tabela 10 nota-se que, apesar de ambos os modelos apresentarem valores de métricas muito bons, o modelo Multinomial apresenta melhor performance tanto no contexto Homogêneo quanto no Heterogêneo, em relação ao modelo Bernoulli.

Comparando o modelo Multinomial Homogêneo e Heterogêneo, é possível identificar que, apesar da pouca diferença entre as performances, o modelo Homogêneo teve maior vantagem.

Tabela 11 – Valores das Métricas para *Label*: Preço

(a) Multinomial-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	100 %	88 %	94 %
Versão 2	97 %	82 %	89 %
Versão 3	100 %	77 %	87 %
Versão 4	100 %	78 %	88 %
Versão 5	97 %	83 %	89 %

(b) Bernoulli-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	0 %	0 %	0 %
Versão 3	33 %	2 %	3 %
Versão 4	40 %	2 %	5 %
Versão 5	80 %	4 %	7 %

(c) Multinomial-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	100 %	76 %	86 %
Versão 2	98 %	73 %	84 %
Versão 3	100 %	78 %	88 %
Versão 4	98 %	78 %	87 %
Versão 5	99 %	80 %	89 %

(d) Bernoulli-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	100 %	67 %	80 %
Versão 2	96 %	64 %	77 %
Versão 3	99 %	73 %	84 %
Versão 4	96 %	69 %	80 %
Versão 5	98 %	78 %	87 %

Fonte – Produzido pelo Autor

Analisando a tabela 11 nota-se que o modelo Multinomial apresentou melhor performance que o modelo Bernoulli, tanto no contexto Homogêneo, quanto no contexto Heterogêneo. Mesmo no contexto Heterogêneo a performance de ambos os modelos tenha apresenta diferença pequena.

Comparando Multinomial Homogêneo e Heterogêneo observa-se que o modelo no contexto Homogêneo apresenta melhor performance de modo geral.

Tabela 12 – Valores das Métricas para *Label*: Entrega

(a) Multinomial-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	100 %	50 %	67 %
Versão 2	100 %	20 %	33 %
Versão 3	50 %	14 %	22 %
Versão 4	100 %	50 %	67 %
Versão 5	100 %	15 %	27 %

(b) Bernoulli-Homogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	0 %	0 %	0 %
Versão 3	0 %	0 %	0 %
Versão 4	0 %	0 %	0 %
Versão 5	0 %	0 %	0 %

(c) Multinomial-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	100 %	20 %	33 %
Versão 3	100 %	14 %	25 %
Versão 4	100 %	20 %	33 %
Versão 5	0 %	0 %	0 %

(d) Bernoulli-Heterogêneo

Versão	Precision	Recall	F1-Score
Versão 1	0 %	0 %	0 %
Versão 2	0 %	0 %	0 %
Versão 3	0 %	0 %	0 %
Versão 4	0 %	0 %	0 %
Versão 5	0 %	0 %	0 %

Fonte – Produzido pelo Autor

Analisando a tabela 12 nota-se que o modelo Bernoulli não conseguiu "aprender"sobre a *label* entrega, onde apresentou resultado de 0% em todas as métricas para todas as versões.

Realizando um comparativo entre o modelo Multinomial Homogêneo e Heterogêneo, observa-se que o modelo Homogêneo apresenta melhor performance do que o Heterogêneo. Onde o modelo Heterogêneo apresentou valor de 0% para todas as métricas na versão 1 e versão 5 da base de dados.

Comparando as performances das *labels* entrega e características, é possível inferir a representatividade de uma *label* na base de dados tem um relação diretamente proporcional aos resultados da performance do modelo na detecção da *label*.

6 CONCLUSÃO E TRABALHOS FUTUROS

O objetivo desse trabalho foi desenvolver um modelo preditivo capaz de identificar, a partir de um comentário, qual o tipo de problema ele se refere. Pra esse trabalho as classes identificadas e utilizadas foram:Informação do Produto; Característica do Produto; Entrega e Preço. Nesse trabalho as tarefas realizadas foram: a seleção de dados, com auxilio da empresa Trustvox; tratamento de dados; comparação entre duas abordagens do Naive Bayes,Bernoulli e Multinomial e a identificação de qual modelo apresenta melhor performance para reviews de *e-commerce*.

O resultado do trabalho foi bastante satisfatório, inclusive os resultados do trabalho foram apresentados para empresa Trustvox que aprovou e adotou esse projeto para seu produto. Abaixo temos um resumo do resultado obtido nesse trabalho.

O primeiro resultado relevante foi identificar que *reviews* com contexto Homogêneo e Heterogêneo afetam a performance do modelo que está sendo utilizado. O segundo resultado relevante foi identificar que para *reviews* de e-commerce o modelo Multinomial com um contexto de *reviews* Homogêneos tem uma maior vantagem em relação ao modelo Bernoulli, tanto no contexto Homogêneo quanto no Heterogêneo, o que para trabalhos futuros que possam utilizar *reviews* de e-commerce já é de grande ajuda, pois essa identificação e tratamento de contexto pode ser incluída na etapa de seleção dos dados e pré-processamento.

Como sugestão de trabalhos futuros relacionados na mesma área, pode ser realizado a comparação com outros modelos de predição como Random Forest, Gaussian Naive Bayes, etc. Também pode ser feita uma analise de sentimentos nos *reviews* para além de identificar sobre o que o *review* se refere também possa ser identificado se é algo positivo ou negativo o que para a empresa Trustvox ajudaria a entender os principais pontos positivos e negativos apontados pelos clientes.

Outra sugestão é a utilização de técnicas de reconhecimento de entidades visando realizar a classificação dos *reviews* o que pode ajudar na classificação de *reviews* que se referem a mais de uma *label*.

REFERÊNCIAS

BERRY, M. J.; LINOFF, G. **Data mining techniques**: for marketing, sales, and customer support. USA: John Wiley & Sons, Inc., 1997.

DOMINGOS, P.; PAZZANI, M. **Beyond independence**: Conditions for the optimality of the simple bayesian classifer. In: Proc. 13th Intl. Conf. Machine Learning. [S.l.: s.n.], 1996. p. 105–112.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification and scene analysis**. [S.l.]: Wiley New York, 1973. v. 3.

EVANGELISTA, T. R.; PADILHA, T. P. P. **Monitoramento de posts sobre empresas de e-commerce em redes sociais utilizando análise de sentimentos**. In: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). [S.l.: s.n.], 2013.

FAYYD, U. M.; SHAPIRO, G. P.; SMYTH, P. **From data mining to knowledge discovery**: an overview. AAAI Press/The MIT Press, 1996.

FELDMAN, R.; SANGER, J. **The text mining handbook**: advanced approaches in analyzing unstructured data. [S.l.]: Cambridge university press, 2007.

HOTHO, A.; NÜRNBERGER, A.; PAASS, G. **A brief survey of text mining**. In: CITESEER. Ldv Forum. [S.l.], 2005. v. 20, n. 1, p. 19–62.

MARTINAZZO, B. **Um Método de Identificação de Emoções em Textos Curtos para o Português do Brasil**. 68 p. Dissertação (Mestrado) — Pontifícia Universidade Católica do Paraná, Curitiba, 2010.

MCCALLUM, A.; NIGAM, K. et al. **A comparison of event models for naive bayes text classification**. In: CITESEER. AAAI-98 workshop on learning for text categorization. [S.l.], 1998. v. 752, n. 1, p. 41–48.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. São Paulo: Editora Saraiva, 2017.

OGURI, P.; LUIZ, R.; RENTERIA, R. **Aprendizado de máquina para o problema de sentiment classification**. 54 p. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2006.

SANTOS, R. S. M. Roney L. de S. **Extração de métricas e analise de sentimentos em comentários web no domínio de hotéis**. In: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). [S.l.: s.n.], 2016. p. 127–138.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining**: mineração de dados. [S.l.]: Ciência Moderna, 2009.