



Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Um Estudo Sobre As Características Que Compõem A Percepção De Utilidade De Comentários Web Sobre Produtos

Jardeson Leandro Nascimento Barbosa

Número de Ordem PPGCC: M001

Teresina-PI, Março de 2019

Jardeson Leandro Nascimento Barbosa

**Um estudo sobre as características que compõem a
percepção de utilidade de comentários web sobre
produtos**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Raimundo Santos Moura

Teresina-PI

Março de 2019

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca Setorial do CCN

B238e Barbosa, Jardeson Leandro Nascimento.

Um estudo sobre as características que compõem a percepção de utilidades de comentários web sobre produtos / Jardeson Leandro Nascimento Barbosa. – Teresina, 2019.
76f.: il.:color.

Dissertação (Mestrado) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Pós-Graduação em Ciência da Computação, 2019.

Orientador: Prof. Dr. Raimundo Santos Moura.

1. Sistemas de Computação. 2. Mineração de Opinião.
3. Redes Neurais Artificiais. I. Título.

CDD 004.21

Bibliotecária: Caryne Maria da Silva Gomes. CRB/3 1461

Jardeson Leandro Nascimento Barbosa

Um estudo sobre as características que compõem a percepção de utilidade de comentários web sobre produtos

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

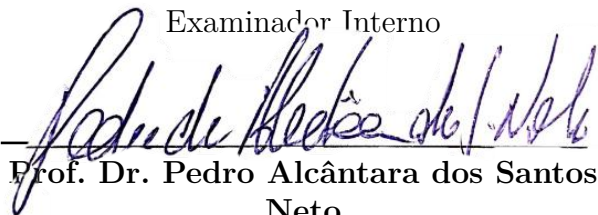
Trabalho aprovado. Teresina-PI, 21 de março de 2019:



Prof. Dr. Raimundo Santos Moura
Orientador



Prof. Dr. Vinícius Ponte Machado
Examinador Interno



Prof. Dr. Pedro Alcântara dos Santos Neto
Examinador Interno



Prof. Dra. Giseli Rabello Lopes
Examinadora Externa

Teresina-PI
Março de 2019

*Aos meus pais Maria de Jesus e José Edimar.
Nada nesta vida seria possível sem o apoio incondicional deles.*

Agradecimentos

Agradeço aos meus pais, Maria de Jesus e José Edimar, pelo imenso investimento ao longo dos anos e por nunca terem desistido de mim.

Agradeço ao meu orientador, Raimundo Santos Moura, pela paciência, suporte e por compartilhar seus conhecimentos durante os anos de Graduação e Mestrado.

Aos amigos que fiz durante essa jornada e a todos os professores que tive o prazer de conhecer nos últimos 26 anos.

*“Remember to look up at the stars
and not down at your feet.”
(Stephen Hawking)*

Resumo

Com o crescimento da Internet, tornou-se comum que as pessoas publiquem suas opiniões online. Compartilhamos diariamente nossas opiniões a respeito de tudo. A partir dessas opiniões, surgiram os *reviews* de produtos online onde usuários podem publicar na Internet textos sobre suas experiências com produtos ou serviços. *Reviews* de usuários online ajudam futuros consumidores a tomarem uma decisão sobre a compra e podem também beneficiar os fornecedores do produto ou serviço. Entretanto, como milhares de *reviews* são publicados diariamente, seria interessante apresentar ao usuário uma lista ordenada com os *reviews* mais importantes. A propagação de *reviews* inúteis ou com informações falsas é outro problema para usuários e fornecedores. Para tornar a tarefa de ler *reviews* online mais agradável, foi desenvolvido um mecanismo em que os próprios usuários votam em *reviews* como úteis ou não úteis no processo de tomada de decisão de compra. Esse sistema está presente atualmente na maioria dos *websites* e é amplamente utilizado, porém não é isento de certos problemas. Algumas tendências a respeito do voto de utilidade de *reviews* online podem prejudicar a visualização de informações relevantes sobre o produto. Para resolver esse problema é necessário entender o que torna um *review* útil e, a partir desse conhecimento, desenvolver um sistema que identifique as características de utilidade de *reviews* de forma automática. Este trabalho descreve um estudo sobre os fatores que contribuem com a percepção de utilidade de *reviews* online, que inclui o desenvolvimento de um modelo de regressão utilizando aprendizado de máquina supervisionado capaz de prever, de forma automática, a utilidade de *reviews* futuros. Diversos experimentos foram realizados para definir melhor topologia das Redes Neurais (MLP e RBF), bem como para determinar as melhores *features*. Os experimentos utilizaram dois Corpora nos domínios de e-jogos e smartphones. Os resultados indicam que a reputação e a expertise do autor, as opiniões e o tamanho do texto são as características mais importantes para a percepção de utilidade de *reviews*.

Palavras-chaves: Mineração de Opiniões. Percepção de Utilidade de *Reviews*. Redes Neurais Artificiais.

Abstract

With the growth of the Internet, it has become common for people to post their opinions online. We share our opinions daily about everything. From these opinions, online user reviews emerged and created a space where Internet users can publish on texts about their experiences with products or services. Online user reviews help future consumers in making a decision about the purchase and can also benefit the product and service suppliers. However, since thousands of reviews are published on a daily basis, it would be interesting to present them to the user in an ordered list with the most important reviews coming first. Another problem is the propagation of useless and fake reviews. To make the task of reading online reviews more enjoyable, a mechanism has been developed in which users themselves vote on reviews as helpful or not helpful in the decision-making process. This system is currently present on most websites and is widely used, but is not exempt from certain problems. Some biases regarding the online review helpfulness vote may hinder the reading of relevant information about the product. To solve this problem it is necessary to understand what makes a review helpful and, from this knowledge, to develop a system that automatically identifies how helpful a review is. This thesis describes a study on the factors that contribute to the helpfulness perception of online reviews, which includes a regression model using supervised machine learning. Several experiments were carried out to define a better topology of the Artificial Neural Networks (MLP and RBF), as well as to determine the best features. The experiments used two corpora in the domains of e-games and smartphones. The results indicate that the author's reputation and expertise, opinions present in the text and text size are the most important features for the perception of review helpfulness reviews.

Keywords: Opinion Mining. Perception of Review Helpfulness. Artificial Neural Networks.

Lista de ilustrações

Figura 1 – Modelo de regressão proposto usando RNA	4
Figura 2 – Processo de decisão de compra: etapas	9
Figura 3 – Diferentes tipos de bens quanto à facilidade de avaliação	12
Figura 4 – Modelo de Neurônio Artificial (SILVA; CARVALHO; SARMENTO, 2012)	16
Figura 5 – Principais arquiteturas de RNAs (SILVA; CARVALHO; SARMENTO, 2012)	18
Figura 6 – Típica RNA MLP de três camadas (SILVA; CARVALHO; SARMENTO, 2012)	19
Figura 7 – Configuração típica da rede RBF(SILVA; SPATTI; FLAUZINO, 2010)	21
Figura 8 – Função de base radial do tipo gaussiana(SILVA; SPATTI; FLAUZINO, 2010)	21
Figura 9 – Topologia de RBF aplicada em aproximação de funções(SILVA; SPATTI; FLAUZINO, 2010)	23
Figura 10 – Exemplo de extração de opiniões em um <i>review online</i>	26
Figura 11 – Etapas da Mineração de Opiniões (BECKER; TUMITAN, 2013)	28
Figura 12 – Representação de documentos (SOARES et al., 2008)	29
Figura 13 – Extração de <i>bag-of-words</i>	29
Figura 14 – Exemplo de <i>review</i> do <i>Steam</i> adaptado	33
Figura 15 – Exemplo de <i>review</i> do <i>Steam</i> em JSON	33
Figura 16 – Exemplo de <i>review</i> do <i>Buscapé</i> adaptado	35
Figura 17 – Exemplo de <i>review</i> do <i>Buscapé</i> em JSON	35
Figura 18 – Review publicado no site Amazon	39
Figura 19 – Distribuição dos Trabalhos Relacionados	43
Figura 20 – Modelo proposto	45
Figura 21 – Vetor de opiniões, sendo o primeiro elemento a quantidade de opiniões e os demais um <i>bag-of-words</i>	47
Figura 22 – Padrões linguísticos (Adaptação dos padrões propostos por Turney (2002))	48
Figura 23 – Fases da extração de opiniões	48
Figura 24 – Rede Neural Artificial do modelo proposto	53
Figura 25 – Rede Neural Artificial do modelo proposto	54
Figura 26 – RMSE para treino com Corpus <i>ReviewsSteam</i> e Rede MLP	58
Figura 27 – RMSE para treino com Corpus <i>ReviewsSteam</i> e Rede RBF	58

Lista de tabelas

Tabela 1 – Resumo dos trabalhos relacionados	44
Tabela 2 – Referências de pontuação da adaptação do teste de legibilidade de Flesch-Kincaid para o Português	51
Tabela 3 – Resumo das <i>features</i> relacionadas ao conteúdo textual do <i>review</i>	52
Tabela 4 – Métricas de avaliação dos modelos MLP e RBF	59
Tabela 5 – R^2 em trabalhos relacionados	59
Tabela 6 – Métricas de avaliação dos modelos MLP e RBF	59
Tabela 7 – Métricas de avaliação das redes MLP e RBF para os Corpora analisados	60
Tabela 8 – Variáveis de entrada dos modelos propostos	61
Tabela 9 – Tabela de correlação da base <i>ReviewsSteam</i>	61
Tabela 10 – Tabela de correlação da base <i>ReviewsBuscape</i>	62
Tabela 11 – Resultados do CFS para cada modelo	62
Tabela 12 – Comparação da performance dos modelos com seleção de <i>features</i>	63

Lista de abreviaturas e siglas

BOW	<i>Bag of Words</i>
eWOM	<i>Electronic Word-Of-Mouth</i>
LSA	<i>Latent Semantic Analysis</i>
MLP	<i>Multi-Layer Perceptron</i>
NLTK	<i>Natural Language ToolKit</i>
PLN	Processamento de Linguagem Natural
POS	<i>Part-of-Speech</i>
PPGCC	Programa de Pós-Graduação em Ciência da Computação
RBF	<i>Radial Basis Function</i>
RNA	Redes Neurais Artificiais
RSO	Redes Sociais Online
SVM	<i>Support Vector Machine</i>
SVR	<i>Support Vector Regression</i>
TF-IDF	<i>Term Frequency–Inverse Document Frequency</i>
UFPI	Universidade Federal do Piauí
WOM	<i>Word-Of-Mouth</i>

Lista de símbolos

H	Medida de utilidade do <i>review</i>
Σ	Combinador linear da Rede Neural Artificial
θ	Limiar de ativação da Rede Neural Artificial
μ	Potencial de ativação da Rede Neural Artificial
Ω	Conjunto de neurônios de uma Rede Neural Artificial
y	Conjunto de conexões entre neurônios de uma Rede Neural Artificial

Sumário

1	INTRODUÇÃO	1
1.1	Motivação	1
1.2	Objetivos	4
1.3	Organização	5
2	REFERENCIAL TEÓRICO	7
2.1	O processo de decisão de compra	7
2.1.1	Fatores que influenciam a decisão de compra	8
2.1.2	Etapas do processo de decisão de compra	8
2.1.3	O impacto de <i>reviews online</i> no Comércio Eletrônico	10
2.1.4	Tipos de Produtos	11
2.2	Percepção de utilidade de <i>reviews online</i>	12
2.3	Aprendizado de Máquina	14
2.3.1	Aplicações e problemas	14
2.3.2	Redes Neurais Artificiais	15
2.3.2.1	Rede Neural Artificial Perceptron Multicamadas (RNA MLP)	18
2.3.2.2	Rede MLP como aproximador universal de funções	19
2.3.2.3	Redes de funções de base radial (RBF)	20
2.3.2.4	Redes RBF como aproximadores universais de funções	22
2.3.2.5	Escolha de topologias das redes neurais artificiais	23
2.3.2.6	Métricas de valiação dos modelos RNA MLP e RBF para regressão	23
2.3.2.7	Validação dos modelos RNA MLP e RBF	24
2.4	Mineração de Opiniões	25
2.4.1	Níveis de análise textual	27
2.4.2	Etapas do processo de mineração de opiniões	27
2.4.3	Modelo de <i>bag-of-words</i>	28
2.4.3.1	Representação booleana	29
2.4.3.2	Representação <i>term-frequency</i>	29
2.4.3.3	Representação <i>TF-IDF</i>	29
2.4.3.4	Análise de Semântica Latente	30
2.5	Ferramentas e Recursos	30
2.5.1	NLTK: Natural Language ToolKit	31
2.5.1.1	Etiquetador MacMorpho	31
2.5.2	<i>Corpus ReviewsSteam</i>	32
2.5.3	<i>Corpus ReviewsBuscapé</i>	34

2.6	Considerações Finais	34
3	TRABALHOS RELACIONADOS E ESTADO DA ARTE	37
3.1	Trabalhos desenvolvidos pelo grupo de PLN da UFPI	37
3.2	Percepção de utilidade de <i>reviews</i> e identificação de <i>reviews</i> úteis no contexto <i>web</i> : revisão da literatura	38
3.2.1	Datasets	38
3.2.2	Avaliação dos modelos	39
3.2.3	Metodologias	39
3.2.4	Principais descobertas na área	40
3.3	Considerações Finais	42
4	MODELO PROPOSTO	45
4.1	Definição de <i>Features</i>	45
4.1.1	Características relativas à autoria	46
4.1.1.1	Reputação do autor	46
4.1.1.2	Expertise	46
4.1.2	Características relativas ao conteúdo textual	46
4.1.2.1	Opiniões do <i>review</i>	47
4.1.2.1.1	Extração de Opiniões	47
4.1.2.1.2	Modelo de <i>bag-of-words</i>	48
4.1.2.2	Elementos estilísticos do <i>review</i>	49
4.1.3	Características referentes aos metadados do <i>review</i>	51
4.2	Análise de Regressão para validar o modelo	52
4.2.1	Modelo utilizando RNA MLP	52
4.2.2	Modelo utilizando RNA RBF	54
4.3	Considerações Finais	54
5	EXPERIMENTOS	57
5.1	Preparação dos dados	57
5.2	Experimento 1: Definição das melhores topologias das RNAs	57
5.3	Experimento 2: Variações do modelo de <i>bag-of-words</i>	59
5.4	Experimento 3: Seleção de <i>features</i> usando CFS	60
5.5	Resultados e Discussão	60
	REFERÊNCIAS	67

1 Introdução

Neste capítulo serão apresentados o contexto do presente trabalho, as motivações da pesquisa, bem como os objetivos buscados através da realização desta pesquisa e, por fim, a organização do presente texto.

1.1 Motivação

Com números que ultrapassam 12 milhões de lojas no planeta, o *e-commerce*¹ é, atualmente, um dos principais serviços da Internet (LEMONSTAND, 2015) e a medida que esse número cresce, torna-se cada vez mais difícil para os consumidores tomar uma decisão sem gerar arrependimentos futuros. Durante o processo de busca de informação, sobre qual produto comprar e onde comprar, consumidores são geralmente submetidos a múltiplas escolhas que podem resultar em esgotamento mental (VOHS et al., 2014).

Um exemplo clássico desse fenômeno vem de um estudo experimental publicado por Iyengar e Lepper (2000). Em um dos experimentos desse estudo, consumidores de um supermercado foram apresentados a um estande de exibição de geleias em dois dias distintos, contendo 24 sabores diferentes de geleia em um dia e 6 sabores diferentes em outro. Dos 242 consumidores que passaram pela mesa no dia em que haviam 24 sabores, 60% provaram das geleias. Em contraste, dos 260 consumidores que passaram pela mesa no segundo dia, apenas 40% provaram. Dessa forma, os consumidores que encontraram a extensiva condição de escolha foram mais atraídos para o estande do que aqueles que encontraram uma limitada condição de escolha, o que sugere que a variedade fornecida na extensiva condição de escolha foi inicialmente mais atraente.

No entanto, a quantidade de sabores provados pelos dois grupos não foi tão diferente. Consumidores da condição de escolha extensiva provaram em média 1,50 geleias, enquanto consumidores do grupo de condição de escolha limitada provaram uma média de 1,38 geleias. A atração inicial ao estande também não se converteu em vendas. Quase 30% (31) dos consumidores que provaram das geleias na condição de escolha limitada subsequentemente compraram pelo menos um pote de geleia, enquanto somente 3% (4) dos consumidores do grupo de condição de escolha extensiva o fizeram. Assim, sugere-se que consumidores inicialmente expostos a escolhas limitadas são consideravelmente mais propensos a comprar produtos do que consumidores que tenham inicialmente encontrado um conjunto muito maior de opções.

Em um cenário onde existem muitas opções, muitos consumidores acabam ficando

¹ Termo em inglês para comércio eletrônico

entediados e desistindo, outros não conseguem comprar com medo de estarem realizando uma escolha errada. Isso acontece porque nosso cérebro não trabalha tão bem quando o número de opções aumenta. Isso é conhecido como “paralisia da escolha” ou “paradoxo da escolha” (SCHWARTZ, 2005).

No *e-commerce*, a paralisia da escolha pode ocorrer no momento de selecionar a loja onde comprar, ou o produto, modelo e marca. Apesar de parecer bem simples, o processo de compra passa por várias etapas e, uma delas, é a etapa de decisão de compra, onde o consumidor deve comprar ou desistir de comprar baseado nas informações adquiridas no processo de busca.

Para auxiliar o processo de decisão dos clientes, empresas de *e-commerce* utilizam diversas táticas para informar melhor os consumidores ou direcioná-los a cenários onde o número de escolhas é mais limitado ou personalizado. Para limitar a quantidade de escolhas, podem ser usados sistemas de buscas com filtros ou sistemas de recomendação automáticos que oferecem recomendações baseadas em informações previamente adquiridas sobre o consumidor, como hábitos de consumo, pesquisas na Internet ou hábitos de consumo de amigos e pessoas próximas. Porém, apesar do bom desempenho, sistemas de recomendação automáticos ainda não são tão eficientes quanto as recomendações feitas por pessoas conhecidas e confiáveis, como familiares e amigos. Sinha e Swearingen (2001) observaram que é mais provável que os consumidores se interessem por produtos recomendados por pessoas que eles confiam do que por aqueles recomendados por sistemas automáticos.

Goldenberg, Libai e Muller (2001) demonstraram que o processo de tomada de decisão dos clientes é fortemente influenciado pela comunicação “boca-a-boca” (do inglês “Word-of-Mouth”, WOM). As opiniões dos consumidores a respeito de um produto influenciam a tomada de decisão de consumidores futuros. Empresas de *e-commerce* encontraram na comunicação “boca-a-boca” uma oportunidade de tornar seus produtos mais visíveis e os consumidores mais bem informados. Na internet, redes sociais *online* concentram a comunicação WOM, uma vez que permitem que consumidores compartilhem suas experiências e recomendações através de opiniões, ou *reviews* – seja escrevendo, votando, comentando ou compartilhando (KIM; SRIVASTAVA, 2007). Por esse motivo, é cada vez maior o número de *sites* de *e-commerce* que incorporam elementos de redes sociais *online* e contam com um sistema de recomendações de usuários.

Em pesquisa conduzida pelo site *Trip Advisor*², estima-se que 80% dos usuários leem, pelo menos, de 6 a 12 *reviews* antes de reservarem um hotel (TRIPADVISOR, 2013). Em um *website* como o *TripAdvisor*, que conta com milhares de *reviews*, pode-se supor que serão lidos os 12 primeiros *reviews* apresentados. Entretanto, dada a enorme quantidade de *spam*, textos de baixa qualidade e informações falsas na Internet, como garantir que os 12 primeiros *reviews* exibidos terão informações confiáveis que afetem a decisão de compra

² <http://www.tripadvisor.com>

dos clientes? Afinal de contas, é de interesse dos *websites* que as visitas sejam convertidas em compras.

Consumidores que leem *reviews* online se atentam não somente às notas dos *reviews*, mas também ao conteúdo textual e informações não-textuais, como a reputação do autor e a data em que o *review* foi escrito (HU; LIU; ZHANG, 2008). Mesmo as avaliações negativas podem ter impacto positivo nas vendas se feitas de maneira educada e legível (HAMILTON; VOHS; MCGILL, 2014). Garantir que esses *reviews* fornecem informações verídicas e de qualidade pode ser a diferença entre uma boa taxa de conversão³ e o fracasso.

Dentre as diversas soluções encontradas pelos *sites* de comércio eletrônico para resolver o problema da qualidade de *reviews online*, uma das mais adotadas foi o estabelecimento de uma métrica chamada de “utilidade do *review*”. Em *sites* como *Amazon*⁴ ou *Americanas*⁵, usuários que leem *reviews* podem votar naqueles que eles consideram úteis ou inúteis no processo de decisão de compra. Esses votos são utilizados para ordenar os *reviews* em função dos mais úteis. Assim, os próprios usuários realizam uma moderação colaborativa, filtrando os textos que devem ser lidos pelos futuros consumidores.

Enquanto essa abordagem é vantajosa por utilizar o *feedback* dos próprios usuários para gerar um *ranking* de melhores *reviews*, ela acaba introduzindo uma série de novos problemas. Um deles é que, para que um *review* seja considerado útil, é necessário que ele possua alguns votos. Consequentemente, publicações mais recentes deixarão de ser lidas por não possuírem votos suficientes, mesmo que sejam melhores ou mais relevantes que as postagens mais votadas (LI et al., 2013). Além disso, certas tendências são naturalmente desenvolvidas, como o “ciclo do vencedor”, em que *reviews* bem votados e bem posicionados (como o mais votado, por exemplo), sejam sempre os *reviews* a serem lidos e, consequentemente, os *reviews* a serem votados. Outra tendência é o “madrugador”, ou seja, os primeiros *reviews* publicados naturalmente serão os mais votados, pois estarão disponíveis por mais tempo. Além disso é difícil ranquear *reviews* que ainda não receberam algum voto. Por esses motivos, seria interessante poder identificar o valor potencial de *reviews online* sem depender exclusivamente do voto dos usuários.

A questão de pesquisa que norteia este trabalho é: Se é possível identificar o grau de percepção de utilidade de um *review*, ou a qualidade desse *review*, sem que ele tenha sido votado pelos usuários?

³ Quantidade de pessoas que visitam a página do produto e efetivam a compra

⁴ <https://www.amazon.com.br/>

⁵ <http://www.americanas.com.br>

1.2 Objetivos

A fim de amenizar os problemas destacados anteriormente, e para responder a questão de pesquisa levantada, este trabalho apresenta um estudo sobre os fatores que contribuem com a percepção de utilidade de *reviews online* que tem como objetivos: a) identificar elementos formativos da percepção de utilidade de *reviews* e b) propor um modelo de regressão para prever automaticamente a utilidade de *reviews online* utilizando aprendizado de máquina. A visão geral do modelo proposto é mostrada na 1, onde X é um vetor de entrada do modelo de regressão composto por diversos elementos formativos da percepção de utilidade de um review; o modelo de regressão faz uso de redes neurais artificiais; e H representa o grau de percepção de utilidade do review.

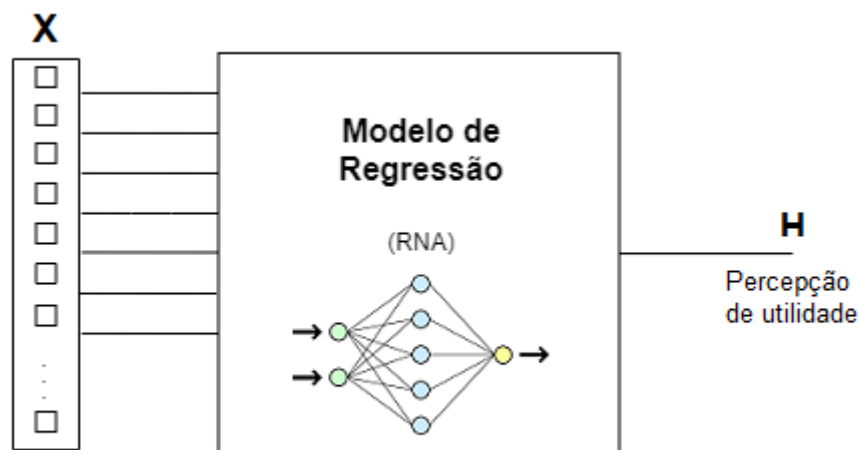


Figura 1 – Modelo de regressão proposto usando RNA

Este trabalho parte do pressuposto de que é possível identificar o grau de percepção de utilidade de um *review* a partir de informações presentes no próprio *review*, como o conteúdo textual e informações sobre o autor. O estudo foi realizado a partir de *reviews* coletados da loja *Steam*⁶, que distribui jogos, filmes, músicas e *software* para computadores pessoais e do *site* Buscapé⁷, um agregador de ofertas e *reviews* de *sites* do comércio varejista brasileiro.

Como contribuições desse trabalho, destaca-se: i) um estudo para investigar, aplicar e validar pesquisas anteriores a respeito de características formativas da utilidade de *reviews* em dois diferentes contextos; e ii) a criação de um modelo para prever a utilidade de *reviews* de forma automática que pode ser aplicado em diferentes *websites*. Este modelo também nos permite identificar através de algoritmos de seleção de *features* o grau de importância das características investigadas na percepção de utilidade de *reviews*. Além disso, as descobertas deste trabalho oferecem orientações de como empresas de *e-commerce* podem

⁶ <http://store.steampowered.com/>

⁷ <https://www.buscape.com.br/>

estruturar seus *websites* a fim de destacar informações mais relevantes aos consumidores através de *reviews*.

1.3 Organização

O restante deste trabalho é organizado da seguinte forma:

O Capítulo 2 trata do referencial teórico onde são discutidos alguns conceitos básicos sobre o processo de decisão de compra, *reviews* de usuários no *e-commerce* e comunicação “boca-a-boca” na *web*. Descreve-se também conceitos básicos da área de Mineração de Opiniões e Aprendizado de Máquina (AM), incluindo modelos computacionais baseados em Redes Neurais Artificiais. Além disso, são discutidas ferramentas e recursos utilizados para a realização deste trabalho.

O Capítulo 3 apresenta uma revisão da literatura, onde são discutidos trabalhos relacionados à mineração de opiniões em *reviews* online e à utilidade do *review*.

O Capítulo 4 descreve as hipóteses levantadas e o modelo proposto.

O Capítulo 5 destaca os experimentos realizados e discute os resultados obtidos.

Por fim, o Capítulo 6 apresenta as considerações finais e trabalhos futuros.

2 Referencial Teórico

Neste capítulo são apresentadas as áreas de interesse desta dissertação, destacando os principais conceitos e definições.

Na primeira seção são apresentados os conceitos relacionados ao processo de decisão de compra e o comportamento de consumidores no comércio *online*, além de conceitos básicos de comércio e marketing.

A segunda seção apresenta o resultado de um estudo feito a respeito da percepção de utilidade de *reviews online*, evidenciando o impacto de *reviews online* no processo de decisão de compra e formas de quantificar essa percepção.

Na terceira seção são apresentados conceitos de Aprendizado de Máquina, que pode ser usado para automatizar processos, detectar padrões ou realizar previsões, por exemplo. Nessa seção nos aprofundamos em um estudo sobre Redes Neurais Artificiais, que são usadas neste trabalho como a base para identificar e prever a percepção de utilidade de *reviews online*.

Na quarta seção falamos de algumas técnicas de Processamento de Linguagem Natural que foram utilizadas neste trabalho, em especial técnicas de Mineração de Opiniões.

A quinta seção é dedicada às ferramentas e recursos utilizados neste trabalho, como bases de dados e módulos.

Por fim, a sexta seção é dedicada às considerações finais a respeito do referencial teórico, bem como uma revisão de tudo o que foi apresentado.

2.1 O processo de decisão de compra

Os profissionais de marketing estudam os padrões de compra dos consumidores para resolver questões como: onde eles compram, o que compram e por que compram. No entanto, o porquê de consumidores adquirirem um produto específico não é fácil de resolver, porque a resposta está na mente dos consumidores. O processo de decisão de compra de um consumidor é uma parte do comportamento humano, sendo influenciado por diversos fatores, incluindo questões sociais, pessoais e psicológicas. Ao estudar o comportamento dos consumidores em compras anteriores, os profissionais de marketing podem estimar o comportamento no futuro das pessoas na tomada de decisão de compra de um produto (KOTLER; ARMSTRONG, 2010).

2.1.1 Fatores que influenciam a decisão de compra

- **Fatores sociais:** cada indivíduo tem alguém influenciando suas decisões de compra, que pode ser um amigo, uma celebridade ou, mais comumente, um membro da família.

A família é o ambiente principal para um indivíduo adquirir valores, desenvolver e moldar a personalidade. Esse ambiente oferece a possibilidade de desenvolver atitudes e opiniões para vários assuntos, como relações sociais, sociedade e política. É no ambiente familiar que se criam as primeiras percepções sobre marcas ou produtos e hábitos de consumo. Por exemplo, os consumidores que criaram percepções de marca quando eram jovens podem realizar essas mesmas seleções de marca na vida adulta sem nem mesmo reconhecer que sua família influenciou essas seleções (KOTLER; ARMSTRONG, 2010).

- **Fatores pessoais:** a idade, a ocupação, a situação econômica, o estilo de vida, a personalidade, o autoconceito, os valores e *hobbies* influenciam as decisões de um indivíduo. Pessoas que compartilham de ocupações semelhantes tendem a ter gostos semelhantes por músicas, roupas e atividades de lazer. O nível de renda afeta o que o consumidor pode pagar e a perspectiva em relação ao dinheiro (SOLOMON, 2004). O estilo de vida dos consumidores diz como a pessoa vive e gasta dinheiro. Além disso, ele define as escolhas por determinados produtos (KHAN, 2007) (KOTLER; ARMSTRONG, 2010).

- **Fatores psicológicos:** motivação, percepção, e crenças também afetam a decisão de compra. Um consumidor é um indivíduo que tem diferentes tipos de necessidades. Essas necessidades podem ser biológicas como a fome, ou psicológicas, decorrentes da necessidade de reconhecimento ou pertencimento. Uma necessidade pode ser despertada para um nível suficiente de intensidade quando altera um motivo. Um motivo é basicamente uma necessidade que leva uma pessoa a buscar satisfação (KOTLER; ARMSTRONG, 2010).

2.1.2 Etapas do processo de decisão de compra

A tomada de decisão é um processo sistemático, que pode ser representado por etapas pelas quais o consumidor passa antes da compra real de um produto. Na verdade, o processo inclui também o momento pós-compra, onde o consumidor avalia o nível de satisfação recebido (KOTLER; ARMSTRONG, 2010). A Figura 2 mostra as cinco etapas deste processo.

A primeira etapa consiste no reconhecimento do problema ou necessidade do consumidor, que pode ser acionado por estímulos internos ou externos. Os estímulos

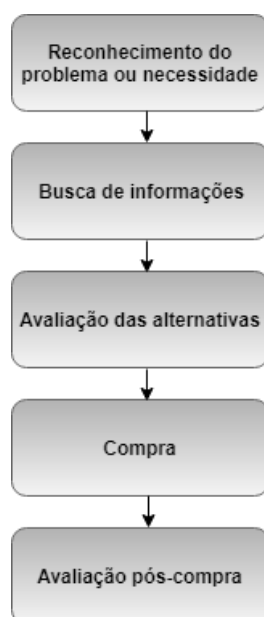


Figura 2 – Processo de decisão de compra: etapas

internos são as necessidades básicas de um ser humano, por exemplo, a fome que faz o consumidor comprar comida. Os estímulos externos incluem uma propaganda na TV ou a recomendação de amigos (KOTLER; ARMSTRONG, 2010).

Na segunda etapa o consumidor se envolve com a busca de informações. As informações podem ser obtidas a partir de várias fontes, como o boca a boca (do inglês: *Word-of-Mouth - WOM*) através da Internet, amigos ou familiares, mas também pode ser obtida experimentando o produto ou através de propagandas. Atualmente, os consumidores obtêm uma enorme quantidade de informações controladas por profissionais de marketing, porém, as fontes mais eficazes tendem a ser pessoais, como familiares ou amigos (KOTLER; ARMSTRONG, 2010).

Na terceira etapa, os consumidores devem avaliar as diferentes alternativas, que variam entre clientes e produtos (WRIGHT, 2006). Em alguns casos, os clientes fazem pouca ou nenhuma avaliação e tomam suas decisões de compra com base no impulso e na intuição. Por outro lado, os consumidores engajados em um processo extenso de resolução de problemas podem avaliar cuidadosamente entre várias marcas ou produtos. As alternativas que são ativamente consideradas durante o processo de seleção são conhecidas como conjunto evocado pelos consumidores, que, geralmente, inclui apenas um pequeno número de alternativas (SOLOMON, 2004).

Na quarta etapa, existem fatores que podem afetar a decisão de compra do consumidor, como as atitudes dos outros ou as crenças sobre a marca criadas pelos profissionais de marketing (KOTLER; ARMSTRONG, 2010). Às vezes, para simplificar a decisão, os consumidores podem criar atalhos mentais que podem levar a ações precipitadas. Esses tipos de atalhos podem variar de pensamentos universais como: “produtos de preço mais

alto são de maior qualidade”, “comprar a mesma marca que a última vez” ou “comprar a mesma marca que minha mãe costumava comprar”. No entanto, esses atalhos podem se revelar desfavoráveis ao consumidor (SOLOMON, 2004).

A última etapa do processo é a avaliação pós-compra. Após o consumidor ter adquirido o produto, será avaliado o nível de satisfação, que pode se referir ao produto em si, bem como à transação financeira ou o transporte. Se o consumidor sentir decepção, as expectativas em relação ao produto não foram atendidas. Se o produto atende às expectativas, o consumidor ficará satisfeito e disposto a gastar mais com essa marca específica no futuro (KHAN, 2007). Além disso, o alto nível de satisfação dos consumidores pode ser traduzido em lealdade à marca. Com a expansão de Redes Sociais Online (RSO), consumidores tendem a expressar suas experiências pós-compra na Internet através de *reviews* de produtos online, posts em *blogs* ou *vlogs*¹. Essas informações *online* podem servir como base para a etapa de Busca de Informações de futuros consumidores. No entanto, filtros devem ser utilizados para considerar apenas comentários sobre o produto.

2.1.3 O impacto de *reviews online* no Comércio Eletrônico

Dois tipos de informações sobre produtos estão disponíveis para um consumidor *online*: a informação criada pelo vendedor através de um canal de comunicação tradicional, como anúncios, e a informação baseada em *reviews online* criados por compradores anteriores com base em suas experiências, avaliações e opiniões de produtos.

Estudos anteriores sugerem que a informação criada pelo consumidor é mais credível do que a informação criada pelo vendedor na perspectiva da confiabilidade (DELLAROCAS, 2003). Quando os vendedores apresentam informações sobre o produto, eles tendem a ocultar aspectos inferiores de um produto e enfatizam os aspectos positivos. Os consumidores, ao contrário, fornecem avaliações honestas dos pontos fortes e fracos de um produto do ponto de vista do usuário. Assim, as revisões de consumidores *online* provavelmente serão mais confiáveis do que as informações criadas pelo vendedor (BICKART; SCHINDLER, 2001).

Reviews online em sites específicos, sites de empresas, blogs e comunidades influenciam várias etapas dos processos de tomada de decisão e compra do consumidor (SCHINDLER; BICKART, 2012). Chevalier e Mayzlin (2006) e Dellarocas, Zhang e Awad (2007) demonstraram que as análises positivas dos consumidores aumentam as vendas, enquanto os *reviews* negativos diminuem. No entanto, alguns pesquisadores descobriram que *reviews* negativos também podem aumentar as vendas (BERGER; SORENSEN; RASMUSSEN, 2010), principalmente porque os produtos analisados têm uma chance maior de pertencer ao conjunto a ser considerado pelos consumidores do que produtos que ainda não foram analisados.

¹ Um blog em que as postagens são feitas principalmente em formato de vídeo

Outro fator considerado em *reviews online* é o consenso entre opiniões positivas e negativas. De fato, se o consumidor percebe um baixo nível de consenso, acredita-se que ele pense que os autores de avaliações negativas são incapazes de usar ou avaliar o produto. No caso de um consenso maior sobre o lado negativo, o cliente desenvolverá inferências negativas em relação ao produto e à marca (LACZNIAK; DECARLO; RAMASWAMI, 2001).

Algumas pesquisas descobriram que comentários de clientes na web são de fato preditores de vendas, mas não as influenciam (CHEN; WU; YOON, 2004). Como forma de ampliar o poder do *eWOM*² como forma de *marketing*, algumas empresas utilizam outros recursos alinhados aos *reviews* escritos, como avaliação do produto sob forma de nota ou estrelas e um sistema de utilidade de *review* em que os próprios consumidores oferecem um *feedback* a respeito dos *reviews* lidos. Todos esses recursos são utilizados para ampliar o poder econômico do comércio eletrônico.

2.1.4 Tipos de Produtos

Um ponto importante em relação aos *reviews online* de produtos ou serviços é a natureza do produto. De acordo com alguns estudiosos, as recomendações na forma de *WOM* desempenham um papel significativo nos processos de tomada de decisão dos consumidores, especialmente quando lidam com **produtos intangíveis** em vez de lidar com os **tangíveis** (MURRAY; SCHLACTER, 1990). Produtos intangíveis não podem ser tocados, sentidos, cheirados e nem vistos, portanto, são particularmente complexos e difíceis de avaliar antes da compra e a percepção de risco é maior porque sua qualidade é muitas vezes desconhecida antes do consumo (MAURI; MINAZZI, 2013).

Nelson (1981) definiu que **bens ou produtos**, podem ser classificados como sendo de **experiência** ou de **busca**. **Bens de experiência** são aqueles dominados por atributos que não podem ser conhecidos até compra e utilização do produto ou para os quais busca de informação é mais cara ou difícil do que a experiência direta com o produto. Exemplos de bens de experiência incluem músicas, filmes, jogos, viagens e jantares. Em contrapartida, os **bens de busca** são aqueles dominados por atributos para o qual a informação completa pode ser adquirida antes da compra, como por exemplo computador, smartphone ou TV. Adicionalmente, Darby e Karni (1973) definiram uma terceira classificação para os bens, aos quais os consumidores não conseguem verificar seus atributos mesmo após a compra, chamados de **bens de crédito**. Exemplos incluem produtos em que é necessário conhecimento técnico, como equipamentos de laboratório. Ford, Smith e Swasy (1988) expandiram o conceito de bens de crédito como sendo aqueles que dependem do nível de conhecimento do consumidor.

² *eWOM* (eletronic Word-Of-Mouth) é qualquer declaração positiva ou negativa feita por clientes atuais ou antigos sobre um produto ou empresa, que é disponibilizada através da Internet

A Figura 3 se refere à facilidade de avaliação de diferentes tipos de bens. Nota-se que a maioria dos produtos é mais fácil de avaliar a partir de observação ou uso, o que não acontece com a maioria dos serviços.

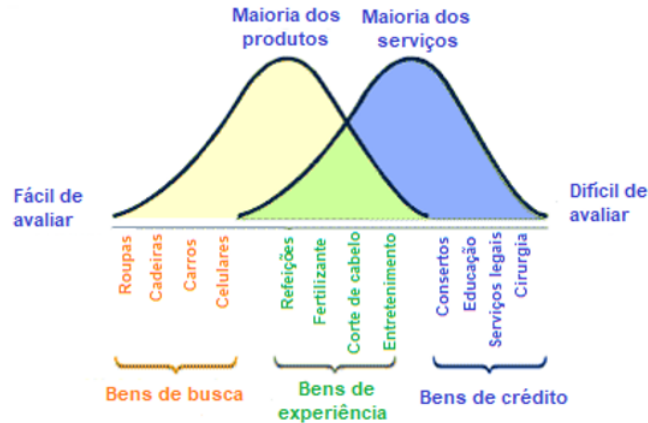


Figura 3 – Diferentes tipos de bens quanto à facilidade de avaliação

2.2 Percepção de utilidade de reviews *online*

A relação entre *reviews online* gerados por consumidores e a venda de produtos é um tema que desperta interesse em vários estudiosos (CHEN; WU; YOON, 2004; FORMAN; GHOSE; WIESENFELD, 2008; YE; LAW; GU, 2009; ZHU; ZHANG, 2010; HU; KOH; REDDY, 2014). Consequentemente, é importante entender quais *reviews* possuem maior impacto no processo de decisão de compra e determinar formas de prever esse comportamento.

Perguntas do tipo “Esse comentário foi útil para você?” têm sido utilizadas para incentivar consumidores a classificar *reviews* publicados por outros usuários.

O voto positivo representa que o consumidor leu e aprovou o conteúdo de um *review*. Assim, em tese, *reviews* mais bem avaliados aparecem primeiro nas listas e o consumidor precisa gastar menos tempo para encontrar as melhores recomendações. Financeiramente falando, o *site Amazon* identificou um aumento de 2,7 bilhões de dólares de lucro desde a introdução dessa questão ao fim de cada *review* (SPOOL, 2009).

Diversos trabalhos se propuseram a definir um conceito para utilidade de *reviews* de produtos. Em um sistema de votos como o proposto pelo *site Amazon*, a utilidade de um *review* pode ser definida como

$$H = \frac{n_p}{n_p + n_n} \quad (2.1)$$

onde n_p representa o número de votos positivos e n_n representa o número de votos negativos (KIM et al., 2006). Entretanto, essa abordagem apresenta algumas tendências desvantajosas como, por exemplo, um cenário em que *reviews* novos ou com poucos votos nunca serão

lidos e, conseqüentemente, votados, por não aparecerem na lista de mais úteis (LI et al., 2013). Além disso, nem todo mundo que lê um *review* necessariamente vota ao final. Na verdade, a quantidade de pessoas que votam *reviews online* é bastante inferior à quantidade de pessoas que compram o produto (KIM et al., 2006; SPOOL, 2009).

Outro problema dessa abordagem é que ela é muito simplista e ineficiente por não levar em consideração a quantidade de votos recebidos por cada *review*. É impossível, por exemplo, diferenciar dois *reviews* com 80% de aprovação, mesmo que um deles possua 10.000 votos e outro apenas 10. Por se tratar de uma distribuição Binomial, uma solução eficiente para contornar esse problema é o uso do limite inferior do intervalo de confiança de Wilson para uma Binomial (WILSON, 1927; AGRESTI; COULL, 1998). Ou seja, assumindo um determinado intervalo de confiança, estima-se a verdadeira fração de votos positivos. Essa equação foi definida por Edwin B. Wilson em 1927 e é apresentado a seguir (2).

$$H = \left(\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{[\hat{p}(1 - \hat{p}) + z_{\alpha/2}^2/4n]/n} \right) / (1 + z_{\alpha/2}^2/n) \quad (2.2)$$

Dada uma situação hipotética em que todos os *reviews* de um produto possam ser lidos e votados igualmente, ao calcular as proporções de votos positivos para cada um desses *reviews*, pode-se ter uma noção exata sobre quais são os *reviews* mais bem avaliados. Essa equação considera a quantidade de votos de um *review* como uma amostra de uma população hipotética em que todos os *reviews* são votados em igual quantidade. Assumindo um determinado nível de confiança (95% neste trabalho), é determinada uma nova probabilidade de sucesso para a distribuição analisada (\hat{p} representa a porcentagem de votos positivos observados, n é o total de votos recebidos e z é o quantil da distribuição normal padronizada para o intervalo de confiança escolhido).

Na prática, quanto mais votos, mais preciso é o resultado da equação. Por exemplo, um *review* com 12 votos positivos e 0 votos negativos (100% de percepção de utilidade se usarmos a equação 2.1) recebe um *score* de 0.757499 utilizando a equação 2.2, já um *review* com 784 votos positivos e 103 votos negativos (88,39% de percepção de utilidade pela equação 2.1) recebe 0.86112 de *score*, utilizando a equação 2.2. O sistema entende que o primeiro *review* teria 75,74% de votos positivos se tivesse uma quantidade grande de votos, e com nível de confiança de 95%, há apenas 5% de chance de ele estar errado. Já o segundo *review* possui votos suficientes para que os dois resultados sejam bem mais próximos (88,39% e 86,11%), demonstrando a eficiência desse cálculo. É interessante notar que esse cálculo já é utilizado em alguns *sites* no processo de ordenação de *reviews*, tópicos e comentários, como *Reddit*³(MUNROE, 2009), *Yelp*⁴ (YELP, 2011) e *Digg*⁵.

³ <https://www.reddit.com/>

⁴ <http://www.yelp.com/>

⁵ <http://digg.com/>

Neste trabalho, sempre que nos referirmos à medida de utilidade de *reviews online*, estamos falando da porcentagem de votos positivos de um *review* com base na equação 2.2. Formalmente, entendemos a medida de utilidade de *reviews* de produtos *online* como a medida na qual consumidores percebem um *review* de produto como sendo capaz de facilitar o julgamento ou decisão de compra (LI et al., 2013).

2.3 Aprendizado de Máquina

O aprendizado de máquina é uma aplicação da inteligência artificial (IA) que fornece aos sistemas a capacidade de aprender e melhorar automaticamente a partir da experiência, sem ser explicitamente programado. O aprendizado de máquina se concentra no desenvolvimento de programas que podem acessar dados e usá-los para aprender por si mesmos. O objetivo principal é permitir que os computadores aprendam automaticamente sem intervenção humana ou assistência e ajustar as ações de acordo.

Nos últimos anos, o Aprendizado de Máquina se tornou um dos pilares da tecnologia da informação e, conseqüentemente, uma parte central da nossa vida. Com o aumento da quantidade de dados gerada diariamente, existem boas razões para acreditar que o uso inteligente de análise de dados se tornará ainda mais difundido como um ingrediente necessário para o progresso tecnológico.

Nesta seção será apresentada uma visão geral sobre o Aprendizado de Máquina, incluindo aplicações e tipos de problema. Eventualmente nos aprofundaremos nos modelos utilizados neste trabalho.

2.3.1 Aplicações e problemas

A aprendizagem de máquina se mostrou uma área fértil de pesquisa, produzindo uma série de problemas tais como regressão, classificação, agrupamento e predição (MITCHELL, 1997). Os métodos de aprendizagem de máquina podem ser divididos em **supervisionados** ou **não supervisionados**.

No aprendizado supervisionado é fornecido ao algoritmo um conjunto de dados rotulados aos quais já são conhecidas as saídas corretas. A partir desse conjunto de dados, os algoritmos de aprendizagem supervisionada têm como objetivo encontrar um modelo que generalize uma função capaz de prever a saída para dados de entrada desconhecidos.

Já a aprendizagem não supervisionada nos permite abordar problemas com pouca ou nenhuma ideia do que nossos resultados devem ser, já que os dados de entrada não são rotulados. Estão entre os possíveis objetivos desses algoritmos agrupar os dados com base em relações entre eles; reduzir o número de dimensões em um conjunto de dados para concentrar somente nos atributos mais úteis; e detectar tendências nos dados.

Existem ainda os métodos de aprendizagem *semi-supervisionada* e *aprendizagem por reforço*. Mais detalhes sobre esses métodos podem ser lidos em (MITCHELL, 1997).

A gama de problemas de Aprendizagem de Máquina é enorme. Por esse motivo os pesquisadores identificaram um número crescente de modelos que podem ser usados para endereçar um grande conjunto de situações. São esses modelos que facilitam a implantação de aprendizado de máquina na prática. A seguir são apresentados alguns dos mais comuns métodos de aprendizado de máquina, destacando suas diferenças e as situações onde são mais adequados:

- **Classificação:** provavelmente o problema mais estudado em aprendizado de máquina e que levou a um grande número de algoritmos importantes e desenvolvimentos teóricos ao longo do século passado. Na classificação, um conjunto de variáveis independentes X é usado para prever o valor que uma variável aleatória discreta y ($y \in \pm 1$) assumirá. Dentre alguns algoritmos de classificação estão as máquinas de vetores de suporte (do inglês: Support Vector Machine - SVM), Árvores de Decisão, Naive Bayes, Regressão Logística e Redes Neurais Artificiais como o Perceptron. Classificar uma notícia como *fake news* ou não é um dos grandes problemas atuais de classificação;
- **Regressão:** tem como objetivo estimar uma variável de valor real (contínuo) y ($y \in \mathbb{R}$) dado um conjunto de variáveis independentes X . Por exemplo, podemos querer estimar a temperatura de uma cidade nos próximos dias utilizando registros anteriores. Dentre alguns algoritmos de regressão estão o LASSO, Regressão Bayesiana e Redes Neurais como Perceptron;
- **Agrupamento:** é a tarefa de particionar o conjunto de dados em grupos, chamados agrupamentos. O objetivo é dividir os dados de forma que os pontos dentro de um único agrupamento sejam muito semelhantes e os pontos em agrupamentos diferentes sejam diferentes. Ele determina o agrupamento entre dados não rotulados. Um exemplo de agrupamento poderia ser determinar o tópico de um texto em um Corpus baseado em seu conteúdo textual. Dentre alguns algoritmos de agrupamento estão o K-Means, Propagação de Afinidades e Mean Shift.

2.3.2 Redes Neurais Artificiais

Redes Neurais Artificiais (RNA) são modelos computacionais inspirados no sistema nervoso de seres vivos. Elas possuem a capacidade de aquisição e manutenção do conhecimento (baseado em informações) e podem ser definidas como um conjunto de unidades de processamento, denominadas por neurônios artificiais, que são interligados por um grande número de interconexões, chamadas de sinapses artificiais, as quais são representadas por vetores ou matrizes de pesos sinápticos (SILVA; CARVALHO; SARMENTO, 2012).

As RNAs foram desenvolvidas a partir de modelos matemáticos e modelos de engenharia de neurônios biológicos. Como qualquer célula biológica, o neurônio é delimitado por uma fina membrana celular que além da sua função biológica normal, é essencial para o funcionamento elétrico da célula nervosa. A partir do corpo celular, ou soma, projetam-se extensões filamentosas, os dendritos e o axônio. Entende-se o neurônio biológico como sendo basicamente o dispositivo computacional elementar do sistema nervoso, que possui entradas e saídas. As entradas ocorrem a partir das conexões sinápticas, que conectam a árvore dendritral aos axônios de outras células nervosas. Os sinais que chegam por estes axônios constituem a informação que o neurônio processará para produzir como saída um impulso nervoso no seu axônio (KOVÁCS, 2002).

Os neurônios artificiais utilizados nos modelos de RNAs não-lineares fornecem saídas tipicamente contínuas e realizam funções simples, como coletar os sinais existentes em suas entradas, agregá-los de acordo com sua função operacional e produzir uma resposta, levando em consideração sua função de ativação inerente (SILVA; CARVALHO; SARMENTO, 2012). A Figura 4 mostra um modelo de neurônio artificial.

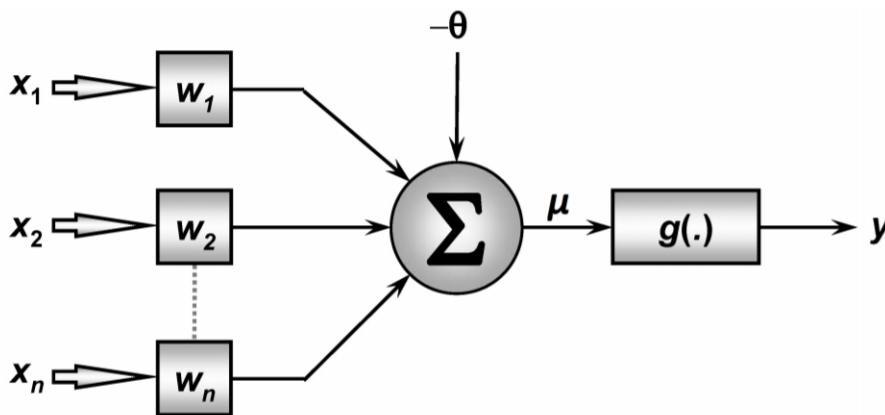


Figura 4 – Modelo de Neurônio Artificial (SILVA; CARVALHO; SARMENTO, 2012)

O funcionamento de um neurônio artificial começa pela apresentação dos sinais de entrada representados pelo conjunto $\{x_1, x_2, \dots, x_n\}$ que são ponderados por meio do conjunto de pesos sinápticos $\{w_1, w_2, \dots, w_n\}$. Depois é obtido o potencial de ativação μ produzido pela soma ponderada dos sinais de entrada através do combinador linear Σ , subtraindo-se ao limiar de ativação θ . Com o potencial de ativação calculado, é aplicada a função de ativação $g(\cdot)$ apropriada, tendo-se como objetivo limitar a saída do neurônio. Finalmente é convertida a saída y a partir da aplicação da função de ativação neural em relação ao seu potencial de ativação.

Uma rede neural artificial é representada matematicamente como uma tripla (Ω, y, w) , contendo dois conjuntos (Ω, y) e uma função (w) , onde Ω é o conjunto de neurônios e y é um conjunto $\{(i, j) | i, j \in N\}$ cujo elementos são chamados de conexões entre o neurônio

i e o neurônio j . A função $w : y \rightarrow R$ define os pesos sinápticos, onde $w((i, j))$ é o peso da conexão entre o neurônio i e o neurônio j , o qual é reduzido para w_{ij} (KRIESEL, 2007).

Uma RNA pode ser dividida em três camadas: camada de entrada, camadas escondidas (ou intermediárias, ocultas ou invisíveis) e camada de saída. A camada de entrada é responsável pelo recebimento das informações externas. As camadas escondidas são compostas por neurônios que extraem as características associadas ao processo ou sistema inferido. Quase todo o processamento interno da rede é realizado nessas camadas. Por fim, a camada de saída é responsável pela produção e apresentação dos resultados finais da rede, resultante dos processamentos efetuados pelos neurônios das camadas anteriores. As funções de ativação $g(\cdot)$ são utilizadas nas camadas escondidas e camadas de saída com o objetivo de limitar a saída do neurônio dentro de um intervalo de valores razoáveis a serem assumidos pela sua própria imagem funcional.

A arquitetura de uma RNA define a forma como os seus diversos neurônios estão arranjados ou dispostos, uns em relação aos outros. Elas podem ser de quatro tipos: *feedforward* de camada simples, *feedforward* de múltiplas camadas, recorrentes ou reticuladas. Entende-se por rede *feedforward* àquela onde o fluxo de informações segue sempre em uma única direção, ou seja, da camada de entrada para a camada de saída. Uma RNA com arquitetura *feedforward* de camada simples tem apenas uma camada de entrada e uma única camada de neurônios, que é a camada de saída. Já uma RNA com arquitetura *feedforward* de múltiplas camadas contém pelo menos uma camada escondida de neurônios. A RNA com arquitetura recorrente tem suas saídas realimentadas como sinais de entrada para outros neurônios. Finalmente, uma RNA com arquitetura reticulada contém a localização espacial dos neurônios diretamente relacionada com o processo de ajuste de seus pesos e limiares. A Figura 5 apresenta a representação gráfica desses quatro tipos de RNAs.

Já a topologia de uma RNA, considerando determinada arquitetura, pode ser definida como sendo as diferentes formas de composições estruturais que esta pode assumir. O treinamento de uma arquitetura específica consiste da aplicação de um conjunto de passos ordenados com o intuito de ajustar os pesos e os limiares de seus neurônios. Dessa forma, tal processo de ajuste (ou algoritmo de aprendizagem) visa sintonizar a rede para que as suas respostas estejam próximas dos valores desejados.

Entre os principais tipos de redes tendo arquitetura *feedforward* de camada simples estão o *Perceptron* (ROSENBLATT, 1958) e o *Adaline* (WIDROW; HOFF, 1988), cujos algoritmos de aprendizado utilizados em seus treinamentos são os baseados na regra de Hebb (HEBB, 1949) e na regra Delta (WIDROW; HOFF, 1988). Em relação às redes com arquitetura *feedforward* de múltiplas camadas, os principais tipos de redes são o *Perceptron* Multicamadas (do inglês: *Multilayer Perceptron*, MLP) e as redes de base radial (do inglês: *Radial Basis Function*, RBF) (BROOMHEAD; LOWE, 1988). Já as principais redes com arquitetura recorrente são as redes de Hopfield (HOPFIELD, 1988) e

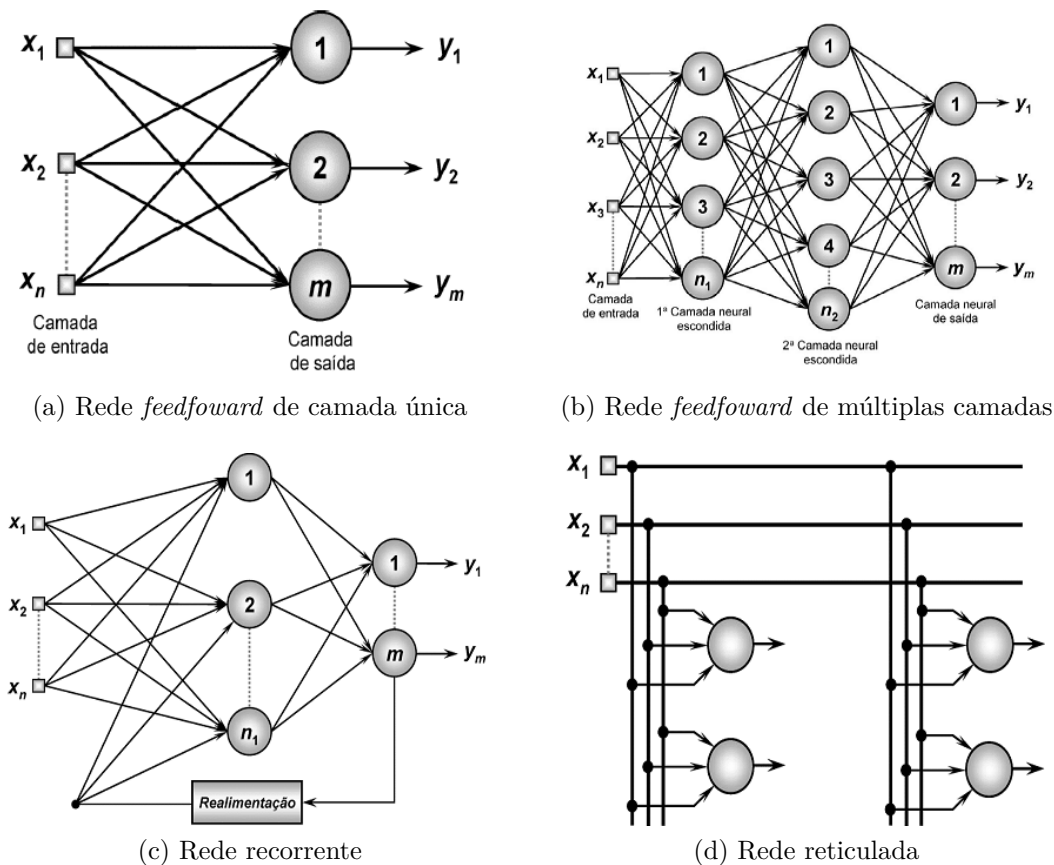


Figura 5 – Principais arquiteturas de RNAs (SILVA; CARVALHO; SARMENTO, 2012)

a rede *Perceptron* com realimentação. Por fim, a rede de Kohonen (KOHONEN, 1988) é a principal representante das redes com arquitetura reticulada.

2.3.2.1 Rede Neural Artificial Perceptron Multicamadas (RNA MLP)

Hornik, Stinchcombe e White (1989) demonstraram que Redes Neurais Artificiais Perceptron Multicamadas (RNA MLP) podem ser treinadas para aproximar virtualmente qualquer função mensurável. Diferente de outras técnicas estatísticas, o MLP não faz suposições prévias sobre a distribuição dos dados. Ele pode modelar funções altamente não-lineares e pode ser treinado para generalizar com precisão quando apresentado com dados novos e não vistos. Essas características do Perceptron Multicamadas fazem dele uma alternativa atraente para o desenvolvimento de modelos numéricos e também quando se escolhe entre abordagens estatísticas (GARDNER; DORLING, 1998).

Redes Neurais Artificiais MLP consistem de unidades organizadas em camadas. Cada camada é composta de neurônios e, em uma rede totalmente conectada, cada neurônio está conectado a todos os neurônios da camada subsequente. Uma RNA MLP é composta de, no mínimo, três camadas, consistindo de uma camada de entrada, uma ou mais camadas escondidas e uma camada de saída (DELASHMIT; MANRY, 2005).

A camada de entrada de uma RNA MLP distribui as entradas para as camadas subsequentes. Os neurônios da camada de entrada possuem funções de ativação linear e não possuem limiar - entrada de controle que determina a mudança dos estados do neurônio. Todos os neurônios das camadas subsequentes possuem limiares associados a eles, em adição aos pesos. Os neurônios da camada escondida possuem funções de ativação não-lineares e os neurônios da camada de saída possuem funções de ativação lineares. Assim, cada sinal de alimentação em um neurônio em uma camada subsequente tem a entrada original multiplicada por um peso com um limiar adicionado que, em seguida, é passado através de uma função de ativação que pode ser linear ou não-linear (unidades ocultas). Uma típica Rede Neural Artificial Perceptron Multicamadas de três camadas é apresentada na Figura 6.

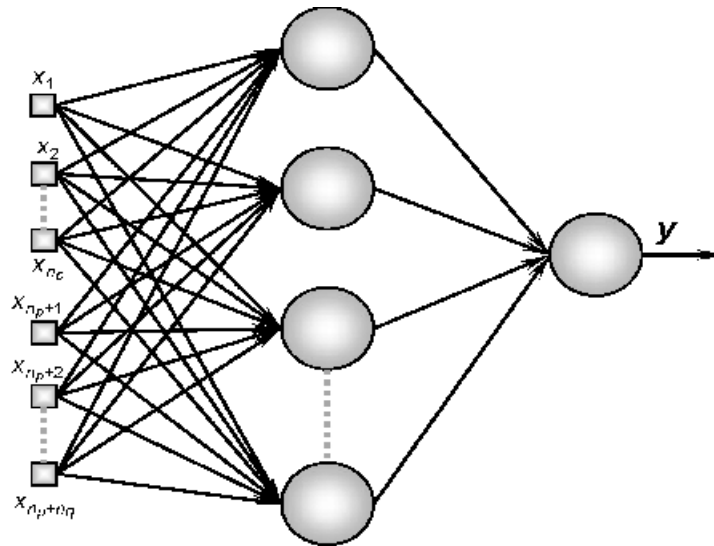


Figura 6 – Típica RNA MLP de três camadas (SILVA; CARVALHO; SARMENTO, 2012)

Cada neurônio j na camada escondida soma seus sinais de entrada x_i , os quais foram multiplicados por seus respectivos pesos de conexão w_{ij} (ORHAN; HEKIM; OZER, 2011). A saída de cada neurônio é descrita como:

$$y_j = f\left(\sum w_{ij}x_i\right) \quad (2.3)$$

onde f é uma função de ativação, que pode ser um simples limiar, função sigmóide ou função tangente hiperbólica (ORHAN; HEKIM; OZER, 2011).

2.3.2.2 Rede MLP como aproximador universal de funções

A RNA MLP para aproximação de funções é composta de três camadas, sendo uma camada de entrada, uma camada escondida e uma camada de saída. A camada escondida processa e transmite a informação de entrada para a camada de saída. Um modelo de Rede Neural MLP com um número insuficiente ou excessivo de neurônios

na camada escondida poderá causar problemas de generalização e *overfitting* (ORHAN; HEKIM; OZER, 2011). A validação cruzada permite definir, através de tentativa e erro, a quantidade mais adequada de neurônios na camada escondida, porém não há método analítico para determinar o número exato (ORHAN; HEKIM; OZER, 2011). Normalmente, utiliza-se a função logística como ativação para os neurônios da camada escondida. Já para a camada de saída, utiliza-se uma função de ativação linear, já que o neurônio de saída realiza apenas uma combinação linear das funções de ativação logística implementadas nos neurônios da camada escondida (SILVA; SPATTI; FLAUZINO, 2010).

Uma RNA MLP para aproximação de funções pode ser descrita da seguinte forma:

$$y(x_1, x_2, \dots, x_n) = \sum_{i=1}^{n_1} \lambda_i g_j(u_i) \quad (2.4)$$

$$u_i = \sum_{j=1}^n W_{ij} x_j - \theta_i \quad (2.5)$$

onde a função y a ser mapeada pelo MLP será constituída pela superposição de funções de ativação do tipo logística, representadas pelos termos $g_i(u_i)$, as quais são ponderadas pelos fatores λ_i (SILVA; SPATTI; FLAUZINO, 2010). A função logística é definida como

$$g(u) = \frac{1}{1 + e^{-\beta * u}} \quad (2.6)$$

onde β é uma constante real associada ao nível de inclinação da função logística frente ao seu ponto de inflexão.

O teorema da aproximação universal aplicado ao MLP fornece as bases necessárias para definir as configurações estruturais dessas redes com a finalidade de mapear funções algébricas (SILVA; SPATTI; FLAUZINO, 2010). Além disso, diferentemente de modelos estatísticos, o modelo de RNA permite que qualquer forma funcional que relacione as variáveis independentes à variável dependente não precise ser imposto pelo modelo, ou seja, é possível capturar relacionamentos não-lineares nos dados sem a necessidade de especificação prévia (LEE; CHOE, 2014).

2.3.2.3 Redes de funções de base radial (RBF)

As funções de base radial, conhecidas como RBF, também podem ser empregadas em problemas tratados pelo MLP, como a aproximação de funções e classificação. A figura 7 apresenta uma configuração típica de rede RBF.

Mas diferente das redes MLP, que podem ser compostas de diferentes camadas intermediárias, a estrutura típica de uma rede RBF é composta por apenas uma, na qual

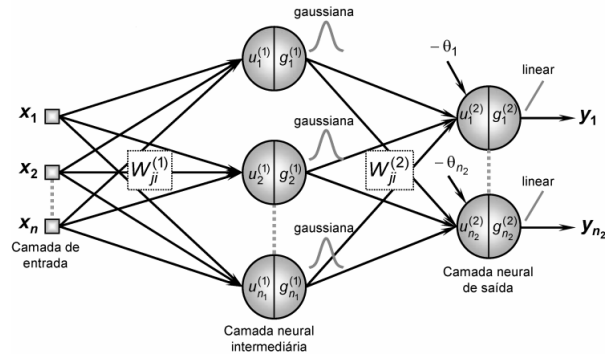


Figura 7 – Configuração típica da rede RBF (SILVA; SPATTI; FLAUZINO, 2010)

as funções de ativação são do tipo gaussiana, como demonstrado na Equação 2.7

$$g(u) = e^{-\frac{(u-c)^2}{2\sigma^2}} \quad (2.7)$$

onde c define o centro da função gaussiana e σ^2 denota a sua variância, a qual indica o quão disperso está o potencial de ativação u em relação ao centro c (SILVA; SPATTI; FLAUZINO, 2010). A Figura 8 ilustra o formato geométrico de uma função gaussiana típica.

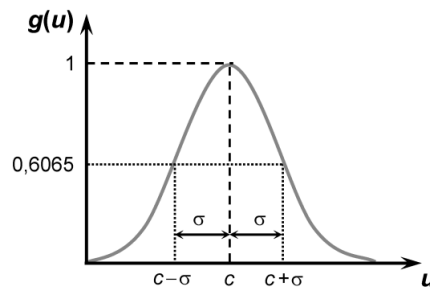


Figura 8 – Função de base radial do tipo gaussiana (SILVA; SPATTI; FLAUZINO, 2010)

O princípio de funcionamento das redes RBF também é similar àqueles do MLP, onde cada uma de suas entradas x_i será propagada pela camada intermediária em direção à camada de saída. Entretanto, a estratégia de treinamento da RBF é constituída de dois estágios bem distintos entre si. O primeiro estágio refere-se ao ajuste dos pesos dos neurônios da camada intermediária, que adota um método de aprendizagem não-supervisionado e que depende apenas das características dos dados de entrada. Já o segundo estágio está vinculado ao ajuste dos pesos dos neurônios da camada de saída e utiliza um critério de aprendizagem similar àquele usado na última camada do MLP, ou seja, a regra delta generalizada (SILVA; SPATTI; FLAUZINO, 2010).

Considerando a expressão fornecida em 2.8, os parâmetros livres a serem ajustados seriam então o centro c e a variância σ^2 . Dessa forma, a entrada $u_j^{(1)}$ de cada neurônio será

o próprio vetor de entrada x , que representa os n sinais externos que são aplicados na rede. Consequentemente, a saída de cada neurônio j da camada intermediária é expressa por:

$$g_j^{(1)}(u_j^1) = g_j^{(1)}(\mathbf{x}) = e^{-\frac{\sum_{i=1}^n (x_i - w_{ji}^{(1)})^2}{2\sigma_j^2}} \quad (2.8)$$

onde $j = 1, \dots, n_1$

A aplicação dos passos de ajuste referentes aos pesos dos neurônios da camada de saída deve ser executada, em sequência, somente após a finalização do primeiro estágio de treinamento. Desta forma, o segundo estágio de treinamento é efetivado utilizando os mesmos procedimentos usados para a camada de saída do MLP.

Dessa forma, o conjunto de treinamento para os neurônios da camada de saída será então constituído por pares de entrada e saída desejada, em que as entradas serão as respostas produzidas pelas funções de ativação gaussianas dos neurônios da camada intermediária frente às respectivas amostras de treinamento, ou seja:

$$(u_j^2) = \sum_{i=1}^{n_1} W_{ji}^{(2)} \cdot g_j^{(1)}(u_i^1) - \theta_j \quad (2.9)$$

onde $j = 1, \dots, n_2$ e $W_{ji}^{(2)}$ e θ são respectivamente os pesos e limiares referentes aos neurônios da camada de saída.

Finalmente, assim como ocorre no MLP, tem-se que os neurônios de saída realizarão somente uma combinação linear das funções de ativação gaussianas produzidas pelos neurônios da camada anterior. Assim, as respostas produzidas pelos neurônios de saída serão dadas por:

$$y_j = g_j^{(2)}(u_j^2) = u_j^2 \quad (2.10)$$

onde $j = 1, \dots, n_2$

2.3.2.4 Redes RBF como aproximadores universais de funções

Park & Sandberg (1957) demonstraram que redes RBF são consideradas aproximadores universais de funções, assim como as redes MLP. De fato, substituindo a expressão 2.9 em 2.10, obtém-se o seguinte relacionamento:

$$y_j = \sum_{i=1}^{n_1} W_{ji}^{(2)} \cdot g_i^{(1)}(u_j^1) - \theta \quad (2.11)$$

onde $j = 1, \dots, n_2$

Através da Figura 7, pode-se concluir que a função y_j a ser mapeada pela RBF é constituída por uma superposição de funções de ativação gaussianas, representadas pelos termos $g_i^{(1)}(u_i^{(1)})$, as quais serão linearmente ponderadas pelos pesos $W_{ji}^{(2)}$ associados aos neurônios de sua camada de saída (SILVA; SPATTI; FLAUZINO, 2010).

A Figura 9 ilustra a RBF a ser utilizada no mapeamento do problema cuja camada escondida é constituída de três neurônios.

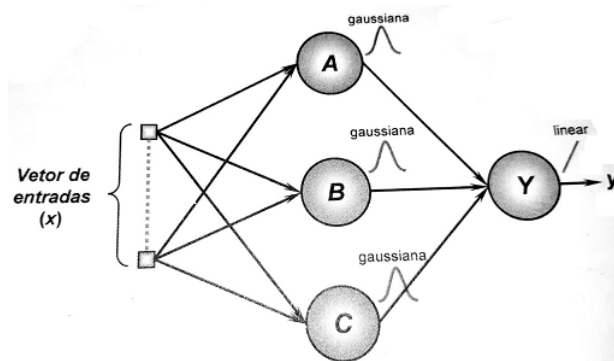


Figura 9 – Topologia de RBF aplicada em aproximação de funções(SILVA; SPATTI; FLAUZINO, 2010)

Assim como ocorrido com a rede MLP, uma RBF pode ser utilizada para mapear qualquer função não-linear contínua, definida num domínio fechado. Entretanto, a quantidade exata de neurônios em sua camada escondida ainda é motivo de pesquisa, pois também depende da complexidade do problema a ser mapeado. Por conseguinte, a técnica de validação cruzada pode também ser utilizada na RBF como estratégia para selecionar a topologia mais apropriada.

2.3.2.5 Escolha de topologias das redes neurais artificiais

A especificação da topologia de uma rede neural artificial mais apropriada para mapear um problema específico é usualmente efetuada de forma empírica (SILVA; SPATTI; FLAUZINO, 2010). O processo de seleção de topologias candidatas geralmente se dá por tentativa e erro (LEE; CHOEH, 2014). Para selecionar a melhor topologia dentre as candidatas, pode-se utilizar a técnica de validação cruzada *k-fold*. Nesse processo, a amostra total é dividida em k partições, sendo $(k - 1)$ delas usadas para compor o subconjunto de treinamento enquanto a partição restante constitui o subconjunto de teste (SILVA; SPATTI; FLAUZINO, 2010). O processo de aprendizado se repete k vezes até que todas as partições tenham sido utilizadas como subconjunto de teste.

2.3.2.6 Métricas de valiação dos modelos RNA MLP e RBF para regressão

Após a construção dos modelos de regressão, existem alguns critérios pelos quais eles podem ser avaliados e comparados. Entre os mais utilizados estão o **Erro Absoluto**

Médio (do inglês *Mean absolute error*, ou MAE), o **Erro Quadrático Médio** (do inglês *Root Mean Square Error*, ou RMSE) e o **Coefficiente de Determinação** (do inglês *R Squared*, ou R^2). Essas métricas levam em consideração o erro da saída do modelo para determinar o quão eficiente esse modelo é na predição de entradas futuras.

- **MAE**: média da diferença absoluta entre o valores previstos e valor observado. O MAE é uma pontuação linear, o que significa que todas as diferenças individuais são ponderadas igualmente na média. Se \hat{y}_i é o valor previsto da i -ésima amostra, e y_i é o valor verdadeiro correspondente, então o MAE é calculado usando a expressão

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.12)$$

- **RMSE**: representa o desvio padrão da amostra das diferenças entre os valores previstos e os valores observados, chamados de resíduos. Resíduos representam a medida de distância dos pontos de dados da linha de regressão. O cálculo do RMSE é uma medida de como se espalham esses resíduos. Em outras palavras, ele informa como os dados estão concentrados em torno da linha de melhor ajuste. Se \hat{y}_i é o valor previsto da i -ésima amostra, e y_i é o valor verdadeiro correspondente, então o RMSE é calculado usando a expressão

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.13)$$

- R^2 : frequentemente usado para fins explicativos, essa métrica demonstra o quanto as variáveis independentes selecionadas explicam a variabilidade em sua variável dependente. Essa métrica fornece uma medida de quão bem as amostras futuras provavelmente serão previstas pelo modelo. A melhor pontuação possível é 1 e pode ser negativa (porque o modelo pode ser arbitrariamente pior do que escolha aleatória). Um modelo constante que sempre prevê o mesmo valor esperado de y , desconsiderando os recursos de entrada, obteria uma pontuação R^2 de 0. Se \hat{y}_i é o valor previsto da i -ésima amostra, e y_i é o valor verdadeiro correspondente, então o R^2 é calculado usando a expressão

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (2.14)$$

onde

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.15)$$

2.3.2.7 Validação dos modelos RNA MLP e RBF

Pode-se calcular o relacionamento entre cada variável de entrada e a variável de saída utilizando o cálculo de *Relative Strength* (RS) (LEE; CHOEH, 2014) (YOON;

GUIMARAES; SWALES, 1994):

$$RS_{ji} = \frac{\sum_{k=0}^n (W_{ki}W_{jk})}{\sum_{i=0}^m \sum_{k=0}^n |W_{ki}W_{jk}|} \quad (2.16)$$

onde W_{ki} indica o peso entre a k -ésima unidade escondida e a i -ésima unidade de entrada. W_{jk} denota o peso entre a j -ésima unidade de saída e a k -ésima unidade escondida. RS_{ji} calcula o peso relativo entre a i -ésima variável de entrada e a j -ésima variável de saída. O numerador do cálculo de RS_{ji} calcula a relação entre a i -ésima variável de entrada e a j -ésima variável de saída e pode ser tanto positivo quanto negativo, dependendo dos pesos. Já o denominador calcula o relacionamento total entre todas as variáveis de entrada e saída.

Para escolher e validar as melhores entradas de uma RNA, pode-se utilizar o algoritmo Correlation-based Feature Selection (CFS) (HALL, 1998). O CFS usa a abordagem de avaliação de subconjunto. As *features* mais adequados são selecionadas levando em consideração a correlação entre todas as *features* do modelo. Um bom modelo possui *features* que não estão correlacionadas entre si, mas estão correlacionadas com a classe. *Features* com alta correlação com a variável de classe são altamente relevantes. A função de avaliação do subconjunto de *features* do CFS é descrita da seguinte forma:

$$M_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \quad (2.17)$$

onde M_s é o “mérito” heurístico de um subconjunto de *features* S contendo k *features*, \bar{r}_{cf} é a correlação média de *feature*-classe ($f \in S$) e \bar{r}_{ff} é a intercorrelação média de *feature*-*feature* (HALL, 1998). O numerador da Equação 2.6 fornece uma indicação de quão preditiva da classe é um conjunto de *features*; o denominador é uma indicação de quanta redundância existe entre as *features* (HALL, 1998).

2.4 Mineração de Opiniões

Mineração de Opiniões é uma sub-área do Processamento de Linguagem Natural que consiste no estudo computacional de opiniões, avaliações, atitudes e emoções das pessoas referenciando entidades, indivíduos, questões, eventos, tópicos e seus atributos (JACKSON; MOULINIER, 2007 apud SANTOS, 2017). O estudo é realizado sobre a emoção do usuário em relação ao item, atribuindo uma classificação a essa opinião.

Independente do tipo de texto analisado, toda opinião é composta de pelo menos dois elementos chave: um **alvo** e um **sentimento** sobre este alvo (LIU, 2012). O alvo pode ser uma entidade, aspecto de uma entidade, ou tópico, representando um produto,

pessoa, organização, marca, evento, etc. Já o sentimento representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. A polaridade de um sentimento corresponde a um ponto em alguma escala que representa a avaliação positiva, neutra ou negativa do significado deste sentimento (BECKER; TUMITAN, 2013).

Liu (2012) define opiniões como sendo de dois tipos: **regulares** e **comparativas**. Uma opinião é regular quando o autor da opinião expressa um sentimento, atitude, emoção ou percepção sobre um alvo. As opiniões comparativas expressam o sentimento baseadas na relação de similaridades ou diferenças entre duas ou mais entidades ou quando há algum aspecto compartilhado.

Formalmente, Liu (2012) define Opinião Regular como sendo uma quintupla do tipo $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, onde:

- e_i é a entidade alvo da opinião (ou sentimento);
- a_{ij} é o aspecto da entidade e_i opinado;
- s_{ijkl} é o sentimento sobre o aspecto a_{ij} ;
- h_k é o autor da opinião;
- t_l é a data ou instante em que a opinião foi expressa.

A Figura 10 apresenta um exemplo de *review* e a extração de uma das opiniões presentes nele.



Figura 10 – Exemplo de extração de opiniões em um *review online*

A existência de todos os componentes da quintupla é importante, uma vez que a falta de algum deles pode trazer problemas. Sem identificar a entidade de uma opinião, por exemplo, pode haver perda de informação, uma vez que várias entidades diferentes podem ter o mesmo aspecto (por exemplo, o aspecto tela pode se referir a uma TV ou a um smartphone). Em complemento, todos os componentes da quintupla devem corresponder um ao outro, por exemplo, o sentimento s_{ijkl} deve ser determinado pelo detentor da opinião

h_k sobre o aspecto a_{ij} da entidade e_i no instante t_l , no qual qualquer incompatibilidade torna-se um erro. Para as opiniões comparativas não se aplicam as mesmas definições.

Segundo Liu, Wu e Yao (2006), opiniões regulares podem ainda ser de duas subcategorias: **diretas** e **indiretas**. Opiniões regulares são chamadas de diretas quando há a referência direta a uma entidade ou um aspecto da entidade, como por exemplo na sentença “Este remédio é muito bom” (sentimento positivo). No caso das opiniões regulares indiretas, a opinião é expressa de forma indireta sobre uma entidade ou aspecto da entidade, como por exemplo em “Depois de tomar este remédio, minha pressão sanguínea subiu”. Neste caso, pode-se inferir que o remédio é ruim (sentimento negativo).

Além disso, Liu, Wu e Yao (2006) sugerem também que opiniões podem ser vistas como **explícitas** e **implícitas**. Nas opiniões explícitas o sentimento está escrito no texto, como por exemplo em “Este jogo é ótimo” (sentimento positivo explícito). Já nas opiniões implícitas o sentimento está expresso indiretamente, como por exemplo em “Este jogo travou meu computador” (sentimento negativo implícito).

2.4.1 Níveis de análise textual

A mineração de opiniões pode ocorrer a diferentes níveis. Liu (2012) define esses níveis como sendo de:

- **Documento:** o objetivo é identificar, como positivo ou negativo, o sentimento expresso por todo o documento. Pode ser utilizado ao minerar opiniões de documentos que se referem a uma única entidade;
- **Sentença:** em documento com mais de uma entidade, pode-se determinar o sentimento de cada sentença ou de sentenças específicas;
- **Entidade ou Aspecto:** um complexo nível de análise que tem como tarefa identificar opiniões a respeito de entidades ou aspectos específicos. É um dos grandes desafios de PLN graças a dificuldade de reconhecimento de entidades nomeadas, e à presença de figuras de linguagens como anáforas, sarcasmo e ironia (LIU, 2010).

2.4.2 Etapas do processo de mineração de opiniões

O processo de mineração de opiniões pode ser caracterizado em três grandes etapas: i) **identificação**; ii) **classificação**; e iii) **sumarização** (TSYTSARAU; PALPANAS, 2012).

A etapa de identificação consiste em encontrar as entidades e aspectos existentes e possivelmente associá-los com o respectivo conteúdo subjetivo. Na etapa de classificação é atribuída ao dado conteúdo subjetivo uma das três classes de sentimento: positivo, negativo

ou neutro. Por fim, a etapa de sumarização cria métricas que representam o sentimento geral, as quais podem ser visualizadas ou servir de entrada para outras aplicações. A Figura 11 apresenta uma visão geral do processo de mineração de opiniões.

A complexidade da identificação do alvo da opinião depende em muito da mídia considerada, e de seu grau de estruturação. A aplicação mais frequente em mineração de opiniões é a de *reviews* de produtos e serviços, porque o alvo pode ser mais facilmente identificado. Para esse tipo de documento, é possível assumir que todo o conteúdo refere-se a uma única entidade, o produto ou serviço sendo avaliado, restando apenas identificar os aspectos desta entidade, se a análise for nesta granularidade (BECKER; TUMITAN, 2013).

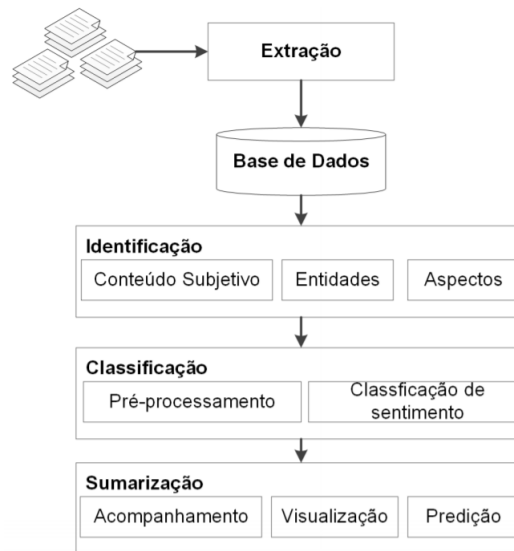


Figura 11 – Etapas da Mineração de Opiniões (BECKER; TUMITAN, 2013)

Para mais detalhes sobre o processo de Mineração de Opiniões, recomenda-se a leitura de Liu, Wu e Yao (2006), Liu (2012) e Becker e Tumitan (2013).

2.4.3 Modelo de *bag-of-words*

A representação computacional de documentos pode ser feita utilizando um modelo chamado de *bag-of-words*. No modelo de *bag-of-words*, cada documento é representado como um vetor das palavras que ocorrem no documento, frases ou sentenças.

A abordagem *bag-of-words* pode utilizar o formato de uma tabela atributo-valor, como representada na Figura 12. As linhas representam os d_i documentos e as colunas t_j representam os termos presentes no Corpus. Dessa forma, cada documento d_i é um vetor $d_i = (a_{i1}, a_{i2}, \dots, a_{iM})$, no qual o valor a_{ij} refere-se ao valor associado ao j -ésimo termo do documento i . O valor a_{ij} pode ser calculado utilizando diferentes medidas, sendo as mais comuns a representação booleana, a frequência absoluta do termo (*term-frequency - tf*) e

a medida que pondera a frequência dos termos na coleção de documentos (*term-frequency - inverse document frequency - TF-IDF*) (SOARES et al., 2008).

	t_1	t_2	...	t_M	C
d_1	a_{11}	a_{12}	...	a_{1M}	c_1
d_2	a_{21}	a_{22}	...	a_{2M}	c_2
...
d_N	a_{N1}	a_{N2}	...	a_{NM}	c_N

Figura 12 – Representação de documentos (SOARES et al., 2008)

A Figura 13 representa o processo de extração de *bag-of-words*, onde cada documento é transformado em uma matriz que pode ter diferentes representações.



Figura 13 – Extração de *bag-of-words*

2.4.3.1 Representação booleana

A medida booleana usa a representação binária para os termos. Quando o termo está presente no documento o valor de a_{ij} é 1, e a_{ij} é 0, caso contrário.

2.4.3.2 Representação *term-frequency*

Muitas vezes medidas estatísticas são empregadas levando em consideração a frequência em que os termos são encontrados nos documentos. Essa medida é chamada de *term-frequency* ou *tf*. A medida *tf* considera o valor de a_{ij} como a frequência em que o termo aparece no documento. Essa medida é definida pela expressão 2.18, na qual $freq(t_j, d_i)$ é a frequência do termo t_j no documento d_i .

$$tf(t_j, d_i) = freq(t_j, d_i) \quad (2.18)$$

2.4.3.3 Representação *TF-IDF*

Quando alguns termos aparecem na maioria dos documentos, esses termos raramente fornecem informações úteis que possam diferenciá-los em uma tarefa de mineração de textos. Dessa forma, é possível utilizar também informações que indiquem a frequência que o termo aparece no *Corpus*. A medida *TF-IDF* oferece um fator de ponderação que é

utilizado para que os termos que aparecem na maioria dos documentos tenha uma “força” de representação menor. A medida *TF-IDF* é definida na expressão 2.19. A frequência do termo no documento é multiplicada por um fator de ponderação. Esse fator varia entre 0 e $\log N$ e é demonstrado na expressão 2.20, na qual N é o número de documentos do *Corpus* e $d(t_j)$ é o número de documentos nos quais o termo t_j ocorre pelo menos uma vez. Assim, quando o termo aparece em todos os documentos o fator de ponderação é igual a 0, quando aparece em apenas 1 documento ele é $\log N$.

$$tf - idf(t_j, d_i) = freq(t_i, d_j) \cdot idf(t_j) \quad (2.19)$$

$$idf(t_j) = \log \frac{N}{d(t_j)} \quad (2.20)$$

2.4.3.4 Análise de Semântica Latente

Como forma de minimizar o tamanho da matriz de *bag-of-words*, que pode ser realmente grande em um *Corpus* com muitos termos únicos, pode se aplicar o algoritmo de Análise Semântica Latente (do inglês *Latent Semantic Analysis - LSA*).

LSA é uma técnica de extração e representação de uso contextual de palavras utilizando computações estatísticas para um grande *Corpus* de texto. O princípio da LSA é que o agregado de todos os contextos de palavras em que uma dada palavra aparece ou não aparece fornece um conjunto de restrições mútuas que determina, em grande parte, a similaridade do significado de palavras e conjuntos de palavras (LANDAUER; FOLTZ; LAHAM, 1998).

É aplicada a uma matriz de *bag-of-words* uma técnica matemática chamada SVD, usada para reduzir o número de linhas enquanto preserva a estrutura de similaridade entre colunas. As palavras são então comparadas tomando o cosseno do ângulo entre os dois vetores (ou o produto de ponto entre as normalizações dos dois vetores) formado por quaisquer duas linhas. Valores próximos de 1 representam palavras muito semelhantes, enquanto valores próximos de 0 representam palavras muito diferentes (DUMAIS, 2004).

2.5 Ferramentas e Recursos

Nesta seção, são descritas as ferramentas e recursos utilizados no desenvolvimento deste trabalho, sendo eles o NLTK, ferramenta do Python utilizada para o Processamento de Linguagem Natural, e os dois *Corpora*⁶ produzidos.

⁶ Entende-se como *Corpus* um conjunto de documentos de um determinado tema, no caso, *reviews* de produtos. *Corpora* é o plural de *Corpus*

2.5.1 NLTK: Natural Language ToolKit

Criado originalmente em 2001 como parte de um curso de linguística computacional do Departamento de Ciência da Computação e Informação da Universidade da Pensilvânia, o *NLTK* é uma plataforma usada para construir programas *Python* que trabalham com dados de linguagem humana para aplicação em PLN. O NLTK define uma infraestrutura que pode ser usada para construir programas de PLN em *Python*; fornece classes básicas para representar dados relevantes para o processamento da linguagem natural; interfaces padrão para executar tarefas como tokenização, *Part-Of-Speech*, análise sintática e classificação de texto; e implementações padrão para cada tarefa que podem ser combinadas para resolver problemas complexos (BIRD; KLEIN; LOPER, 2009).

2.5.1.1 Etiketador MacMorpho

Um dos módulos presentes no NLTK é a etiquetagem (Part-of-Speech Tagger ou POS Tagger). Segundo Bird, Klein e Loper (2009), um etiquetador é responsável pelo processo de definição da classe gramatical das palavras, de acordo com as funções sintáticas. As principais classes no Português são os substantivos, pronomes, adjetivos, verbos, advérbios, preposições, conjunções, numerais e interjeições. Geralmente o resultado do processamento de um etiquetador é dado por conjunto de etiquetas (*tags*). No NLTK, o etiquetador usado para processar palavras em Português é o MacMorpho, proposto por (ALUÍSIO et al., 2003). Eles apresentam um Corpus de 1.1 milhões de palavras validadas manualmente com anotação morfossintática. O conjunto de texto base foi retirado do jornal Folha de São Paulo, onde de acordo com os autores, retorna uma alta qualidade contemporânea pelo fato de englobar diferentes autores e domínios. Sua documentação está disponível para livre acesso no site do NILC/USP⁷.

Dentre o conjunto de etiquetas presentes no Mac-Morpho, algumas variações sintáticas estão explicitadas para melhores resultados, como a possibilidade de identificar tipos diferentes de pronomes (substantivos, pessoais, conectivos), advérbios (conectivos, relativos), conjunções (coordenativa, subordinativa), tipos de verbos, além de extras como estrangeirismos, apostos, datas, entre outros.

Como exemplo, quando a frase “O hotel é bom” é submetida para análise, primeiramente o NLTK separa as palavras em tokens para depois dar uma etiqueta para a palavra. No caso da frase descrita, o conjunto de tags resultante do Mac-Morpho seria [(‘O’, ‘ART’), (‘hotel’, ‘N’), (‘é’, ‘V’), (‘bom’, ‘ADJ’)]. A lista de tags se torna manipulável para executar as técnicas para extrair as características e palavras opinativas (SANTOS, 2017).

⁷ <http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf>

2.5.2 *Corpus ReviewsSteam*

O Corpus *ReviewsSteam* foi construído com dados do *site Steam* coletados em fevereiro de 2016, em um período arbitrário (dias 26, 27 e 28). A coleta foi realizada de forma automática através de um *scraper*⁸ escrito em *Python* utilizando a biblioteca *Scrapy*⁹.

Definiu-se como a população de interesse, todos os *reviews* de jogos digitais do *site Steam* escritos em Português Brasileiro com pelo menos 1 voto de utilidade. Como os *reviews* do *Steam* são ordenados de maneira decrescente por nota recebida, a coleta concentrou-se apenas nos 100 *games* mais jogados no *Steam* no dia 26 de fevereiro. Essa lista é atualizada diariamente no próprio *site*¹⁰. Dos 100 *games* mais jogados no dia 26 de fevereiro, cinco não possuíam *reviews* em português brasileiro com pelo menos 1 voto.

Inicialmente coletou-se uma amostra de 12.168 *reviews*. Optou-se por eliminar *reviews* com menos de três votos para evitar desbalanceamento e tendências nos dados, já que a maioria dos *reviews* possui poucos votos. Como pré-processamento dos dados, que serve para eliminar *reviews* de baixa qualidade, foram removidos *reviews* com menos de 3 votos e os *reviews* com baixa corretude ortográfica. Para tal, foi utilizado como *threshold* a média de corretude do *dataset* original (76%, o média de palavras corretas por *review* na base original). Ao final obteve-se um *dataset* contendo 5.576 *reviews*.

O *site Steam* possui um detalhado sistema de *reviews* que inclui, além do texto do usuário, diversas outras informações. A Figura 14 mostra um exemplo típico de *review* do *Steam*. A avaliação final do usuário é apresentada em destaque acima do texto. Em um *review* também se encontra informações sobre o autor, a quantidade de horas gastas com o produto, a quantidade de pessoas que consideraram a análise útil e um sistema de votos que inclui, além dos tradicionais “sim” e “não”, a opção “engraçada” que pode ser votada de forma independente do voto de utilidade.

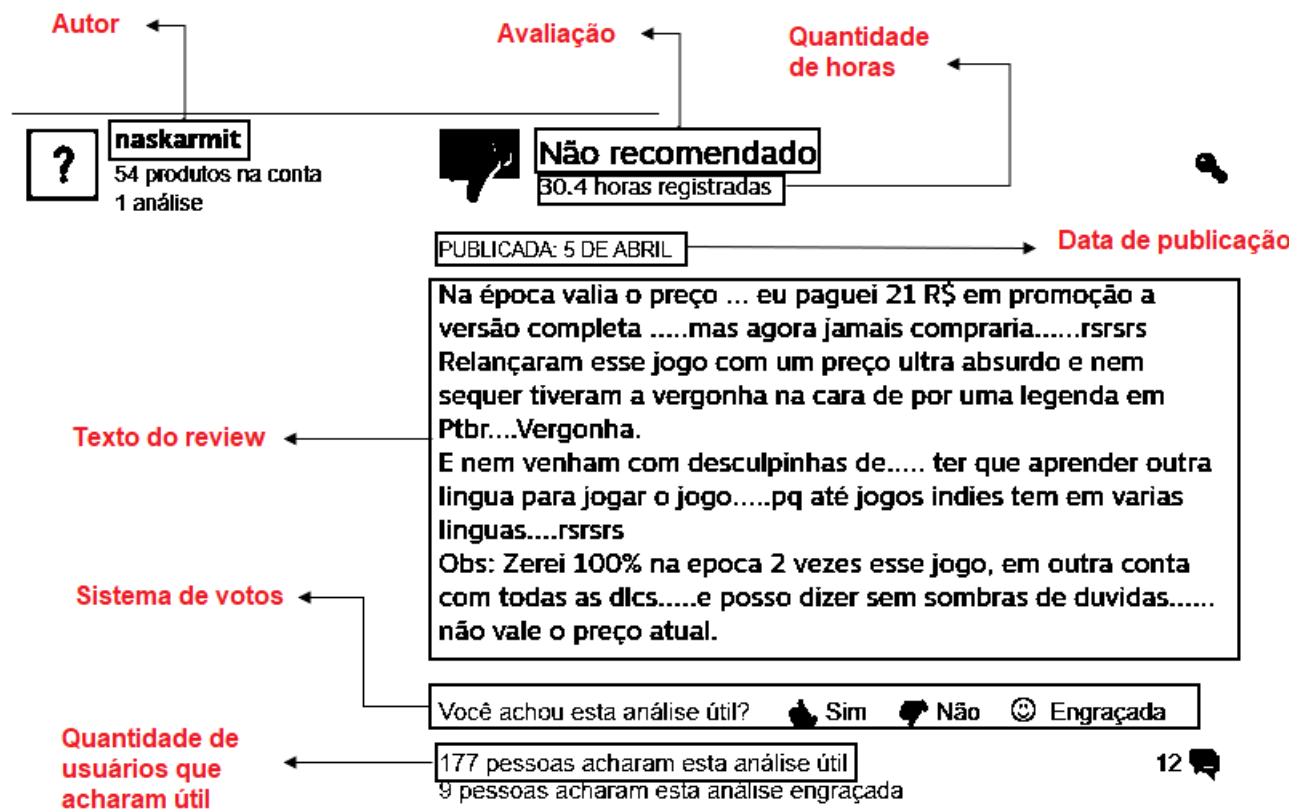
A Figura 15 apresenta um exemplo de *review* no formato *JSON*¹¹. No modelo proposto, um *review* consiste de: avaliação final, quantidade de produtos que o autor possui em sua conta, nome do autor do *review*, *url* do perfil do autor, nome do produto avaliado, data de postagem, quantidade de votos positivos e negativos, quantidade de horas que o autor gastou com o jogo até o momento da postagem, porcentagem de votos de utilidade, valor de utilidade do *review* (calculada a partir da expressão 2.2), texto do *review*, quantidade de *reviews* publicados pelo autor e quantidade de amigos que o autor possui dentro da comunidade de usuários do *Steam*. Adicionalmente, coletou-se a avaliação média dos usuários para cada um dos jogos presentes na amostra.

⁸ Termo usado para representar técnicas de extração de dados utilizadas para coletar dados de *websites*

⁹ <http://scrapy.org/>

¹⁰ <http://store.steampowered.com/stats/>

¹¹ Acrônimo de *JavaScript Object Notation*. É um formato compacto e de padrão aberto para troca de dados entre sistemas

Figura 14 – Exemplo de *review* do *Steam* adaptado

```

{
  "rating": "Recommended",
  "productAmount": 251,
  "name": "Luis Gustavo da Luz",
  "profileUrl": "http://steamcommunity.com/id/OPunkreas/",
  "productName": "Dragon Quest Heroes",
  "date": "Publicada: 10 de dezembro de 2015",
  "positive": 12.0,
  "negative": 0.0,
  "hours": 14.4,
  "helpfulness": "(100%)",
  "wilsonEq": 0.7574992425007574,
  "text": " Dragon Quest chegou ao PC, What a Glorious day, to be alive !!",
  "reviewAmount": 24,
  "friends": 99
}

```

Figura 15 – Exemplo de *review* do *Steam* em JSON

É importante notar, também, algumas limitações do processo de coleta. O *Steam* permite que os usuários tornem seus perfis protegidos, podendo apenas ser acessados por amigos, como uma forma de garantir privacidade. Por esse motivo, alguns dados de *reviews* tornam-se invisíveis para *crawlers*¹² e não podem ser coletados. Sempre que isso aconteceu, optou-se por remover o *review* da base. Ao fim, a amostra foi reduzida para 5.823 *reviews*.

2.5.3 Corpus *ReviewsBuscapé*

O Corpus *ReviewsBuscapé* foi construído com dos dados do *site Buscapé* coletados no dia 15 de novembro de 2017 de forma automática através de um *scraper* similar ao utilizado anteriormente na coleta de dados do *Steam*. Definiu-se como a população de interesse, todos os *reviews* de smartphones do *site Buscapé* disponíveis na data da coleta, resultando em uma amostra de 26.781 *reviews*. Assim como feito no *Steam*, optou-se por eliminar *reviews* com menos de três votos e com corretude abaixo de 76%, obtendo uma amostra final com 4.519 *reviews*.

O sistema de *reviews* do *site Buscapé* é similar ao do *Steam*. A Figura 16 mostra um exemplo típico de *review* do *Buscapé*. A avaliação final do usuário é apresentada em destaque acima do texto. Em um *review* encontram-se o título do *review*, o conteúdo textual, a data de publicação, o nome do autor, a avaliação em estrelas, a quantidade de pessoas que consideraram a análise útil e o veredito do *review*, que pode ser “Recomento este produto” ou “Não recomendo este produto”. Porém, diferente do *Steam*, nem todos os *reviews* possuem nome do autor e nem todos os autores possuem perfil público, impossibilitando a coleta das informações necessárias para uma análise mais aprofundada.

A Figura 17 apresenta um exemplo de *review* no formato *JSON*. No modelo proposto, um *review* consiste de: avaliação final, nome do autor do *review*, *url* do perfil do autor, nome do produto avaliado, data de postagem, quantidade de votos positivos e negativos, confirmação de compra do produto, porcentagem de votos de utilidade, valor de utilidade do *review* utilizando a equação 2.8, texto do *review* e quantidade de *reviews* publicados pelo autor. Adicionalmente, coletou-se a avaliação média dos usuários para cada um dos produtos presentes na amostra.

2.6 Considerações Finais

Este capítulo apresentou os principais conceitos que servem como base para a pesquisa discutida nesta Dissertação.

Inicialmente discutiu-se alguns conceitos chave de comércio eletrônico e marketing, explicitando as etapas pelas quais os consumidores passam durante o processo de decisão

¹² Programa de computador que navega pela Internet de uma forma metódica e automatizada geralmente com propósito de indexação

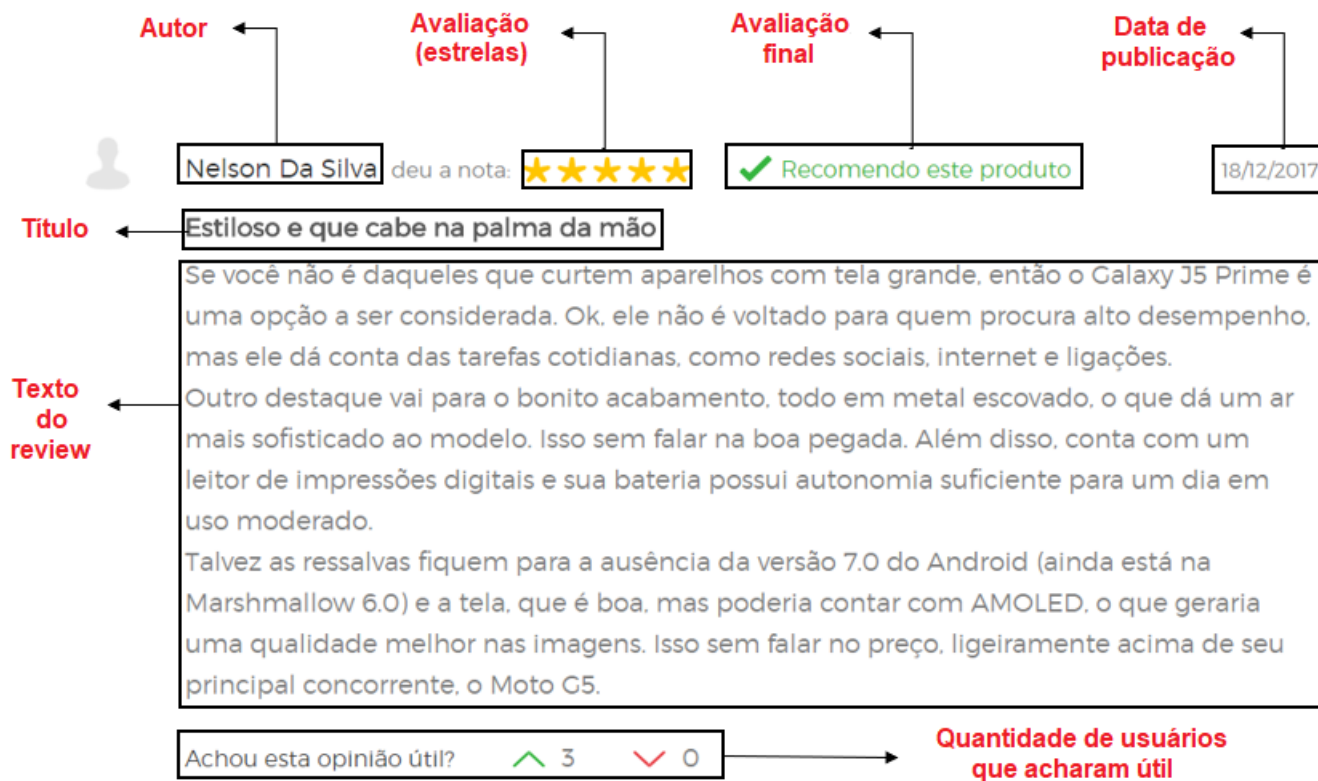


Figura 16 – Exemplo de *review* do *Buscapé* adaptado

```
{
  "rating": "Eu recomendo!",
  "name": "Jonatha Fernandes",
  "profileUrl": "https://www.buscape.com.br/usuario/13214688/jonatha-fernandes",
  "productName": "Smartphone Samsung Galaxy J7 Prime SM-G610M",
  "date": "19/12/2016",
  "positive": 82.0,
  "negative": 43.0,
  "confirmedBuyer": 0,
  "helpfulness": "65%",
  "wilsonEq": 0.6342214,
  "text": "Quero comprar o produto que esta por 1279 nas lojas C&A e quando entra no site esta por 1599 !!!",
  "reviewAmount": 1,
}
```

Figura 17 – Exemplo de *review* do *Buscapé* em JSON

de compra e discutindo as principais influências exercidas sobre ele. Estabelecemos também a importância da comunicação WOM no processo de decisão de compra de consumidores e discutimos sobre *reviews online* como WOM.

Outro conceito apresentado foi o de **percepção de utilidade de *reviews online***, que nada mais é do que o valor a ser previsto. Foram demonstradas formas de calcular esse valor utilizando os dados disponíveis aos usuários nos *sites* de *e-commerce*, bem como os principais problemas dessa abordagem.

Nas seções seguintes exploramos os possíveis modelos para o nosso método de predição, mais especificamente um modelo de regressão. Diante dos principais métodos de regressão, apresentamos as Redes Neurais Artificiais, em específico MLP e RBF, que são aproximadores de funções universais, um outro termo para regressão. Demonstramos o funcionamento dessas redes, bem como formas de validar o treinamento feito por elas.

Por fim, apresentamos a área de Mineração de Opiniões, usada em nossa proposta para extrair informação dos textos dos *reviews*, assim como alguns dos recursos utilizados na produção neste trabalho.

3 Trabalhos Relacionados e Estado da Arte

Neste capítulo será feita uma breve revisão de alguns dos últimos trabalhos publicados pelo grupo de PLN da UFPI que estão, de certa forma, relacionados a este trabalho; bem como serão apresentados os principais trabalhos recentemente publicados sobre a percepção de utilidade de *reviews online*. Também é proposta uma definição preliminar do estado da arte dessa subárea de pesquisa.

3.1 Trabalhos desenvolvidos pelo grupo de PLN da UFPI

Esta seção serve como retrospectiva para os trabalhos recentemente publicados pelo grupo de PLN da UFPI que, de certa forma, serviram como base para este trabalho.

[Sousa, Rabelo e Moura \(2015\)](#) propuseram uma abordagem, denominada Top(X) para estimar o grau de importância de comentários sobre produtos e serviços utilizando um Sistema Fuzzy com três variáveis de entrada: reputação do autor, quantidade de tuplas <característica, palavra opinativa> e analisador de riqueza e uma variável de saída: grau de importância do comentário. Os experimentos realizados mostraram que essa abordagem melhorou os resultados do método de orientação semântica, com menos de 10% dos textos do corpus.

[Santos et al. \(2016\)](#) exploraram duas das variáveis de entrada da abordagem Top(X): quantidade de tuplas <característica, palavra opinativa> e analisador de riqueza aplicadas no domínio de hotéis. Além disso, também foram apresentadas abordagens utilizando Sistemas Fuzzy e Redes Neurais Artificiais (RNA) a partir das adaptações propostas. Os experimentos mostraram a eficácia na precisão de *reviews* com polaridade negativa e *recall* de *reviews* com polaridade positiva em 84,93% e 94,33%, respectivamente, e os comentários mais importantes foram encontrados em uma medida F considerada satisfatória de 50% em polaridades positivas e neutras.

[Sá, Vieira e Moura \(2017\)](#) também utilizaram a abordagem Top(X) como base para propor melhorias, dessa vez, na variável de entrada reputação do autor. Foi feita uma abordagem para definir quais as medidas mais importantes para avaliar a reputação dos autores de comentários Web usando uma RNA. Em uma avaliação preliminar, a RNA apresentou uma precisão de 91,01% no processo de classificação do autor. Adicionalmente, foi conduzido um experimento com o objetivo de comparar as duas abordagens, e os resultados mostram que a abordagem adaptada teve melhor desempenho na classificação da importância dos comentários.

Assim como a abordagem Top(X), proposta inicialmente por [Sousa, Rabelo e Moura](#)

(2015), o modelo proposto no nosso trabalho também pode ser utilizado para estimar o grau de importância de *reviews online*, aqui chamado de percepção de utilidade do *review*, porém utilizando uma metodologia diferente, mas ainda partindo do princípio de que é possível estimar esse valor a partir de variáveis como a reputação do autor e o conteúdo textual do *review*.

3.2 Percepção de utilidade de *reviews* e identificação de *reviews* úteis no contexto *web*: revisão da literatura

A percepção de utilidade de *reviews* é foco de diversos trabalhos recentes. Evidências suportam a conceitualização teórica de utilidade de *reviews* como uma construção formativa (LI et al., 2013). Assim, entende-se que existe uma série de características que influenciam a percepção de utilidade de *reviews*.

As pesquisas nessa subárea podem ser divididas em dois grandes grupos. O primeiro está interessado na natureza das características de *reviews* que influenciam humanos a votarem em *reviews* como úteis ou não-úteis e no processo de avaliação de *reviews* em geral. O segundo grupo diz respeito à previsão automática da utilidade de *reviews* e está mais relacionado a métodos de aprendizagem de máquina e *deep learning*, utilizando técnicas de PLN para identificar *features* de *reviews* que são úteis para a previsão automática de utilidade. Este trabalho situa-se no segundo grupo.

Apesar da grande quantidade de trabalhos publicados sobre o tema, é difícil determinar o estado da arte na previsão automática de utilidade. Isso se deve ao fato de que esses trabalhos apresentam modelos treinados em *datasets* diferentes, sendo muitos deles não disponíveis publicamente. Além disso, muitos trabalhos foram desenvolvidos em paralelo e não compartilham das mesmas *baselines*, sendo difícil a comparação entre eles. Dessa forma, torna-se difícil criar uma linha de progresso entre os trabalhos, já que a maioria é independente entre si.

Esse problema também se deve à grande quantidade de dados independentes em diversos domínios para a realização de tarefas de aprendizagem de máquina, em especial para o Português. Além disso, os avanços da área tecnológica (novas redes sociais e/ou aplicativos de comunicação) também ocorrem a uma velocidade superior aos avanços realizados nas pesquisas. Por fim, é necessário concentrar esforços dos grupos de pesquisas para gerar Corpora para treinamento e teste de modelos que possam ser comparados.

3.2.1 Datasets

A maioria dos trabalhos anteriores foi desenvolvida com base em *datasets* do *site Amazon*. O modelo de *reviews* do *Amazon* é também seguido por diversos sites, que

geralmente possuem apenas uma opção de voto que classifica o *review* como útil, podendo também possuir um voto de não-útil. A Figura 18 apresenta um exemplo de *review* do site *Amazon*.

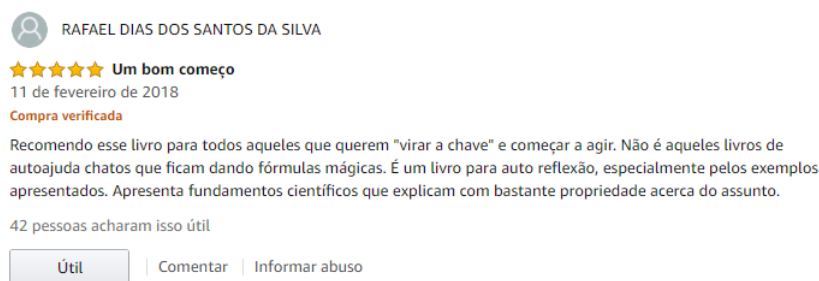


Figura 18 – Review publicado no site Amazon

Alguns trabalhos variam dessa norma, utilizando primordialmente *datasets* oriundos de *sites* de reservas de restaurantes e hotelaria, mas a maioria dos pesquisadores simplesmente usa a fração de usuários que votam em um *review* como útil na forma em que aparecem nos sites durante o treinamento e validação de seus modelos. Para resolverem problemas como a quantidade variável de votos por *review*, os pesquisadores frequentemente filtram aqueles que não têm um número mínimo de votos (ex.: eliminar *reviews* com menos de três votos).

3.2.2 Avaliação dos modelos

Em relação às medidas de desempenho, as tarefas de classificação geralmente usam métricas de avaliação da área de recuperação de informação, a saber: Precision, Recall e F-measure como métricas de avaliação. As tarefas de regressão, por sua vez, usam principalmente o erro quadrático médio (MSE), que mede a média da soma do erro quadrático, e a raiz erro quadrático médio (RMSE), que é definido como a raiz quadrada do MSE. Os sistemas de classificação usaram o *Normalized Discounted Cumulative Gain* (NDCG), que é usado popularmente para medir a relevância dos resultados de pesquisa na recuperação de informações e NDCG@k, uma versão especial do NDCG que só leva em conta os principais k itens em um ranking (isso é usado porque os usuários só lêem um número limitado de comentários). Os pesquisadores também usam as correlações de *Pearson* e *Spearman* para medir o desempenho do modelo e o desempenho do ranking (DIAZ; NG, 2018).

3.2.3 Metodologias

Muitos dos trabalhos propuseram modelos baseados em regressão, utilizando modelos de fatoração extensiva (MOGHADDAM; JAMALI; ESTER, 2012), regressão linear (LU et al., 2010), SVM (KIM et al., 2006) (ZHANG; VARADARAJAN, 2006) (YANG et al., 2015), Redes Neurais (LEE; CHOEH, 2014), entre outros.

As abordagens de classificação foram baseadas principalmente em SVMs (HONG et al., 2012) (KRISHNAMOORTHY, 2015) (LIU et al., 2017), mas modelos de Random Forest (GHOSE; IPEIROTIS, 2011), Naive Bayes, J48 e JRip também foram utilizados (KRISHNAMOORTHY, 2015). Trabalhos recentes também abordaram a tarefa de classificação com redes neurais (MALIK; HUSSAIN, 2017) (CHEN et al., 2018).

Dentre as *features* utilizadas nos modelos propostos, podemos dividi-las em quatro grandes grupos: *features* sintáticas, *features* semânticas, *features* relacionadas ao autor do *review* e *features* relacionadas aos metadados dos *reviews*. Dentre as *features* sintáticas, estão o tamanho do *review* (KIM et al., 2006), baseado na ideia de que avaliações mais longas têm mais informações e, portanto, são mais úteis; a legibilidade, que se baseia na hipótese de que, se uma revisão for mais fácil de ler, ela será considerada útil por mais usuários (OTTERBACHER, 2009); Dentre as *features* semânticas, estão as opiniões do autor, baseado na ideia de que quanto mais opinativo, mais útil é o *review* (LIU et al., 2017); e o sentimento expresso no texto, baseado na hipótese de que textos que evoquem sentimentos mais fortes podem ser mais úteis (HONG et al., 2012). As *features* baseadas no autor se baseiam em informações históricas do autor para prever pontuações futuras de utilidade (SCHINDLER; BICKART, 2012); além de *features* que tentam capturar a similaridade entre leitores e autores (CONNORS; MUDAMBI; SCHUFF, 2011) (FORMAN; GHOSE; WIESENFELD, 2008). Já *features* baseadas em metadados podem incluir a data de publicação do *review* ou a quantidade de estrelas dadas pelo autor, por exemplo (DANESCU-NICULESCU-MIZIL et al., 2009).

3.2.4 Principais descobertas na área

Alguns trabalhos identificaram a influência direta do conteúdo do *review* na percepção de sua utilidade. Com a ajuda de um modelo de aprendizado de máquina supervisionado, Kim et al. (2006) descobriram utilizando SVM que os pesos TF-IDF de unigramas lematizados, o tamanho do *review* e a classificação do produto (número de estrelas) são os recursos mais importantes. Danescu-Niculescu-Mizil et al. (2009) realizaram um estudo sobre um Corpus do *site Amazon* e identificaram que a percepção de utilidade de *reviews* depende não somente do conteúdo descritivo do *review*, mas também do relacionamento entre a avaliação expressa no *review* e as demais avaliações do mesmo produto, ou seja, um *review* é avaliado como mais útil quando a sua nota (ou avaliação) é próxima ao consenso das demais notas para o mesmo produto.

Ainda nessa mesma linha, Malik e Hussain (2017) propuseram cinco *features* linguísticas baseadas em elementos semânticos para prever utilidade de *reviews*: substantivos singulares, substantivos comuns, preposições, pronomes pessoais e advérbios. Os achados indicam que substantivos singulares, substantivos comuns, preposições e pronomes pessoais são as *features* mais eficazes para a utilidade de *reviews*.

Schindler e Bickart (2012) examinaram características do conteúdo do *review*, como declarações descritivas a respeito do produto e do autor, a precisão gramatical e a polaridade. Foi descoberto que o tamanho do texto e a quantidade de informações expressas influenciam positivamente a percepção de utilidade de *reviews*, mas somente até certo ponto. Além disso, percebeu-se que *reviews* com quantidade moderada de declarações positivas e alta proporção de declarações que descrevam o produto possuem maior impacto. Já Otterbacher (2009) analisou dados do *site Amazon* e determinou que a utilidade de *reviews* pode ser manifestada pela relevância do tópico, facilidade de compreensão, credibilidade e objetividade. Além disso, foi descoberto um forte relacionamento entre a ordem cronológica dos *reviews* e a utilidade. Já Huang et al. (2015) descobriram que a contagem de palavras tem um limiar em seus efeitos na utilidade do *review*. Além desse limite, seu efeito diminui significativamente ou se torna quase inexistente.

Liu et al. (2017) hipotetizaram que a utilidade de um *review* é diretamente relacionada com a argumentação do texto, i.e.: porcentagem de setenças argumentativas. Eles encontraram indícios de que *features* baseadas em argumentos podem melhorar a performance de identificação de utilidade de *reviews*, bem como ser usadas para interpretar o que faz um *review* útil.

Alguns pesquisadores também exploraram a influência da declaração de autoria na percepção de utilidade de *reviews*. Connors, Mudambi e Schuff (2011) examinaram alguns fatores básicos associados com a utilidade de *reviews*. Eles perceberam que a *expertise* do autor, ou seja, o conhecimento do autor a respeito do domínio do produto, possui influência positiva na percepção de utilidade de *reviews*. Já características como a similaridade entre o autor e o leitor não aparentam influenciar a utilidade de um *review*. Forman, Ghose e Wiesenfeld (2008) descobriram que a presença de informações descritivas sobre a identidade do autor de *reviews online* possui impacto positivo na percepção de utilidade de *reviews* e conseqüentemente na venda de produtos *online*. Em contraste, Huang et al. (2015) descobriram que a experiência do autor não é preditor estatisticamente significativo de utilidade, mas as avaliações de utilidade de *reviews* anteriores tendem a prever avaliações futuras.

Ngo-Ye e Sinha (2014) propuseram uma diferente abordagem que analisa as características de engajamento do autor e suas relações com a utilidade do *review*. Eles definiram um padrão chamado de RFM (Recency, Frequency, Monetary Value) que analisa três características dos autores (recência, frequência e valor monetário) e descobriram, utilizando uma abordagem híbrida que combina as características textuais do modelo de *bag-of-words* e as dimensões do RFM, que tanto o texto do *review* quanto as características do envolvimento do autor ajudam a prever a utilidade do *review*.

O tipo do produto avaliado também possui alguma influência na percepção de utilidade de *reviews*. Sen e Lerman (2007) investigaram como os consumidores avaliam a

utilidade de *reviews* positivos e negativos e descobriram que o tipo do produto modera o efeito de valência do *review*. Observou-se que para produtos utilitários, leitores são mais propícios a relacionar a opinião negativa expressa no *review* às experiências do autor sobre o produto, já para produtos hedônicos, avaliações negativas são atribuídas a motivos não relacionados ao produto. Usando modelagem analítica, [Mudambi e Schuff \(2010\)](#) identificaram que, para bens de experiência, *reviews* com avaliações extremas são menos úteis do que *reviews* com avaliações moderadas. A profundidade do *review* também exerce influência positiva na percepção de utilidade, sendo mais forte para bens de busca.

Talvez a observação mais importante seja que os escores de utilidade podem não estar fortemente correlacionados à qualidade da revisão ([LIU et al., 2007](#)) ([DANESCU-NICULESCU-MIZIL et al., 2009](#)) ([GHOSE; IPEIROTIS, 2011](#)) ([YANG et al., 2015](#)). Em pelo menos um estudo, os anotadores independentes concordaram com mais frequência (85%) com um ranking de utilidade alternativo do que com um baseado em pontuações de utilidade ([TSUR; RAPPOPORT, 2009](#)).

3.3 Considerações Finais

O trabalho proposto nesta Dissertação apresenta uma abordagem que complementa descobertas anteriores e será apresentado no próximo Capítulo. Propomos utilizar como *features*, as características textuais do *review*, o engajamento do autor, tempo de publicação do *review* e *expertise* do autor. Como em [Ngo-Ye e Sinha \(2014\)](#), também utilizamos Redes Neurais Artificiais ao invés de outras técnicas já estabelecidas, como Regressão por Vetores de Suporte (do inglês: *Support Vector Regression*, SVR).

Destaca-se como um diferencial do trabalho a avaliação do modelo em dois domínios de aplicação: *reviews* de jogos e *reviews* de produtos eletrônicos. Além disso, as *features* foram extraídas diretamente de indicadores em ambos os *datasets* e podem servir como base para a definição de modelos que possam ser seguidos por outros *websites*, ao invés de *features* que foram definidas arbitrariamente pelos pesquisadores. Como métrica para a utilidade de *reviews*, utilizamos a equação definida em (2.2), ao invés de simplesmente utilizar a fração de votos positivos. Isso nos permite comparar melhor *reviews* com diferentes quantidades de votos.

A principal diferença entre este trabalho e os demais citados está na forma de mensurar variáveis como engajamento e *expertise* do autor, qualidade do conteúdo textual e argumentos do *review*. Acreditamos que a base de *reviews* do *Steam* oferece métodos mais adequados para definir essas variáveis do que os tradicionais *datasets* do *site Amazon*, graças ao formato de Rede Social *Online* proposto. Como método de comparação, o modelo também foi aplicado em um *site* com menos recursos, o Buscapé.

A Figura 19 apresenta uma representação geral de como esses trabalhos estão

distribuídos em termos de *features* e modelos de aprendizado. Adicionalmente, a Tabela 1 mostra um resumo dos trabalhos que serviram como base para esta Dissertação, detalhando as abordagens, os Corpora e as principais características. O modelo utilizado neste dissertação está em destaque.

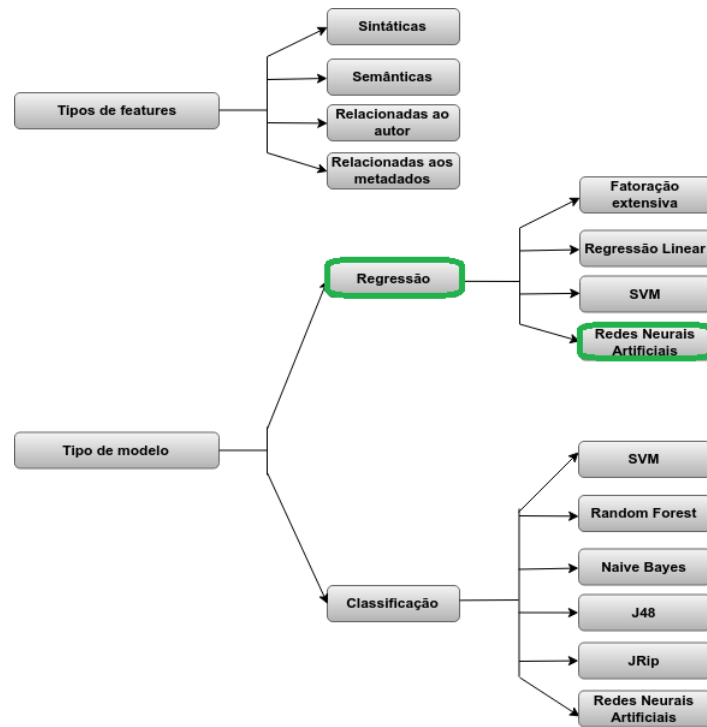


Figura 19 – Distribuição dos Trabalhos Relacionados

Tabela 1 – Resumo dos trabalhos relacionados

Trabalho	Abordagem	Diferenças	Base
(MOGHADDAM; JAMALI; ESTER, 2012)	Modelos gráficos probabilísticos baseados em Fatoração Matricial e Fatoração Tensorial	Características textuais baseadas em contagens, como número de tokens e sentenças.	Epinions.com
(LU et al., 2010)	Teste de hipóteses usando regressão	Explora informações contextuais sobre as identidades dos autores em redes sociais para melhorar a previsão da utilidade do review.	Ciao UK
(KIM et al., 2006)	Análise de regressão usando SVM	Limitado a características textuais.	Amazon.com
(ZHANG; VARADARAJAN, 2006)	Dois modelos de regressão distintos	Limitado a características textuais.	IMDB
(YANG et al., 2015)	Análise de regressão usando SVM	Limitado a duas características semânticas do texto.	Amazon.com
(LEE; CHOE, 2014)	Modelo de regressão usando RNA MLP	Determina a importância das variáveis de entrada usando o cálculo de Força Relativa dos Pesos.	Amazon.com
(HONG et al., 2012)	Modelo de classificação usando SVM	Utiliza features que representam as preferências dos usuários, mas que podem ser geradas sem qualquer intervenção humana.	Amazon.com
(KRISHNAMOORTHY, 2015)	Modelo de classificação usando SVM	Limitado a características sintáticas do texto do review.	Amazon.com
(LIU et al., 2017)	Modelo de classificação usando SVM	Baseado em elementos textuais dos reviews e em features chamadas de argumentos que foram manualmente anotadas na base de dados.	TripAdvisor
(GHOSE; IPEIROTIS, 2011)	Modelo de classificação usando Random Forest	Relaciona utilidade de reviews e características de autores de reviews com vendas do produto.	Amazon.com
(MALIK; HUSSAIN, 2017)	Modelo de classificação usando Deep Learning	Baseado em características de emoções extraídas diretamente do texto e divididas em positivas e negativas.	Amazon.com
(OTTERBACHER, 2009)	Análise de Regressão	Determina as variáveis mais importantes na percepção de utilidade de reviews através de análises de regressão e correlação. Trabalho não propõe modelo automático.	Amazon.com
(SCHINDLER; BICKART, 2012)	Análise de hipóteses	O objetivo é estimar a relação entre elementos de WOM do texto do review com a percepção de utilidade. Trabalho não propõe modelo automático.	Amazon.com
(CONNORS; MUDAMBI; SCHUFF, 2011)	Análise de hipóteses	O objetivo do trabalho é identificar características importantes para a percepção de utilidade do review, sem propor um modelo de análise automático.	Amazon.com
(FORMAN; GHOSE; WIESENFELD, 2008)	Análise de hipóteses	O objetivo do trabalho é identificar a relação entre características relacionadas a autores de reviews, como sua localização geográfica e descrição de identidade na percepção de utilidade de reviews. Trabalho não propõe modelo automático.	Amazon.com
(DANESCU-NICULESCU-MIZIL et al., 2009)	Análise de hipóteses	Utiliza uma técnica de plágio de reviews (publicando o mesmo review em diferentes produtos) para identificar elementos que possam afetar a utilidade de reviews e propõe um modelo matemático que utiliza essas descobertas.	Amazon.com
Huang et al. (2015)	Análise de hipóteses	Utiliza uma técnica de plágio de reviews (publicando o mesmo review em diferentes produtos) para identificar elementos que possam afetar a utilidade de reviews e propõe um modelo matemático que utiliza essas descobertas.	Amazon.com

4 Modelo Proposto

O modelo proposto neste trabalho é um sistema que tem como entrada um vetor X de características de *reviews online* e como saída um valor real H na faixa de 0 a 1, que representa a utilidade de *review* prevista. A Figura 20 apresenta uma visão geral desse modelo.

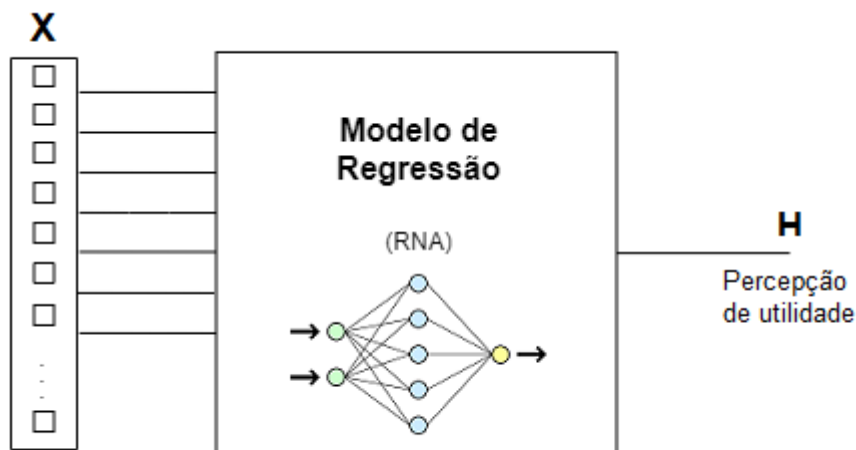


Figura 20 – Modelo proposto

Esse modelo foi desenvolvido em duas etapas distintas: i) definição das *features* que formam a percepção de utilidade de *reviews* de produtos online; e ii) análise de regressão para validar as *features* escolhidas.

Para um dado *review*, o modelo proposto deve encontrar o grau de percepção de utilidade H , onde H é um número real na faixa de 0 a 1. Por se tratar de um modelo de aprendizagem supervisionado, definimos nossa classe H a partir da equação definida na seção 2.2 por Edwin B. Wilson em 1927 e transcrita a seguir:

$$H = \left(\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{[\hat{p}(1 - \hat{p}) + z_{\alpha/2}^2/4n]/n} \right) / (1 + z_{\alpha/2}^2/n) \quad (4.1)$$

4.1 Definição de *Features*

No desenvolvimento do modelo para prever o valor de H , estudou-se as características formativas da percepção de utilidade de *reviews*. Esse estudo baseou-se nas definições propostas no estado da arte, mas tomou algumas liberdades no que se refere à forma como modelar essas informações.

Com base no estudo, separamos as *features* em três grupos: i) características a respeito do autor do *review*; ii) características textuais; e iii) características relativas aos metadados do *review*. O modelo proposto é dado por uma função de regressão que tem como entrada um vetor X contendo essas características e como saída um valor escalar H , que é a medida de percepção de utilidade. Para aproximação da função utilizou-se as Redes Neurais Artificiais MLP e RBF.

4.1.1 Características relativas à autoria

Definimos duas características sobre o autor que servem como *features* de entrada no nosso modelo: a reputação e a *expertise*. Estudos sugerem que *reviews* submetidos por autores com histórico positivo são vistos como mais úteis (WATHEN; BURKELL, 2002). A reputação também incorpora elementos referentes à credibilidade do autor. A *expertise* indica o nível de conhecimento de um autor a respeito do tópico abordado.

4.1.1.1 Reputação do autor

Para os *reviews* extraídos do *site Steam*, modelou-se a reputação de um autor através de três variáveis: a média de votos de *reviews* pela quantidade de *reviews* publicados pelo usuário; a razão dos votos positivos pela quantidade total de votos do usuário; e o número de amigos que o usuário possui na comunidade *Steam*. *A priori*, as mesmas variáveis foram definidas para a base do *Buscapé*, porém, como usuários não possuem amigos no *site Buscapé*, essa última variável não pôde ser utilizada e foi mapeada com o valor 0 para todas as amostras.

4.1.1.2 Expertise

Por *expertise*, no *site Steam*, resolveu-se explorar a quantidade de horas que o autor do *review* dedicou-se ao jogo analisado. Um autor que jogou apenas 2 horas de um determinado jogo pode ter menos proficiência no tópico que foi analisado do que aquele que jogou 20 horas. Para o *Buscapé* podemos considerar como *expertise* um valor binário que depende da confirmação de compra, sendo 1 para usuários em que a compra foi confirmada e 0 caso contrário. Como apenas uma pequena quantidade de *reviews* são feitos por compradores confirmados, essa variável acaba sendo irrelevante no *site Buscapé*, mas para efeitos de generalização, optamos por não ignorá-la.

4.1.2 Características relativas ao conteúdo textual

Fundamentalmente, pesquisas anteriores se dedicaram a dois tipos de análise textual em *reviews online*: análise das características semânticas e análise das características estilísticas do texto (KIM et al., 2006; CAO; DUAN; GAN, 2011). Ao explorar o conteúdo

textual de um *review*, pode-se atentar também às opiniões e sentimentos expressos no texto.

4.1.2.1 Opiniões do *review*

Acredita-se que *reviews online* com uma grande quantidade de opiniões são mais informativos. Há uma grande dificuldade, porém, em determinar a melhor forma de mapear essa informação em modelos de aprendizagem de máquina.

Propomos duas alternativas neste trabalho: i) modelar opiniões de forma numérica. É importante notar que a polaridade dessas opiniões não é importante para este problema, especialmente porque tem-se acesso à avaliação final de cada *review*. Logo, é interessante saber apenas a quantidade de opiniões expressas em cada *review*. Para tal, realizou-se um processo de extração de opiniões segundo uma adaptação do modelo proposto por (SOUSA; RABELO; MOURA, 2015); e ii) modelar um *bag-of-words* utilizando TF-IDF. Espera-se que palavras com grande poder opinativo sejam importantes em um *Corpus* e essa relação entre as palavras pode ser facilmente capturada utilizando um modelo de *bag-of-words* como TF-IDF.

As características de opiniões e palavras importantes foram modeladas, então, como um vetor onde o primeiro elemento é referente à quantidade de opiniões, e os demais n elementos são referentes às n palavras presentes em todo o *Corpus*, como pode ser observado na Figura 21.

(n. opin.)	jogo	diversão	legal	...	gráficos
4	0	0	0.89	...	0

Figura 21 – Vetor de opiniões, sendo o primeiro elemento a quantidade de opiniões e os demais um *bag-of-words*

4.1.2.1.1 Extração de Opiniões

Na extração de opiniões, os textos são analisados de acordo com a estrutura frasal das sentenças. Adotou-se o padrão "*sujeito + verbo + predicado*", onde o núcleo do sujeito representa a característica qualificada e o núcleo do predicado representa a palavra qualificadora. Por exemplo, "esse jogo é divertido" é extraído como (*jogo, divertido*). Além disso, foram incluídos outros padrões linguísticos adaptados dos padrões especificados por (TURNERY, 2002).

Os padrões linguísticos foram definidos através de estudos empíricos realizados sobre uma outra amostra aleatória de 385 *reviews* coletados do *site Steam* em dezembro de 2015 (quantidade definida para um nível de confiança de 95% e margem de erro de

5%). Nota-se que os *reviews* dessa amostra não necessariamente estão presentes na base utilizada no treinamento do nosso modelo (mas podem estar) e foram coletados apenas para esse estudo de definição dos padrões. Durante a definição desses padrões, foram ignoradas classes gramaticais que não contém valor semântico, como artigos, pronomes, conjunções e numerais. O resultado é apresentado na Figura 22, onde o símbolo ? indica que o elemento é opcional, | indica uma alternativa entre os elementos e o símbolo * indica que o grupo pode repetir zero ou n vezes.

SUBST ADV? V? (ADV? ADJ | ADV V)*
 ADJ SUBST
 ADV ADJ

Figura 22 – Padrões linguísticos (Adaptação dos padrões propostos por [Turney \(2002\)](#))

Com os padrões definidos é possível capturar automaticamente a quantidade de opiniões em cada texto anotado com as classes gramaticais, usando expressões regulares.

Com exceção do último padrão, esses padrões linguísticos não se aplicam a orações com sujeito oculto, um caso bem comum em Português, mas difícil de mapear utilizando análise sintática, sem que seja aceita uma grande quantidade de erros. Por exemplo, a sentença “muito bom” pode se referir a um sujeito oculto que é caracterizado pelo adjetivo “bom” ou fazer parte de uma sentença maior. Como alternativa, aplicamos a Análise Semântica Latente (do inglês: Latent Semantic Analysis, LSA) na matriz de *bag-of-words* com o intuito de encontrar essas relações.

O processo de extração de padrões é realizado em quatro fases. Primeiramente, todos os textos são *tokenizados* em unigramas. Em seguida, os *tokens* são etiquetados pelo *POS tagger MacMorpho* ([ALUÍSIO et al., 2003](#)) da plataforma NLTK, que associa cada palavra a uma classe gramatical. A terceira etapa consiste na remoção de símbolos e palavras das classes gramaticais sem valor semântico. Na quarta e última etapa, todos os textos são processados novamente a fim de determinar os padrões presentes, que são contabilizados. A Figura 23 apresenta as fases executadas na extração de padrões.



Figura 23 – Fases da extração de opiniões

4.1.2.1.2 Modelo de *bag-of-words*

Para utilizar o texto dos *reviews* como entrada nas RNA, propusemos um modelo de *bag-of-words* que utiliza como representação o TF-IDF. A ideia é que palavras mais

importantes recebam maiores escores e sirvam como melhores indicadores para a utilidade de *reviews*. Uma alternativa seria utilizar a frequência em que os termos aparecem nos *reviews*. Para isso, precisamos inicialmente pré-processar o texto.

O tratamento do texto dos *reviews* foi dividido em duas partes. A primeira inclui o pré-processamento, normalização e limpeza do texto. A segunda é a transformação em uma *bag-of-words*.

O pré-processamento foi feito seguindo as seguintes etapas:

1. Capitalização do texto: Todo o *Corpus* foi reduzido para letras minúsculas com o objetivo de reduzir o vocabulário.
2. Tokenização e anotação: O texto foi dividido em tokens e anotadas as classes gramaticais de cada token.
3. Limpeza: Remoção de *stopwords*¹.
4. Normalização: Todas as variações de palavras foram reduzidas para uma única palavra (ex.: um verbo e todas as suas conjugações tornam-se uma única palavra).

Para a transformação do texto, foi feita a geração da *bag-of-words* utilizando TF-IDF, que foi posteriormente reduzida utilizando LSA.

Na aplicação do LSA podemos escolher a quantidade de componentes que queremos na matriz. Essa quantidade seria mais ou menos a quantidade de termos que queremos na matriz após a aplicação do algoritmo. Definimos o número padrão de 200 componentes arbitrariamente para a escolha da topologia da Rede, mas também executamos o modelo final com variações desse número e uma versão sem LSA.

4.1.2.2 Elementos estilísticos do *review*

Foram considerados como elementos estilísticos a riqueza de vocabulário, a legibilidade do texto e o tamanho do texto. A riqueza do vocabulário nada mais é do que o grau de correte de um texto, ou seja, o quão ortograficamente correto ele é. Tecnicamente, é importante analisar a correte dos textos porque os *POS taggers* podem encontrar dificuldades para etiquetar palavras que não pertencem ao Português (por exemplo, *bugar*). Em um nível social, a relação entre a qualidade do texto e o quão gramaticalmente correto ele é não é interessante, principalmente na era da Internet.

A língua é uma forma de representação cultural e acima de tudo, mutável (COSTA; PEREIRA, 2004). Gírias, abreviações e estrangeirismos são perfeitamente aceitáveis em

¹ Stopword ou palavra vazia é uma palavra que é removida antes ou após o processamento de um texto em linguagem natural. Qualquer grupo de palavras pode ser escolhido como grupo de stopwords de acordo com o objetivo do processamento. Em alguns trabalhos, as stopwords são as palavras mais comuns da língua e palavras funcionais, como *o*, *a*, *em* e *no*.

comunidades *online*, especialmente porque cada comunidade define a sua norma linguística. Assim, ainda que um texto esteja incorreto de um ponto de vista ortográfico e gramatical, ele não pode ser simplesmente descartado, pois pode ser capaz de transmitir a mensagem originalmente pretendida pelo emissor. A gramática não é a verdade absoluta da língua (DANTAS, 2004).

Como neste trabalho objetiva-se uma análise automática, optou-se por eliminar todos os textos com grau de corretude abaixo de um *threshold* definido arbitrariamente a fim de remover ruídos dos dados (como textos em outros idiomas). Modelou-se a corretude do texto como o número de palavras corretas dividido pelo número total de palavras em um texto. Para determinar se a palavra pertence ou não ao Português, usou-se o dicionário *br.ispell* desenvolvido por Ricardo Ueda Karpiscek do Instituto de Matemática e Estatística da USP². Esse dicionário possui um total de 260.795 termos do Português. Como no domínio de jogos (do *site Steam*) é comum encontrar textos com termos estrangeiros, abreviações e neologismos, optou-se pela criação de uma base de termos *gamers*, com 166 termos observados em uma amostra aleatória do *site Steam* e termos encontrados na Internet³. Essa base inclui palavras estrangeiras como *action*, acrônimos como *fps* (ou *frames per second*), e neologismos como *bugado* (um *software* que possui *bugs*).

Com relação à legibilidade, que indica a facilidade em que um texto pode ser compreendido, para textos em inglês pode ser analisada utilizando o teste de Flesch-Kincaid, um método matemático que avalia o quão legível é um texto por meio da sua média de palavras por sentenças e média de sílabas por total de palavras (KINCAID et al., 1975). Uma adaptação para o Português foi feita pelo jornalista Alberto Dines (SQUARISI; SALVADOR, 2005), de acordo com a fórmula:

$$\left(\frac{n_w}{n_s} + n_p\right) * 0.4 \quad (4.2)$$

onde n_w é o número total de palavras do texto, n_s é o número total de sentenças e n_p é o número total de polissílabos. A constante 0,4 representa a média de letras da palavra em frases do Português.

A Tabela 2 contém as referências de pontuação dessa medida. Quanto maior a pontuação no teste de legibilidade, mais difícil de ser lido é o texto. Os textos presentes em ambos os Corpora desta pesquisa possuem pontuações presentes em todas as faixas. Neste trabalho, escolheu-se a legibilidade como uma característica formativa da percepção de utilidade de *reviews* e modelou-se a legibilidade de um *review* utilizando o algoritmo de legibilidade para o Português, descrito pela equação 4.2.

A relação entre tamanho do texto e utilidade de *reviews* já foi estabelecida em

² <http://www.ime.usp.br/~ueda/br.ispell/>

³ <http://www.tecmundo.com.br/dicionario/14762-video-games-o-dicionario-de-a-a-z.htm>

Tabela 2 – Referências de pontuação da adaptação do teste de legibilidade de Flesch-Kincaid para o Português

Nota	Observações
1 a 7	história em quadrinhos
8 a 10	excepcional
11 a 15	ótimo
16 a 19	pequena dificuldade
20 a 30	muito difícil
31 a 40	linguagem técnica
>41	nebulosidade

trabalhos anteriores (KIM et al., 2006). Descobriu-se que à medida que o tamanho do *review* aumenta, a percepção de utilidade de *reviews* também aumenta, porém, em certo ponto o tamanho do texto passa a exercer influência negativa na percepção de utilidade, pois textos muito grandes provavelmente não serão lidos completamente. Lee e Choeh (2014) descobriram que o número de palavras com apenas uma letra era um bom indicativo da percepção de utilidade do review. Baseado nessas descobertas, modelou-se o tamanho do texto a partir do número de palavras e número de sentenças e a quantidade de monossílabos do texto, já que o Português não possui grande quantidade de palavras semanticamente relevantes com apenas uma letra.

Como a divisão silábica não é uma tarefa trivial e depende de uma série de regras, criou-se uma base de divisão silábica da Língua Portuguesa. Essa base consiste de 181.119 termos do Português, suas classes gramaticais e suas formas divididas silabicamente. Esses termos foram extraídos do Portal da Língua Portuguesa⁴. Para palavras não presentes na base, aplicou-se um algoritmo que leva em consideração os princípios fonéticos das palavras. Nesse algoritmo, definiu-se três grupos básicos de fonemas vocálicos: ditongos, tritongos e hiato. Para descobrir quantas sílabas uma palavra w possui, conta-se a quantidade de tritongos, em seguida conta-se a quantidade de ditongos e, por fim, conta-se a quantidade de hiatos, já que no Português, não existem sílabas sem vogais. Entretanto, a cobertura do algoritmo não é completa, porque em alguns casos é impossível distinguir o hiato de encontros vocálicos considerando apenas a forma escrita da palavra.

A Tabela 3 apresenta um resumo de todas as *features* relativas ao conteúdo textual que foram descritas nesta seção.

4.1.3 Características referentes aos metadados do *review*

Considera-se como metadados a avaliação final do *review* e a data de postagem do *review* em dias. Avaliação final refere-se à avaliação de produto binária presente em cada *review*, que pode ser “recomendado” ou “não recomendado”. Apesar de evidências indicarem

⁴ <http://www.portaldalinguaportuguesa.org/>

Tabela 3 – Resumo das *features* relacionadas ao conteúdo textual do *review*

Feature	Representação
<i>Opiniões</i>	Vetor com quantidade de opiniões e palavras do Corpus
<i>Riqueza do vocabulário</i>	Porcentagem de palavras corretas no texto
<i>Legibilidade</i>	Grau de legibilidade
<i>Tamanho do texto em palavras</i>	Quantidade de palavras total por texto
<i>Tamanho do texto em sentenças</i>	Quantidade de sentenças por texto
<i>Tamanho do texto e monossílabos</i>	Quantidade de monossílabos por texto

que a avaliação de *reviews* não influencia sua percepção de utilidade, é importante notar que leitores podem ser influenciados pela média das avaliações dos produto. Um *review* negativo para um produto em que a maioria dos *reviews* são positivos, pode ser visto como menos útil. Esse ideia foi explorada por Danescu-Niculescu-Mizil et al. (2009) e é discutida com mais profundidade na psicologia social de conformidade (BOND; SMITH, 1996).

Considerando uma avaliação de *review* como sendo 1, para “recomendado” (ou positivo), e 0, para “não recomendado” (ou negativa), modelou-se a avaliação como uma expressão, onde p é porcentagem de *reviews* com avaliações positivas (em representação decimal) para um produto e x é a avaliação individual do *review* para o mesmo produto:

$$x * p + (1 - x) * (1 - p) \quad (4.3)$$

Partindo das descobertas de Otterbacher (2009), modelou-se o tempo de postagem de um *review* como a diferença entre a data de lançamento do produto (presente na página do produto) e a data de postagem do *review* em dias.

4.2 Análise de Regressão para validar o modelo

Para a treinar o modelo proposto, foram desenvolvidas duas redes neurais, uma RNA MLP e uma RNA RBF. Para definição das topologias, testou-se um conjunto de topologias candidatas utilizando validação cruzada com 10 partições (do inglês *10-fold cross-validation*).

4.2.1 Modelo utilizando RNA MLP

O modelo proposto tem como entrada as 11 *features* discutidas na seção anterior e sumarizadas a seguir. Após o processo de treinamento dessa rede, a matriz de pesos referentes ao neurônio de saída corresponderá aos próprios parâmetros λ_i da equação 2.4, isto é $\lambda_i = W_{1,i}$. A ilustração dessa rede é apresentada na Figura 24.

As variáveis de entrada são:

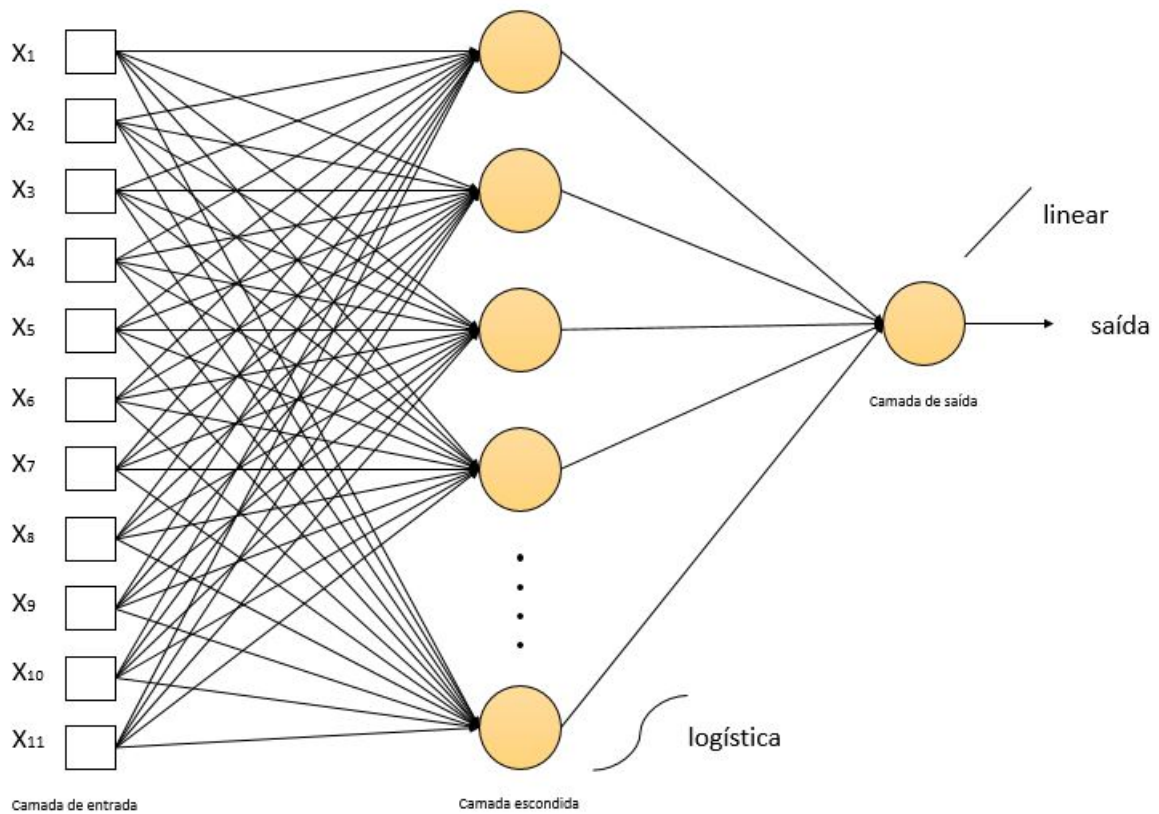


Figura 24 – Rede Neural Artificial do modelo proposto

x_1 a quantidade de votos de *reviews* pela quantidade total de *reviews* de um usuário;

x_2 a razão dos votos positivos pela quantidade total de votos de um usuário;

x_3 o número de amigos do usuário na comunidade *Steam* (automaticamente 0 para todos os *reviews* do *Buscapé*);

x_4 a quantidade de horas que o autor dedicou ao jogo analisado para *reviews* do *Steam* ou a confirmação de compra para *reviews* do *Buscapé* (binário);

x_5 um vetor contendo a quantidade de opiniões expressas no texto e as palavras mais importantes do Corpus;

x_6 a legibilidade do texto;

x_7 o tamanho do texto em palavras;

x_8 o número de sentenças no texto;

x_9 o número de palavras monossílabas no texto;

x_{10} a diferença entre a avaliação média do produto e a avaliação do usuário;

x_{11} a idade do *review* em dias.

A variável de saída utilizada no treinamento foi a percepção de utilidade do *review* H , descrita anteriormente.

Para validação dos resultados foram utilizados os cálculos de *Relative Strength* (RS) e o algoritmo CFS.

4.2.2 Modelo utilizando RNA RBF

O modelo que utiliza RNA RBF é uma alternativa ao modelo que utiliza RNA MLP com as mesmas variáveis de entrada e saída e processo de validação, porém possui uma topologia diferente, como ilustrado pela Figura 25.

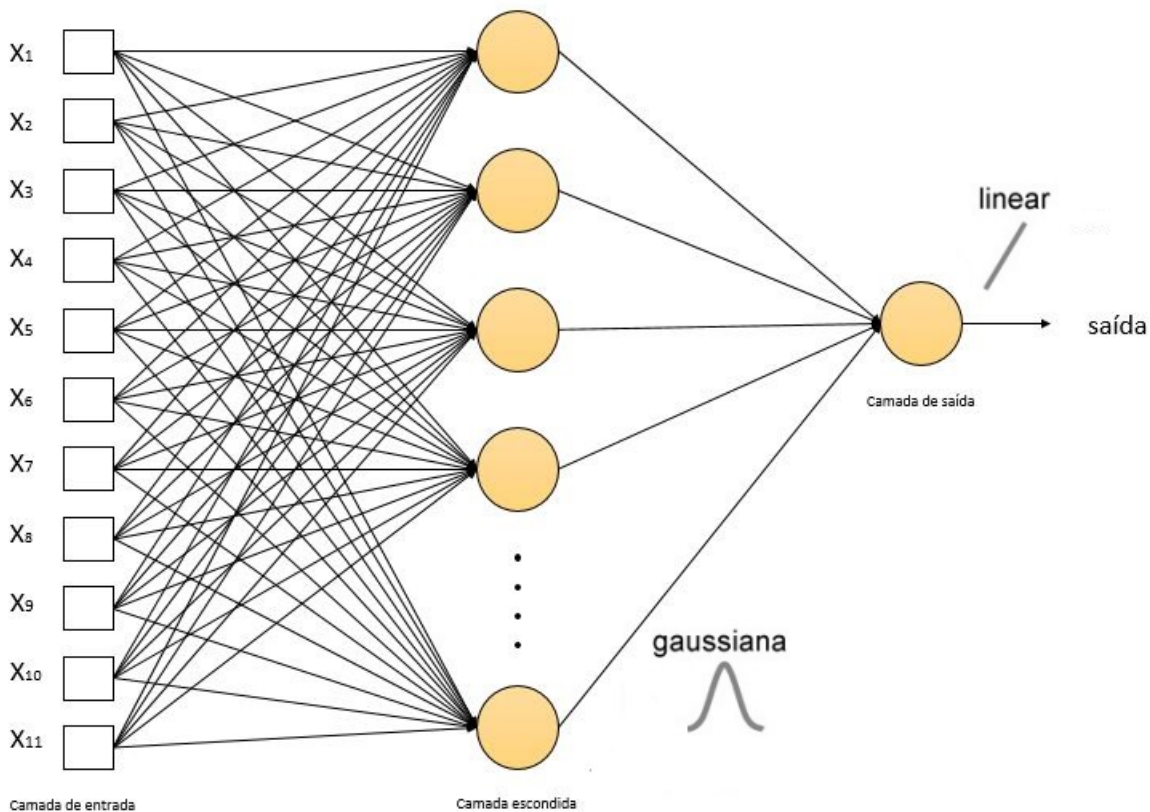


Figura 25 – Rede Neural Artificial do modelo proposto

4.3 Considerações Finais

Este Capítulo apresentou detalhes do modelo proposto nesta Dissertação que utiliza Redes Neurais Artificiais para identificar automaticamente a utilidade de *reviews online*. Foram definidas *features* com base nas descobertas de pesquisas anteriores e proposto um modelo de regressão que pode ser utilizado tanto para *reviews* do *site Steam*, quanto do *site Buscapé*. Nesse modelo são analisadas três categorias de características de *reviews*: características relacionadas ao conteúdo textual, características relacionadas ao autor e

metadados dos *review*. O modelo de regressão proposto analisa duas redes neurais, uma RNA MLP e uma RNA RBF. Os experimentos realizados pra definir a melhor topologia das redes, bem como os resultados produzidos com a execução dos modelos serão apresentados no próximo Capítulo.

5 Experimentos

Nesta seção, serão detalhados os experimentos executados sobre o modelo proposto e resultados obtidos.

5.1 Preparação dos dados

Todas *features* das bases *ReviewsSteam* e *ReviewsBucapé* foram pré-processadas de acordo com o descrito no Capítulo 4. Porém, como os dados estão em diferentes escalas, aplicamos um processo de padronização, ou remoção da média e escalonamento de variância. Dessa forma, todas as nossas *features* estão em uma mesma escala entre 0 e 1.

5.2 Experimento 1: Definição das melhores topologias das RNAs

Para a escolha da melhor topologia para as redes desenvolvidas neste trabalho, utilizou-se a validação cruzada *k-fold*, como descrita na seção 2.3.2.6 e definiu-se *k* com o valor 10. O desempenho global de cada topologia candidata é obtido em função da média entre os desempenhos individuais observados quando da aplicação das *k* partições.

As RNAs implementadas nesta pesquisa têm valores aleatórios como pesos iniciais, taxa de aprendizado 0.1 e *momentum* 0.3. Optou-se por valores pequenos para garantir a convergência do modelo, mesmo que o processo necessite de mais épocas. Para evitar *overfitting*¹, cada topologia foi treinada até que fosse encontrado o menor RMSE local.

Em relação aos dados usados no treinamento, a *bag-of-words* dos *reviews* da base *ReviewsSteam* foi gerada a partir do TF-IDF de todos os termos do Corpus. Optou-se por gerar o *bag-of-words* considerando colocações de até três termos (trigramas).

A Figura 26 apresenta um gráfico com os RMSE para cada uma das topologias testadas para a RNA MLP utilizando a base *ReviewsSteam*. Iniciamos os testes com uma topologia de 1 neurônio na camada escondida e aumentamos essa camada em 1 neurônio a cada nova topologia. O mesmo procedimento foi feito para RNA RBF utilizando a base *ReviewsSteam* e os resultados são apresentados na Figura 27.

Executando o treinamento da Rede MLP com uma variação de 1 a 300 neurônios, encontramos o menor RMSE ao atingir 201 neurônios, com valor de **0.1901**. Executando o mesmo experimento para a Rede RBF, encontrou-se **0.1915** como menor RMSE para uma topologia de 120 neurônios na camada escondida. A Tabela 4 apresenta os valores

¹ Em Mineração de Dados, é um termo usado para descrever quando um modelo se ajusta muito bem ao conjunto de treino, mas se mostra ineficaz para prever novos resultados

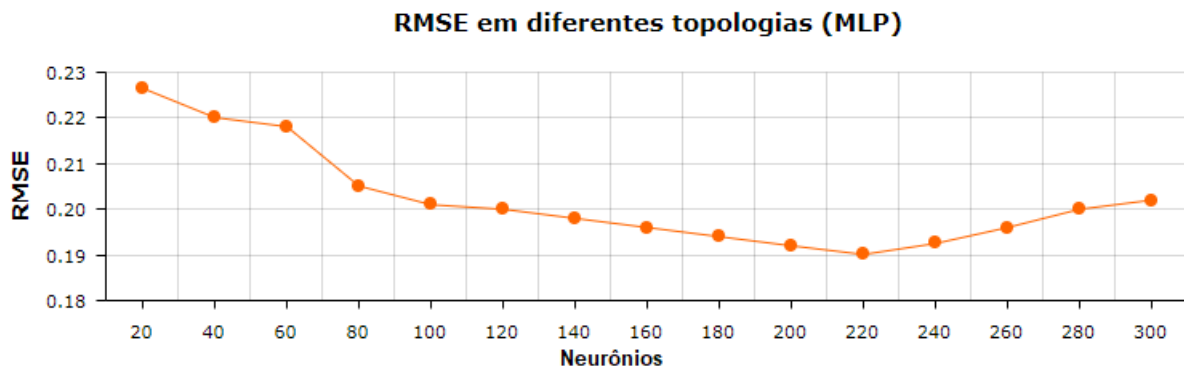


Figura 26 – RMSE para treino com Corpus *ReviewsSteam* e Rede MLP

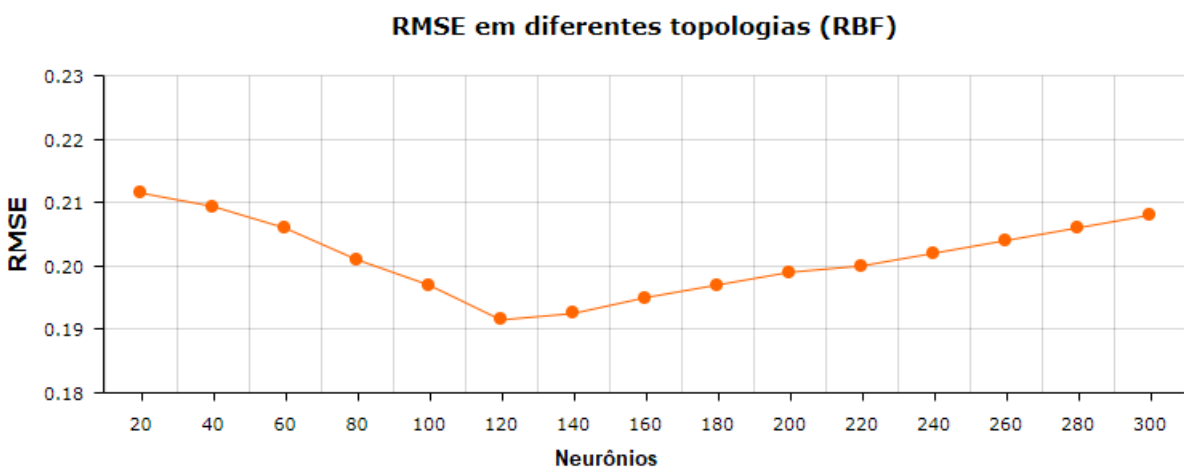


Figura 27 – RMSE para treino com Corpus *ReviewsSteam* e Rede RBF

de MAE, RMSE e R^2 para ambas as topologias. Os valores MAE e RMSE representam os erros médios do treinamento e o valor de R^2 dá indícios do quão preditivo é o nosso modelo. Ele fornece uma medida de quão bem os resultados observados são replicados, com base na proporção da variação total de resultados explicada pelo modelo (GLANTZ; SLINKER, 2000). Em estudos que analisam comportamento humano, valores bem abaixo de 0.5 para essa medida ainda são consideráveis aceitáveis, dada a complexidade de se realizar essa tarefa (KVALSETH, 1985). A Tabela 4 apresenta os valores obtidos por nossos experimentos e, como referência, a Tabela 5 apresenta os valores de R^2 obtidos por alguns trabalhos relacionados citados no Capítulo 3. Note que o valor R^2 é bem mais alto para o caso de *reviews* em que a utilidade foi anotada manualmente, ou seja, não depende do voto dos usuários, como é o caso deste trabalho, e sim da definição predeterminada por especialistas. Porém, não necessariamente esse modelo terá uma melhor performance na prática. É importante também destacar que esses trabalhos utilizaram algoritmos, Corporas e abordagens diferentes, então essa comparação não é direta ou conclusiva.

Para os próximos experimentos, o modelo MLP escolhido é uma rede de três camadas que consiste em 11 nós de entrada, 201 nós na camada escondida e 1 nó de saída.

Tabela 4 – Métricas de avaliação dos modelos MLP e RBF

Modelo	MAE	RMSE	R^2
MLP	0.1543	0.1901	0.2940
RBF	0.1542	0.1915	0.3031

Tabela 5 – R^2 em trabalhos relacionados

Modelo	R^2
(ZHANG; VARADARAJAN, 2006) <i>Reviews</i> de câmeras	0.3028
(YANG et al., 2015) <i>Reviews</i> anotados automaticamente	0.498
(YANG et al., 2015) <i>Reviews</i> anotados por humanos	0.702

Já o modelo RBF escolhido é uma rede de três camadas que consiste em 11 nós de entrada, 120 nós na camada escondida e 1 nó de saída.

5.3 Experimento 2: Variações do modelo de *bag-of-words*

Com as topologias definidas, foram feitas variações no modelo de *bag-of-words* de entrada a fim de determinar qual modelo nos dá o menor RMSE utilizando a base *ReviewsSteam*. Foram testadas *bag-of-words* geradas a partir da frequência dos termos (TF), TF-IDF considerando apenas unigramas, TF-IDF considerando colocações de até dois termos (bigramas), TF-IDF considerando colocações de até três termos (trigramas, a mesma utilizada no Experimento 1) e LSA com 50, 100 e 200 componentes respectivamente, utilizando TF-IDF com colocações de até três termos. Os resultados para ambas as redes e ambas as bases são apresentados na Tabela 6.

Tabela 6 – Métricas de avaliação dos modelos MLP e RBF

Modelo	TF	TF-IDF (1)	TF-IDF (2)	TF-IDF (3)	LSA (50)	LSA (100)	LSA (200)
MLP	0.2526	0.1994	0.1961	0.1901	0.1589	0.1574	0.1564
RBF	0.2440	0.1958	0.1919	0.1915	0.1609	0.1594	0.1578

Identificamos que, para ambas as redes, o menor RMSE é aquele dado pela *bag-of-words* que utiliza LSA com 200 componentes. Esse será o modelo que servirá de entrada para os próximos experimentos. A Tabela 7 apresenta as métricas de avaliação para essa topologia utilizando os Corpora *ReviewsSteam* e *ReviewsBuscape*. É importante notar que para em ambas as redes, o modelo com o menor RMSE também nos dá um valor menor MAE e, conseqüentemente, um valor mais alto para o coeficiente de determinação (R^2) se comparados com o modelo obtido no Experimento 1. Isso nos dá indícios de que o modelo que utiliza LSA com 200 componentes possui maior poder preditivo para *reviews* futuros.

Tabela 7 – Métricas de avaliação das redes MLP e RBF para os Corpora analisados

	ReviewsSteam			ReviewsBuscapé		
	MAE	RMSE	R^2	MAE	RMSE	R^2
MLP	0.1273	0.1564	0.5195	0.1521	0.1886	0.3120
RBF	0.1285	0.1578	0.5075	0.1508	0.1834	0.3433

5.4 Experimento 3: Seleção de *features* usando CFS

Com as melhores topologias e melhores *bag-of-words* selecionadas, definimos o melhor modelo para análise automática da utilidade de *reviews online*. Porém, é importante também saber qual das características de *review* oferecem o maior impacto na predição de utilidade. Para isso, foram feitas análises de correlação entre as todas as entradas, exceto a *bag-of-words*, dada a enorme dimensão e dificuldade de visualização; além de experimentos de seleção de *features* utilizando o algoritmo CFS. A partir desses experimentos, é possível limitar nosso modelo a certas variáveis de entrada que produzem o melhor RMSE.

As Tabelas 9 e 10 apresentam os valores de correlação linear (Pearson) entre as variáveis de entrada do modelo e a correlação entre as variáveis de entrada e a classe. Note que é importante que haja uma baixa correlação entre as variáveis independentes, mas alta correlação entre as variáveis independentes e a classe. Os valores podem variar de -1 a 1 e serão interpretados na próxima seção. Esses valores nos dão um direcionamento inicial sobre quais variáveis estão linearmente correlacionadas com a classe e entre si. Esse valores não são, no entanto, definitivos sobre essas relações entre as variáveis, já que esses relacionamentos podem não ser lineares. Como referência, a Tabela 8 apresenta o mapeamento das variáveis de entrada do modelo e da classe.

Aplicando o algoritmo CFS para dados discretos, podemos encontrar as *features* mais importantes do nosso modelo. Os resultados utilizando dois diferentes métodos de busca são apresentados na Tabela 11. O valor de Mérito é obtido pelo algoritmo CFS de acordo com a equação 2.17 e determina a utilidade de *features* individuais para prever a classe junto com o nível de intercorrelação entre eles (HALL, 1998), porém não há um consenso na literatura sobre quais valores representam um bom Mérito.

Realizando o treinamento do nosso modelo utilizando apenas as *features* selecionadas pelo algoritmo CFS nos dá os resultados apresentados na Tabela 12.

5.5 Resultados e Discussão

Em termos de RMSE, os modelos MLP e RBF obtiveram um valor semelhante para a base *ReviewsSteam*. Ambos RMSE e R^2 foram menores no treinamento dos modelos MLP e RBF para a base *ReviewsBuscapé*. É de se esperar que os resultados sejam diferentes entre as bases já que elas fazem parte de diferentes domínios, porém é possível que neste

Tabela 8 – Variáveis de entrada dos modelos propostos

Variável	Descrição
x_0	Quantidade de votos de reviews pela quantidade total de reviews de um usuário
x_1	Razão dos votos positivos pela quantidade total de votos de um usuário
x_2	Número de amigos do usuário na comunidade Steam/Buscapé
x_3	Quantidade de horas (em horas) que o autor dedicou ao jogo analisado para reviews do Steam; Confirmação de compra para reviews do Buscapé
x_4	Vetor contendo a quantidade de opiniões expressas no texto e as palavras do Corpus anotadas por TF ou TF-IDF (palavras não utilizadas nas análises de correlação e covariância)
x_5	Legibilidade do texto
x_6	Tamanho do texto em palavras
x_7	Número de sentenças no texto
x_8	Número de palavras monossílabas no texto
x_9	Diferença entre a avaliação média do produto e a avaliação do usuário
x_{10}	Idade do review em dias
H	Classe - percepção de utilidade de reviews

Tabela 9 – Tabela de correlação da base *ReviewsSteam*

	x_0	x_1	x_2	x_3	x_4^*	x_5	x_6	x_7	x_8	x_9	x_{10}	H
x_0	1.000											
x_1	0.211	1.000										
x_2	0.255	0.274	1.000									
x_3	-0.001	-0.086	0.040	1.000								
x_4^*	0.119	0.155	0.087	-0.010	1.000							
x_5	0.090	0.130	0.071	-0.001	0.716	1.000						
x_6	0.119	0.145	0.075	0.010	0.911	0.766	1.000					
x_7	0.122	0.133	0.083	-0.008	0.806	0.501	0.867	1.000				
x_8	0.114	0.139	0.072	0.009	0.899	0.759	0.996	0.858	1.000			
x_9	0.111	0.223	0.162	0.059	0.154	0.101	0.099	0.096	0.085	1.000		
x_{10}	0.050	0.011	0.010	0.196	-0.061	-0.049	-0.054	-0.061	-0.055	0.044	1.000	
H	0.140	0.232	0.133	0.090	0.236	0.189	0.227	0.219	0.215	0.669	0.029	1.000

estudo as diferenças se dão pela forma como as variáveis são mapeadas. No caso da base de dados *ReviewsBuscapé*, por exemplo, em que autores não possuem amigos, a variável que mapeia a quantidade de amigos na rede (x_2) foi mapeada como zero em todas as amostras e acabou não sendo utilizada no treinamento. Isso pode ter causado um grande impacto no RMSE, já que utilizando o algoritmo CFS, essa foi uma das variáveis selecionadas para o Corpus *ReviewsSteam* como uma das mais importantes.

Em termos de correlação, a reputação do autor (representada pela porcentagem de votos positivos do autor), as opiniões do texto, o tamanho do texto (em palavras, sentenças e monossílabas) e, principalmente, a diferença entre a avaliação média do produto e a avaliação do usuário são as variáveis mais correlacionadas com a classe na base *SteamReviews*, dando indícios de que usuários veem *reviews* que concordam com a

Tabela 10 – Tabela de correlação da base *ReviewsBuscape*

	x_0	x_1	x_2	x_3	x_4^*	x_5	x_6	x_7	x_8	x_9	x_{10}	H
x_0	1.000											
x_1	0.227	1.000										
x_2	-	-	-									
x_3	-0.010	-0.070	-	1.000								
x_4	0.146	0.151	-	-0.013	1.000							
x_5	0.111	0.125	-	-0.011	0.698	1.000						
x_6	0.153	0.154	-	-0.001	0.905	0.750	1.000					
x_7	0.154	0.148	-	-0.009	0.798	0.479	0.869	1.000				
x_8	0.149	0.151	-	-0.002	0.892	0.744	0.995	0.860	1.000			
x_9	0.060	0.138	-	0.086	0.083	0.016	0.025	0.032	0.014	1.000		
x_{10}	0.024	0.025	-	0.195	-0.052	-0.050	-0.050	-0.050	-0.051	0.173	1.000	
H	0.204	0.231	-	0.098	0.256	0.197	0.265	0.251	0.256	0.397	0.048	1.000

Tabela 11 – Resultados do CFS para cada modelo

Base	BestFirst	GreedyStepwise
ReviewsSteam	Mérito: 0.613	Mérito: 0.613
	Features:	Features:
	x_1	x_1
	x_2	x_2
	x_3	x_3
	x_4	x_4
	x_8	x_8
	x_9	x_9
Buscapé	Mérito: 0.456	Mérito: 0.456
	Features:	Features:
	x_1	x_1
	x_3	x_3
	x_4	x_4
	x_9	x_9
	x_{10}	x_{10}

opinião média dos demais usuários como sendo úteis.

Para a base *ReviewsBuscape*, além da razão da quantidade de votos pela quantidade de reviews de um usuário, as variáveis com maior correlação com a classe são as mesmas encontradas na base *ReviewsSteam*. Novamente a relação entre a avaliação final do *review* e avaliação média dos usuários possui o maior valor de correlação com a classe, mas em uma proporção bem menor.

Apesar disso, não é possível dizer que essas são as variáveis mais importantes do modelo, porque a análise de correlação leva em consideração apenas a correlação linear entre essas variáveis. Além disso, as variáveis independentes também se correlacionam entre si. Um exemplo são as variáveis x_6 , x_7 e x_8 que são fortemente correlacionadas entre

Tabela 12 – Comparação da performance dos modelos com seleção de *features*

	MAE	RMSE	R^2
MLP ReviewsSteam sem CFS	0.1273	0.1564	0.5195
MLP ReviewsSteam com CFS	0.1293	0.1597	0.5063
RBF ReviewsSteam sem CFS	0.1285	0.1578	0.5075
RBF ReviewsSteam com CFS	0.1280	0.1576	0.5116
MLP ReviewsBuscape sem CFS	0.1521	0.1886	0.3120
MLP ReviewsBuscape com CFS	0.1576	0.1951	0.2571
RBF ReviewsBuscape sem CFS	0.1508	0.1834	0.3433
RBF ReviewsBuscape com CFS	0.1589	0.1988	0.2564

si. Uma possível explicação é que elas podem representar a mesma informação (tamanho do texto). Uma forma de ponderar a relação entre as variáveis independentes e a classe com a relação das variáveis independentes entre si é o uso do algoritmo CFS.

A partir do algoritmo CFS, encontrou-se que, para a base *ReviewsSteam* sete variáveis independentes (mais especificamente $x_1, x_2, x_3, x_4, x_8, x_9, x_{10}$) são responsáveis pelas variações na classe H. Já para a base *ReviewsBuscape*, apenas cinco (especificamente $x_1, x_3, x_4, x_9, x_{10}$). Note que o tamanho do texto é representado apenas uma vez, na base *ReviewsSteam* como a contagem de palavras monossílabas.

Executando um treinamento das redes utilizando apenas as variáveis selecionadas pelo CFS nos dá os resultados obtidos na Tabela 12. Nota-se que os resultados não melhoraram com a exclusão das demais variáveis, porém tiveram uma mudança negligenciável, o que nos indica que, mesmo perdendo informações, é possível reduzir a quantidade de variáveis de entrada do modelo, o que pode, em alguns casos, representar um ganho de poder computacional.

Conclusões e Trabalhos Futuros

Este trabalho teve como objetivo estudar formas de quantificar e prever a percepção de utilidade de *reviews* de usuários *online* automaticamente através de um modelo que utiliza Redes Neurais Artificiais. Os resultados deste estudo foram positivos e dão indícios sobre quais características formam a percepção de utilidade de *reviews*.

Determinamos que é possível identificar a percepção da utilidade de *reviews* através de um modelo de regressão que tem como entrada características desses *reviews*. Porém, identificamos que as variáveis independentes propostas por nosso modelo não são suficientes para explicar totalmente a variância na utilidade de *reviews*, o que indica que há outros fatores envolvidos além daqueles estudados. Isso nos permite concluir que, apesar da resposta positiva à questão de pesquisa deste trabalho, não podemos determinar que o modelo proposto nesta Dissertação é o melhor para esta tarefa, já que existe outras características de *reviews* que não foram exploradas.

A partir dos resultados do treinamento do modelo, pode-se dizer que é possível prever a utilidade de *reviews online* a partir de características desses *reviews*, como informações sobre o autor, conteúdo textual e metadados. A proposta dessa dissertação oferece um modelo que apresenta coeficiente de determinação 0.51 na melhor configuração. Porém, esse número indica que esse provavelmente não é o melhor modelo e que ainda existem explicações para a variância da classe H que não foram exploradas por nossa proposta. Entre essas explicações, podem haver fatores psicológicos ou motivações implícitas, difíceis de serem mapeadas em variáveis numéricas. Sem uma *baseline* definida, é difícil estimar o quanto que o modelo proposto representa de progresso em relação ao que já existe na área.

De certa forma, os resultados desta análise corroboram pesquisas anteriormente realizadas por (DANESCU-NICULESCU-MIZIL et al., 2009), (SCHINDLER; BICKART, 2012) e (KIM et al., 2006). Notou-se que a avaliação do usuário e o tamanho do texto são características que impactam positivamente a percepção de utilidade de *reviews*. Já a data de postagem dos *reviews* aparenta não ter forte impacto na percepção utilidade nos dados analisados, o que contraria descobertas realizadas por (OTTERBACHER, 2009). Uma das possíveis justificativas poderia ser o fato de que nestes domínios (jogos e Smartphones) essa não é uma característica relevante para consumidores. Outra possível explicação é que, em ambos *Steam* e *Buscapé*, os *reviews* não são ordenados em função da data de postagem por padrão, o que torna essa informação irrelevante para a maioria dos usuários.

Entre as novas descobertas, tem-se que a reputação do autor é fator importante na percepção de utilidade de *reviews*. Descobriu-se que autores com histórico positivo de votos recebidos têm *reviews* melhor avaliados. A *expertise* do autor também aparenta ser

uma característica importante. No caso do *Steam*, a quantidade de horas jogadas por um autor no momento da publicação do *review* tem impacto positivo sobre a percepção de utilidade. Por fim, a quantidade de amigos do usuário na comunidade também possui algum impacto positivo sobre a percepção de utilidade de *reviews*.

Para o setor de *e-commerce*, estes resultados dão pistas sobre quais estruturas de *website* podem ser utilizadas para que os melhores *reviews* sejam lidos pelos consumidores e, conseqüentemente, maximizar as vendas.

Outra contribuição deste trabalho está na disponibilização dos artefatos gerados (dicionários e bases de dados), para que as pesquisas da área feitas em Português Brasileiro possam ser exploradas a partir de um mesmo ponto, facilitando as comparações entre os trabalhos.

Como principais problemas, encontrou-se: i) grande dificuldade em realizar análise semântica em textos escritos em Português na Internet; ii) diversas funções de linguagem em vigor na Internet; e iii) a quantidade de ferramentas ainda é bastante escassa. Esses problemas refletem a análise de padrões linguísticos de baixa qualidade. Para suavizar, fez-se o uso de *bag-of-words*.

Os resultados também estão alinhados com o que foi descoberto até então por outros pesquisadores, porém a diferença entre abordagens e bases utilizadas torna difícil comparar este trabalho com os já existentes.

Trabalhos Futuros

Como trabalhos futuros, é interessante utilizar métodos de análise de texto e mineração de opiniões mais robustos, como *Word Embeddings*, por exemplo utilizá-las no lugar de *bag-of-words* e tentar capturar a relação entre as palavras de um texto e seu significado semântico.

Outra proposta é fazer análise de regressão utilizando outros algoritmos e comparar os resultados com o que foi obtido através de Redes Neurais Artificiais. Além disso, o trabalho pode ser replicado utilizando bases de dados mais tradicionais como TripAdvisor e Amazon.

Também é importante validar o modelo com um experimento que simule um ambiente real de *reviews online*, com usuários reais votando nos *reviews* e comparando os resultados com aqueles previstos por nosso modelo. Isso pode ser feito em *sites* de *e-commerce* já existentes. No entanto, esses *sites* devem se adequar ao modelo de redes sociais existente no *Steam* ou incorporar dados de Redes Sociais Online já existentes, como o *Facebook*, permitindo que sejam mapeadas as relações entre os usuários.

Referências

- AGRESTI, A.; COULL, B. A. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, American Statistical Association, v. 52, n. 2, p. 119–126, maio 1998. ISSN 00031305. Disponível em: <<http://dx.doi.org/10.2307/2685469>>. Citado na página 13.
- ALUÍSIO, S. et al. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: *6th Int. Conf. on Computational Processing Of Portuguese Language (PROPOR)*. [S.l.: s.n.], 2003. p. 110–117. Citado 2 vezes nas páginas 31 e 48.
- BECKER, K.; TUMITAN, D. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. In: *Anais do 28 Simpósio Brasileiro de Banco de Dados*. [S.l.: s.n.], 2013. Citado 3 vezes nas páginas 15, 26 e 28.
- BERGER, J.; SORENSEN, A. T.; RASMUSSEN, S. J. Positive effects of negative publicity: when negative reviews increase sales. *Marketing Science*, 29 (5), p. 815–827, 2010. Citado na página 10.
- BICKART, B.; SCHINDLER, R. M. Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*, John Wiley and Sons, Inc., v. 15, n. 3, p. 31–40, 2001. ISSN 1520-6653. Disponível em: <<http://dx.doi.org/10.1002/dir.1014>>. Citado na página 10.
- BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2009. ISBN 0596516495, 9780596516499. Citado na página 31.
- BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. [S.l.]: O'Reilly, 2012. Nenhuma citação no texto.
- BOND, R.; SMITH, P. B. Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, v. 119, n. 1, p. 111–137, 1996. Citado na página 52.
- BROOMHEAD, D.; LOWE, D. Radial basis functions, multi-variable functional interpolation and adaptive networks. RSRE-MEMO-4148, 03 1988. Citado na página 17.
- CAO, Q.; DUAN, W.; GAN, Q. Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach. *Decision Support Systems*, v. 50, n. 2, p. 511 – 521, 2011. ISSN 0167-9236. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167923610001909>>. Citado na página 46.
- CHEN, C. et al. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018. p. 602–607. Disponível em: <<http://aclweb.org/anthology/N18-2095>>. Citado na página 40.

CHEN, P.-Y.; WU, S.-y.; YOON, J. The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings. Paper 58*, 2004. Citado 2 vezes nas páginas 11 e 12.

CHEVALIER, J. A.; MAYZLIN, D. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research Vol. 43, No. 3*, p. 345–354, 2006. Citado na página 10.

CONNORS, L.; MUDAMBI, S. M.; SCHUFF, D. Is it the review or the reviewer? a multimethod approach to determine the antecedents of online review helpfulness. In: *In 44th Hawaii International Conference on System Sciences*. [S.l.: s.n.], 2011. Citado 3 vezes nas páginas 40, 41 e 44.

COSTA, S. R.; PEREIRA, A. P. M. S. O teclar e o escrever: conceitos e preconceitos linguísticos, culturais e sociais. *Estudos em avaliação educacional*, v. 29, p. 87–110, 2004. Disponível em: <<http://publicacoes.fcc.org.br/ojs/index.php/eae/article/view/2160>>. Citado na página 49.

DANESCU-NICULESCU-MIZIL, C. et al. How opinions are received by online communities: A case study on amazon.com helpfulness votes. *18th International Conference on World Wide Web*, 2009. Citado 5 vezes nas páginas 40, 42, 44, 52 e 65.

DANTAS, J. C. A norma oculta: língua e poder na sociedade brasileira. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, scielo, v. 20, p. 181 – 184, 06 2004. ISSN 0102-4450. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502004000100009&nrm=iso>. Citado na página 50.

DARBY, M. R.; KARNI, E. Free competition and the optimal amount of fraud. *The Journal of Law and Economics*, [University of Chicago Press, Booth School of Business, University of Chicago, University of Chicago Law School], v. 16, n. 1, p. 67–88, 1973. ISSN 00222186, 15375285. Disponível em: <<http://www.jstor.org/stable/724826>>. Citado na página 11.

DASGUPTA, D. Advances in artificial immune systems. *Comp. Intell. Mag.*, IEEE Press, Piscataway, NJ, USA, v. 1, n. 4, p. 40–49, nov. 2006. ISSN 1556-603X. Disponível em: <<http://dx.doi.org/10.1109/MCI.2006.329705>>. Nenhuma citação no texto.

DELASHMIT, W. H.; MANRY, M. T. Recent developments in multilayer perceptron neural networks. In: *Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC*. [S.l.: s.n.], 2005. Citado na página 18.

DELLAROCAS, C. The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, v. 49, n. 10, p. 1407–1424, 2003. Citado na página 10.

DELLAROCAS, C.; ZHANG, X. M.; AWAD, N. F. Exploring the value of online product ratings in revenue forecasting: The case of motion pictures. *Journal of Interactive Marketing*, v. 21, n. 4, p. 23–45, 2007. Citado na página 10.

DIAZ, G. O.; NG, V. Modeling and prediction of online product review helpfulness: A survey. In: *Proceedings of 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2018. p. 698–708. Citado na página 39.

- D'INVERNO, M.; LUCK, M. *Understanding Agent Systems, second edition*. [S.l.]: Springer-Verlag Berlin Heidelberg, 2004. (Springer Series on Agent Technology). Nenhuma citação no texto.
- DUMAIS, S. T. Latent semantic analysis. *Annual Review of Information Science and Technology*, v. 38, n. 1, p. 188–230, 2004. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440380105>>. Citado na página 30.
- FORD, G. T.; SMITH, D. B.; SWASY, J. L. An empirical test of the search, experience and credence attributes framework. *ACR North American Advances*, 1988. Citado na página 11.
- FORMAN, C.; GHOSE, A.; WIESENFELD, B. *Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets*. [S.l.]: INFORMS, 2008. 291—313 p. Citado 4 vezes nas páginas 12, 40, 41 e 44.
- GARDNER, M.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, v. 32, n. 14, p. 2627 – 2636, 1998. ISSN 1352-2310. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1352231097004470>>. Citado na página 18.
- GHOSE, A.; IPEIROTIS, P. G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, v. 23, n. 10, p. 1498–1512, Oct 2011. ISSN 1041-4347. Citado 3 vezes nas páginas 40, 42 e 44.
- GLANTZ, S.; SLINKER, B. *Primer of Applied Regression & Analysis of Variance*. McGraw-Hill Education, 2000. ISBN 9780071360869. Disponível em: <<https://books.google.com.br/books?id=fzV2QgAACAAJ>>. Citado na página 58.
- GOLDENBERG, J.; LIBAI, B.; MULLER, E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters, Volume 12, Issue 3*, p. 211–223, 2001. Citado na página 2.
- HALL, M. A. *Correlation-based Feature Subset Selection for Machine Learning*. Tese (Doutorado) — University of Waikato, Hamilton, New Zealand, 1998. Citado 2 vezes nas páginas 25 e 60.
- HAMILTON, R.; VOHS, K. D.; MCGILL, A. L. We'll be honest, this won't be the best article you'll ever read: The use of dispreferred markers in word-of-mouth communication. *Journal of Consumer Research, Inc.*, v. 4, p. 197–212, 2014. Citado na página 3.
- HEBB, D. O. *The organization of behavior: A neuropsychological theory*. [S.l.]: Wiley, 1949. Hardcover. ISBN 0-8058-4300-0. Citado na página 17.
- HONG, Y. et al. What reviews are satisfactory: Novel features for automatic helpfulness voting. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2012. (SIGIR '12), p. 495–504. ISBN 978-1-4503-1472-5. Disponível em: <<http://doi.acm.org/10.1145/2348283.2348351>>. Citado 2 vezes nas páginas 40 e 44.

HOPFIELD, J. J. Neurocomputing: Foundations of research. In: ANDERSON, J. A.; ROSENFELD, E. (Ed.). Cambridge, MA, USA: MIT Press, 1988. cap. Neural Networks and Physical Systems with Emergent Collective Computational Abilities, p. 457–464. ISBN 0-262-01097-6. Disponível em: <<http://dl.acm.org/citation.cfm?id=65669.104422>>. Citado na página 17.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, v. 2, n. 5, p. 359 – 366, 1989. ISSN 0893-6080. Disponível em: <<http://www.sciencedirect.com/science/article/pii/0893608089900208>>. Citado na página 18.

HU, N.; KOH, N. S.; REDDY, S. K. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision Support Systems*, v. 57, p. 42–53, 2014. Disponível em: <<http://dblp.uni-trier.de/db/journals/dss/dss57.html#HuKR14>>. Citado na página 12.

HU, N.; LIU, L.; ZHANG, J. Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Information Technology and Management Vol. 9 No. 3*, p. 201–214, 2008. Citado na página 3.

HUANG, A. H. et al. A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, Elsevier, v. 48, p. 17–27, 2015. Citado 2 vezes nas páginas 41 e 44.

IYENGAR, S. S.; LEPPER, M. R. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, Vol 79(6), p. 995–1006, 2000. Citado na página 1.

JACKSON, P.; MOULINIER, I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins Pub., 2007. (Natural language processing). ISBN 9789027249920. Disponível em: <<https://books.google.com.br/books?id=WzJa1Wpv8UwC>>. Citado na página 25.

KENNEDY, J.; EBERHART, R. C.; SHI, Y. Swarm intelligence / j. kennedy, r.c. eberhart, y. shi. 12 2018. Nenhuma citação no texto.

KHAN, M. *Consumer Behaviour*. New Age International (P) Limited, 2007. ISBN 9788122415315. Disponível em: <<https://books.google.com.br/books?id=BxGw27yTVHIC>>. Citado 2 vezes nas páginas 8 e 10.

KIM, S.-M. et al. Automatically assessing review helpfulness. In: JURAFSKY, D.; GAUSSIÉ, E. (Ed.). *EMNLP*. ACL, 2006. p. 423–430. ISBN 1-932432-73-6. Disponível em: <<http://dblp.uni-trier.de/db/conf/emnlp/emnlp2006.html>>. Citado 8 vezes nas páginas 12, 13, 39, 40, 44, 46, 51 e 65.

KIM, Y. A.; SRIVASTAVA, J. Impact of social influence in e-commerce decision making. In: *Proceedings of the Ninth International Conference on Electronic Commerce*. New York, NY, USA: ACM, 2007. (ICEC 07), p. 293–302. ISBN 978-1-59593-700-1. Disponível em: <<http://doi.acm.org/10.1145/1282100.1282157>>. Citado na página 2.

KINCAID, J. P. et al. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*.

- [S.l.], 1975. Disponível em: <<http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED108134>>. Citado na página 50.
- KOHONEN, T. Neurocomputing: Foundations of research. In: ANDERSON, J. A.; ROSENFELD, E. (Ed.). Cambridge, MA, USA: MIT Press, 1988. cap. Self-organized Formation of Topologically Correct Feature Maps, p. 509–521. ISBN 0-262-01097-6. Disponível em: <<http://dl.acm.org/citation.cfm?id=65669.104428>>. Citado na página 18.
- KOTLER, P.; ARMSTRONG, G. *Principles of Marketing*. Pearson, 2010. (The Prentice-Hall series in marketing). ISBN 9780137006694. Disponível em: <<https://books.google.com.br/books?id=ZW2u5LOmbs4C>>. Citado 3 vezes nas páginas 7, 8 e 9.
- KOVÁCS, Z. *Redes Neurais Artificiais*. LIVRARIA DA FISICA, 2002. ISBN 9788588325142. Disponível em: <<https://books.google.com.br/books?id=O0nLxR67wmUC>>. Citado na página 16.
- KRIESEL, D. *A Brief Introduction to Neural Networks*. Zeta2. [s.n.], 2007. Disponível em: <http://www.dkriesel.com/science/neural_networks>. Citado na página 17.
- KRISHNAMOORTHY, S. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, v. 42, n. 7, p. 3751 – 3759, 2015. ISSN 0957-4174. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0957417414008239>>. Citado 2 vezes nas páginas 40 e 44.
- KVALSETH, T. O. Cautionary note about r 2. *The American Statistician*, Taylor and Francis, v. 39, n. 4, p. 279–285, 1985. Disponível em: <<https://doi.org/10.1080/00031305.1985.10479448>>. Citado na página 58.
- LACZNIAK, R. N.; DECARLO, T. E.; RAMASWAMI, S. N. Consumers' responses to negative word-of-mouth communication: An attribution theory perspective. *Journal of Consumer Psychology*, v. 11, n. 1, p. 57 – 73, 2001. ISSN 1057-7408. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1057740801703117>>. Citado na página 11.
- LANDAUER, T.; FOLTZ, P.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, ABLEX PUBLISHING CO, v. 25, p. 259–284, 1998. Citado na página 30.
- LEE, S.; CHOE, J. Y. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Syst. Appl.*, v. 41, n. 6, p. 3041–3046, 2014. Disponível em: <<http://dblp.uni-trier.de/db/journals/eswa/eswa41.html#LeeC14a>>. Citado 6 vezes nas páginas 20, 23, 24, 39, 44 e 51.
- LEMONSTAND. *Global eCommerce Market Size*. <http://blog.lemonstand.com/just-how-big-is-the-ecommerce-market-youll-never-guess>. 2015. [Online; accessed 25-September-2015]. Disponível em: <<http://blog.lemonstand.com/just-how-big-is-the-ecommerce-market-youll-never-guess>>. Citado na página 1.

LI, M. et al. Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, v. 17, p. 101–136, 2013. Citado 4 vezes nas páginas 3, 13, 14 e 38.

LIU, B. Sentiment analysis and subjectivity. In: *Handbook of Natural Language Processing, Second Edition*. Taylor and Francis Group, Boca. [S.l.: s.n.], 2010. Citado na página 27.

LIU, B. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. (Synthesis Lectures on Human Language Technologies). Disponível em: <<http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016>>. Citado 4 vezes nas páginas 25, 26, 27 e 28.

LIU, H. et al. Using argument-based features to predict and analyse review helpfulness. *CoRR*, abs/1707.07279, 2017. Disponível em: <<http://arxiv.org/abs/1707.07279>>. Citado 3 vezes nas páginas 40, 41 e 44.

LIU, J. et al. Low-quality product review detection in opinion summarization. In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. [S.l.: s.n.], 2007. p. 334–342. Poster paper. Citado na página 42.

LIU, J.; WU, G.; YAO, J. Opinion searching in multi-product reviews. In: *The Sixth IEEE International Conference on Computer and Information Technology (CIT'06)*. [S.l.: s.n.], 2006. p. 25–25. Citado 2 vezes nas páginas 27 e 28.

LU, Y. et al. Exploiting social context for review quality prediction. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 691–700. ISBN 978-1-60558-799-8. Disponível em: <<http://doi.acm.org/10.1145/1772690.1772761>>. Citado 2 vezes nas páginas 39 e 44.

MALIK, M.; HUSSAIN, A. Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, v. 73, p. 290–302, 08 2017. Citado 2 vezes nas páginas 40 e 44.

MAURI, A. G.; MINAZZI, R. Web reviews influence on expectations and purchasing intentions of hotel potential customers. *International Journal of Hospitality Management*, v. 34, p. 99 – 107, 2013. ISSN 0278-4319. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0278431913000236>>. Citado na página 11.

MICHALEWICZ, Z. *Genetic Algorithms + Data Structures = Evolution Programs (3rd Ed.)*. Berlin, Heidelberg: Springer-Verlag, 1996. ISBN 3-540-60676-9. Nenhuma citação no texto.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. Citado 2 vezes nas páginas 14 e 15.

MOGHADDAM, S.; JAMALI, M.; ESTER, M. Etf: Extended tensor factorization model for personalizing prediction of review helpfulness. In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2012. (WSDM '12), p. 163–172. ISBN 978-1-4503-0747-5. Disponível em: <<http://doi.acm.org/10.1145/2124295.2124316>>. Citado 2 vezes nas páginas 39 e 44.

MUDAMBI, S. M.; SCHUFF, D. What makes a helpful online review? a study of customer reviews on amazon.com. *MIS Quarterly*, v. 34, n. 1, p. 185–200, 2010. Disponível em: <<http://dblp.uni-trier.de/db/journals/misq/misq34.html#MudambiS10>>. Citado na página 42.

MUNROE, R. *Reddit's new comment sorting system in* <http://www.redditblog.com/2009/10/reddits-new-comment-sorting-system.html>. 2009. Disponível em: <<http://www.redditblog.com/2009/10/reddits-new-comment-sorting-system.html>>. Citado na página 13.

MURRAY, K. B.; SCHLACTER, J. L. The impact of services versus goods on consumers' assessment of perceived risk and variability. *Journal of the Academy of Marketing Science*, v. 18, n. 1, p. 51–65, Dec 1990. ISSN 1552-7824. Disponível em: <<https://doi.org/10.1007/BF02729762>>. Citado na página 11.

NELSON, P. J. Economics of information. In: _____. Dordrecht: Springer Netherlands, 1981. cap. Consumer Information and Advertising, p. 42–77. ISBN 978-94-009-8168-3. Disponível em: <http://dx.doi.org/10.1007/978-94-009-8168-3_5>. Citado na página 11.

NGO-YE, T. L.; SINHA, A. P. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, v. 61, p. 47 – 58, 2014. ISSN 0167-9236. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167923614000128>>. Citado 2 vezes nas páginas 41 e 42.

ORHAN, U.; HEKIM, M.; OZER, M. Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, v. 38, n. 10, p. 13475 – 13481, 2011. ISSN 0957-4174. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0957417411006762>>. Citado 2 vezes nas páginas 19 e 20.

OTTERBACHER, J. 'helpfulness' in online communities: A measure of message quality. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. [S.l.]: ACM, 2009. (CHI '09), p. 955–964. Citado 5 vezes nas páginas 40, 41, 44, 52 e 65.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Nenhuma citação no texto.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, v. 65 6, p. 386–408, 1958. Citado na página 17.

ROSS, T. *Fuzzy logic with engineering applications*. McGraw-Hill, 1995. ISBN 9780070539174. Disponível em: <<https://books.google.com.br/books?id=XJxRAAAAMAJ>>. Nenhuma citação no texto.

Sá, C. A. de; VIEIRA, J. P. A.; MOURA, R. S. Approach to define author reputation in web product reviews using artificial neural networks. In: *2017 XLIII Latin American Computer Conference (CLEI)*. [S.l.: s.n.], 2017. p. 1–10. Citado na página 37.

SANTOS, R. et al. Evaluating the importance of web comments through metrics extraction and opinion mining. In: . [S.l.: s.n.], 2016. p. 1–11. Citado na página 37.

SANTOS, R. L. de S. *Um Estudo Comparativo entre Abordagens baseadas em Sistemas Fuzzy e Redes Neurais Artificiais para Estimar a Importância de Comentários Web*. Dissertação (Mestrado) — Universidade Federal do Piauí, 2017. Citado 2 vezes nas páginas 25 e 31.

SCHINDLER, R.; BICKART, B. Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour*, v. 11, p. 234–243, 2012. Citado 5 vezes nas páginas 10, 40, 41, 44 e 65.

SCHWARTZ, B. *The Paradox of Choice: Why More Is Less*. Harper Perennial, 2005. Paperback. ISBN 0060005696. Disponível em: <<http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0060005696>>. Citado na página 2.

SEN, S.; LERMAN, D. Why are you telling me this? an examination into negative consumer reviews on the web. *Journal of Interactive Marketing*, v. 21, p. 76–96, 2007. Citado na página 41.

SILVA, I. N. da; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais artificiais: para engenharia e ciências aplicadas. In: _____. [S.l.]: Artliber, 2010. cap. Redes Perceptron Multicamadas. Citado 4 vezes nas páginas 15, 20, 21 e 23.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*. [S.l.: s.n.], 2012. p. 218–228. Citado 4 vezes nas páginas 15, 16, 18 e 19.

SINHA, R. R.; SWEARINGEN, K. Comparing recommendations made by online systems and friends. In: *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*. [s.n.], 2001. Disponível em: <<http://dblp.uni-trier.de/db/conf/delos/delos2001.html#SinhaS01>>. Citado na página 2.

SOARES, M. V. B. et al. *Pretext ii: descrição da reestruturação da ferramenta de pré-processamento de textos*. São Carlos, SP, Brasil., 2008. Citado 2 vezes nas páginas 15 e 29.

SOLOMON, M. *Consumer Behavior: Buying, Having, and Being*. Pearson/Prentice Hall, 2004. ISBN 9780131404069. Disponível em: <<https://books.google.com.br/books?id=keC7AAAAIAAJ>>. Citado 3 vezes nas páginas 8, 9 e 10.

SOUSA, R. F. d.; RABELO, R. A. L.; MOURA, R. S. A fuzzy system-based approach to estimate the importance of online customer reviews. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015. Citado 3 vezes nas páginas 37, 38 e 47.

SPOOL, J. M. *The Magic Behind Amazon's 2.7 Billion Dollar Question*. 2009. Disponível em: <<https://www.uie.com/articles/magicbehindamazon>>. Citado 2 vezes nas páginas 12 e 13.

SQUARISI, D.; SALVADOR, A. A arte de escrever bem: um guia para jornalistas e profissionais do texto. In: . [S.l.]: Editora Contexto, 2005. p. 50–52. ISBN 85-7244-279-0. Citado na página 50.

TRIPADVISOR. *Two in three more likely to book hotel with management responses to reviews*. *www.tripadvisor.co.uk*. 2013. Disponível em: <http://www.tripadvisor.co.uk/PressCenter-i6603-c1-Press_Releases.html>. Citado na página 2.

TSUR, O.; RAPPOPORT, A. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In: ADAR, E. et al. (Ed.). *ICWSM*. The AAAI Press, 2009. ISBN 978-1-57735-421-5. Disponível em: <<http://dblp.uni-trier.de/db/conf/icwsm/icwsm2009.html#TsurR09>>. Citado na página 42.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 3, p. 478–514, maio 2012. ISSN 1384-5810. Disponível em: <<http://dx.doi.org/10.1007/s10618-011-0238-6>>. Citado na página 27.

TURNEY, P. D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ACL '02), p. 417–424. Disponível em: <<http://dx.doi.org/10.3115/1073083.1073153>>. Citado 3 vezes nas páginas 15, 47 e 48.

VOHS, K. D. et al. Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *Motivation Science*, Vol 1(S), p. 19–42, 2014. Citado na página 1.

WATHEN, C. N.; BURKELL, J. Believe it or not: Factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology*, John Wiley Sons, Inc., v. 53, n. 2, p. 134–144, 2002. Citado na página 46.

WIDROW, B.; HOFF, M. E. Neurocomputing: Foundations of research. In: ANDERSON, J. A.; ROSENFELD, E. (Ed.). Cambridge, MA, USA: MIT Press, 1988. cap. Adaptive Switching Circuits, p. 123–134. ISBN 0-262-01097-6. Disponível em: <<http://dl.acm.org/citation.cfm?id=65669.104390>>. Citado na página 17.

WILSON, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, American Statistical Association, v. 22, n. 158, p. 209–212, 1927. ISSN 01621459. Disponível em: <<http://dx.doi.org/10.2307/2276774>>. Citado na página 13.

WRIGHT, R. *Consumer Behaviour*. Thomson Learning, 2006. ISBN 9781844801381. Disponível em: <https://books.google.com.br/books?id=R329I_4nQOgC>. Citado na página 9.

YANG, Y. et al. Semantic analysis and helpfulness prediction of text for online product reviews. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2015. p. 38–44. Disponível em: <<http://aclweb.org/anthology/P15-2007>>. Citado 4 vezes nas páginas 39, 42, 44 e 59.

YE, Q.; LAW, R.; GU, B. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, v. 28, n. 1, p. 180 – 182, 2009. ISSN 0278-4319. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0278431908000546>>. Citado na página 12.

YELP. *The Most Romantic City on Yelp is... in* <http://officialblog.yelp.com/2011/02/the-most-romantic-city-on-yelp-is.html>. 2011. Disponível em: <<http://officialblog.yelp.com/2011/02/the-most-romantic-city-on-yelp-is.html>>. Citado na página 13.

YOON, Y.; GUIMARAES, T.; SWALES, G. Integrating artificial neural networks with rule-based expert systems. *Decis. Support Syst.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 11, n. 5, p. 497–507, jun. 1994. ISSN 0167-9236. Disponível em: <[http://dx.doi.org/10.1016/0167-9236\(94\)90021-3](http://dx.doi.org/10.1016/0167-9236(94)90021-3)>. Citado na página 25.

ZHANG, Z.; VARADARAJAN, B. Utility scoring of product reviews. In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2006. (CIKM '06), p. 51–57. ISBN 1-59593-433-2. Disponível em: <<http://doi.acm.org/10.1145/1183614.1183626>>. Citado 3 vezes nas páginas 39, 44 e 59.

ZHU, F.; ZHANG, X. M. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, v. 74, n. 2, p. 133–148, 2010. Citado na página 12.