UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Classificação da utilidade de opiniões em português brasileiro

**Rogério Figueredo de Sousa**

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

**Rogério Figueredo de Sousa**

# Classificação da utilidade de opiniões em português brasileiro

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo

**USP – São Carlos**
**Fevereiro de 2023**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S725c

Sousa, Rogério Figueredo de
Classificação da utilidade de opiniões em português brasileiro / Rogério Figueredo de Sousa; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2023.
132 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Processamento de Língua Natural. 2. Utilidade de Opiniões. 3. Mineração de Opiniões. 4. Aprendizado de Máquina. I. Pardo, Thiago Alexandre Salgueiro, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

**Rogério Figueredo de Sousa**

# Helpfulness classification of online reviews in brazilian portuguese

Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo

**USP – São Carlos**
**February 2023**

*Este trabalho é dedicado à minha linda e amada esposa, Milena Carvalho, minhas filhinhas Yasmin e Serena, aos meus pais, Rubem e Rosa, aos meus irmãos, Rafael, Rebeca e Rubem Filho, aos meus amigos, colegas, professores e todas as pessoas que pelo menos por algum momento estiveram presentes no decorrer dessa jornada.*

# AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, pois Ele me deu vida e capacidade durante toda essa jornada. Sem a graça dEle eu nada poderia fazer.

Sou muito grato também à minha amada esposa, Milena, por todo carinho e atenção durante esses anos, e pelo seu amor incondicional ao me acompanhar para outra cidade até então desconhecida e sempre estar ao meu lado, independentemente da situação. A minha princesinha linda Yasmin, "produção" do doutorado, mesmo pequenininha, entendeu os momentos em que eu não podia dar a atenção que ela merece. A Serena que, no momento em que escrevo estas palavras, ainda está no "forninho", pois ela já teve participação indispensável nesse fim de doutorado.

Agradeço aos meus pais, Rubem e Rosa, que dispensaram incontáveis esforços para que eu pudesse ter uma boa educação dentro e fora de casa. Além de proverem todos os recursos necessários para meu crescimento saudável, tanto físico quanto mental e por serem meus exemplos de vida. Aos meus irmãos, Rafael, Rebeca e Rubem Filho pela amizade e companheirismo que só eles poderiam me proporcionar.

Aos meus professores da Universidade Federal do Piauí, especialmente ao Prof. Dr. Raimundo Santos Moura, por aceitar me orientar como aluno de iniciação científica, trabalho de conclusão de curso e de mestrado e me acompanhado com excelentes conselhos e ensinamentos que me ajudaram em todas as etapas da minha vida acadêmica.

Aos meus queridos orientadores, Profa. Dra. Maria das Graças Volpe Nunes e Prof. Dr Thiago Alexandre Salgueiro Pardo. A professora Graça foi fundamental no início do doutorado, pelos seus valiosos conselhos e conhecimentos compartilhados; sempre atenciosa e disposta a ajudar para me encaminhar para os melhores caminhos. Ao professor Thiago, por aceitar ser meu orientador, sempre com paciência e zelo. Da mesma forma, agradeço pela parceria, ensinamentos e pelos muitos conhecimentos que compartilhamos. Sem os dois, eu não teria chegado até esse momento, ambos são incríveis e são exemplos a serem seguidos, pessoal e profissionalmente.

Aos meus companheiros de NILC, Henrico, Ana Carolina, Ana Caroline, Sidney, Edresson, Marco, Márcio, Renata, Laura e João Paulo, com quem compartilhei vários momentos de descontração e de trabalho que levarei por toda vida.

Aos amigos que fiz em São Carlos, principalmente da Igreja Presbiteriana de São Carlos, especialmente o Rafael da Rosaine, Rosaine do Rafael, Pastor Will, Carol e Gabriel, sei que nos acompanharão para sempre. Eles são um presente para nós, amigos mais chegados que

irmãos. Aos meus amigos que já levei do Piauí, Roney, Rafael e Alissa, sempre companheiros, atenciosos, prontos a ajudar, também mais que amigos, irmãos. Amo todos vocês!

E por fim, agradeço à Universidade de São Paulo, ao Instituto de Ciências Matemáticas e de Computação, ao *Center for Artificial Inteligence* (C4AI) e ao Instituto Federal de Ciência e Tecnologia do Piauí por todo suporte financeiro e estrutural dispensados durante o desenvolvimento dessa pesquisa.

# RESUMO

SOUSA, R. F. **Classificação da utilidade de opiniões em português brasileiro**. 2023 . 132 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023 .

A quantidade de dados gerados pelos usuários na *Web* está aumentando. Entre esses conteúdos, as opiniões são consideradas um caso especial. Esse tipo de texto geralmente inclui coloquialismos, ruídos, erros, gírias, abreviações, etc. Portanto, eles são difíceis de serem processados por máquinas e podem ser difíceis de serem lidos até por seres humanos, em alguns casos. Particularmente, para um consumidor que procura conteúdo útil e de qualidade para ajudar nas suas decisões, como escolher um produto para comprar ou um filme para assistir, esta tarefa está se tornando cada vez mais complicada, devido aos problemas mencionados anteriormente e a grande oferta de opiniões na *Web*. Nesse contexto, surgiu a tarefa de Modelagem e Predição da Utilidade de Opiniões, cujo principal objetivo é estudar, modelar e processar opiniões geradas por usuários, a fim de selecionar automaticamente as mais úteis e destacá-las para ajudar outros usuários. Prever a utilidade das opiniões não é uma tarefa simples. Muita informação é necessária para caracterizar a utilidade das opiniões e, além disso, a utilidade é considerada um critério subjetivo, dependente de fatores extra-textuais, como a necessidade de informações do próprio leitor e o tempo disponível para leitura e avaliação de opiniões suficientes. Muitos trabalhos foram realizados desde a origem da área, mas, para a língua portuguesa, poucos avanços foram realizados até o momento. Este trabalho de doutorado teve como objetivo investigar e propor métodos para a tarefa de classificação automática da utilidade de opiniões para a língua portuguesa, utilizando informações linguísticas e de metadados disponíveis. Para atingir esse objetivo, um córpus de dois domínios, aplicativos para smartphones e filmes, foi coletado e anotado. Avaliou-se qualitativamente e quantitativamente uma ampla gama de atributos e técnicas que pudessem caracterizar a utilidade das opiniões e, dessa forma, foram descobertos fatores relevantes para a discriminação das opiniões úteis das não úteis. Nesta tese de doutorado, foram discutidos os principais desafios da área de pesquisa, e foi estabelecido um *benchmark* para a tarefa na língua portuguesa. Além disso, desenvolveu-se um novo método baseado em grafos que pode ser usado como alternativa para classificação da utilidade de opiniões. Por fim, elaboramos um método que pode classificar com excelente acurácia as opiniões de aplicativos e com boa acurácia as opiniões de filmes.

**Palavras-chave:** Processamento de Línguas Naturais, Mineração de Opiniões, Utilidade de Opiniões.

# ABSTRACT

SOUSA, R. F. **Helpfulness classification of online reviews in brazilian portuguese**. 2023 . 132 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023 .

The amount of user-generated data on the *Web* is increasing. Among these contents, opinions are considered a special case. This type of text usually includes colloquialisms, noise, errors, slangs, abreviations, etc. Therefore, they are difficult for machines to process and can be difficult for humans to read in some cases. Particularly, for a consumer looking for useful and quality content to help with their decisions, such as choosing a product to buy or a movie to watch, this task is becoming more and more complicated, due to the previously mentioned problems and the great offer of opinions on the *Web*. The task of Modeling and Prediction of Opinion Helpfulness aims at studying, modelling and processing user-generated opinions in order to automatically select the most useful ones and highlight them to assist other users. Predicting the usefulness of opinions is not a simple task. Much information is needed to characterize the helpfulness of opinions and, moreover, helpfulness is considered a subjective criterion, dependent on extra-textual factors, such as the reader's own information needs and the time available for reading and evaluating sufficient opinions. Many works have been done since the origin of the area, but, for the Portuguese language, few works have been published so far. This PhD work aimed to investigate and propose methods for the task of automatic classification of the helpfulness of opinions for the Portuguese language, using linguistic information and available metadata. To achieve this goal, a corpus of two domains, smartphone apps and movies, was collected and annotated. A wide range of attributes and techniques that could characterize the helpfulness of the opinions were evaluated qualitatively and quantitatively and, in this way, relevant factors were discovered for the discrimination of useful opinions from those that were not. In this doctoral thesis, the main challenges of the research area were discussed, and a benchmark for the task in the Portuguese language was established. In addition, a new graph-based method was developed that can be used as an alternative for classifying the helpfulness of opinions. Finally, we developed a method that can classify app reviews with excellent accuracy and movie reviews with good accuracy.

**Keywords:** Natural Language Processing, Opinion Mining, Opinion Helpfulness.

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# SUMÁRIO

CAPÍTULO

# 1

# INTRODUÇÃO

## 1.1 Contextualização e Motivação

Adquirir produtos e contratar serviços é uma ação comum na vida das pessoas. E a *Web* facilitou ainda mais o acesso delas aos seus desejados produtos e/ou serviços. Além disso, é inerente ao ser humano buscar por melhores produtos e querer concretizar bons negócios. Por consequência, as pessoas tentam coletar impressões e opiniões de outras tentando criar um panorama do(s) item(ns) desejado(s), visando fundamentar as suas escolhas. Segundo Liu (2012), as opiniões são elementos centrais para a maioria das atividades humanas e são capazes de influenciar o comportamento humano. Com a popularização da *Web*, essa ação também foi transportada para o comércio eletrônico. Não só as páginas da Web de comércio eletrônico, mas diversos outros *websites* permitiram que os clientes deixassem suas opiniões sobre os produtos que adquiriram e isso aumentou muito a exposição das pessoas a diversos tipos de opiniões, com diversos pontos de vista diferentes, o que permitiu que as pessoas pudessem nortear as suas escolhas.

É evidente que a facilidade de geração, difusão e acesso às opiniões possui diversas vantagens, no entanto, ela trouxe várias desvantagens também. Os conteúdos gerados por usuários, UGC (do inglês, *User Generated Content*), fazem parte das principais fontes de conteúdo na *Web*. Os comentários sobre produtos e serviços formam uma grande parcela desses conteúdos. Contudo, parte desse conteúdo pode ser considerado indesejado. Kim *et al.* (2006) mencionam que o conteúdo indesejado contempla textos mal escritos, opiniões vagas, textos com conteúdo duvidoso, etc., ou seja, os conteúdos gerados por usuários variam muito em qualidade e tais textos não ajudam na tomada de decisão dos leitores. Além disso, a grande oferta de comentários para os produtos e serviços populares torna ainda mais complicado encontrar conteúdo relevante, pois é impossível que os usuários leiam todo o conteúdo de boa qualidade disponível. E, além de tudo isso, as informações existentes nos comentários podem não ser úteis para a tomada de decisão das pessoas.

As desvantagens mencionadas anteriormente instigaram a comunidade de pesquisadores em Processamento de Línguas Naturais (PLN) a estudarem mais profundamente as opiniões. Consequentemente, uma nova área foi criada em PLN, tendo as opiniões como objeto de estudo. Diversos nomes foram atribuídos no decorrer dos anos, sendo que os mais proeminentes são Análise de Sentimentos e Mineração de Opiniões (LIU, 2012) (Optamos, neste trabalho, por utilizar Mineração de Opiniões). Os recentes avanços da *Web* permitiram que a Mineração de Opiniões ganhasse ainda mais evidência.

Entre as possíveis tarefas da área de Mineração de Opiniões (Classificação de Polaridade, Classificação de Subjetividade, Sumarização de Opiniões, etc. (LIU, 2012)), está a tarefa de Modelagem e Predição da Utilidade de Opiniões, responsável por descobrir e descrever os fatores que caracterizam e influenciam a utilidade das opiniões (DIAZ; NG, 2018), possibilitando a predição automática de sua utilidade. Seu objetivo é, portanto, identificar conteúdo relevante (útil) em comentários de usuários. Por exemplo, nas Figuras de 1 a 4 são apresentados exemplos de comentários considerados úteis e não úteis em dois domínios: Aplicativos Móveis e Filmes. Nesses comentários, é possível perceber que os comentários considerados úteis trazem muitas informações que podem auxiliar aos usuários nas suas decisões. Já os não úteis trazem informações mais vagas. Entretanto, apesar de isso ser comum, nem sempre é verdade.

A Predição da Utilidade de Opiniões é considerada por muitos autores como uma tarefa muito subjetiva (TSUR; RAPPOPORT, 2009; YANG *et al.*, 2015; DIAZ; NG, 2018; GAMZU *et al.*, 2021). Existem indícios que o processo de avaliação é pessoal e depende de fatores inerentes a quem está avaliando. E essa afirmação vem da percepção que, em muitos momentos, um comentário não faz jus à sua utilidade. Por exemplo, os comentários das Figuras 2 e 4 não são considerados úteis, em primeira instância, dado o critério de utilidade adotado ao anotá-los. No entanto, algumas pessoas podem, facilmente, considerá-los úteis. No primeiro caso (Figura 2), o comentário apresenta informações relevantes sobre o aplicativo avaliado; no segundo caso (Figura 4), o comentário, apesar de vago, expõe uma opinião forte sobre o filme. Esse tipo de avaliação levanta questionamentos importantes com relação à utilidade dos comentários e os fatores externos que influenciam a percepção de utilidade do leitor. Muitos autores, inclusive, contestam o modo como a utilidade é popularmente calculada e preferem usar formas alternativas para definir a utilidade, por exemplo, usando *crowdsourcing* (YANG *et al.*, 2015), ou, ainda, o modo mais comum, por meio de anotações convencionais, selecionando comentários e solicitando que um grupo de anotadores os avalie em ambiente controlado (TSUR; RAPPOPORT, 2009). Entretanto, mesmo adotando um método alternativo, reconhecem que existem perdas (YANG *et al.*, 2015). O contexto no qual o comentário e o leitor estão inseridos, junto a necessidade de informação[1] (GAMZU *et al.*, 2021) e intenção do leitor, podem ser decisivos para julgar um comentário como útil ou não, e isso não é plenamente replicável em laboratório. Essa subjetividade é um fator importante e torna a tarefa ainda mais desafiadora.

[1] Qual a necessidade do consumidor? Quais informações ele precisa? Qual a intenção dele ao procurar informações sobre o produto?

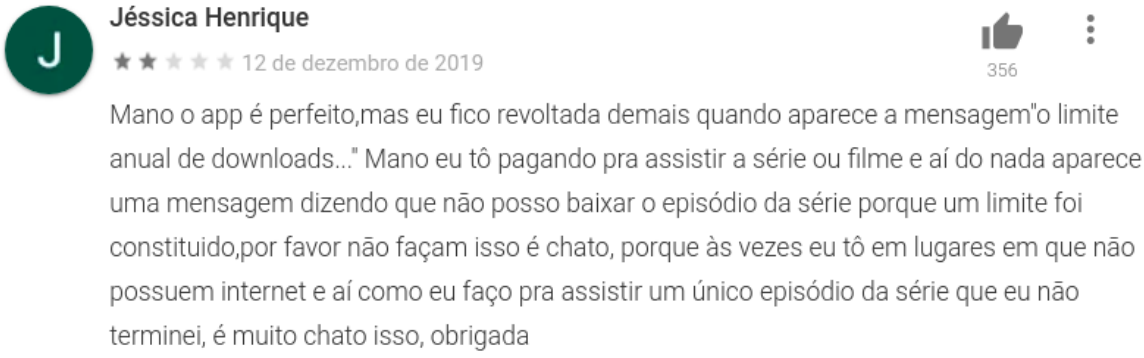



Figura 1 – Exemplo de comentário **útil** sobre um aplicativo móvel

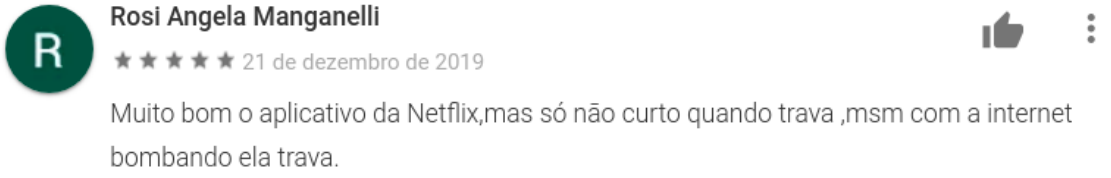



Figura 2 – Exemplo de comentário **não útil** sobre um aplicativo móvel

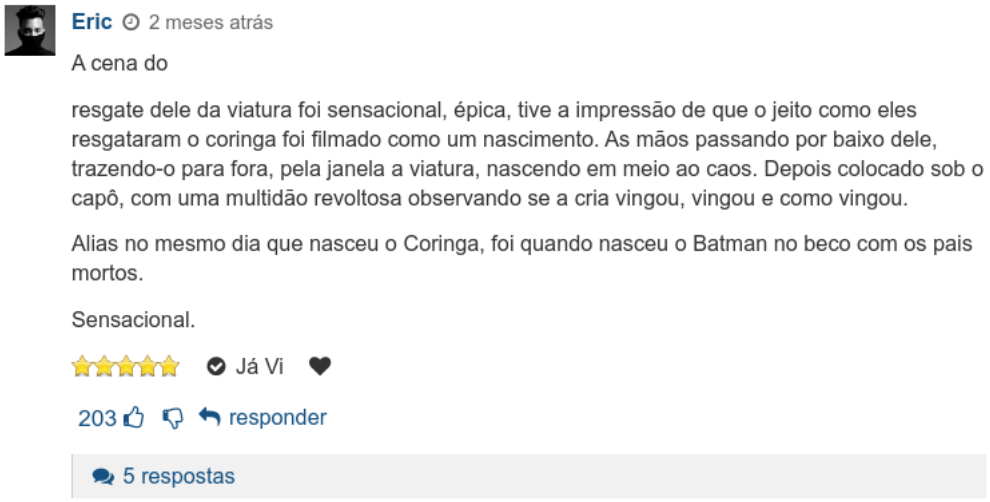



Figura 3 – Exemplo de comentário **útil** sobre um filme

Há diversas aplicações da predição da utilidade de opiniões. A aplicação mais comum é compartilhada por muitos *websites* sendo proposta para auxiliar seus usuários. A preocupação dos *sites* de comércio eletrônico em apresentar conteúdo útil é grande, e por isso alguns deles pedem um *feedback* explícito ao usuário: *esse comentário é útil ou não?*. A Figura 5 ilustra esse fato. Esse sistema de votação manual é usado para ordenar os comentários conforme os votos que receberam, os mais úteis primeiro. Os seguintes problemas podem ser observados nesse método de aquisição manual (LIU *et al.*, 2007; KIM *et al.*, 2006; PANG; LEE, 2008):

1. Novos comentários úteis dificilmente estarão no topo do ranque. É necessário algum tempo para que várias pessoas votem neles e assim ganhem a devida visibilidade.


Figura 4 – Exemplo de comentário **não útil** sobre um filme

2. Itens que possuem baixo tráfego de visitas podem não possuir votos suficientes para gerar um ranque confiável.

3. Algumas pessoas mal intencionadas podem fazer avaliações falsas da utilidade de comentários. Estes são conhecidos como *Spammers*, eles avaliam desonestamente alguns comentários para que estes subam ou desçam no ranque.

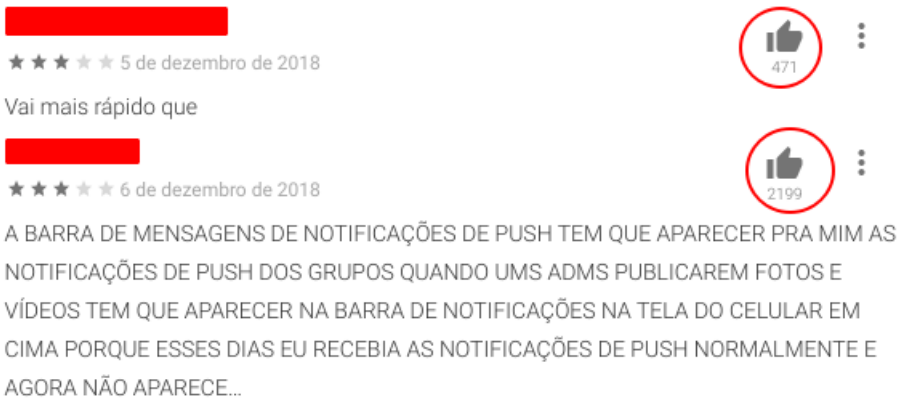



Figura 5 – Exemplo de votos manuais em um comentário sobre um aplicativo móvel

A tarefa de predição da utilidade de opiniões visa resolver esses problemas automatizando a avaliação de utilidade dos comentários gerados por usuários. Segundo Susan e David (2010), um comentário que facilita a tomada de decisão dos consumidores é considerado útil. Sabendo disso, é necessário que haja uma métrica capaz de dar valor à utilidade dos comentários. Para isso, os pesquisadores da área utilizam os valores dados pelos usuários para a pergunta evidenciada anteriormente: *esse comentário é útil ou não?*. Dessa forma, o valor de utilidade de um comentário é dado pela fórmula 1.1, sendo chamado de escore de utilidade (do inglês: *helpfulness score*):

$$h = \frac{votos\ uteis}{votos\ uteis + votos\ nao\ uteis} \tag{1.1}$$

Usando o valor da utilidade calculada pela Equação 1.1, os trabalhos, usualmente, abordam a tarefa por meio de regressão, classificação binária ou *ranking*. A maioria dos trabalhos utiliza regressão, e nesse caso os métodos buscam inferir o escore de utilidade de cada comentário. Já para a classificação, um limiar é aplicado ao escore de utilidade e os comentários com escore

de utilidade superior ao limiar são considerados úteis, e os que possuem valor inferior ao limiar são considerados não úteis. Na abordagem de ranque, os comentários são ordenados conforme o escore de utilidade e os comentários do topo do ranque são avaliados: quanto mais próximo um ranque gerado estiver de uma referência, melhores são os resultados. A principal diferença entre as abordagens de regressão e ranque é a forma de avaliação dos resultados: a primeira usa métricas de erro (MSE [2], RMSE [3]) e a segunda utiliza medidas de relevância populares usadas na recuperação de informação (NDCG, NDCG@k [4]) (DIAZ; NG, 2018).

Finalmente, vale mencionar que além da aplicação prática apresentada anteriormente, a qual é diretamente ligada ao usuário final, a predição automática da utilidade de opiniões pode ser utilizada como ferramenta para diversas aplicações. Há pesquisadores utilizando a utilidade como filtro de opiniões em tarefas de classificação de polaridade, o que é o caso do trabalho de Liu *et al.* (2007), que antes de aplicar seus métodos de classificação de polaridade realiza uma filtragem dos comentários com baixa qualidade. Outra aplicação beneficiada pela classificação da utilidade de opiniões é a de sumarização de opiniões como fazem Anchiêta *et al.* (2017), que conseguem resultados substanciais com a filtragem de conteúdo sem utilidade. Portanto, a modelagem e predição da utilidade de opiniões é uma tarefa com diversos desafios a serem superados e, além disso, possui diversas aplicações que podem beneficiar tanto usuários finais diretamente quanto aplicações que recorram a opiniões de usuários.

## 1.2 Objetivo e Hipóteses

Este trabalho de doutorado se insere na área de Processamento de Línguas Naturais (PLN), que é a área de pesquisa que busca investigar, propor e desenvolver formalismos, modelos, técnicas, métodos e sistemas computacionais que têm a língua natural como objeto primário. Mais especificamente, este trabalho se alinha à Mineração de Opiniões ao ter como objeto de estudo as opiniões de usuários. Todos os dias a quantidade de opiniões emitidas na Web cresce muito, e, dessa forma, a necessidade de analisá-las automaticamente cresce na mesma proporção. Encontrar conteúdo útil é uma necessidade crescente e diversos pesquisadores têm se debruçado sobre essa tarefa. Sabendo disso, o PLN tem desenvolvido técnicas e métodos dedicados a essa tarefa.

A tarefa de modelagem e predição da utilidade de opiniões é relativamente recente e, apesar disso, muitos trabalhos têm sido desenvolvidos nos últimos anos. Várias técnicas e métodos foram propostos, porém, a maioria é aplicada para outras línguas, predominantemente o inglês. Alguns dos trabalhos brevemente introduzidos neste capítulo consideram a tarefa muito dependente da língua, apesar de possivelmente conter características em comum entre os idiomas e muito provavelmente características bem específicas da língua alvo. Além disso, a tarefa alvo

[2] *Mean Square Error*
[3] *Root Mean Square Error*
[4] *Normalized Discounted Cumulative Gain*

deste trabalho de doutorado parece requerer conhecimento extra-textual, dependendo de fatores de várias naturezas, como a reputação do opinador e da intenção do leitor, entre outras, o que a torna muito desafiadora.

Nesse contexto, o objetivo geral deste trabalho é **investigar e propor métodos para a tarefa de classificação automática da utilidade de opiniões para a língua portuguesa, por meio de informações linguísticas e de metadados disponíveis, usando técnicas de PLN**. Foram utilizadas várias técnicas de PLN, clássicas e modernas, além da representação por meio de redes complexas, para gerar modelos baseados em diversos algoritmos de aprendizado de máquina, para classificar a utilidade das opiniões em úteis ou não úteis.

A principal tese que norteia esse trabalho é que existem fatores linguísticos e contextuais que podem ser usados para distinguir as opiniões úteis das não úteis. O embasamento dessa tese vem da confirmação da outra hipótese definida para esse trabalho que afirma que as pessoas conseguem decidir a utilidade das opiniões com consistência observando apenas o texto e os metadados explícitos aos quais elas são expostas.

Apesar da tarefa permitir várias abordagens diferentes, neste trabalho, decidiu-se trabalhar com a classificação da utilidade de opiniões e deixar as abordagens de regressão e ranking para trabalhos posteriores. Percebeu-se que a tarefa é complicada e desafiadora, portanto, tratar as três abordagens simultaneamente seria inviável. E, além disso, considerou-se que a abordagem de classificação é suficientemente ampla para que os métodos e conhecimentos gerados por este trabalho sejam relevantes para a tarefa.

O objetivo geral deste projeto de doutorado pode ser desmembrado nos seguintes objetivos específicos:

- Criar um córpus multidomínio de opiniões em português contendo informações de suas respectivas utilidades;
- Caracterizar linguisticamente as opiniões e, a partir de sua análise, levantar possíveis fatores capazes de diferenciar a utilidade das opiniões;
- Desenvolver métodos para classificar a utilidade das opiniões considerando múltiplos domínios;
- Realizar a experimentação dos métodos e técnicas propostos neste trabalho e, ainda, comparar com os principais trabalhos do estado-da-arte da área.

Este trabalho planejou responder as seguintes perguntas:

1. É possível identificar automaticamente opiniões úteis ou não úteis escritas em português por meio do uso de fatores textuais e\ou contextuais?

2. As pessoas concordam entre si ao avaliarem a utilidade de opiniões?

3. Os fatores que podem ser identificados, capazes de distinguir as opiniões úteis das não úteis, podem ser extraídos por meio do uso das técnicas, ferramentas e recursos existentes em PLN?

4. É possível adaptar os métodos e atributos desenvolvidos para outras línguas para a classificação de utilidade das opiniões escritas em português?

## 1.3 Contribuições

As principais contribuições resultantes desta tese estão elencadas a seguir:

- A geração e a disponibilização de um córpus de opiniões escritas em português, multi-domínio, contendo informações da utilidade das opiniões. Este córpus é composto por milhões de comentários dos domínios de aplicativos para dispositivos móveis e filmes. Espera-se que ele possa ser usado para apoiar pesquisas futuras, não só para a tarefa de classificação da utilidade de opiniões, mas para quaisquer outras tarefas que tenham as opiniões como objeto de pesquisa.

- A proposição de um modo de anotação automática da utilidade de opiniões usando apenas os votos úteis e a data de publicação da opinião, considerando que muitos *websites* não estão disponibilizando os votos não úteis dados aos comentários.

- A identificação de atributos (*features*) relevantes para a classificação de utilidade das opiniões em português, bem como a especificação dos atributos mais relevantes para a tarefa.

- Os métodos de classificação da utilidade das opiniões.

- A constatação de que há concordância entre as pessoas ao avaliarem a utilidade de opiniões, e que existe um padrão não aleatório nesse processo.

## 1.4 Estrutura do Texto

Além deste capítulo introdutório, este trabalho é composto de outros seis capítulos. Os dois próximos capítulos são necessários para melhor entendimento do conteúdo principal, o qual é coberto pelos artigos publicados durante a execução do trabalho de doutorado. Os capítulos são detalhados como segue.

No Capítulo 2, são discutidos conceitos e definições importantes para a pesquisa, sendo apresentadas as ferramentas e os recursos utilizados neste trabalho.

No Capítulo 3, apresenta-se uma revisão da literatura com os principais trabalhos da tarefa de modelagem e predição automática da utilidade de opiniões.

No Capítulo 4, descreve-se a geração do córpus utilizado no trabalho (SOUSA; BRUM; NUNES, 2019). O processo de anotação é detalhado e são apresentados os primeiros resultados.

No Capítulo 5, apresenta-se uma avaliação qualitativa do córpus (SOUSA; PARDO, 2021). São explicitados alguns grandes desafios que a predição da utilidade de opiniões possui.

No Capítulo 6, apresenta-se uma alternativa de método para a classificação da utilidade de opiniões, usando um método baseado em grafos (SOUSA; ANCHIêTA; NUNES, 2020). Mostra-se a viabilidade da utilização de métodos de grafos para classificar a utilidade de opiniões. Além disso, um *benchmark* é estabelecido para a tarefa na língua portuguesa (SOUSA; PARDO, 2022). Apresentam-se os resultados de vários métodos de classificação clássicos e atuais, superficiais e profundos.

No Capítulo 7, as contribuições para a área são sumarizadas. Seus impactos e limitações são discutidos. Além disso, são elencadas possíveis direções para pesquisas futuras.

CAPÍTULO

# 2

# FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados alguns conceitos teóricos importantes utilizados nesta Tese. Os primeiros são referentes à Mineração de Opiniões, pois é a área na qual este trabalho de doutorado está inserida (Seção 2.1). Em seguida, na Seção 2.2, é descrita a tarefa principal deste trabalho e, após isso, a Seção 2.3 apresenta os recursos e ferramentas utilizadas neste trabalho. Finalmente, na Seção 2.4, destacam-se as principais métricas de avaliação da tarefa de Modelagem e Predição da Utilidade de opiniões.

## 2.1 Mineração de Opiniões

Segundo Liu (2010), de maneira geral, as informações textuais encontradas no mundo podem ser classificadas em dois tipos principais: fatos e opiniões. Fatos são expressões objetivas sobre entidades, pessoas, eventos e suas propriedades, já as opiniões são expressões subjetivas que descrevem o sentimento das pessoas sobre essas entidades, eventos e suas propriedades. Com a evolução das Redes Sociais Online (RSOs), o número de opiniões emitidas por usuários na Web cresceu e fez surgir o campo de estudos em "Análise de Sentimentos". Ainda, de acordo com Liu (2012), a Análise de Sentimentos "é o campo de estudo que analisa opiniões de pessoas, sentimentos, avaliações, atitudes e emoções a respeito de entidades, como produtos, serviços, organizações, indivíduos, acontecimentos, eventos, tópicos e seus atributos".

A Análise de Sentimentos é conhecida na literatura por diversos nomes: Mineração de Opiniões, Extração de Opiniões e Mineração de Sentimentos, entre outros. Apesar da abundância de nomes, os mais comuns são Análise de Sentimentos ou Mineração de Opiniões (neste trabalho, optou-se por usar Mineração de Opiniões por acreditar que esse termo abrange melhor a tarefa abordada nesta tese). Em se tratando de campo de pesquisa, a Mineração de Opiniões é recente, tendo seu início por volta dos anos 2000, o que coincide com o primeiro grande período de crescimento da Web. Ela é uma subárea de PLN que compartilha características com diversas outras áreas de pesquisa, como Recuperação de Informações, Mineração de Textos, Aprendizado

de Máquina, etc. Uma comprovação para essa afirmação é o aparecimento de trabalhos de Mineração de Opiniões em publicações das áreas correlatas (LIU, 2012).

A Mineração de Opiniões é aplicada sobre qualquer porção de texto de qualquer tamanho e formato, como páginas web, comentários em sites de vendas, *tweets* e *posts* em *blogs*, entre outros. Desses textos, busca-se extrair e analisar conteúdo subjetivo que representa opiniões de pessoas sobre um alvo (LIU, 2012), ou ainda conteúdo objetivo que gera opiniões até mesmo sem palavras que possuam sentimentos.

Formalmente, Liu (2010) define opinião como uma quíntupla $\left(e_i, a_{ij}, s_{ijkl}, h_k, t_l\right)$, onde:

- $e_i$: é o nome de uma entidade;
- $a_{ij}$: é um aspecto da entidade $e_i$;
- $s_{ijkl}$: é a polaridade do sentimento sobre o aspecto $a_{ij}$ que tem como alvo a entidade $e_i$, emitido por $h_k$ no instante $t_l$;
- $h_k$: é o detentor do sentimento (isto é, quem expressou o sentimento), também chamado de fonte de opinião (do inglês, *opinion holder*);
- $t_l$: é o instante no qual a opinião foi expressa por $h_k$.

É importante destacar que nem sempre todos os elementos estão presentes em um texto. Basicamente, uma opinião é composta por pelo menos dois elementos que são indispensáveis: o alvo da opinião e o sentimento sobre ele.

O alvo pode ser uma entidade ou um aspecto de uma entidade, um produto, pessoa, organização, marca ou evento, entre outros. Com relação ao sentimento, ele representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. O sentimento expresso pelo autor da opinião define o seu ponto de vista que pode ser representado por alguma escala, revelando, por exemplo, um sentimento positivo, negativo ou neutro sobre o alvo da opinião. A Figura 6 mostra um exemplo de comentário extraído do site *www.buscape.com.br*. Como exemplo, a partir deste comentário foram extraídas as seguintes tuplas:

- (Galaxy A20, geral, positivo, Francisco, 02/09/2019)
- (Galaxy A20, display, positivo, Francisco, 02/09/2019)
- (Galaxy A20, desempenho, negativo, Francisco, 02/09/2019)
- (Galaxy A20, preço, positivo, Francisco, 02/09/2019)
- (Galaxy A20, bateria, positivo, Francisco, 02/09/2019)
- (Galaxy A20, design, positivo, Francisco, 02/09/2019)

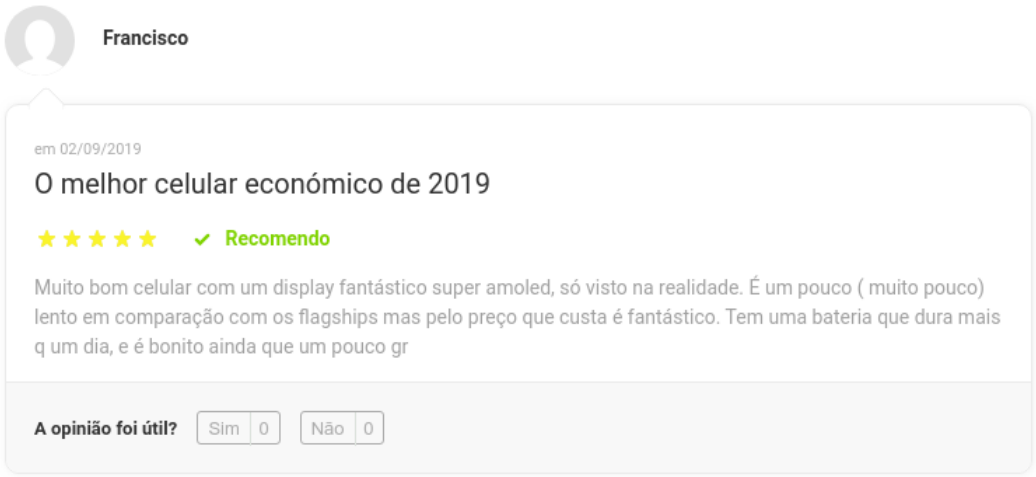



Figura 6 – Exemplo de comentário no buscapé sobre o Galaxy A20

É perceptível que esse formalismo encaixa-se perfeitamente no tratamento de opiniões baseando-se nos seus aspectos. Entretanto, é possível considerar diferentes níveis de granularidade para a análise de opiniões:

- **Documento**: nesse nível de granularidade, cada texto extraído representa uma opinião e o objetivo desse nível de análise é avaliar o texto como um todo e categorizá-lo com seu respectivo sentimento. A análise no nível de documento não consegue capturar as especificidades do texto e, se houver mais de uma opinião no texto analisado, todas serão tratadas como uma só.

- **Sentença**: a tarefa nesse nível de granularidade tem o objetivo de analisar as sentenças do texto e extrair a polaridade de cada sentença. Essa análise pode ser considerada como um passo intermediário para a análise no nível de documento, pois a composição das sentenças pode indicar a polaridade predominante do texto. Tal nível está muito relacionado com a classificação de subjetividade, que tem em vista definir se uma sentença é subjetiva ou objetiva.

- **Aspecto ou Característica**: tanto a análise no nível de documento quanto no nível de sentença não descobrem realmente de quais aspectos de um produto as pessoas gostam ou não gostam. O nível de aspecto ou característica oferece uma granularidade mais fina. Baseado neste nível de análise, um sumário de opiniões sobre as entidades e suas características pode ser criado, promovendo a estruturação de dados que naturalmente são não-estruturados, e, além disso, eles podem ser usados para todo tipo de análises qualitativas e quantitativas.

Dos níveis apresentados, o mais desafiador é o nível de aspecto, pois envolve vários problemas de PLN considerados não totalmente resolvidos, tais como reconhecimento de entidades nomeadas, resolução de anáfora, escopo de negação e desambiguação de sentidos (LIU,

2010). Além de considerar os níveis de granularidade, as opiniões podem ainda ser divididas em tipos que definem o quão difícil é a análise destas, pois o modo como as opiniões são expressas influencia diretamente na habilidade de processá-las corretamente.

De modo geral, as opiniões podem ser:

- **Regulares ou Comparativas**: em opiniões regulares, o autor da opinião expressa o seu sentimento, emoção, atitude ou percepção sobre um alvo. Já nas comparativas o sentimento é expresso com base na relação de similaridades ou diferenças entre duas ou mais entidades, ou ainda preferência quanto a algum aspecto compartilhado. Como exemplo, a Figura 7 mostra um comentário que possui ambos os tipos de opiniões. Quando o autor do comentário usa as expressões "*Celular fenomenal*", "*tela de altíssima definição*" e "*Ótima memória*", ele fala diretamente sobre o produto avaliado e seus aspectos, porém, em determinado local no comentário, quando o consumidor diz "*Muito melhor que o Iphone 5*", o autor está utilizando na sua argumentação uma comparação entre o aparelho da *Apple* e o aparelho da *Samsung*.

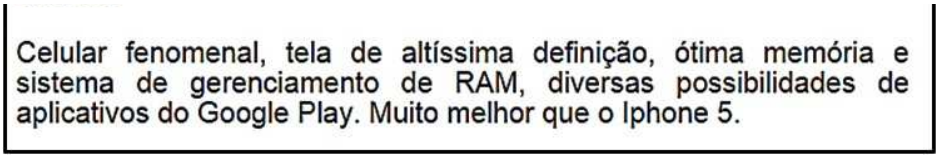

Celular fenomenal, tela de altíssima definição, ótima memória e sistema de gerenciamento de RAM, diversas possibilidades de aplicativos do Google Play. Muito melhor que o Iphone 5.

Figura 7 – Um exemplo de opinião regular e comparativa

- **Diretas ou Indiretas**: as opiniões podem ainda ser diretas quando se referem diretamente a uma entidade ou um aspecto da entidade. No caso das indiretas, uma opinião é expressa indiretamente sobre uma entidade ou um aspecto baseando-se nos seus efeitos sobre outras entidades. A Figura 8 mostra um comentário com opiniões diretas e também indiretas. O autor do comentário, ao argumentar sobre o tamanho do aparelho, cita diretamente o aspecto, dando o seu sentimento sobre ele (*Pessoalmente acho o tamanho incômodo*). Quando o mesmo autor trata, no início do comentário, sobre o aparelho ser ágil e fluido, ele indiretamente está se referindo ao processamento do aparelho, ou um conjunto de características que tornam um aparelho ágil e fluido (processamento, memória, animações, etc.).

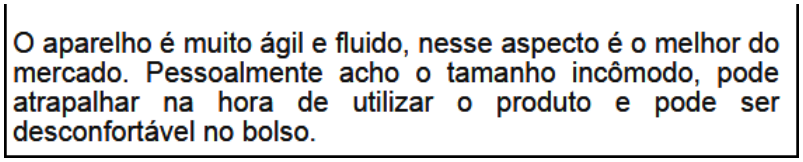

O aparelho é muito ágil e fluido, nesse aspecto é o melhor do mercado. Pessoalmente acho o tamanho incômodo, pode atrapalhar na hora de utilizar o produto e pode ser desconfortável no bolso.

Figura 8 – Um exemplo de opinião direta e indireta

- **Implícitas ou Explícitas**: as opiniões explícitas expressam diretamente o sentimento, enquanto as implícitas sugerem-no indiretamente. No comentário da Figura 9, podem ser

observados estes dois tipos de características. No início do comentário, o autor utiliza o termo "*caro*" para tratar explicitamente do preço alto do produto. Porém, no final do comentário, ao tratar da tela do aparelho, ele emite uma opinião implícita: "*(...) é que os produtos vendidos no Brasil nunca utilizam a tecnologia Gorilla Glass em suas telas.*". Neste trecho, a referência está na durabilidade da tela do aparelho, dado que a tecnologia *Gorilla Glass*, citada no texto, permite uma rigidez e tolerância maior às telas de aparelhos. É perceptível a busca de informações para associar uma dada opinião ao aspecto citado pelo autor, bem como à sua possível orientação semântica.

Como é lançamento, ele ainda é caro. Melhor comprar o SII ou o Razr i que têm hardwares superiores e é possível encontrar pelo mesmo preço. Um contra que está sempre presente nos smartphones da samsung é que os produtos vendidos no Brasil nunca utilizam a tecnologia Gorilla Glass em suas telas.

Figura 9 – Um exemplo de opinião implícita e explícita

Além da Classificação de Polaridade, ilustrada nesta seção, a área de Mineração de Opiniões engloba diversas outras tarefas, como, por exemplo, a Sumarização de Opiniões, a Classificação de *Spams*, a Classificação de Subjetividade e a tarefa deste projeto, de Predição da Utilidade de Opiniões, que é detalhada na seção seguinte.

## 2.2 Modelagem e Predição da Utilidade de Opiniões

A tarefa de modelagem e predição da utilidade de opiniões é uma tarefa relativamente recente. Como mencionado na seção anterior, ela faz parte do conjunto de tarefas englobadas pela área de Mineração de Opiniões. Os primeiros trabalhos preocupados com a utilidade das opiniões emitidas por usuários na *Web* surgiram por volta de 2006 (KIM *et al.*, 2006; ZHANG; VARADARAJAN, 2006; GHOSE; IPEIROTIS, 2007). Eles perceberam o aumento da quantidade de conteúdo opinativo gerado por usuários na *Web* e que esse conteúdo variava grandemente em qualidade e, portanto, era necessário determinar quais comentários os usuários considerariam úteis (KIM *et al.*, 2006).

Kim *et al.* (2006) mencionam alguns problemas que justificam a relevância da tarefa. Na época, os principais sites de comércio eletrônico (*Amazon.com, Overstock.com, Epinions.com*, etc.) já estavam preocupados em destacar os comentários mais úteis e já permitiam aos usuários votarem nos comentários que acreditavam terem sido úteis para eles. Porém, essa forma de votação não resolve o problema e gera outros. O primeiro é que não há como avaliar a utilidade dos comentários pouco votados, o que inclui os mais recentes, uma vez que muitas avaliações são necessárias para se estimar a utilidade dos comentários. Outro problema é que alguns itens com baixa popularidade podem nunca atingir um valor suficiente de votos. Os trabalhos de Liu

*et al.* (2007) e Pang e Lee (2008) corroboram essa avaliação e ainda acrescentam a atuação dos *spammers*, pessoas mal intencionadas que, desonestamente, fazem avaliações falsas para que comentários forjados sejam mais bem ranqueados. Diante disso, é relevante que a utilidade dos comentários seja avaliada automaticamente assim que eles sejam submetidos, permitindo acelerar a consolidação dos ranques de comentários e consequentemente o *feedback* aos usuários.

Nesse contexto, Liu (2012), afirma que objetivo da tarefa é determinar automaticamente a utilidade (do inglês, *quality, helpfulness, usefulness ou utility*) de cada opinião. Mais atualmente, o trabalho de Diaz e Ng (2018) define a tarefa mais amplamente da seguinte forma: "A modelagem e predição da utilidade de opiniões é uma tarefa que estuda os fatores que determinam a utilidade das opiniões e tenta corretamente predizê-la".

O principal conceito referente à utilidade das opiniões é o escore de utilidade (do inglês, *helpfulness score*), que vem sendo usado na literatura como quantificador de utilidade. O escore de utilidade de um comentário é definido como apresentado na Equação 2.1.

$$h = \frac{votos\ uteis}{votos\ uteis + votos\ nao\ uteis} \tag{2.1}$$

O escore de utilidade é um valor pertencente ao intervalo $[0,1]$, que indica o quão útil é um comentário. Utilizando esse valor, a tarefa é geralmente modelada de três formas diferentes: Regressão, Classificação Binária ou *Ranking*.

- Regressão - Em se tratando de regressão, o modelo desenvolvido tenta prever o valor de $h$ para cada comentário. Esse valor é utilizado para ranquear o comentário ou recomendá-lo (KIM *et al.*, 2006; ZHANG; VARADARAJAN, 2006; LEE; CHOEH, 2014; YANG *et al.*, 2015; MUKHERJEE; POPAT; WEIKUM, 2017; SAUMYA; SINGH; DWIVEDI, 2019).
- Classificação Binária - Nesse caso, um limiar é aplicado ao escore de utilidade. Caso o valor seja superior ao limiar, o comentário é considerado útil; se o valor for inferior, o comentário não é útil (ZENG *et al.*, 2014; KRISHNAMOORTHY, 2015; MALIK; HUSSAIN, 2017; BAOWALY; TU; CHEN, 2019).
- *Ranking* - Na abordagem de *Ranking*, os comentários são ordenados de acordo com as suas respectivas utilidades, e a avaliação é realizada por meio de comparação com ranques de referência (TSUR; RAPPOPORT, 2009; TSAPARAS; NTOULAS; TERZI, 2011; SAUMYA *et al.*, 2018; CHUNLI; WENJUN, 2018).

Para cada uma das abordagens, existe um conjunto de métricas de avaliação comumente utilizadas. Para a regressão, os pesquisadores tendem a utilizar métricas capazes de medir o erro do modelo gerado; as mais comuns são *RMSE* (*Root Mean Square Error*) ou ainda *MSE* (*Mean Square Error*). Na classificação, são utilizadas as métricas clássicas de Aprendizado de Máquina: Precisão, Cobertura e Medida-F, além de ser comumente utilizada a métrica de

Acurácia para os *datasets* balanceados. Em se tratando de *Ranking*, as métricas são ainda mais específicas, geralmente utiliza-se a métrica NDCG (*Normalized Discounted Cumulative Gain*) (JÄRVELIN; JÄRVELIN; KEKÄLÄINEN, 2000), sendo muito comum em sistemas de recuperação de informação. Uma variação da NDGC também é utilizada, conhecida por NDGC@k (WANG *et al.*, 2013). Essa é uma versão especial da anterior, que dá mais ênfase aos *top k* itens do *ranking*, observando se os itens realmente relevantes foram recuperados logo no início da lista. As métricas de correlação de *Pearson* (BENESTY *et al.*, 2009) e *Spearman* (ZAR, 1972) também podem ser utilizadas. Todas essas métricas serão detalhadas, posteriormente, na Seção 2.4.

Muitos atributos são utilizados nos trabalhos da literatura. Normalmente eles são divididas em grandes categorias. Alguns autores sugerem diferentes possibilidades de categorizações, mas normalmente são equivalentes, diferenciando-se apenas quanto ao nome da categoria. Por exemplo, no trabalho de Almutairi, Abdullah e Alahmadi (2019), os atributos são separadas em atributos textuais e atributos do revisor (autor do comentário), as quais, considerando a definição, são as mesmas apresentadas em Diaz e Ng (2018), porém, com nomes diferentes. Neste trabalho de doutorado, será considerada a nomenclatura utilizada em Diaz e Ng (2018): atributos de conteúdo e atributos de contexto. Os atributos de conteúdo englobam todas as informações que podem ser retiradas diretamente do comentário postado em algum *website*, por exemplo, a quantidade de palavras, quantidade de sentenças, número de estrelas, etc. Já os atributos de contexto são aqueles que não são derivados diretamente do comentário, por exemplo, informações do revisor ou ligações (interações) entre revisores (em caso de haver uma rede social inerente).

Normalmente, os atributos representam hipóteses que os autores acreditam influenciar a utilidade das opiniões. Por exemplo, muitos autores utilizam atributos de conteúdo como a quantidade de palavras de um comentário, pois, intuitivamente, acreditam que o tamanho dos textos seja relevante para a utilidade deles (KIM *et al.*, 2006; SUSAN; DAVID, 2010), entre outras possibilidades. A Tabela 1 apresenta uma lista de atributos utilizados na literatura e alguns trabalhos dos quais eles foram retirados.

Posteriormente, na Seção 2.3, são apresentados alguns córpus nos quais os autores têm baseado suas pesquisas. A maioria são córpus com comentários na língua inglesa e será mencionado o único que havia para a língua portuguesa até o início deste doutorado. No Capítulo 4, será apresentado o UTLCorpus, contendo comentários em português, fruto deste projeto.

Tabela 1 – Tabela de Atributos (Adaptado de (DIAZ; NG, 2018; ALMUTAIRI; ABDULLAH; ALAHMADI, 2019))

| ***Tipos*** | ***Atributo*** | **Referências** | **Descrição** |
|---|---|---|---|
| | | **Atributos de Conteúdo** | |
| **Tam. Textual** | Tam. Médio de Sentenças | Liu *et al.* (2007), Lu *et al.* (2010), Yang *et al.* (2015), Salehan e Kim (2016), Zhang e Zhang (2014), Gao, Hu e Bose (2017), Karimi e Wang (2017), Zhang e Lin (2018) | Razão entre a quantidade de palavras pela quantidade de sentenças. |
| | Quant. de Sentenças | Liu *et al.* (2007), Lu *et al.* (2010), Yang *et al.* (2015), Kwok e Xie (2016), Li, Pham e Chuang (2019) | - |
| | Quant. de Palavras | Susan e David (2010), Kim *et al.* (2006), Chua e Banerjee (2016), Li, Pham e Chuang (2019) | - |
| **Legibilidade Textual** | Legibilidade | Ghose e Ipeirotis (2011), Korfiatis, García-Bariocanal e Sánchez-Alonso (2012), Wu (2017), Zhang e Zhang (2014) | Uso de métricas de legibilidade para descobrir a facilidade de leitura dos comentários. |
| | Erros Ortográficos | Ghose e Ipeirotis (2011), Li, Pham e Chuang (2019) | Contagem de erros de ortografia |
| | Métricas de Parágrafos | Kim *et al.* (2006), Tanaka *et al.* (2012) | Quantidade de parágrafos, média de parágrafos |
| **Nível de Palavras** | Unigramas TF-IDF | Kim *et al.* (2006) | Cálculo de TF-IDF sobre tokens individuais |
| | Termos Dominantes | Tsur e Rappoport (2009) | Descoberta de tokens mais representativos de domínios. |

*Continua na próxima página*

Tabela 1 – *Continuação da página anterior*

| ***Tipos*** | ***Atributo*** | **Referências** | **Descrição** |
|---|---|---|---|
| **Palavras Específicas** | Aspectos de Produtos | Kim *et al.* (2006), Liu *et al.* (2007), Krestel e Dokoohaki (2011), Hong *et al.* (2012), Yang, Chen e Bao (2016) | - |
| | *Tokens* Subjetivos | Zhang e Varadarajan (2006), Ghose e Ipeirotis (2011) | Tokens que denotem subjetividade. Comparação com um léxico de sentimentos. |
| | Palavras de Sentimento | Kim *et al.* (2006), Yang *et al.* (2015), Chua e Banerjee (2016), Ren e Hong (2019) | Tokens que demonstrem sentimentos no texto. |
| | *Tokens* Sintáticos | Kim *et al.* (2006), Liu *et al.* (2007), Yang *et al.* (2015) | Classes morfossintáticas específicas. |
| **Diferença de Conteúdo** | Opinião x Descrição Produto | Zhang e Varadarajan (2006) | Comparação entre o conteúdo da opinião com a descrição do produto |
| | Diferença de Sentimento | Kim *et al.* (2006), Liu *et al.* (2007), Hong *et al.* (2012) | Diferença entre o sentimento geral sobre o item e o sentimento expresso pelo autor do comentário. |
| | Diferença Textual | Lu *et al.* (2010) | Divergência entre o modelo de língua do comentário e o modelo de língua agregado de todos os comentários. |

*Continua na próxima página*

Tabela 1 – *Continuação da página anterior*

| *Tipos* | *Atributo* | **Referências** | **Descrição** |
|---|---|---|---|
| **Diversas** | Quant. Estrelas | Danescu-Niculescu-Mizil *et al.* (2009), Susan e David (2010), Huang *et al.* (2015) | - |
| | Subjetividade | Ghose e Ipeirotis (2011), Zhang e Zhang (2014), Chen *et al.* (2015), Gao, Hu e Bose (2017), Hong *et al.* (2017), Kaushik *et al.* (2018) | Probabilidade de um comentário ser subjetivo. |
| | Tempo do comentário | Wu (2017), Hong *et al.* (2017), Chen *et al.* (2015) | Há quanto tempo o comentário foi postado. |
| | Extremismo | Siering, Muntermann e Rajagopalan (2018) | Diferença absoluta entre as estrelas do comentário pela média de estrelas do produto. |
| | Profundidade | Wu (2017), Zhang e Zhang (2014), Siering, Muntermann e Rajagopalan (2018) | Quão longo é um comentário. |
| **Atributos de Contexto** | | | |
| **Revisor** | Quant. Reviews Escritos | Ghose e Ipeirotis (2011), Ngo-Ye e Sinha (2014), Huang *et al.* (2015) | - |
| | Quant. Votos de Utilidade | Huang *et al.* (2015) | Total de votos de utilidade recebidos pelo autor. |
| | Média de utilidade | Huang *et al.* (2015), Ngo-Ye e Sinha (2014) | Razão entre a quantidade de votos recebidos pelo total de comentários publicados. |

*Continua na próxima página*

Tabela 1 – *Continuação da página anterior*

| ***Tipos*** | ***Atributo*** | **Referências** | **Descrição** |
|---|---|---|---|
| | Credibilidade | Wu (2017) | Se o comentário foi feito devido a uma compra verificada. |
| | Foto do Revisor | Karimi e Wang (2017) | - |
| | Expertise do Revisor | Gao, Hu e Bose (2017), Barbosa, Moura e Santos (2016) | Conhecimento do autor sobre o item. (Por exemplo, em jogos, a quantidade de horas jogadas pelo autor do comentário.) |
| **Usuário-Revisor** | Força de Conexão | Lu *et al.* (2010), Tang *et al.* (2013) | Uso de análise de rede social para calcular a força de conexão entre o autor e o leitor. |
| | Similaridade de Revisões | Tang *et al.* (2013) | Similaridade entre os históricos do autor e do leitor. |

## 2.3 Recursos e Ferramentas Utilizadas na Área

### *2.3.1 NLPNet*

Para a extração de muitas dos atributos apresentadas neste trabalho, é necessário determinar as classes morfossintáticas das palavras existentes nos textos. Para isso, é necessário o uso de um etiquetador morfossintático. Várias bibliotecas e ferramentas estão disponíveis, sendo que uma delas é a NLPNet (FONSECA; ROSA, 2013), que, na época em que esse trabalho iniciou, era a ferramenta estado-da-arte em português brasileiro.

A NLPNet é uma biblioteca baseada em Redes Neurais desenvolvida em *Python*. Segundo os desenvolvedores, a maioria das suas funcionalidades é independente de língua, mas algumas funções são moldadas especialmente para o Português Brasileiro. A NLPNet possui funcionalidades para realizar anotação de papéis semânticos, *parsing* de dependência e etiquetagem morfossintática (*Part of Speech*). Para este trabalho, o principal interesse é na etiquetagem morfossintática. Nessa tarefa, atingiram-se 97,33% de acurácia no córpus de referência

Mac-Morpho [1] (ALUÍSIO *et al.*, 2003; FONSECA; ROSA, 2013).

O conjunto de rótulos morfossintáticos utilizados pelos autores da NLPNet é derivado do maior córpus manualmente anotado com etiquetas morfossintáticas para o português, o Mac-Morpho (ALUÍSIO *et al.*, 2003). Originalmente, o Mac-Morpho possui 22 *tags* (além de 19 referentes a pontuações). Em Fonseca, Rosa e Aluísio (2015), os autores propuseram revisões no *tagset* do Mac-Morpho baseando-se no *Penn Treebank* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993). Após as revisões, o conjunto de rótulos da NLPNet ficou com um total de 20 rótulos, apresentados na Tabela 2.

| Etiqueta | Classe Gramatical | Etiqueta | Classe Gramatical |
|---|---|---|---|
| ADJ | Adjetivo | NUM | Numeral |
| ADV | Advérbio | PCP | Particípio |
| ADV-KS | Advérbio Conectivo Subordinado | PDEN | Palavra Denotativa |
| ART | Artigo | PREP | Preposição |
| CUR | Moeda | PROADJ | Pronome Adjetivo |
| IN | Interjeição | PRO-KS | Pronome Conectivo Subordinativo |
| KC | Conjunção Coordenativa | PROPESS | Pronome Pessoal |
| KS | Conjunção Subordinativa | PROSUB | Pronome Substantivo |
| N | Substantivo | V | Verbo |
| NPROP | Nome Próprio | PU | Pontuação |

Tabela 2 – Tagset da NLPNet

### 2.3.2 LIWC

Muitos trabalhos têm procurado descobrir os impactos da semântica na predição da utilidade de opiniões. E dentre as diversas ferramentas e recursos o LIWC [2](PENNEBAKER; FRANCIS; BOOTH, 2001) (do inglês, *Linguistic Inquiry and Word Count*) tem sido bastante utilizado (YANG *et al.*, 2015; WANG; TANG; KIM, 2019; BAOWALY; TU; CHEN, 2019). O LIWC é um programa de análise textual capaz de detectar e medir vários componentes emocionais, cognitivos e estruturais contidos em informações textuais. E ele consegue realizar esse tipo de análise usando como base em dicionários que relacionam as palavras com suas respectivas categorias. O programa recebe um texto e compara cada palavra do texto com a lista de palavras dos seus dicionários e calcula o percentual de palavras no texto que combinam com cada uma das categorias dos dicionários. Seu funcionamento, aparentemente simples, se baseia na tese de que a linguagem das pessoas pode fornecer percepções extremamente ricas sobre seus estados psicológicos, incluindo suas emoções, estilos de pensamento e preocupações sociais.

O LIWC já passou por diversas atualizações, a sua versão mais atual foi lançada em 2022 (BOYD *et al.*, 2022). Nessa versão, o dicionário interno contém cerca de 12.000 palavras,

[1] <http://nilc.icmc.usp.br/macmorpho/>
[2] <http://www.liwc.net/>

radicais de palavras, expressões e emoções selecionadas. Com relação às categorias, o LIWC é organizado de forma hierárquica consistindo de 111 sub-categorias distribuídas em quatro super-categorias: Variáveis de Resumo, Dimensões Linguísticas, Processos Psicológicos e Dicionário Expandido.

Ao longo dos anos, os dicionários do LIWC tem sido traduzidos para várias línguas em colaboração com pesquisadores ao redor do mundo. O português brasileiro já conta com duas versões traduzidas. A primeira delas foi traduzida em 2013, Filho, Pardo e Aluísio (2013) apresentam uma avaliação da tradução da versão lançada em 2007 em comparação com outros léxicos que contém sentimentos. Já em 2019, (CARVALHO *et al.*, 2019) introduzem a nova versão traduzida do LIWC, dessa vez a de 2015. Os autores comparam as duas versões em português brasileiro, mostrando que versão de 2015 conseguiu obter melhores resultados que a versão de 2007.

### 2.3.3 NLTK

Embora muitas linguagens de programação (Python, por exemplo) já tenham muitas funcionalidades necessárias para executar tarefas simples de PLN. Elas ainda não são poderosas o suficiente para a maioria das tarefas padrão de PLN. Foi pensando nisso que o NLTK (*Natural Language Toolkit*) foi desenvolvido. O NLTK[3], é uma biblioteca de código aberto desenvolvida em Python que provê muitos tipos de dados, tarefas de processamento, amostras de córpus, entre outras funcionalidades úteis para trabalhar com PLN. Segundo os desenvolvedores (BIRD; KLEIN; LOPER, 2009), a biblioteca foi projetada para ser simples, consistente, extensível e modular.

De acordo com Bird, Klein e Loper (2009), o NLTK foi originalmente criado em 2011 como parte do curso de linguística computacional do Departamento de Computação e Ciência de Informação da Universidade da Pensilvânia. Desde então ele tem sido desenvolvido e expandido com a ajuda de dezenas de colaboradores. E, atualmente, tem sido adotado em cursos de diversas universidades do mundo, e serve como base de muitos projetos de pesquisa.

Uma das grandes vantagens em usar o NLTK é que ele é inteiramente auto-contido (MADNANI, 2007). Ele fornece funções convenientes e *wrappers*[4] que podem ser usados como blocos de construção para tarefas comuns de PLN, e, além disso, fornece versões brutas e pré-processadas de córpus padrão, usados em literatura, cursos de PLN e como base para diversas tarefas, como, por exemplo, o *Brown* córpus (MAVERICK, 1969), o *Peen Treebank* (MARCINKIEWICZ, 1994) e o MacMorpho (FONSECA; ROSA, 2013). Uma lista das principais funcionalidades disponibilizadas pelo NLTK é apresentada na Tabela 3.

[3] <https://www.nltk.org/>

[4] Em Ciência da Computação, um *wrapper* é qualquer entidade que encapsula (envolve) outro item. - <https://techlib.wiki/definition/wrapper.html>

| Tarefa | Módulo do NLTK | Funcionalidade |
|---|---|---|
| Acesso aos córpus | `nltk.corpus` | Interface padrão para acessar córpus e léxicos |
| Processamento de *Strings* | `nltk.tokenize`, `nltk.stem` | Tokenizadores, tokenizadores sentenciais e stemizadores |
| Descoberta de Colocações | `nltk.collocations` | *t-test, chi-squared, point-wise mutual information* |
| Etiquetagem Morfossintática | `nltk.tag` | Atribuir classes gramaticas para os *tokens* |
| Classificação | `nltk.classify`, `nltk.cluster` | Uso de métodos de classificação de texto |
| Fragmentação (*Chunking*) | `nltk.chunk` | Anotação de partes de palavras e sentenças |
| *Parsing* | `nltk.parse` | Analisadores sintáticos, semânticos e outros |
| Métricas de Avaliação | `nltk.metrics` | Várias métricas de avaliação: precisão, cobertura, coeficientes de concordância |
| Probabilidade e estimação | `nltk.probability` | Distribuições de frequência, distribuições de probabilidade suavizadas |

Tabela 3 – Tarefas de processamento de língua e os módulos do NLTK correspondentes (Adaptado de (BIRD; KLEIN; LOPER, 2009))

### 2.3.4 Córpus

Na literatura da área de Modelagem e Predição da Utilidade de Opiniões, são reportados diversos *datasets*. Diaz e Ng (2018) argumentam que, apesar de existirem diversos trabalhos, os autores ainda não possuem a prática de construir seus trabalhos sobre os trabalhos anteriores já realizados. Por esse motivo, a maioria dos trabalhos não utiliza córpus de referência comuns e isso prejudica a comparação de resultados. Portanto, surgiu dos pesquisadores da tarefa, a necessidade da utilização de córpus de referência pré-coletados, que são córpus públicos, consolidados e utilizados em vários trabalhos. O uso desses córpus permite uniformizar os resultados atingidos pelos trabalhos. Vale ressaltar que essa é uma boa prática a ser incorporada pelas pesquisas nacionais.

Nesse sentido, merecem destaque alguns córpus públicos internacionalmente reconhecidos pela área, que possuem grande quantidade de dados e que contenham a informação de utilidade dos comentários. Para o domínio de produtos, os três principais córpus pré-coletados mencionados nos principais trabalhos da área de predição da utilidade de comentários que satisfazem esses objetivos são os seguintes: *Multi-Domain Sentiment Dataset* [5] (MDSD) (BLITZER; DREDZE; PEREIRA, 2007), *Amazon Review Dataset* [6] (ARD) (HE; MCAULEY, 2016;

[5] <https://www.cs.jhu.edu/~mdredze/datasets/sentiment/>
[6] <http://jmcauley.ucsd.edu/data/amazon/>

MCAULEY *et al.*, 2015; NI; LI; MCAULEY, 2019) e *Ciao Dataset* [7] (TANG *et al.*, 2013). Suas características específicas são apresentadas a seguir:

- *Multi-Domain Sentiment Dataset* (MDSD): é um *dataset* coletado da *Amazom.com*. Ele possui 1.422.530 comentários divididos em 25 categorias diferentes. Já está em sua segunda versão (2009).

- *Amazon Review Dataset* (ARD): também foi coletado da *Amazon.com*, porém, contém muito mais textos. Está em sua terceira versão. A versão mais utilizada foi disponibilizada em 2014 [8], contendo 142.8 milhões de comentários publicados de Maio de 2016 a Julho de 2014, divididos em 24 categorias. Em 2018, foi lançada sua terceira versão. A quantidade de comentários aumentou para 233.1 milhões. Nessa versão, foram coletados comentários publicados no intervalo de Maio de 2016 a Outubro de 2018, divididos em 29 categorias. Ele possui mais metadados disponíveis que o MDSD.

- *Ciao Dataset*: esse *dataset* foi criado por meio da coleta de um extinto site de comércio eletrônico chamado de *ciao.com*. Ele possui 302.232 comentários. O seu diferencial é que, além dos metadados comuns, ele possui informações de relacionamentos sociais entre os usuários com seus respectivos identificadores, totalizando 43.666 usuários.

Como a maioria dos trabalhos desenvolvidos na tarefa é do domínio de produtos, há um destaque especial para os *datasets* desse domínio, portanto, já são comuns os *datasets* pré-coletados. No entanto, há outros *datasets* que possuem informações de utilidade disponíveis para a tarefa, porém, eles foram coletados especificamente para trabalhos desenvolvidos específicos. Até o momento, foi encontrado um único *dataset* que não é do domínio de produtos que se encontra público e disponível para uso, o *Yelp Public Dataset* [9]. Ele contém atualmente 6.685.900 comentários referentes a 192.609 restaurantes localizados em 10 áreas metropolitanas (*Edinburgh, Karlsruhe, Montreal, Waterloo, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas e Madison*) de quatro países (Reino Unido, Estados Unidos, Canadá e Alemanha). Além de ser utilizado como *dataset* para desafios propostos pelo próprio *Yelp*, alguns autores o utilizaram para seus trabalhos, por exemplo, Ngo-Ye e Sinha (2014) e Zhang e Lin (2018).

Outros trabalhos coletaram *datasets* de outros domínios, mas esses não estão disponíveis publicamente. É o caso da *Steam* no domínio de jogos (EBERHARD *et al.*, 2018; BAOWALY; TU; CHEN, 2019), da *Google Play* no domínio de aplicativos móveis (KARIMI; WANG, 2017) e *TripAdvisor* no domínio de hotéis (GAO; HU; BOSE, 2017).

Já para a língua portuguesa, no início deste projeto de doutorado, existia apenas um *dataset* público com informações de utilidade, o Buscapé (HARTMANN *et al.*, 2014). O córpus

[7] <https://www.cse.msu.edu/~tangjili/trust.html>
[8] <https://nijianmo.github.io/amazon/index.html>
[9] <https://www.yelp.com/dataset>

Buscapé [10] foi criado em setembro de 2013. Foram coletados 85.910 comentários públicos do *buscape.com*. O córpus possui cerca de 681 Mb de tamanho, perfazendo um total de 4.097.905 *tokens* e 68.633 *types*. Porém, apesar da quantidade, apenas 33% do córpus, aproximadamente, é passível de uso para utilidade de opiniões, pois apenas 28.774 comentários possuem informações de utilidade.

Durante o desenvolvimento deste trabalho, Jorge (2022) criou e disponibilizou um córpus, do domínio de jogos, com comentários em Português Brasileiro coletados da *Steam* [11]. Ele coletou 2.789.893 comentários de 12.872 jogos distribuídos em 10 categorias. Além disso, foi realizada uma filtragem pela quantidade de votos recebidos, gerando um subcórpus com 233.824 comentários com mais de três votos de utilidade. A Tabela 4 apresenta um panorama do córpus.

| **Gênero** | **Base de dados sem filtro** | **Base de dados filtrada (pelo menos 3 votos)** |
|---|---|---|
| | n. de comentários | n. de comentários |
| Ação | 734.894 | 65.164 |
| Indie | 504.648 | 39.442 |
| RPG | 469.548 | 38.658 |
| Aventura | 366.078 | 33.088 |
| Estratégia | 189.073 | 17.842 |
| Simulação | 157.536 | 11.164 |
| Terror | 148.510 | 12.880 |
| FPS | 102.368 | 7.714 |
| Corrida | 93.743 | 7.166 |
| Esportes | 23.495 | 1.966 |
| Total | 2.789.893 | 233.824 |

Tabela 4 – Número de comentários por gênero de jogo do córpus da Steam. Fonte: (JORGE, 2022)

Um dos produtos gerados por esse trabalho de doutorado foi a criação de um córpus multidomínio com informações de utilidade. No Capítulo 4, é apresentado o UTLCorpus, que contém comentários de dois domínios: Filmes e Aplicativos Móveis.

## 2.4 Avaliação de Resultados

Os trabalhos da literatura de modelagem e predição da utilidade de opiniões usam diversas métricas de avaliação de resultados. Elas variam conforme o tipo de abordagem utilizada. As principais são listadas a seguir.

[10] http://www.buscape.com.br
[11] <store.steampowered.com/>

### 2.4.1 Métricas para Classificação: Precisão, Cobertura e Medida-F

Essas são as métricas de avaliação mais amplamente utilizadas pelos pesquisadores (POWERS, 2011). A Precisão (P) é calculada conforme a Equação 2.2, onde, $VP$ é o total de Verdadeiros Positivos (itens da classe "positivo" corretamente previstos) e $FP$ é o total de Falsos Positivos (itens previstos como da classe "positivo", porém são da classe "negativo"), e representa a taxa de itens classificados corretamente por algum método, pela quantidade possível de elementos classificáveis.

$$precisao = \frac{VP}{VP+FP} \tag{2.2}$$

Já o *Recall* (R) é calculado segundo a Equação 2.3, onde, $FN$ é o total de Falsos Negativos (itens classificados como sendo da classe "negativo", no entanto, são da classe "positivo"), e representa a taxa de itens que foram corretamente classificados por algum método, pela quantidade de elementos corretos.

$$recall = \frac{VP}{VP+FN} \tag{2.3}$$

Por fim, a *F-Measure* (F1) é uma média harmônica ponderada da Precisão (RIJSBERGEN, 1979) (Equação 2.4) e do *Recall*, também muito utilizada na literatura.

$$f-measure = 2 * \frac{P*R}{P+R} \tag{2.4}$$

### 2.4.2 Métricas para Regressão

Para a regressão, também são utilizadas métricas bem conhecidas e bem fundamentadas. Normalmente os trabalhos utilizam as seguintes métricas: *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE) ou ainda as correlações de *Pearson* (BENESTY *et al.*, 2009) ou de *Spearman* (ZAR, 1972). As métricas MAE, MSE e RMSE são calculadas de acordo com as Equações 2.5 a 2.7.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{2.5}$$

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \tag{2.6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{2.7}$$

Nas Equações 2.5, 2.6 e 2.7, os termos $y_j$ representam os valores calculados pelo método avaliado e os termos $\hat{y}_j$ são os valores esperados (corretos) que o método deveria acertar. As

métricas calculam a diferença entre os valores $(y_j - \hat{y}_j)$ e, dessa forma, calculam o erro reportado pelo método avaliado. Essas três métricas de erro possuem equações similares, e se diferenciam na penalização dos erros calculados. Dentre as três métricas, a RMSE é a que mais penaliza os erros cometidos pelo método, seguida pela MSE e pela MAE, que pode ser considerada a mais branda das três.

Já as métricas de correlação de *Pearson* e *Spearman* são calculadas de acordo com as Equações 2.8 e 2.9.

$$Pearson = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}} \tag{2.8}$$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2.9}$$

onde $x$ e $y$ são os vetores a serem comparados, $m_x$ e $m_y$ são médias dos vetores $x$ e $y$, $d$ é a distância em pares dos *ranks* das variáveis $x_i$ e $y_i$ e $n$ é a quantidade de observações.

Essas duas métricas medem a correlação entre duas variáveis quantitativas. A correlação de *Pearson* é também conhecida como Coeficiente de Correlação Linear e, portanto, calcula o grau de correlação linear entre as variáveis avaliadas. O valor da correlação de *Pearson* sempre será um valor no intervalo $[-1, 1]$. Sendo que $-1$ indica correlação negativa perfeita (se uma aumenta e a outra sempre diminui), o 1 indica correlação positiva perfeita (se uma aumenta e a outra sempre aumenta) e o 0 indica não haver correlação linear entre as duas variáveis, entretanto, alguma correção não-linear pode existir. Nesses casos, a correlação de *Spearman* pode ser utilizada, pois, diferentemente da correlação de *Pearson*, não requer a suposição que a correlação entre as variáveis seja linear. A correlação de *Spearman* também varia no intervalo $[-1, 1]$ e possui as mesmas interpretações da correção de *Pearson*, mas sua maior diferença está no fato dela avaliar relações monótonas (uma relação é considerada monótona se ela preservar a relação de ordem no seu domínio, isto é, sempre será crescente ou decrescente no domínio) sejam elas lineares ou não.

### 2.4.3 Métricas para Ranking

Em se tratando de *ranking*, a métrica mais utilizada é a *Discounted Cumulative Gain* (DCG), que calcula a influência do *ranking* gerado sobre a acurácia do método. Ela não possui limitação de valores e, por esse motivo, os trabalhos tendem a utilizar sua versão normalizada, a *Normalized Discounted Cumulative Gain* (NDCG). A NDCG é aplicada sobre o *ranking* completo, porém, em muitos casos, é mais interessante calcular os resultados sobre uma pequena porção do *ranking*. Nesses casos, os autores costumam utilizar um corte sobre o *ranking*, considerando apenas os valores do topo. Essa versão que considera os $top\ k$ itens do *ranking* é

conhecida por NDCG@k, que é uma versão da DCG@k normalizada. A DCG@k é calculada conforme a Equação 2.10.

$$DCG@k = \sum_{i=1}^{k} \frac{2^{r(u_i)} - 1}{log(1+i)} \tag{2.10}$$

onde $i$ é a posição do elemento no *ranking* e $r(u_i)$ é igual a 1 se o *rank* do elemento estiver correto e 0, caso contrário.

O NDCG@k é calculado considerando o DCG@k conforme a Equação 2.11. De forma geral, o NDCG@k é a razão entre o valor encontrado no DCG@k e o valor ideal possível para o *ranking* calculado ($Z_n$).

$$NDCG@k = \frac{DCG@k}{Z_n} \tag{2.11}$$

Dessa forma, a Equação 2.12 apresenta a fórmula completa que é usada para calcular o NDCG@k.

$$NDCG@k = \frac{\sum_{i=1}^{k} \frac{2^{r(u_i)} - 1}{log(1+i)}}{Z_n} \tag{2.12}$$

## 2.5 Considerações Finais

Neste capítulo, foram apresentados os principais conceitos teóricos relacionados ao conteúdo deste trabalho, além de ferramentas, córpus e métricas que foram utilizados. É necessário, no entanto, destacar alguns pontos importantes com relação aos métodos, ferramentas, recursos e abordagens utilizadas neste trabalho, fazendo os devidos recortes. Primeiro, com relação aos atributos e abordagens apresentadas na Seção 2.2, nesta tese, focou-se na abordagem de classificação de utilidade usando atributos de conteúdo 5. Em segundo lugar, com relação aos córpus, foram tratados apenas os domínios de aplicativos e filmes usando o UTLCorpus. Em terceiro lugar, quanto às ferramentas e recursos, todas foram usadas em alguns dos trabalhos apresentados nos capítulos 5 e 6. E, por fim, como o foco deste trabalho foi em classificação, foram usadas as métricas referentes à classificação para avaliação dos resultados dos experimentos.

CAPÍTULO

# 3

# TRABALHOS RELACIONADOS

Neste capítulo, são apresentados os principais trabalhos da tarefa de modelagem e predição da utilidade de opiniões. Como mencionado no Capítulo 2, a área é relativamente nova; o primeiro trabalho foi publicado em 2006. Apesar disso, muitos trabalhos já foram realizados, porém, somente os principais trabalhos relevantes ao objetivo deste trabalho são detalhados neste capítulo. Optou-se por dividi-los segundo as principais abordagens da tarefa: Regressão, Classificação e Ranking. Para cada abordagem, também são descritos os trabalhos realizados para a língua portuguesa. Vale ressaltar também que alguns trabalhos tratam duas abordagens simultaneamente, nesses casos eles possuem uma abordagem predominante, logo, são apresentados na respectiva seção da sua abordagem principal.

## 3.1 Regressão

Kim *et al.* (2006) utilizam *Support Vector Machine* (SVM) como modelo de regressão para ranquear os comentários de acordo com sua utilidade. Para isso, eles usam um conjunto de atributos que variam desde estruturais a semânticos. Os atributos utilizados são:

- **Atributos Estruturais**: Tamanho do comentário, número de sentenças, percentual de sentenças interrogativas e exclamativas, número de tags HTML de negrito (<b>) e quebras de linha (<br>);
- **Atributos Léxicos**: Unigramas e Bigramas com pesos TF-IDF;
- **Atributos Sintáticos**: Percentual de *tokens* de classe aberta (substantivos, verbos, adjetivos e advérbios), percentual de substantivos, percentual de verbos, percentual de verbos conjugados na primeira pessoa e percentual de adjetivos ou advérbios;
- **Atributos Semânticos**: aspectos de produtos e palavras com sentimento (possui orientação semântica);

- **Atributos de metadados**: número de estrelas.

Recorrendo aos atributos mencionados anteriormente, eles coletam um conjunto de comentários da *Amazon.com* para servir como referência para o treinamento do modelo. Para construir o corpus, houve a necessidade de criar um índice de utilidade que pudesse ser calculado pela observação dos comentários (*helpfulness score*). Dessa forma, eles usam a quantidade de votos positivos e negativos de cada comentário, ao aplicar na fórmula:

$$h(r \in R) = \frac{rating_{+}(x)}{rating_{+}(x) + rating_{-}(x)} \tag{3.1}$$

onde $rating_{+}(x)$ é o número de pessoas que consideraram o comentário útil e $rating_{-}(x)$ é o número de pessoas que não consideraram o comentário útil, implicando em uma dependência do modelo com relação ao *ranking* gerado manualmente nos sites. Eles realizaram suas experimentações sobre um córpus que continha comentários dos domínios de MP3 *players* e de câmeras digitais, obtendo os seguintes resultados na métrica de *spearman*: 0,656 de correlação para o domínio de MP3 *players* e 0,604 de correlação para o domínio de câmeras digitais.

Similarmente, o trabalho de Zhang e Varadarajan (2006) utiliza um conjunto de atributos parecido, porém, eles não incluem nenhum atributo referente aos metadados dos comentários. Apesar disso, os autores assumem que um comentário de qualidade deve abordar muitos aspectos do produto alvo. Dessa forma, eles fazem uma comparação entre o comentário e as especificações do produto analisado. Outro atributo adicionado foi a similaridade com as revisões de especialistas do editorial do site. No entanto, esses atributos não melhoraram o desempenho do sistema. Eles realizaram sua experimentação sobre comentários dos domínios de produtos e filmes. Eles atingiram 0,056 de MSE (*Mean Squared Error*) no domínio de livros e 0,082 de MSE no domínio de filmes.

Ao analisar seus resultados, os autores afirmam que a utilidade dos comentários depende grandemente do seu estilo linguístico (isto é, quantidade de palavras, comparações e superlativos, nomes próprios, etc). Além disso, em contraste com Kim *et al.* (2006), existe uma fraca correlação entre o tamanho do comentário e o valor de utilidade. Porém, de acordo com Pang e Lee (2008), essa diferença se deve ao domínio de estudo dos trabalhos. O trabalho de Kim *et al.* (2006) utiliza comentários do domínio de produtos eletrônicos, os quais não contêm uma linguagem sofisticada como nos comentários do domínio de livros e filmes utilizados por Zhang e Varadarajan (2006). Vale ressaltar que, em ambos os trabalhos, os autores dispuseram de córpus anotados automaticamente em função dos votos de utilidade dos usuários.

Outro trabalho importante a ser citado é o trabalho de Liu *et al.* (2008). Diferentemente dos trabalhos citados anteriormente, eles observam o problema de outra perspectiva, introduzindo três novos fatores que podem afetar a qualidade de comentários: perícia (*Expertise*) do autor, a linha de tempo do comentário (função que relaciona a data de publicação do comentário com a data de lançamento do filme, aplicando um decaimento exponencial conforme a diferença entre

as datas), e o estilo do comentário baseado em classes morfossintáticas. Os autores utilizam um modelo de regressão não-linear (*Radial Basis Function* - RBF) para integrar os atributos. Após a realização de muitos experimentos em um conjunto de comentários sobre filmes coletados do IMDB, os seus resultados demonstram boa eficiência do modelo proposto. Suas experimentações atingiram um resultado de 0,033 de MSE. Os autores acrescentam ainda que o modelo criado por eles é geral o suficiente para ser aplicado em outros domínios, bastando apenas realizar a substituição do gênero dos filmes pelas categorias de produtos, além da modelagem das linhas de tempo e o estilo de escrita.

Os trabalhos anteriores focam somente na extração de metadados dos comentários e no conteúdo textual para analisar várias propriedades de comentários e assim predizer a qualidade deles. Porém, diversos outros estudos tentam abordar o problema de outras perspectivas.

Huang *et al.* (2009) usam outra fonte de informação: o comportamento do autor nos sites de comércio eletrônico, por exemplo, informações coletadas de suas transações podem ajudar a avaliar a qualidade de comentários e identificar os possíveis *spams*. Os autores definem ainda três outros atributos: reputação pessoal, grau de contribuição como vendedor (quanto um usuário contribuiu como vendedor em todas as transações) e a experiência do usuário com os produtos comprados ou vendidos. Com essas informações, os autores calculam uma correlação entre cada atributo e os votos de utilidade usando uma análise de regressão linear. Eles experimentaram suas hipóteses em um córpus coletado do eBay.com, especificamente usando comentários de Telefones Celulares, DVD, câmeras digitais e MP3 players, totalizando, aproximadamente, 155.000 comentários. Seu método alcançou o valor de 0,326 de RMSE.

Outro trabalho que merece atenção especial é o de Lu *et al.* (2010), onde os autores investigam o uso do contexto social na predição de qualidade. Eles acreditam que o contexto social dos autores possui informações importantes que afetam a qualidade dos autores e dos comentários escritos por eles. Especificamente, a abordagem deles é baseada nas seguintes hipóteses: (i) Hipótese de consistência do autor, que considera que comentários do mesmo autor são similares em qualidade; (ii) Hipótese de consistência de confiança, que supõe que uma ligação entre dois autores é uma declaração de confiança; (iii) Hipótese de correlação de consistência, que prevê que pessoas são consistentes quando confiam em outras pessoas e, além disso, se dois autores possuem um mesmo autor em comum, esse terceiro autor deve ter comentários de mesma qualidade; (iv) Hipótese de consistência de relação, que supõe que se dois autores estão conectados em uma rede social, a qualidade de seus comentários deve ser similar.

Os autores utilizam um modelo de regressão linear sobre comentários dos domínios de telefones celulares, beleza e câmeras digitais existentes no córpus *Ciao*, obtendo 0,085 de MSE nas suas experimentações. Eles ainda afirmam que a abordagem pode ser aplicada e generalizada para avaliação da qualidade de outros tipos de conteúdos gerados por usuários. Porém, sua aplicação depende da existência de alguma forma de rede social interna ao site, ou ainda a conexão segura com uma rede social externa.

Mais atualmente, Qazi *et al.* (2016), com o objetivo de ampliar a pesquisa atual em qualidade de comentários, argumentaram que não se devem considerar apenas os fatores quantitativos dos comentários (ex., quantidade de palavras), mas também fatores qualitativos (ex., o tipo de opinião, ou seja, opiniões regulares, comparativas e sugestivas). Os autores reportam ainda que o estudo contribui para o desenvolvimento conceitual e entendimento dos componentes da utilidade de comentários de uma perspectiva no nível de conceitos. Para isso, eles propõem quatro hipóteses:

1. A quantidade média de conceitos por sentença influencia a utilidade de comentários;
2. A quantidade de conceitos por comentário influencia a utilidade;
3. O tipo do comentário controla o efeito da quantidade de conceitos na utilidade do comentário (ex., para comentários comparativos, comentários longos têm um efeito positivo na utilidade);
4. O tipo de comentário controla o efeito da quantidade de conceitos por sentença na utilidade do comentário.

Para a modelagem do problema, eles utilizam diversas variáveis (métricas de legibilidade, tamanho médio de sentenças, etc.). A novidade é que, com o uso do recurso *SenticNet* 3 (CAMBRIA; OLSHER; RAJAGOPAL, 2014), eles coletam a quantidade de conceitos presentes nos textos, por exemplo, "*living room*" e "*reserve restaurant table*". Além disso, eles manualmente anotam um conjunto de 1.336 comentários sobre hotéis coletados do *TripAdvisor.com*. A anotação serve para indicar quais comentários são regulares, comparativos ou sugestivos para uso na abordagem criada. Seus resultados indicam que a verbosidade (quantidade de palavras) e o tamanho médio de sentenças diferem significativamente entre os tipos de comentários (regulares, comparativos e sugestivos), confirmando suas hipóteses. É importante mencionar que seu método atingiu 0,167 de MSE.

Alguns trabalhos mais atuais aproveitam o poder das abordagens profundas para a predição da utilidade das opiniões. Por exemplo, os trabalhos de Saumya, Singh e Dwivedi (2020) e Xu, Barbosa e Hong (2020) usam Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN) e BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN *et al.*, 2018), respectivamente, para inferir o escore de utilidade dos comentários de seus domínios.

O trabalho de Saumya, Singh e Dwivedi (2020) treina um modelo de CNN com duas camadas de convolução sobre dois conjuntos de dados extraídos de *websites* de língua indiana: 29.215 comentários da *Amazon* e 12.886 comentários do *Snapdeal*. Ambos são *websites* do domínio de produtos, sendo que foram selecionadas cinco categorias de produtos para utilizar em seu trabalho: *power banks*, telefones celulares, cartões de memória, produtos de bebês e

livros. Eles realizam experimentos com *embeddings* pré-treinadas com o GloVe (PENNINGTON; SOCHER; MANNING, 2014) e com Word2Vec (MIKOLOV *et al.*, 2013) para representação dos comentários e, assim, usá-los como entrada para a CNN, obtendo melhores resultados com o GloVe. Os melhores resultados reportados são similares para os dois websites, atingindo 0,213 de MSE na *Amazon* e 0,220 de MSE no *Snapdeal*. Vale ressaltar que esse trabalho é o primeiro que tem em vista predizer a utilidade de opiniões de comentários usando a estratégia de aprendizado de representação usando convoluções profundas.

O trabalho de Xu, Barbosa e Hong (2020) segue a mesma ideia do trabalho apresentado anteriormente ao evitar o uso de atributos manualmente elaborados (*hand-crafted features*), pois considera que esses atributos podem ser tendenciosos e caros (SAUMYA; SINGH; DWIVEDI, 2020). Entretanto, os autores usam o BERT para predizer a utilidade dos comentários, visto que é um modelo de linguagem com resultados do estado-da-arte em diversas tarefas de PLN. Para seus experimentos, eles fazem o ajuste fino do modelo BERT pré-treinado sobre comentários da *Amazon* das categorias de câmeras e video games. Após o processo de ajuste fino, eles utilizam o modelo ajustado como extrator de atributos vetoriais dos comentários e concatenam os vetores extraídos com o número de estrelas e o tipo de produto. O vetor resultante é aplicado a uma MLP (*Multilayer Perceptron*) para, por fim, calcular os escores de utilidade. Os autores reportam um valor médio de 0,0556 de MSE para a categoria de video games e de 0,03689 para a categoria de câmeras.

Nosso levantamento revelou dois trabalhos que tratam a predição da utilidade de opiniões para a língua portuguesa como um problema de regressão: o de Barbosa e Moura (2016) e o de Martins e Tacla (2015). No trabalho de Barbosa e Moura (2016), os autores buscam avaliar a utilidade de opiniões no domínio de jogos. Para essa tarefa eles coletam comentários da *Steam*[1] e extraem características relativas à autoria da opinião, características textuais e metadados existentes no site. Um fato interessante a ser mencionado é com relação à autoria da opinião. Os autores utilizam a quantidade de horas jogadas pelo revisor (informação disponibilizada pelo site), além de utilizar informações como o número de amigos que o autor possui, o que deve demonstrar a credibilidade do autor. Como características textuais, eles utilizam a estrutura frasal da sentença e a métrica de inteligibilidade de *Flesch-Kincaid* (KINCAID *et al.*, 1975) com adaptações feitas por (SQUARISI; SALVADOR, 2008), entre outras. Uma rede neural artificial do tipo MLP foi usada para mapear os atributos para a utilidade dos comentários. Após os experimentos, eles reportam bons resultados (0,167 de MSE) e indicam que as métricas relativas à autoria foram mais relevantes com as métricas que indicam o tamanho do texto. Já a data de postagem dos comentários não apresentou forte impacto na avaliação.

O trabalho de Martins e Tacla (2015) apresenta uma metodologia de avaliação de utilidade de opiniões com foco na identificação de atributos que exercem maior influência sobre os votos de utilidade. Os experimentos são aplicados em opiniões escritas em português e do domínio

[1] http://store.steampowered.com

de serviços (hotéis). Os autores propõem diversos atributos que caracterizem os comentários. Os atributos são divididos em duas categorias: textuais e semânticos. Os atributos textuais são compostos principalmente por métricas de inteligibilidade. Para sua extração, eles usam uma versão adaptada por eles do *Coh-Metrix-Port* (SCARTON; ALUíSIO, 2010). Alguns exemplos incluem, além do índice de inteligibilidade, a contagem de sentenças, palavras, sílabas, etc. Para os atributos semânticos, é utilizado o *Latent Semantics Analysis* (LSA) (DUMAIS, 2004). Regressão Logística Ordinal (OLR) é a técnica utilizada para investigar as relações entre os atributos propostos. As contribuições incluem a confirmação do impacto positivo dos atributos semânticos na avaliação de utilidade das opiniões também na língua portuguesa; O índice de inteligibilidade revelou que opiniões mais complexas e com mais palavras são mais úteis que as opiniões com menos palavras e mais fáceis de ler. Além dos achados reportados, a experimentação realizada pelos autores atingiu um MSE de 0,1723.

## 3.2 Classificação

Como mencionado anteriormente, uma grande parte dos trabalhos relacionados na área utilizam regressão como método de avaliação da qualidade de comentários, porém, muitos autores usam outros métodos para essa tarefa, sendo que o mais comum é o uso de classificadores. Alguns trabalhos são citados a seguir.

Os trabalhos de Ghose e Ipeirotis (2007) e Ghose e Ipeirotis (2011) acrescentam três novos conjuntos de atributos: (i) atributos referentes ao perfil do autor do comentário; (ii) atributos referentes ao histórico do autor, que consideram os votos recebidos pelo autor no passado; e (iii) atributos referentes à legibilidade do texto, isto é, erros ortográficos e índices de legibilidade estudados pela área de pesquisa em legibilidade textual (FLESCH, 1948). No primeiro trabalho, Ghose e Ipeirotis (2007) estudam o relacionamento entre a subjetividade de um comentário e sua utilidade. Um classificador binário decide se a informação contida em cada sentença do comentário é subjetiva (o autor fornece uma descrição pessoal do produto) ou objetiva (o autor busca seguir a descrição dada pelo vendedor, confirmando ou negando as características). Em seguida, calcula-se uma média que indica a probabilidade de um comentário ser subjetivo e, sabendo que um comentário é uma mistura de sentenças subjetivas e negativas, eles calculam o desvio padrão do valor de subjetividade. Os resultados indicam que o desvio padrão e a legibilidade possuem uma forte influência na avaliação da utilidade. Os resultados reportados pelos autores mostram que eles atingiram uma F1 de 0,85 ao analisar seu método sobre comentários de produtos das categorias de áudio-vídeo e câmeras digitais.

No segundo trabalho, Ghose e Ipeirotis (2011) expandem sua pesquisa e examinam múltiplas características de produtos, e adicionam outros atributos textuais mencionadas anteriormente (perfil e histórico do autor e legibilidade). De acordo com seus resultados, eles afirmam que a utilidade dos comentários e as vendas dos produtos são influenciados pela subjetividade do texto.

Os comentários que contêm uma mistura de informações subjetivas e objetivas são mais úteis e, assim, influenciam as vendas. Adicionalmente, atributos de legibilidade e informatividade correlacionam-se positivamente com vendas e utilidade. Uma contribuição importante deste trabalho é que o tipo de produto afeta a utilidade de um comentário. Para produtos baseados em características (ex., eletrônicos), comentários que incluem mais informações objetivas possuem mais utilidade, enquanto, em produtos baseados em experiência (ex., DVDs), a subjetividade é mais importante. Eles reportaram 0,89 de F1 e 0,94 de ROC (*Receiver Operating Characteristic*) nas suas experimentações.

O trabalho de Chua e Banerjee (2016) demonstra ter encontrado as seguintes relações: entre a utilidade de um comentário e o sentimento presente nele; entre a utilidade e o tipo de produto; e entre a utilidade e a qualidade de informação. O sentimento do comentário foi classificado em três categorias: favorável, desfavorável e misto. Os tipos de produtos foram categorizados em *search products* e *experience products*. A qualidade de informação tem três dimensões: compreensividade, especificidade e confiabilidade. A compreensividade é referente à facilidade para entender um comentário; a especificidade refere-se à adequação da informação dada nos comentários; e a confiabilidade é referente à confiança dos consumidores nos comentários, isto é, referente à imparcialidade do autor. O trabalho de experimentação que os autores realizaram foi norteado estatisticamente por meio da tentativa de correlação entre os atributos propostos com a utilidade usando um método chamado de Fatorial ANOVA. Seus resultados foram em termos qualitativos, e ao analisar um conjunto de dados da *Amazon.com*, eles concluem que a utilidade variou conforme o sentimento do comentário, sendo independente do tipo de produto. No entanto, o relacionamento entre qualidade de informação e a utilidade dos comentários variou como uma função do sentimento do comentário e do tipo de produto.

Da mesma forma que os trabalhos apresentados para regressão, os pesquisadores que abordam a tarefa por meio de classificação, voltaram suas atenções para os classificadores profundos e modelos de linguagem mais robustos. Nesse sentindo valem ser mencionados os trabalhos de Du *et al.* (2020), Kong *et al.* (2020) e Bilal e Almazroi (2022).

Du *et al.* (2020) descrevem o uso de um tipo específico de rede neural recorrente (*Recurrent Neural Network*, do inglês) para classificação da utilidade de comentários sobre produtos postados na *Amazon.com*. Eles estudam a interação de *embeddings* textuais com *embeddings* de estrelas (*Text-Rating Interaction*). Eles geram uma arquitetura específica para geração de ambos os tipos de *embeddings* usando GLUs (*Gated-Linear Units*). Os melhores resultados de seus experimentos são de aproximadamente 0,87 de F1 na categoria de CDs e Discos de Vinil. Os resultados variam de 0,72 até 0,87 de F1.

Já o trabalho de Kong *et al.* (2020) usa uma abordagem híbrida para classificar os comentários. Eles combinam Redes Neurais Convolucionais com uma estratégia de extração de atributos em grafos de conhecimento, conhecida como *TransE* (BORDES *et al.*, 2013). A *TransE* é usada para capturar os relacionamentos entre diferentes entidades mencionadas no comentário.

E, além dessas duas estratégias, eles definem e extraem alguns atributos manualmente. Em seus experimentos, eles estudam o impacto de cada estratégia individualmente e, por fim, mesclam todas as estratégias, atingindo um valor de Macro-F1 máximo de 0,73 em um *dataset* da *Amazon*.

Em Bilal e Almazroi (2022), é apresentada uma estratégia de classificação de comentários usando BERT. Os autores usam um dataset coletado do *Yelp* [2] (avaliações de empresas como restaurantes, shoppings, salões de beleza, etc.), contendo 48.442 comentários sobre shoppings, no entanto, eles usaram apenas 10.000 (50% úteis e 50% não úteis) comentários devido ao processo de seleção empregado. Foi realizado o ajuste fino do modelo BERT com diversas configurações diferentes e os modelos gerados foram comparados com classificadores *baseline*, como *k-nearest neighbours, naive bayes e support vector machine*. O modelo BERT obteve os melhores resultados, atingindo 0,71 de F1.

Finalmente, vale mencionar o trabalho de (BAOWALY; TU; CHEN, 2019), que pode ser considerado o estado-da-arte da tarefa na abordagem de classificação. Os autores geram um modelo de classificação baseado no algoritmo GBM (*Gradient Boosting Machine*) (FRIEDMAN, 2001), que é um *ensemble* de classificadores baseados em árvores de decisão com grande poder preditivo. Apesar dos autores também apresentarem resultados de regressão, os resultados de classificação são os mais notórios. Os autores coletaram um corpus do site da *Steam*[3] e especificaram os seguintes intervalos de escore de utilidade: Comentários úteis $\in$ [0,90, 0,95]; Comentários não úteis $\in$ [0,05, 0,10]. Eles extraíram diversos atributos do *dataset*, por exemplo, metadados dos comentários (recomendação, data de postagem, número de caracteres, etc...), metadados dos revisores (tempo de jogo, quantidade de comentários e quantidade de jogos), atributos semânticos (LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001), LDA), TF-IDF e vetores Word2Vec. Os resultados das experimentações atingiram 0,94 de F1 para a classe dos úteis e 0,83 de F1 para a classe dos não úteis. E, por fim, foi apresentado um ranque dos atributos mais importantes, mostrando que os metadados e os *embeddings* foram responsáveis por mais de 90% dos resultados apresentados.

Para a língua portuguesa, existem alguns trabalhos que tratam da predição da utilidade de opiniões por meio de classificação, sendo apresentados a seguir.

Sousa, Rabêlo e Moura (2015) apresentam uma abordagem diferente para classificar a importância de comentários para a língua portuguesa. De um corpus de comentários coletados do *buscapé.com.br*, os autores extraem três características: reputação do autor, quantidade de pares do tipo *(característica, palavra opinativa)* e riqueza do vocabulário. Por meio de experimentações, os autores configuram um Sistema de Inferência *Fuzzy* para classificar os comentários em 4 classes: Insuficiente, Suficiente, Bom e Excelente. Após a aplicação do Sistema *Fuzzy*, os comentários são ranqueados de acordo com o valor expresso pelo sistema.

Um método de classificação da polaridade de comentários foi implementado e aplicado

[2] <https://www.yelp.com/dataset>
[3] <https://store.steampowered.com/>

em duas etapas no conjunto de comentários. A primeira etapa aplica o método de classificação de polaridade sobre todos os comentários e calcula as métricas de Precisão, Cobertura e Medida-f. A segunda etapa aplica o método de classificação de polaridade somente após a classificação do Sistema *Fuzzy*. Eles buscam definir um ponto de corte que contenha os melhores resultados com a menor quantidade de comentários possível. O melhor resultado é encontrado com menos de 10% dos comentários. O objetivo dos autores é demonstrar que utilizar um conjunto de comentários menor, porém de maior qualidade, melhora a avaliação de orientação semântica. Além da experimentação extrínseca proposta, os autores realizaram uma avaliação sobre um córpus anotado e obtiveram 0,5326 de F1.

Seguindo a mesma linha, o trabalho de Santos *et al.* (2016) estende o anterior. Eles melhoram as definições de algumas características e propõem um estudo experimental para comparar a abordagem anterior de sistemas *fuzzy* com a sua proposta com Redes Neurais Artificiais. Foram propostas duas topologias de redes neurais artificiais e os autores obtiveram uma média de 0,3984 de F1 nos seus experimentos. Após os experimentos, os autores reportam que não conseguiram melhorar os resultados anteriores e que isso se deve aos seguintes fatores: as amostras obtiveram resultados esperados dispersos, o que tornou difícil a generalização da rede; e nenhuma das topologias candidatas atingiu a acurácia mínima. A acurácia mínima serve como limite mínimo para considerar a rede treinada.

## 3.3 *Ranking*

Como apresentado no Capítulo 2, o *ranking* de comentários tem o objetivo de selecionar os comentários mais úteis de um conjunto de comentários. De preferência, dado um *ranking* de referência, os comentários retornados deverão ser os do topo do *ranking*. Apesar de interessante, essa abordagem é pouco explorada na literatura, mas possui alguns trabalhos recentes. A seguir são apresentados alguns trabalhos que utilizam essa abordagem.

O trabalho de Wu, Xu e Li (2011) apresenta um algoritmo não supervisionado para ranquear automaticamente comentários de um site de comércio eletrônico. Especificamente, os autores coletaram comentários sobre livros e sobre *MP3 Players*. Eles propõem o uso de três métodos de *ranking* diferentes para extração de atributos: o algoritmo *PageRank* (PAGE *et al.*, 1999), o *HITS (Hypertext Induced Topic Search)* (KLEINBERG, 1999) e o tamanho do texto. E, além de aplicá-los individualmente, eles realizam uma combinação dos três. Para utilizar os algoritmos *PageRank* e *HITS*, eles modelam um grafo com as sentenças dos comentários. Os melhores resultados são atingidos utilizando todos os atributos propostos. Para o domínio de livros, os autores reportaram uma média de 0,77 na *NDCG@k* com o $k$ variando no intervalo de $[1, 10]$, da mesma forma no domínio de MP3 *players*, em que alcançou média de 0,88 de NDCG@k.

Seguindo uma linha diferente, Hong *et al.* (2012) propuseram um sistema de *ranking* de

comentários usando um algoritmo supervisionado. Usando o SVM para *ranking*, eles inicialmente classificam os comentários em úteis e não úteis e, em seguida, realizam o ranqueamento dos comentários úteis. O foco do trabalho é em três atributos específicos, denominados de preferências dos usuários: informatividade, credibilidade e popularidade. Na primeira, quanto mais informativo o texto é, considerando a menção de atributos e funções dos produtos, mais útil ele será. Na segunda, quanto menos incerteza o revisor demonstrar em seu texto, mais útil ele será. E, por fim, na terceira, eles calculam a diferença de sentimento entre a média dos comentários dos produtos com o comentário avaliado. Os resultados são reportados usando a métrica *NDCG@k* com $k = 10$, atingindo um valor de 0,75 sobre o Corpus MDSD da *Amazon*.

Saumya *et al.* (2018) tentam gerar um *ranking* de comentários por meio da sua utilidade. Eles coletam um grande conjunto de atributos dos comentários (por exemplo, quantidade de substantivos, total de palavras, quantidade de types, etc.) e dos produtos (por exemplo, similaridade entre a descrição do produto e o texto do comentário) de comentários extraídos de dois sites de produtos (*Amazon.com* e *Snapdeal.com*). Inicialmente, eles usam a mesma ideia de Hong *et al.* (2012), classificando os comentários em úteis e não úteis, e apenas depois tentam calcular o valor da utilidade dos comentários para então ordená-los. Em termos de classificação, eles atingem um F1 de 0,86. Para calcular o valor de utilidade dos comentários, eles aplicam uma abordagem de regressão que consegue 0,267 de MSE no *dataset* extraído da *Amazon* e 0,623 no dataset da *Snapdeal*. Somente com o modelo de regressão gerado, eles passam a ordenar os textos. Eles aplicam uma abordagem de conjuntos para validar o ranque gerado. Eles limitam o tamanho do ranque em dez comentários e calculam o valor da intersecção para o conjunto de referência, atingindo uma acurácia de aproximadamente, 56% no *dataset* da *Amazon* e de 61% no *dataset* da *Snapdeal*.

Mais recentemente, Melleng, Jurek-Loughrey e P (2021) descrevem um método não supervisionado para ranquear comentários sobre produtos da *Amazon*. Eles usam três atributos dos comentários: relevância (se um comentário discute aspectos relevantes de um produto específico), intensidade emocional (o nível de emoções em um comentário) e especificidade (nível de detalhes discutidos em um comentário). A combinação desses três atributos é usada para ranquear os comentários. O método desenvolvido não depende da quantidade de votos de utilidade e usa apenas o conteúdo e as estrelas atribuídas aos comentários. Os resultados dos experimentos realizados mostram que o método atingiu um máximo de 0,98 de *NDGC@k*, sendo considerada o estado-da-arte da abordagem de ranking, atualmente.

Para a língua portuguesa, não foram encontrados trabalhos que tenham utilizado a abordagem de *ranking*. Muito embora os trabalhos que utilizaram a abordagem de regressão possam ser expandidos para a abordagem de ranking.

## 3.4 Considerações Finais

Neste capítulo, foram apresentados os principais trabalhos da literatura referentes à tarefa de modelagem e predição automática da utilidade de opiniões. Existem muitos outros trabalhos não relacionados aqui, mas, com os trabalhos listados, é possível perceber a diversidade de atributos e abordagens que os autores desenvolveram. Vale ressaltar também que muitos outros domínios estão sendo propostos, por exemplo: médicos (SHAH; MUHAMMAD; LEE, 2022), mercado automotivo (CAO; YANG, 2022), turismo (XIA, 2023) e remédios (ZHENG *et al.*, 2021), entre outros. E, além dos diferentes domínios, muitos trabalhos estão buscando encontrar diferentes fatores que influenciem a utilidade: imagens de produtos e de usuários (multimodalidade) (WU; WU; WANG, 2021; HAN *et al.*, 2022; LIU *et al.*, 2021), gênero (SHEN *et al.*, 2022), emoções negativas e preços de produtos (XU; ZHENG; YANG, 2023) e retórica (MORADI; DASS; KUMAR, 2023), entre outros.

Este trabalho de doutorado, inicialmente, se diferencia da maioria dos trabalhos apresentados nessa seção, principalmente devido à língua alvo, baseando-se na tese de que a língua influencia na forma como a utilidade é percebida. No entanto, podem haver elementos utilizados por outros pesquisadores de outras línguas que são aplicáveis ao português, e isso é um dos pontos de ligação com os trabalhos que foram apresentados aqui. Além disso, foi necessário evoluir a área para a língua portuguesa, e diversas técnicas usadas mundialmente são aplicáveis independentemente do idioma, além do fato de tentar entender a percepção de utilidade para os leitores brasileiros.

Finalmente, a Tabela 5 relaciona e resume todos os trabalhos apresentados nesta seção, e a Figura 10 apresenta uma linha do tempo mostrando os marcos da história da área de modelagem e predição da utilidade de opiniões em geral e para a língua portuguesa.

Tabela 5 – Resumo de Trabalhos Relacionados

| Título | Abordagem | Método | Avaliação | Idioma |
|---|---|---|---|---|
| Automatically assessing review helpfulness (KIM *et al.*, 2006) | Regressão e Ranking | SVR | 0,656 e 0,604 de *spearman* para MP3 *players* e cameras digitais | Inglês |
| Utility Scoring of Product Reviews (ZHANG; VARADARAJAN, 2006) | Regressão | SVR e SLR | 0,056 e 0,082 de MSE livros e filmes, respectivamente | Inglês |
| Designing novel review ranking systems: predicting the usefulness and impact of reviews (GHOSE; IPEIROTIS, 2007) | Classificação | Regressão Logística | F1: 0,85 | Inglês |
| Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics (GHOSE; IPEIROTIS, 2011) | Classificação | Random Forest | F1: 89,04 em produtos; ROC: 0,94 | Inglês |
| Modeling and Predicting the Helpfulness of Online Reviews (LIU *et al.*, 2008) | Regressão | RBF | MSE: 0,033 | inglês |
| Discovering clues for review quality from author's behaviors on e-commerce sites (HUANG *et al.*, 2009) | Regressão | Regressão Linear | RMSE: 0,326 | Inglês |

*Continua na próxima página*

Tabela 5 – *Continuação da página anterior*

| Título | Abordagem | Método | Avaliação | Idioma |
|---|---|---|---|---|
| Exploiting social context for reviewquality prediction (LU *et al.*, 2010) | Regressão | Regressão Linear | MSE: 0,085 | Inglês |
| Avaliação automática da utilidade de reviews usando redes neurais artificiais no corpus do steam (BARBOSA; MOURA; SANTOS, 2016) | Regressão | MLP | RMSE: 0,1929 | Português |
| Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality (CHUA; BANERJEE, 2016) | Classificação | ANOVA (Regressão Logística) | Valores de tolerância maiores que 0,2 indicando falta de multicolinearidade | Inglês |
| A concept-level approach to the analysis of online review helpfulness (QAZI *et al.*, 2016) | Regressão | Tobit | MSE: 0,167 | Inglês |
| Assessement of features influencing the voting for opinions helpfulness about services in portuguese (MARTINS; TACLA, 2015) | Regressão | Regressão Logística Ordinal e SVD | MSE : 0,1723 | Português |

*Continua na próxima página*

Tabela 5 – *Continuação da página anterior*

| Título | Abordagem | Método | Avaliação | Idioma |
|---|---|---|---|---|
| A fuzzy system-based approach to estimate the importance of online customer reviews (SOUSA; RABÊLO; MOURA, 2015) | Classificação | Fuzzy | F1: 53,26 | Português |
| An experimental study based on fuzzy systems and artificial neural networks to estimate the importance of reviewsabout product and services (SANTOS *et al.*, 2016) | Classificação | Fuzzy | F1: 39,84 | Português |
| An Unsupervised Approach to Rank Product Reviews (WU; XU; LI, 2011) | Ranking | Link Analysis | NDCG@k: 0,77 no domínio de livros e 0,88 em MP3 players | Inglês |
| What Reviews are Satisfactory: Novel Features for Automatic Helpfulness Voting (HONG *et al.*, 2012) | Ranking | SVM | NDCG@k: 0,75 | Inglês |
| Ranking online consumer reviews (SAUMYA *et al.*, 2018) | Ranking | SVM | Acurácia: 56% no *dataset* da *Amazon* e 61% no *dataset* da *Snapdeal* | Inglês |

*Continua na próxima página*

Tabela 5 – *Continuação da página anterior*

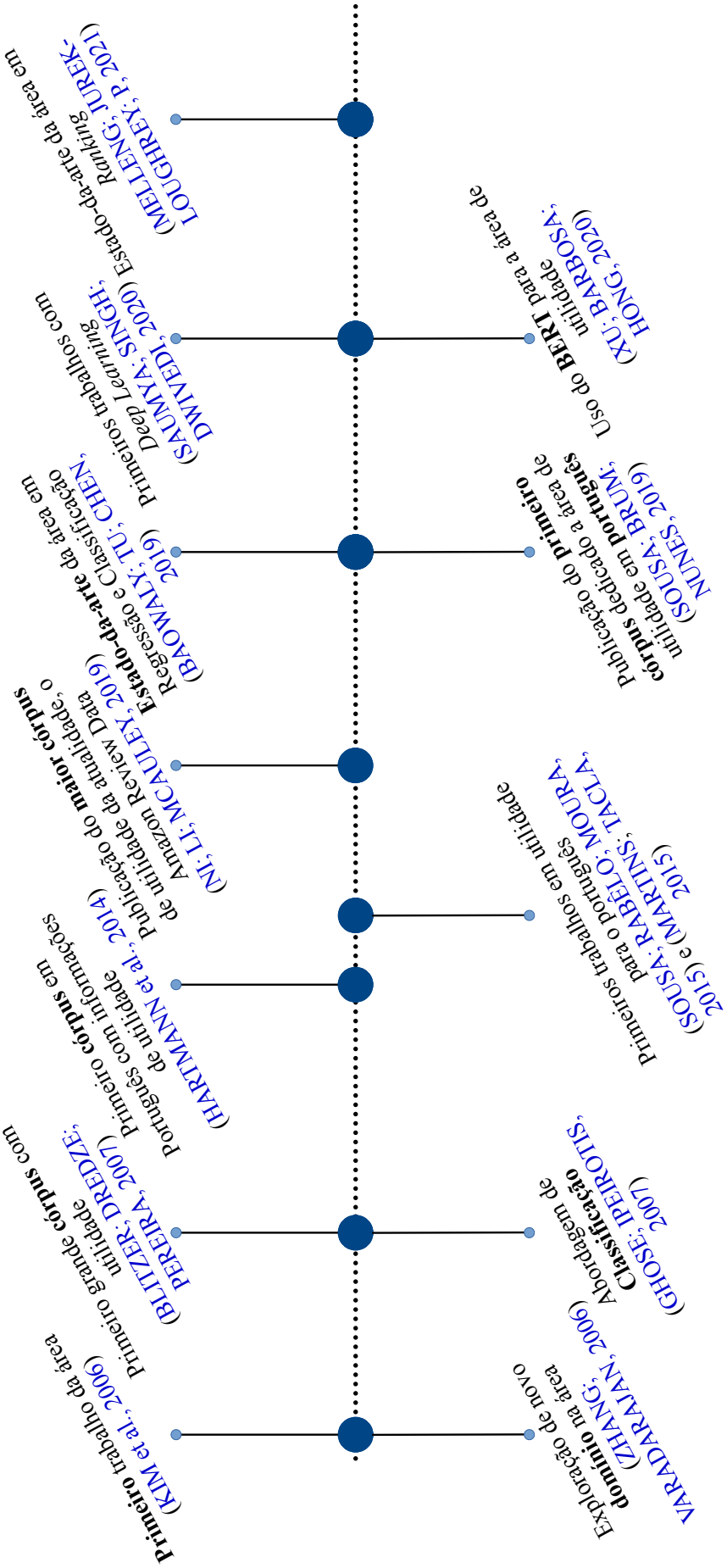
| Título | Abordagem | Método | Avaliação | Idioma |
|---|---|---|---|---|
| Predicting the Helpfulness Score of Online Reviews Using Convolutional Neural Network (SAUMYA; SINGH; DWIVEDI, 2020) | Regressão | CNN | MSE: 0,213 no *dataset* da *Amazon* e 0,220 no *dataset* da *Snapdeal* | Inglês |
| Bert Feature Based Model for Predicting the Helpfulness Scores of Online Customers Review (XU; BARBOSA; HONG, 2020) | Regressão | BERT | MSE: 0,0556 na categoria de video games e 0,03689 na categoria de câmeras da *Amazon* | Inglês |
| Interactive Network for End-to-end Review Helpfulness Modeling (DU *et al.*, 2020) | Classificação | RNN | F1: 0,87 na categoria de CDs e discos de vinil da *Amazon* | Inglês |
| Predicting Product Review Helpfulness a Hybrid Method (KONG *et al.*, 2020) | Classificação | CNN e Redes Complexas | Macro-F1: 0,73 em um dataset da *Amazon* | Inglês |
| Effectiveness of Fine-tuned Bert Model in Classification of Helpful and Unhelpful Online Customer Reviews (BILAL; ALMAZROI, 2022) | Classificação | BERT | F1: 0,71 em um dataset do *Yelp* | Inglês |
| Predicting the Helpfulness of Game Reviews: A Case Study on the Steam Store (BAOWALY; TU; CHEN, 2019) | Classificação e Regressão | GBM | F1: 0,94 para a classe de úteis e 0,83 para a classe dos não úteis no dataset da *Steam* | Inglês |

*Continua na próxima página*

Tabela 5 – *Continuação da página anterior*

| Título | Abordagem | Método | Avaliação | Idioma |
|---|---|---|---|---|
| Ranking Online Reviews Based on their Helpfulness: An Unsupervised Approach (MELLENG; JUREK-LOUGHREY; P, 2021) | Ranking | Roberta | NDCG@k: 0,98 | Inglês |

Figura 10 – Linha do tempo da Modelagem e Predição da Utilidade de Opiniões.




Fonte: Elaborada pelo autor.

CAPÍTULO

# 4

# CONSTRUÇÃO DO CÓRPUS DE TRABALHO: O UTLCORPUS

Este capítulo apresenta o artigo publicado que descreve o processo de construção do UTLCorpus, um córpus de opiniões em português do Brasil dos domínios de aplicativos para android e filmes, criado para dar suporte a este trabalho de doutorado e fomentar pesquisas da tarefa na língua portuguesa.

Nesse artigo foram apresentados os primeiros resultados de experimentações deste trabalho de doutorado. O trabalho apresentado neste capítulo é o seguinte:

*Rogério Figueredo de Sousa, Henrico Bertini Brum e Maria das Graças Volpe Nunes. 2019. "A Bunch of Helpfulness and Sentiment Corpora in Brazilian Portuguese". In Proceedings of the XII Symposium in Information and Human Language Technology - STIL. Salvador, BA: SBC, 2019. p. 209–218.*

## Declaração de Contribuição

R. F. Sousa realizou a coleta do córpus, realizou a anotação de utilidade dos comentários, participou do processo de experimentação e contribuiu com a escrita do manuscrito. H. B. Brum realizou a anotação de polaridade dos comentários, contribuiu com o processo de experimentação e escrita do manuscrito. M. G. V. Nunes auxiliou na escrita do manuscrito e supervisou o projeto.

# A bunch of helpfulness and sentiment corpora in Brazilian Portuguese

**Rogério Figueredo de Sousa, Henrico Bertini Brum, Maria das Graças Volpe Nunes**

[1]Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
Av. Trabalhador São Carlense, 400 – 13.566-590 – São Carlos – SP – Brazil

{rogerfig,henrico.brum}@usp.br, gracan@icmc.usp.br

***Abstract.*** *This paper presents the UTLcorpus, a novel corpus in Brazilian Portuguese for helpfulness classification of online reviews. There is a lack of corpora in Brazilian Portuguese annotated with helpfulness information, therefore there are also few works on modeling and predicting helpfulness of online reviews in this language. Moreover, there is no reference corpus to ground those results. This work tries to partially solve this problem by presenting UTLcorpus, a huge amount of annotated online reviews regarding helpfulness. Since the source data also contain star score labels, this paper also explores polarity labels in the data set. Some experiments show that both tasks of predicting helpfulness and polarity are benefited by the use of this corpus.*

## 1. Introduction

Navigating websites for buying clothes, picking travel locations or choosing a movie can be a hard task considering the number of choices we can find. It may be even harder if we are looking for user reviews to support our final decision. The high amount of comments in such websites can be a hold back for user looking for opinions that may help them to choose products/services, or even alert them for flaws already known to previous acquirers. Helpfulness Prediction (HP) is the task that aims to correctly predict whether a review or opinion is helpful for a user to read before acquiring a product or hiring a service.

Prediction models are usually based on supervised learning, therefore, demanding for linguistic resources (corpora) of labeled reviews. One of the challenges for helpfulness research in Brazilian Portuguese is the few number of available data sets. We present UTLcorpus, a data set composed of two automatic annotated corpora for helpfulness in that language. The data were extracted from two different domains: movie reviews from a Brazilian social network for movies[1] and app reviews from Google App Store[2].

The data was anonymized and preprocessed. Evaluations were carried out to bootstrap the corpora for the HP task and the results were compared to other literature corpora in Brazilian Portuguese. Since UTLcorpus also contains labels for binary polarity classification (star score indicative of positive and negative reviews) we also used literature methods for evaluating the data for this task.

[1]www.filmow.com. Accessed in May 19th, 2019.
[2]play.google.com. Accessed in May 19th, 2019.

The main contribution of this work is the creation of resources mainly for helpfulness prediction, but also for the polarity classification task. The corpus created in this work should be useful to increase the research in these areas and help to find the particularities of the helpfulness modeling task, thus enabling the understanding of this phenomenon in Brazilian Portuguese.

The paper is organized as follows. Section 2 presents an overview of Helpfulness Modeling and Prediction Task. The UTLcorpus is presented in Section 3. Experiments with the corpus in the tasks of helpfulness prediction and polarity classification are presented in Section 4. In Section 5 some important literature works on Polarity Classification and Helpfulness Prediction are discussed. Finally in Section 6 some conclusions and future works are presented.

## 2. Helpfulness Prediction Task

Modeling and prediction online reviews helpfulness (quality, usefulness or utility [Liu 2012]) are relevant for ranking and displaying comments to users who search comments on products or services. Most e-commerce websites present the most useful ones first and delegate to the users the task of evaluating whether they are helpful or not. Questions like "Was this review helpful to you?" are presented to the users and the feedback allows the system to re-rank eventually the set of reviews.

The drawback of this functionality is that the reviews can take a long time to accumulate a good number of user feedback. This is especially noticeable in new reviews, which can even be useful, but because of their low posting time, they can not get sufficient votes to achieve the top of the ranking. This fact demonstrates one of the advantages of automating the task. Websites that do not have ranking systems can benefit as well as the rankings themselves can be improved by the use of helpfulness prediction. In addition, the prediction of helpfulness can be used to filter off low-quality reviews, which can improve other tasks, such as the reviews summarization [Anchiêta et al. 2017].

Helpfulness prediction tasks mainly include score regression, binary review classification and review ranking. These three methods depend on the helpfulness score which is usually calculated for each review by the Equation 1. The score regression aims to predict the helpfulness score $h \, \epsilon \, [0, 1]$. The binary review classification seeks to decide whether comments are helpful or not based on a specific threshold (e.g. $h > 0.5$). And the review ranking needs to order the reviews by their helpfulness according to a reference ranking.

$$h = \frac{helpful\ votes}{helpful\ votes + unhelpful\ votes} \tag{1}$$

Several features have been used to characterize helpfulness in the literature. Usually they are split in two categories: Content and Context features [Diaz and Ng 2018]. The content features are related to the information that can be extracted directly from the review, such as the text and the stars given by the author. And the context features are those extracted from outside the review, such as reviewer information. In the survey of [Diaz and Ng 2018] one can find the most important features of the literature.

Contrary to what occurs for Portuguese, some large English corpora with utility annotations are available:

- Multi-Domain Sentiment Dataset (MDSD) [He and McAuley 2016] [3]: Collected from Amazon.com. Contains 25 product categories and 1.422.530 reviews.
- Amazon Review Dataset (ARD) [Blitzer et al. 2007, McAuley et al. 2015] [4]: Also collected from Amazon.com, contains 24 product categories and 142.8 million reviews and includes more metadata information than MDSD.
- Ciao Dataset [Tang et al. 2013] [5]: Was collected from an extinct e-commerce website and contains 302.232 reviews. The main difference from the previous ones is that it contains a social network between their users.

For the best of our knowledge, there is only one available corpus in Brazilian Portuguese, the *Buscapé* [Hartmann et al. 2014], containing $28.774$ product reviews annotated with information that can be used for calculating helpfulness. Therefore, this work presents a new *corpus* containing information of helpfulness to promote researches in this task.

## 3. The UTLcorpus

The data set is a collection of reviews extracted from two domains: movies and apps. $2.881.589$ reviews ($1.839.851$ of movies and $1.041.738$ of apps) were collected using two web crawlers. The domains were chosen for the popularity, the high amount of data and the presence of a public "like" counter in each review, which makes possible to infer a helpfulness label. Besides the "like" counter, the data also contains scores given by users to the movie/app they are evaluating. We used the later for inferring positive and negative labels.

The methodology for labeling the polarity was proposed in [Avanço 2015]. Each review has a 5-star score according to the author's evaluation of the related movie/app. Reviews with $0$ and $5$ stars are ignored to avoid those cases in which the users stars are not coherent with the review text. Also the $3$ star reviews are discarded because they usually contain positive and negative sentiment about the entity.

In order to label the data we looked for the utility labels in the data set. Both domains provide the number of "likes" a review received (indicating it was helpful for other users) and the main issue we faced was the lack of a counterpart indicating the number of "dislikes" were attributed to the review. The majority of works in the literature [Kim et al. 2006, Malik and Hussain 2017] divide the number of positive likes by the sum of likes and dislikes to obtain a value and determine a threshold of helpfulness for a data set.

Since we can not count on dislikes, we define helpfulness in UTLcorpus as following. First, we group the data by category (movie titles and app names). This is performed because more popular apps/movies aggregate more likes by review than the less popular ones. Then we sort the reviews by the number of likes each of them has received (ignoring the ones with zero likes, since we can not identify if they have anything helpful in

[3]https://www.cs.jhu.edu/ mdredze/datasets/sentiment/

[4]http://jmcauley.ucsd.edu/data/amazon/

[5]https://www.cse.msu.edu/ tangjili/trust.html

| | Movie review subset | App review subset |
|---|---|---|
| **# documents** | 1.834.702 | 921.257 |
| **# types** | 1.828.647 | 419.713 |
| **# tokens** | 60.177.264 | 11.919.636 |
| **Avg. token per doc.** | 32.7994 | 12.9384 |
| **Helpfulness Labeled** | 1.833.691 | 898.847 |
| | *helpful:* 381.083 *(20%)* | *helpful:* 50.166 *(5%)* |
| **Sentiment Labeled** | 862.768 | 320.255 |
| | *positive:* 702.720 *(81%)* | *positive:* 113.351 *(35%)* |

**Table 1. UTLcorpus information**

| Subset | Movie Review | | App Review | |
|---|---|---|---|---|
| | Positive | Negative | Positive | Negative |
| **Helpful** | 155.672 | 37.699 | 6.533 | 18.473 |
| **Unhelpful** | 546.357 | 122.029 | 98.374 | 174.465 |

**Table 2. Intersection between classes**

it). Next, we remove replications and then consider helpful any comment with more likes then the first percentile of the distribution. To determine the negative samples we consider any review with fewer likes than the threshold previously defined and crawled at least five days after the review was published; the observation of the timespan is important since recent reviews take longer to achieve higher like counts thus becoming a false negative noise in the data set.

Online reviews are classified as User Generated Content (UGC) [Krishnamoorthy 2015], a type of text which carries many noisy linguistic phenomenons such as typos and Internet slangs. In order to reduce the noise in UTLcorpus the data was normalized using Enelvo [Bertaglia 2017], a tool for normalizing UGC in Brazilian Portuguese[6].

The data set is composed of two subsets representing two different domains:

**Movie review corpus:** The movie reviews corpus was obtained by crawling Filmow, a popular film social network containing reviews, scores, evaluations and general movie information. We crawled reviews from the 4.283 most popular movies in the platform (we stopped for storage reasons) and the reviews generally represent opinions about films, actors/actresses and all kind of experience the users can have in watching a movie. Reviews can be directed to recent blockbusters as well as classics and we made available the movie titles which the reviews are about. The class distribution is skewed and the majority class for helpfulness is of unhelpful (80% of the documents) and positive (around 80% of the documents). Even though the helpfulness and polarity tags come from the

[6]Available in `https://github.com/tfcbertaglia/enelvo`. Visited in March 19th, 2019.

same data set, some reviews do not have enough information to be part of both of the subsets (972.945 documents do not the overlap).

The most frequent terms in the corpus (ignoring stop-words such as demonstratives pronouns, conjunctions and punctuation) are: *movie*, *well*, *good*, *story* and *best*.

**App review corpus:** The App Review corpus was obtained by crawling Google Play Store[7]. The corpus contains app reviews and one important feature of this data set is the absence of reviews with zero stars, since it is mandatory to evaluate with a star score any review. We gathered reviews from 243 apps (the most popular ones) and the whole corpus contains 921.257 reviews. The data is also skewed in both labels, being unhelpful the majority class for helpfulness (95%), and negative the majority class for polarity (65%). 371.651 reviews have only one label and do not overlap.

The most frequent terms in the corpus (also ignoring stop-words) are: *app*, *best*, *great*, *can* and *cool*. It is interesting no notice that several words are more frequent in both corpora even though they are skewed for different polarity classes.

Table 1 contains detailed data set information. The skewing of the data is a challenge for machine learning classification methods and we address this issue in section 4. Table 2 presents the intersection between the classes (Helpfulness and Polarity) for both datasets. It is possible to see that, in the movie reviews subcorpus, 80% of the helpful comments have positive polarity, and so are 81% of the unhelpful comments. Moreover, in the app reviews subcorpus, most of the helpful reviews (73%) as well as of the unhelpful ones are negative (63%).

## 4. Corpus Evaluation

In order to observe and evaluate the characteristics of the corpora on classification tasks we performed experiments in both subsets of UTLcorpus, Movie Review corpus (MR) and App Review corpus (AR), using a baseline and machine learning classifiers for comparison purposes. Helpfulness Prediction can be seen as a task very similar to polarity classification thus we defined a baseline and also used three machine learning classifiers (Support Vector Machines[8], Multi-layer Perceptron[9] and Random Forest[10]) following the work of [Brum and Nunes 2018], originally proposed for polarity classification of sentences. The work of [Brum and Nunes 2018] used a grid search technique to set the hyper parameters.

For baseline purposes we represented each sentence using a 2-dimensional vector with the number of positive and negative terms using a Brazilian Portuguese sentiment lexicon – Sentilex [Silva et al. 2012], which contains Portuguese terms (eg. *bom*, *ruim*, *péssimo*) and their respective polarity label. We trained a SVM model using this feature representation and evaluated the data sets in a 10-fold cross validation scheme. Furthermore, we used three classifiers trained and evaluated on a 10-fold cross validation. To avoid the skewing of the majority class, the data were balanced randomly (by the minority class), thus reducing the data sets considerably. The final sizes of the corpora are

[7] `https://play.google.com/store`. Visited in March 19th, 2019.

[8] Hyper parameters – C: 1; *alpha*: 0.1; linear kernel.

[9] Hyper parameters – Activation: $tanh(x)$; learning rate: 0.001; alpha: 0.0001; neurons: 200; layers: 2.

[10] Hyper parameters – number of estimators: 200.

762.078 documents in the MR corpus and 100.322 in the AR corpus, nevertheless, the final corpus is still larger than the Brazilian Portuguese corpus *Buscapé*.

To the other three classifiers, differently from the baseline method, the data was represented using pre-trained 600-dimensional word2vec embeddings trained in more than 1 billion Portuguese Brazilian user-generated content (tweets and forums). The representation is described in [Corrêa et al. 2017] and has also been used for polarity classification in [Brum and Nunes 2018].

| Classifier | Movie Review | | | App Review | | | Buscapé | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1-Help | F1-No-Help | F1-Measure | F1-Help | F1-No-Help | F1-Measure | F1-Help | F1-No-Help | F1-Measure |
| **Baseline** | 0.4499 | 0.6493 | 0.5496 | 0.5896 | 0.6617 | 0.6256 | 0.4967 | 0.6343 | 0.5655 |
| **Linear SVM** | 0.6341 | 0.6039 | 0.6189 | 0.7115 | 0.6436 | 0.6775 | 0.6072 | 0.6142 | 0.6107 |
| **MLP** | 0.6387 | 0.6118 | **0.6252** | 0.7082 | 0.6516 | 0.6799 | 0.6114 | 0.5983 | 0.6048 |
| **Random Forest** | 0.6220 | 0.5920 | 0.6072 | 0.7267 | 0.7182 | **0.7224** | 0.6361 | 0.6436 | **0.6398** |

**Table 3. Helpfulness detection results**

The results obtained in the classification are shown in Table 3. The F1 values presented in the table are acquired with 10-fold cross-validation technique (Mean of 10 executions). Before classifying the data the class distribution was balanced by using *undersampling*, in other words, we removed samples of the majority class before performing the cross-validation. Experiments with the unbalanced corpora resulted in F1 far below the ones presented in Table 3, even the best results had the minority class F1 below 0.1.

The baseline worked pretty well in relation to F1-Measure, but one can see that it does not handle well the positive class. The best results for helpfulness detection in both corpora were obtained using Random Forest classifier which predicts the class based on several estimators (Decision Trees). It is still uncertain if our method for defining the helpful class is reliable enough, but with this methodology it is still possible to predict the correct label in 60% of the time for movie reviews and 70% of the time for app reviews ($std.dev. = 0.004$). We believe that one possible explanation for the results is that the prediction of the utility does not depend on text only. We understand that helpfulness may be affected by the context (domain, category, website, etc.) in which the comment is inserted, as well as by the intention with which the reader is reading the comment.

Since UTLcorpus also has polarity labels we were able to perform experiments using them. The main difference was the baseline used: for polarity classification we represented the data similarly (positive and negative term frequency) but predicted as positive any sentence with more positive terms than negative ones. In Table 4 we can see the results for polarity classification in the corpora.

| Classifier | Movie Review | | | App Review | | | Buscapé | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1-Pos | F1-Neg | F1-Measure | F1-Pos | F1-Neg | F1-Measure | F1-Pos | F1-Neg | F1-Measure |
| **Baseline** | 0.6167 | 0.1675 | 0.3920 | 0.6378 | 0.1541 | 0.3959 | 0.5467 | 0.2031 | 0.3748 |
| **Linear SVM** | 0.6878 | 0.6517 | 0.6697 | 0.7687 | 0.7710 | 0.7698 | 0.8106 | 0.8243 | 0.8174 |
| **MLP** | 0.6602 | 0.6843 | **0.6722** | 0.7818 | 0.7814 | **0.7815** | 0.8146 | 0.8121 | 0.8133 |
| **Random Forest** | 0.6528 | 0.6644 | 0.6586 | 0.7588 | 0.7786 | 0.7686 | 0.8310 | 0.8128 | **0.8218** |

**Table 4. Polarity classification results**

Finally, we merged the corpora and perform experiments using the two domains. The choices of candidates reviews are the same for each domain and all selected by the criteria above mentioned are used this time. The results are presented on Table 5.

| Classifier | Helpfulness Prediction | | | Polarity Classification | | |
|---|---|---|---|---|---|---|
| | F1-Help | F1-No-Help | F1-Measure | F1-Pos | F1-Neg | F1-Measure |
| **Baseline** | 0.6708 | 0.4583 | 0.5645 | 0.1570 | 0.6179 | 0.3874 |
| **Linear SVM** | 0.6299 | 0.6759 | 0.6529 | 0.7011 | 0.7578 | 0.7294 |
| **MLP** | 0.6383 | 0.6863 | **0.6623** | 0.7304 | 0.7646 | 0.7475 |
| **Random Forest** | 0.6398 | 0.6834 | 0.6616 | – | – | – |

**Table 5. Helpfulness prediction and Polarity classification results using the whole UTLcorpus**

For polarity classification we are able to better compare the results than for helpfulness since the literature contains more works relating that task. The results of Table 4 reached almost $0.8$ F1 and one of the reasons may be that polarities are easier to separate from each other – usually people use different expressions and different words when evaluating positively or negatively a movie or app, the same does not always apply for helpfulness. Even though the best result obtained in *Buscapé* corpus was $0.8174$ in F1, other authors achieved $0.8935\%$ in the same corpus [Avanço et al. 2016].

One of the reasons for the low results is that the representation used (pre-trained word embeddings) usually works well with neural models since they basically rearrange the data using weights. It may explain why the best results for the whole corpus were obtained using MLP (Table 5), which follows the same principle with less layers. Another limitation was the size of the dataset. Linguistic approaches (which use n-grams for example) demand more resources for storage and processing. The $t\text{-}value$ between results was measured in order to calculate the significance of the differences and all of them were significant at $p < 0.05$.

## 5. Related Work

This paper relates to several other works both in helpfulness detection and sentiment analysis due to its similarities between fields.

For helpfulness prediction as a classification task, [Krishnamoorthy 2015] examines the impact of some specific linguistic features based on a model named Linguistic Category Model (LCM) [Semin 2011], on helpfulness prediction task. The author builds three machine learning methods for helpfulness binary classification, using a threshold $h = 0.60$.

Using a corpus extracted from Amazon.com (MDSD), the Random Forest method achieved the best result reaching an average of 84% of F-measure using all features. Individually the LCM features obtained the best results.

[Zeng et al. 2014] addressed the helpfulness prediction problem as a three-class classification problem. The classes are (1) Helpful positive reviews (star rating $\epsilon$ [4,5] and helpfulness score $h > threshold$); (2) Helpful negative reviews (star rating $\epsilon$ [1,2]

and helpfulness score $h > threshold$), and (3) Unhelpful reviews (helpfulness score $h < threshold$). They collected $8.690$ reviews from Amazon.com. The experimentation included an empirical test to decide the helpfulness score threshold. The best value obtained 72.82% of accuracy on ten-fold cross-validation. Specifically, regarding each class, the helpful positives reached 69% in macro-f1; the helpful negatives, 79,5% in macro-f1 and the unhelpful ones, 80% in macro-f1.

[Malik and Hussain 2017] used an emotion score of reviews (confidence, surprise, anger, etc.) as a feature to predict helpfulness. The authors modeled and evaluated a set of learning methods on Amazon.com corpus (MDSD), and they achieved 89% of f-measure, using emotion features as input for a deep neural network method.

In [Hartmann et al. 2014] the authors introduce *Buscapé*, a corpus for user-generated content research constructed using product reviews from an e-commerce website in Brazilian Portuguese. The authors extracted $85.910$ documents and annotated typos and Internet slangs for normalisation task. The corpus also contains a 5-star-based score and "like" votes that we used in this paper (section 4) for comparison with our own results in UTLcorpus.

This data set has been used several times in literature [Avanço et al. 2016, Brum and Nunes 2018, Bertaglia 2017]. We emphasize the work of [Avanço et al. 2016] because the authors classified the data using machine learning classifiers (SVM and Naive Bayes), lexical-based classifiers and ensemble of classifiers (both machine learning-based and lexical-based) and achieved the state-of-the-art for the corpus, $0.8935$ in f1. The main difference of this paper with ours is that we only used machine learning classifiers and we used embeddings for data representation, whilst those authors used a combination of *bag-of-words* and linguistic features such as number of sentiment words and PoS tags.

We can also compare our work with [Corrêa et al. 2017] since they also annotated a large corpus for semantic purposes (polarity classification). In this paper the authors crawled Twitter for Brazilian Portuguese posts and used Distant Supervision, automatically labeling documents based on semantic clues, to form a large corpora for sentiment analysis. Pelesent is composed of $980.067$ tweets that contained emojis and/or emoticons indicating negative or positive polarity (eg. ":)" for positive and ":(" for negative).

## 6. Discussion and future work

In this paper we presented UTLcorpus, a review corpus of two domains (movies and apps) with automatic labels for helpfulness detection and polarity classification. We proposed an automatic label methodology for helpfulness using "like" votes and used a literature inspired method for attaching a polarity (positive or negative) to $2.755.959$ forming two corpora – one of Movie Reviews ($1.834.702$ documents) and one of App Reviews ($921.257$). The methodology was replicated in a similar corpus (Buscapé) in order for comparing sizes and results obtained in classification experiments.

UTLcorpus is one of the first data sets for helpfulness detection in Brazilian Portuguese, but it can also be used for sentiment analysis (polarity classification, aspect extraction or else) and for others NLP tasks such as language modeling, normalization, discourse analysis or semantic parsing, for example.

We evaluated the corpus using machine learning methods from the literature and

obtained results up for replication and comparison with other models. The dataset is available in the github `github.com/RogerFig/UTLCorpus` already pre-processed, normalized with Enelvo and anonymised in order to be used for research purposes.

One of the future works to be conduct is the exploration of state-of-the-art models for classification such as convolutional neural networks [Kim 2014] and Long-short Term Memory architectures as well as the investigation of different representation models – morphological-based embeddings [Bojanowski et al. 2017] or context embeddings such as Elmo [Gardner et al. 2017]. Another future work is to expand the usefulness prediction for handling comment rating so that a list of best comments can be presented to users. Finally, manual evaluation of helpfulness can be performed on reviews, although we believe that a reader who is not interested in a product will handle a review differently from an interested one.

## References


Anchiêta, R., Sousa, R. F., Moura, R., and Pardo, T. (2017). Improving opinion summarization by assessing sentence importance in on-line reviews. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 32–36.

Avanço, L. V. (2015). Sobre normalização e classificação de polaridade de textos opinativos na web.

Avanço, L. V., Brum, H. B., and Nunes, M. d. G. V. (2016). Improving opinion classifiers by combining different methods and resources. *XIII Encontro Nacional de Inteligência Artificial e Computacional*.

Bertaglia, T. F. C. (2017). Normalização textual de conteúdo gerado por usuário.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brum, H. B. and Nunes, M. d. G. V. (2018). Building a sentiment corpus of tweets in brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resources Association.

Corrêa, E. A., Marinho, V. Q., dos Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., and Brum, H. B. (2017). Pelesent: Cross-domain polarity classification using distant supervision. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 49–54. IEEE.

Diaz, G. O. and Ng, V. (2018). Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 698–708.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform.

Hartmann, N., Avanço, L., Balage Filho, P. P., Duran, M. S., Nunes, M. D. G. V., Pardo, T. A. S., Aluísio, S. M., et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *LREC*, pages 3865–3871.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Malik, M. and Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, 73:290–302.

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA. ACM.

Semin, G. R. (2011). The linguistic category model. *Handbook of theories of social psychology*, 1:309–326.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. *International Conference on Computational Processing of the Portuguese Language*, pages 218–228.

Tang, J., Gao, H., Hu, X., and Liu, H. (2013). Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 1–8, New York, NY, USA. ACM.

Zeng, Y.-C., Ku, T., Wu, S.-H., Chen, L.-P., and Chen, G.-D. (2014). Modeling the helpful opinion mining of online consumer reviews as a classification problem. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014*, 19(2).

CAPÍTULO

5

# AVALIAÇÃO DE ATRIBUTOS

Neste capítulo, é apresentado o artigo publicado que descreve e avalia os atributos gerados manualmente utilizados neste trabalho de doutorado.

Nesse artigo é descrito um processo de avaliação qualitativa da tarefa de modelagem e predição da utilidade de opiniões por meio de um processo de anotação manual e da análise de atributos relevantes da tarefa. Também discutem-se no artigo alguns desafios da tarefa, entre os quais a capacidade de avaliação da utilidade por parte das pessoas.

*Rogério Figueredo de Sousa e Thiago Alexandre Salgueiro Pardo. 2021. "The challenges of modeling and predicting online review helpfulness". In Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). 2021. p. 727–738.*

## Declaração de Contribuição

R.F. Sousa desenvolveu a pesquisa, contribuiu com a elaboração, execução e avaliação do processo de anotação e colaborou com a escrita do manuscrito. T.A.S. Pardo contribuiu com a elaboração do processo de anotação, auxiliou na escrita do manuscrito e supervisionou o projeto.

# The Challenges of Modeling and Predicting Online Review Helpfulness

**Rogério Figueredo de Sousa, Thiago Alexandre Salgueiro Pardo**

[1]Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
Av. Trabalhador São Carlense, 400 – 13.566-590 – São Carlos – SP – Brazil

rogerfig@usp.br, taspardo@icmc.usp.br

***Abstract.*** *Predicting review helpfulness is an important task in Natural Language Processing. It is useful for dealing with the huge amount of online reviews on varied domains and languages, helping and guiding users on what to read and consider in their daily decisions. However, there are limited initiatives to investigate the nature of this task and how hard it is. This paper aims to fulfill this gap, providing a better understanding of it. Two complementary experiments are performed in order to uncover patterns of usefulness evaluation as performed by humans and relevant features for machine prediction. To assure our results, we run the experiments for two different domains: movies and apps. We show that humans agree on the process of assigning helpfulness to reviews, despite the difficulty of the task. More than this, people perform this process systematically and consistently. Finally, we empirically identify the most relevant content features for machine learning prediction of review helpfulness.*

## 1. Introduction

Web popularized access to large sets of information. Frequent actions as buying products and purchasing services may be done more consciously, as there are millions of reviews about products, movies, apps, and so forth. Unfortunately, such amount of information is a double-edged sword. On one hand, it provides valuable material to the users, but, on the other hand, it contains more information than a person can handle. This is a problem that is the subject of several areas. One of them is Natural Language Processing (NLP). In this paper, we are particularly interested in the subtask of Modeling and Predicting Online Review Helpfulness.

Among the large amount of data on the Web, User-Generated Content (UGC) is a major source, and product and service comments form a great portion of that content. However, not every comment (or opinion or review) is considered useful or relevant by other users. Indeed, some of this content may be considered unwanted, such as poorly written texts, vague opinions, texts with questionable content, etc [Kim et al. 2006]. This shows that user-generated content varies a lot in quality and such texts do not necessarily help readers' decision-making. A helpful review, according to [Mudambi and Schuff 2010] is a "peer-generated product evaluation that facilitates the consumer's purchase decision process". In such situation, modeling and predicting review helpfulness comprise the definition of models for characterizing good quality content and the proposition of methods for classifying opinions regarding their helpfulness degree.

Despite the importance of such research line, few studies have focused on the nature of this task and on determining how systematic and difficult it may be. The purpose of this paper is to bring some understanding on what influences people perception on review helpfulness and which features are more relevant for machines to automatically deal with online reviews. We run two complementary experiments on two different domains (movies and apps). We show that humans agree on the process of assigning helpfulness to reviews, despite the difficulty of the task. Moreover, we show that people perform this process systematically and consistently. Finally, we also identify the most relevant content features for machine learning prediction of review helpfulness.

The paper is organized as follows. Section 2 presents the main definitions about the task and also describes the main related work. Section 3 details the corpus that is used in this work. Section 4 describes the adopted methodology. In Section 5, we report the achieved results. Finally, Section 6 concludes the paper, indicating future research.

## 2. Related Work

Modeling and prediction of online review helpfulness are part of a task that studies the factors that determine review helpfulness and attempts to accurately predict it [Diaz and Ng 2018].

Helpfulness is relevant for ranking and displaying content to users who search comments on products or services on e-commerce websites. These websites usually present the most helpful ones first and delegate to the users the task of evaluating whether they are helpful or not. Questions like "Was this review helpful to you?" are presented to the users, and the feedback allows the system to eventually re-rank the set of reviews. However, some reviews can take a long time to accumulate a good number of user feedback. Recent reviews and the product with low user traffic are more affected by this fact. Therefore, automating the task is very beneficial. The automatic helpfulness prediction can benefit the websites that do not have ranking systems as well as can improve the manual rankings. In addition, the prediction of helpfulness can be used to filter off low-quality reviews, which can improve other tasks, such as review summarization [Anchiêta et al. 2017].

The main works in helpfulness prediction attempt to perform one of these three tasks: score regression, binary review classification, or review ranking methods. They depend on the helpfulness score that is usually calculated for each review by Equation 1. Score regression aims to predict the helpfulness score $h \, \epsilon \, [0, 1]$. Binary review classification seeks to decide whether comments are helpful or not based on a specific threshold (e.g., $h > 0.5$). Review ranking needs to order the reviews by their helpfulness according to a reference ranking.

$$h = \frac{helpful\ votes}{helpful\ votes + unhelpful\ votes} \tag{1}$$

Several features have been used to characterize helpfulness in the literature. They are usually split in two categories: content and context features [Diaz and Ng 2018]. The content features are related to the information that can be extracted directly from the review, such as the text and the "stars" given by the author. Context features are those extracted from outside the review, such as reviewer information. For the more interested reader, we recommend the survey of [Diaz and Ng 2018].

Most of the works try to generate a model using a set of those features. For instance, [Kim et al. 2006] used structural features (length of the review, number of sentences, etc.), lexical features (Term Frequency - Inverse Document Frequency (*TF-IDF*) statistic of words) or even syntactic and meta-data features (number of stars) in order to predict the helpfulness of reviews. They generated a regression model using a dataset extracted from *Amazon.com* and achieved their best results with the combination of all the features, obtaining 0.656 and 0.604 on Spearman correlation coefficient. More recently, [Baowaly et al. 2019] achieved the state of the art results for helpfulness classification. They used a dataset collected from the *Steam* game database and generated a model with the Gradient Boosting Machine algorithm. Some of their features were categorized in metadata features (e.g., recommendation, posting date of a review, etc.), reviewer features (e.g., number of reviews, number of acquired games, etc.), semantic features, TF-IDF and Word2Vec features, among others. Their model achieved more than 0.99 of f-measure in several categories of games, such as action, survival and RPG.

Some works attempted to understand the impact of the features in the task. [Mudambi and Schuff 2010] investigated what makes reviews helpful to a consumer. They evaluated three features: review extremity, review depth, and product type. Using a dataset from *Amazon.com*, they found out that the product type ("experience" or "search") influences the effect of the review extremity and the review depth over users. For experience goods, the extreme reviews are less helpful than moderate ones. The review depth has a positive influence on both product types, but has a bigger influence on search goods than for experience goods. [Tsur and Rappoport 2009], generated an algorithm to classify the reviews and, in addition, attempted to understand the nature of book review evaluation. Three human annotators evaluated 360 reviews and the authors concluded that review evaluation is subjective, but people still get a high agreement, achieving a Fleiss' kappa value of 73.3%.

Table 1. UTLCorpus numbers.

| | **Movies** | **Apps** |
|---|---|---|
| **# texts** | $1,833,691$ | $898,847$ |
| **# objects** | $4,283$ | $243$ |
| **# types** | $1,828,647$ | $419,713$ |
| **# tokens** | $60,177,264$ | $11,919,636$ |
| **Avg. of Tokens p/ doc** | $32.7994$ | $12.9384$ |
| **Helpfulness Label** | *helpful:* $381,083$ *(20%)* | *helpful:* $50,166$ *(5%)* |

In this paper, inspired by the previous initiatives, we present a deeper investigation of human behavior on evaluating helpfulness and of useful features for machine learning-based helpfulness prediction. We start by briefly describing in the next section the corpus that we use for our experiments.

## 3. The UTLCorpus

In this paper, we use the UTLCorpus [Sousa et al. 2019] as our dataset. This corpus is composed by reviews written in Portuguese for two domains: movies and apps. An amount of $2,732,538$ reviews ($1,833,691$ for movies and $898,847$ for apps) were collected using two web crawlers.

The authors of UTLCorpus anonymized the dataset and made it publicly available. They preserved important metadata fields from the original reviews, such as star rating, publication date, and, specifically in the movie domains, information on whether a reviewer saw a movie or whether the movie is a favorite.

Table 1 synthesizes the basic statistics of the corpus and shows some interesting information. One may see that the average size of movie reviews is much higher than that of apps. The information of helpfulness label shows that the corpus is highly unbalanced, mainly for the apps domain, which can be a problem in some cases. It is worth mentioning that this unbalancing problem does not interfere with the results presented here. The correlation experiments were performed on the balanced (with undersampling) and on the original (unbalanced) datasets, and the results were similar.

## 4. Research Methodology

Trying to understand the textual and non-textual features that characterize the helpfulness of online reviews, this work proposes a study of review helpfulness modeling and prediction. In this section, we present the proposed configuration of our study.

We investigated two complementary questions to guide our study, each one trying to understand a specific property of the helpfulness of reviews on apps and movies. In summary, the questions are as follows:

1. How difficult is the task for humans?
2. Which features are relevant for the task of helpfulness prediction?

Answering such questions may drive research in the area and foster the development of better systems in the future. In the following subsections, we explore each of the questions.

### 4.1. Helpfulness Evaluation is Difficult for Humans?

To answer this question, we need to discover if humans agree with each other while evaluating the helpfulness of reviews. For this purpose, we conduct a manual annotation process, counting with some annotators to accomplish this task.

The annotation process was to read and evaluate the helpfulness of 24 reviews extracted from the UTLCorpus, 12 from each domain, equally distributed in helpful and not helpful categories. These reviews were selected from only a movie and an app, randomly. The respondents needed only to choose among three options: *The review is helpful*, or *the review is unhelpful*, or *I don't know*.

To approximate the annotation process to that found in the ordinary process of evaluating the helpfulness of reviews, we decided to add an "information need" for annotators. Looking at the ordinary process of voting on the helpfulness of reviews on websites, we have found that users do not arbitrarily decide on the helpfulness of reviews. If they are reading reviews about a product, they are concerned with getting some relevant information about it. And because of their interest in the product, they can be more critical when evaluating reviews. This "information need" was specified to the annotators through the following sentences: "*You are deciding whether to download the app [app name] (to watch the movie [movie name]), and you have come across these reviews. You must answer*

*the following question for each review: "Is this opinion helpful to you?" Evaluate whether the review helps you to decide to download or not the app (to watch or not the movie).*" Note the underlined excerpts, they vary for each domain as highlighted in brackets. Figure 1 shows an example of a review in the form with an "information need" text.

We distributed a form to fourteen annotators, and they had a few days to accomplish the task. By the end of the deadline, only ten annotators completed the process.

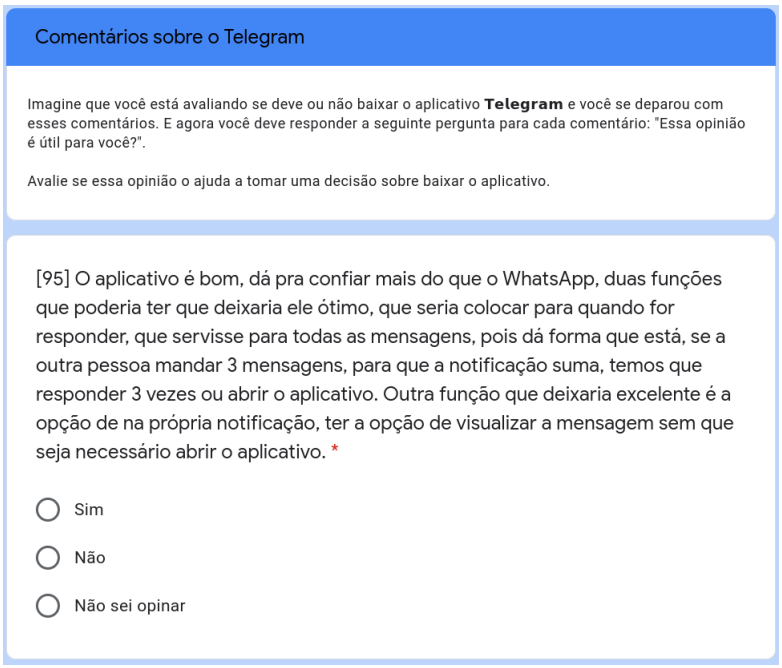



Figure 1. An example review (in Portuguese) on the form distributed to the annotators. It also shows the "information need" provided to annotators.

Although the main objective of the annotation process is to evaluate the agreement of annotators, we aggregate some other side objectives that could help us to understand the evaluation process of helpfulness by humans. We randomly selected the 24 reviews, but in sets with specific conditions. The first condition is the domain, 12 of each domain, as commented before. The second condition is the helpfulness category, being six of each class (helpful or not helpful), and, finally, the last condition is the length of review: three reviews are long and three are short. The short ones have 30 words at most, while the long ones have more than 60 words. Figure 2 helps to illustrate the subset we ended up for human evaluation. The decision to select 24 reviews for manual annotation was due to the nature of the comments. [Liu et al. 2007] and [Tsur and Rappoport 2009] show that the domains where characteristics are not so well-defined generate more open reviews, making evaluation difficult and expensive. Another reason is that this approach brings the process closer to the real voting conditions, where customers typically rate few comments.

We expect with this configuration to get some additional information about what influences people perceived helpfulness, more specifically, if the length of review influences the evaluation of review helpfulness.

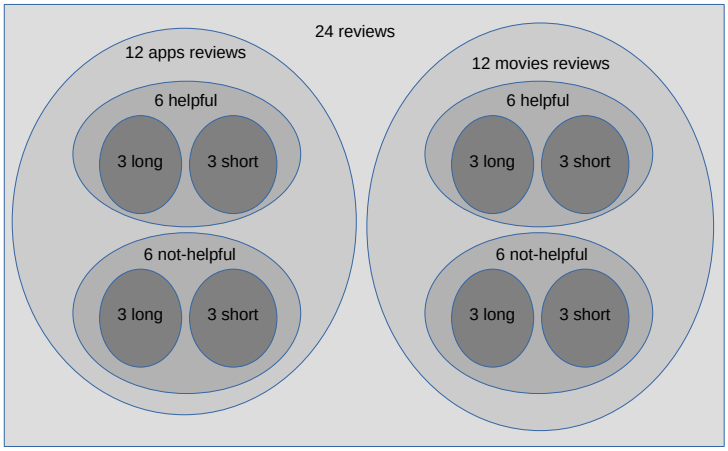



Figure 2. A graphical representation of the subsets of reviews for the evaluation.

### 4.2. Which features are relevant for the task?

The researchers in helpfulness modeling and prediction, along the years, developed many types of features trying to characterize the helpfulness of reviews in different domains and languages.

To answer the question "*Which features are relevant for the task?*", we followed a tiny pipeline:

1. To select the relevant candidate features from the literature in the area.
2. To select the necessary resources to implement and adapt the features to the language of the corpus (which is in Portuguese).
3. To implement the selected features.
4. To calculate the contribution of the features for the target task.

The first step of our pipeline revealed many features in many works. Considering that the features can be classified into different categories, we decided to limit the selection to content features only. The content features extract information directly from the reviews, such as review text and star rating. Most of these features are simple and easy to understand and to replicate, therefore, we were able to adapt and evaluate more features. And it is worth to mention that we selected and adapted the most common features of helpfulness prediction literature.

The second step shows us the necessary resources and tools to adapt the features to our language. Despite the differences in accuracy of many tools between languages, we choose the equivalent resources for each selected feature.

In the third step, we try to adapt the features as accurately as possible[2], considering the particularities of the language

The last step is the most important in our pipeline. In this step, we calculate the impact of the features, individually comparing to the helpfulness of the reviews. We decided to compute the correlation between feature values and the helpfulness class (not

[2]Our entire adaptation code of features is available at `https://github.com/RogerFig/features_experiments`

Table 2. List of Features.

| Feature | Description |
|---|---|
| Average Sentence Length (Avg-SL) | Ratio between the number of words and the number of sentences in the review [Liu et al. 2007, Lu et al. 2010] |
| Number of Sentences (Num-S) | Total of sentences in the review [Liu et al. 2007, Lu et al. 2010] |
| Number of Words (Num-W) | Total of words in the review [Kim et al. 2006, Mudambi and Schuff 2010] |
| Star Rating (Star-R) | The review-assigned product star rating [Huang et al. 2015] |
| Readability Features (READ) | Measures how easy a text is to read and contains the following features: Automated Readability Index (ARI), Coleman-Liau Index (CLI), Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning fog index (GFI) and SMOG [DuBay 2004, Ghose and Ipeirotis 2011] |
| Spelling errors (SPELL) | Total words not found in a lexicon composed of words from the Wiktionary[1] and the Unitex-PB lexicon [Ghose and Ipeirotis 2011, Muniz 2004] |
| Dominant Terms (Dom-Terms) | Presence of important terms in reviews, considering their specificity for the domain and movie/app [Tsur and Rappoport 2009] |
| Product features (Prod-Feat) | Presence of product features in the reviews [Kim et al. 2006, Hong et al. 2012, Liu et al. 2007] |
| Sentiment Words (SENT) | Word count that may reflect opinions, analyses, emotions etc. [Kim et al. 2006]. We use some categories of LIWC dictionary [Balage Filho et al. 2013, Pennebaker et al. 2001] to calculate these features. The categories are: Negate, Swear, Affect, Posemo, Negemo, Anxiety, Anger and Sad. |
| Sentiment divergence (Sent-Div) | Difference between the general sentiment about the movie/app and the sentiment expressed by the author of a review [Hong et al. 2012]. We used the Sentilex sentiment lexicon [Silva et al. 2012] to calculate this feature. |
| Subjectivity (SUB) | The probability of a review and its sentences being subjective [Ghose and Ipeirotis 2011] |
| Syntactic tokens (SYN) | Number of tokens with the following Part-of-Speech tags: Noun (N), Verb (V), Adverb (ADV) and Adjective (ADJ). It also includes counting for open class words (Open) [Kim et al. 2006] |
| Star Deviation (Star-Dev) | Difference between the amount of stars in a review and the average star rating for the movie/app [Hong et al. 2012] |

helpful: 0 and helpful: 1) of reviews using the correlation coefficients of Pearson and Spearman. All features have been normalized and Section 5 presents the correlation results.

With this process, we expect to find clues about the impact of features in helpfulness definition, determining which features are more or less relevant to the task. Table 2 presents and describe all features used in this work, including citations to some of the main previous works that used them.

## 5. Results and Discussion

Considering the methodology described in Section 4, we present in this section the results achieved in the annotation process and the correlation study between features and helpfulness.

### 5.1. The Annotation Process and Evaluation of the Lexical Similarity

In order to evaluate the annotation process, we used a well-known inter-annotators agreement metric: Krippendorff Alpha [Krippendorff 1970]. For the sake of better visualization, the results are divided into some groups.

Figure 3 shows the results of the annotation process. It is worth to remember that we impose some conditions to select the reviews. We split the reviews on these three groups: length (short, long), domain (movies, apps) and helpfulness (helpful, not helpful). Hence, the figure presents the inter-annotator agreement considering the combination of groups. The first part of the figure shows the agreement for bigger groups. The second part of the figure presents the agreement values for composition of two groups: *length* X *domain*. The third part of the figure presents the agreement results considering all three groups: *domain* X *class* X *length*.

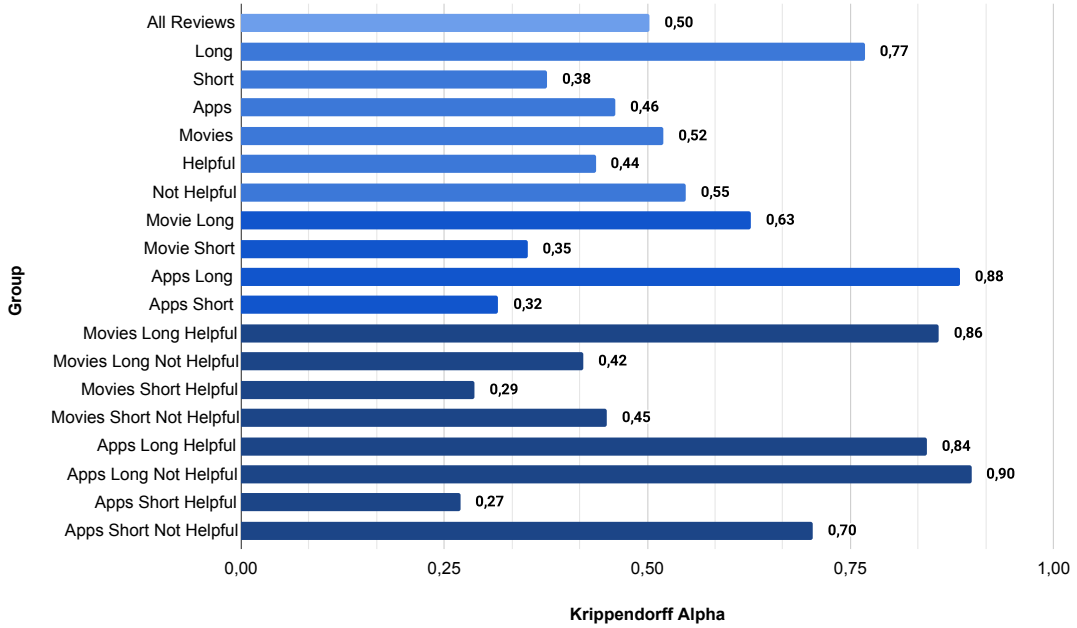



Figure 3. General results for agreement.

There are some information that stands out in the table. In the first place, we can observe an agreement pattern. The long reviews produce a better agreement than the short ones, probably because the longer ones tend to include more information to support the user decision (considering the information need). Apps also produce better agreement values, which may be possibly explained by the less subjective reviews (as they frequently comment on technical aspects of the apps). The best agreement results were achieved by apps' long reviews for the helpful category. It is also interesting how short reviews (for both domains) do not produce good agreement results for the helpful category. Overall, the high agreement results achieved for some cases show that the task is clear enough for humans under certain circumstances, as enough amount of available information (as provided by the longer reviews).

We proposed an additional experiment, which consists of evaluating the lexical similarity of reviews and comparing their categories. If humans are consistent in their annotation, we expect to see higher helpfulness agreement as the lexical similarity increases.

For this experiment, we use the training part of the UTLCorpus, which contains 80% of reviews (1,466,952 movie reviews and 719,077 app reviews). The process was conducted as follows:

1. For each domain, the reviews were split in long and short ones;

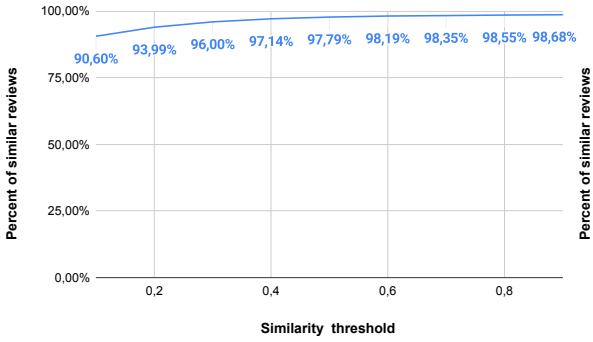



Figure 4. Short apps reviews.

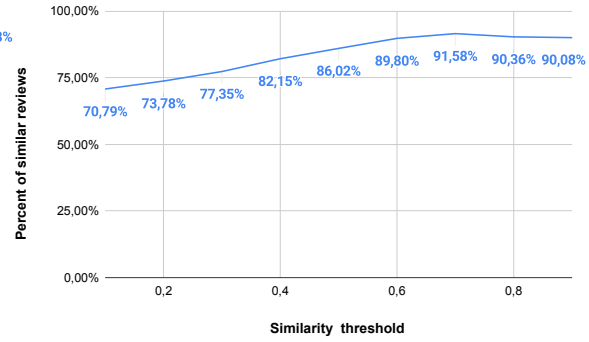



Figure 5. Short movies reviews.

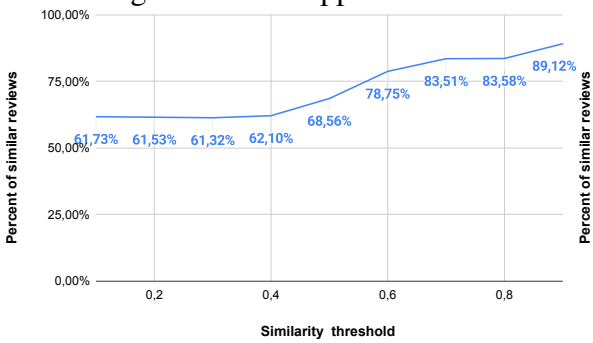



Figure 6. Long apps reviews.

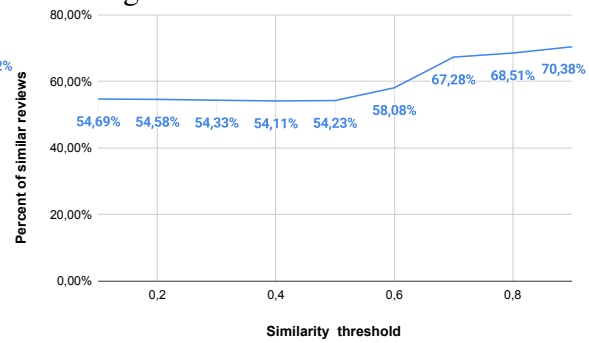



Figure 7. Long movies reviews.

Figure 8. Similarity Experiment.

2. Each review was represented by a Bag-of-Words Vector;
3. The Cosine similarity was calculated among all vectors (all vs. all);
4. We calculate the percentage of the reviews that have a cosine similarity above a threshold and have the same helpfulness category.

Several similarity thresholds have been considered, ranging from 0.1 to 0.9 and the results are presented in Figure 8. The X axis shows the similarity thresholds and the Y axis shows the percentage of reviews with the same helpfulness category. As expected, we may see that the proportion of reviews with the same category grows with the increase of the lexical similarity. The short reviews have a higher proportion of similarity than the long ones. One possible explanation is that users have a tendency to use less diverse vocabulary to write shorter comments. On the other hand, the authors need to use a diversified vocabulary to write the long ones.

Taken together, these results suggest that there is strong evidence that people agree with each other on the process of assigning helpfulness to reviews in domains of movies and apps, and they perform this process systematically and consistently. Moreover, the lexical similarity curves support the evidence that human judgment is not aleatory.

### 5.2. Correlation of features with helpfulness

For the purpose of finding relevant features for determining the helpfulness of reviews, we calculate the correlation coefficients of Pearson and Spearman for all features in Table 2 in relation to the helpfulness class (not helpful: 0 and helpful: 1). For this experiment,

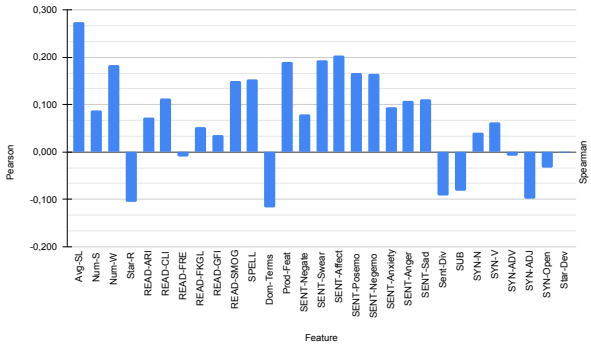



Figure 9. Pearson for Apps Domain.

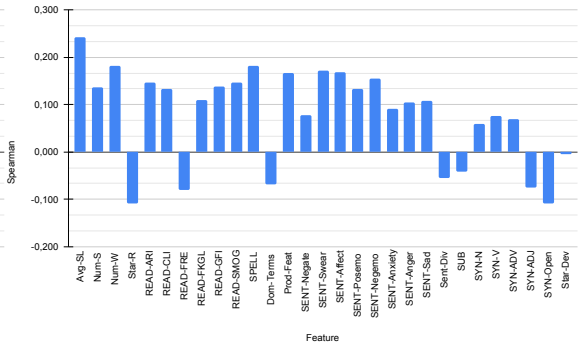



Figure 10. Spearman for Apps Domain.

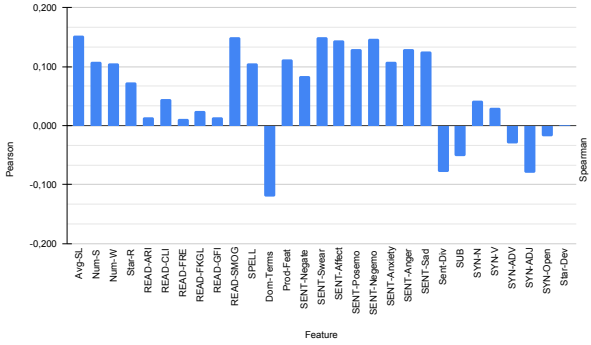



Figure 11. Pearson for Movies Domain.

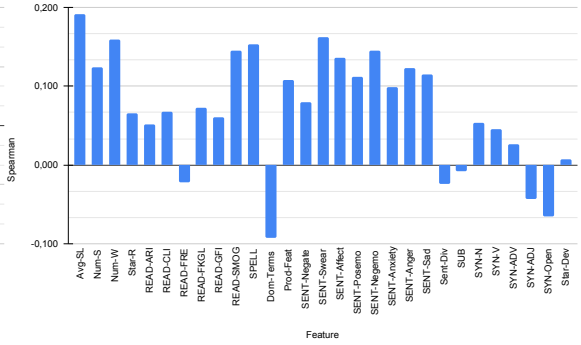



Figure 12. Spearman for Movies Domain.

Figure 13. Feature correlation results.

we used the training subset of UTLCorpus. As explained in the previous subsection, the training subset contains 80% of all the reviews of UTLCorpus.

Figure 13 summarize the results. Figure 11 and Figure 12 provides the results for the movies domain, and Figure 9 and Figure 10 presents the results for the apps domain.

An inspection of the figures shows that in both domains the simple features are among the most positively correlated features, for example, *Average Sentence Length, Number of Sentences, Number of Words, and Spelling Errors*. Some readability scores and the LIWC [Silva et al. 2012] features also showed a noticeable positive correlation. Each of the features in the sentiment words category refers to the category of the same name in the LIWC ("negate", "swear", "affect", "posemo", "negemo", "anxiety", "anger" and "sad"). In the opposite direction, we can highlight some features with inverted correlation, for example, *dominant terms* in both domains and *star rating* for apps domain. Most of the remaining features have not achieved important values of correlation, with intermediate results.

Being more specific, among the content features presented in this subsection, the most correlated ones with movie review helpfulness are (according to the two used correlation measures): Average Sentence Length, Readability-SMOG, and some Sentiment Features. Exclusively for Apps, we have: Average Sentence Length, Number of Words, Readability-SMOG, Spelling Errors, Product Features, and some Sentiment Features. It is interesting to notice that some of the features are relevant for both domains, indicating that they might be useful for building general domain classifiers.

The presence of common relevant features in the two domains is specially important for the area of sentiment analysis, as it is widely known that the domain usually makes a lot of difference in the performance of systems. More experiments must be carried out for obtaining irrefutable conclusions, but our domains (movies and apps) are different enough to allow us to infer that such features might be also relevant for other domains. Some evidence of the domain differences come from some researches that have shown that reviews on topics like movies and books tend to be more "passionate", while reviews on electronic devices and apps tend to be more "technical" (see, e.g., [Vargas and Pardo 2018] for some interesting discussion on this).

## 6. Final Remarks

In this paper, we presented a study of review helpfulness, trying to answer how hard the task is and which features appear to be more useful for prediction. We show that people agree with each other in the task of evaluating the helpfulness of reviews for movie and app domains (specially for longer texts). Moreover, through lexical similarity, we show that people are consistent in the task. We also evidence that some features are clearly correlated to task of helpfulness prediction, independently of the domain, which might help producing better general domain helpfulness classifiers. To the best of our knowledge, the work reported here is the most comprehensive one on such topics. The interested reader may find more information at the web portal of the POeTiSA project[3].

Future work includes generating machine learning classification models with the best features and testing context features, as these new features may bring more understanding about the task.

## Acknowledgments

The authors are grateful to the USP/IBM/FAPESP Center for Artificial Intelligence (C4AI, grant #2019/07665-4) and *Instituto Federal do Piauí* (IFPI).

## References

Anchiêta, R., Sousa, R. F., Moura, R., and Pardo, T. (2017). Improving opinion summarization by assessing sentence importance in on-line reviews. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 32–36.

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Baowaly, M. K., Tu, Y.-P., and Chen, K.-T. (2019). Predicting the helpfulness of game reviews: A case study on the steam store. *Journal of Intelligent & Fuzzy Systems*, 36(5):4731–4742.

Diaz, G. O. and Ng, V. (2018). Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 698–708.

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

[3] https://sites.google.com/icmc.usp.br/poetisa

Ghose, A. and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Hong, Y., Lu, J., Yao, J., Zhu, Q., and Zhou, G. (2012). 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*, SIGIR '12, pages 495–504, New York, NY, USA. ACM.

Huang, A. H., Chen, K., Yen, D. C., and Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, 48:17–27.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. (2010). Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700. ACM.

Mudambi, S. M. and Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200.

Muniz, M. C. M. (2004). *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB*. PhD thesis, Universidade de São Paulo.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language*, pages 218–228. Springer.

Sousa, R. F., Brum, H. B., and Nunes, M. d. G. V. (2019). A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Proceedings of the 12th Brazilian Symposium in Information and Human Language Technology*, pages 209–218. Sociedade Brasileira de Computação.

Tsur, O. and Rappoport, A. (2009). Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *ICWSM*.

Vargas, F. and Pardo, T. (2018). Hierarchical clustering of aspects for opinion mining: a corpus study. In Finatto, M., Rebechi, R., Sarmento, S., and Bocorny, A., editors, *Linguística de Corpus: Perspectivas*, pages 69–91. Porto Alegre: Instituto de Letras da UFRGS.

CAPÍTULO

# 6

# MÉTODOS

Neste capítulo, são apresentados os métodos para classificação da utilidade de opiniões propostos neste trabalho de doutorado. Primeiramente, na Seção 6.1, é descrito um método que se baseia em grafos, considerando que existem informações importantes nos relacionamentos entre tokens, estrelas e comentários que podem auxiliar na classificação da utilidade das opiniões. Já na Seção 6.2, é apresentado um *benchmark* para a tarefa, considerando diversos métodos, clássicos e modernos de aprendizado de máquina.

## 6.1 Método Baseado em Grafos

Neste artigo foi apresentado um vasto conjunto de experimentações com o objetivo de encontrar as melhores combinações de parâmetros para atingir os melhores resultados com a abordagem de grafos, em um cenário de escassez de dados. O trabalho apresentado neste capítulo é o seguinte:

*Rogério Figueredo de Sousa, Rafael Torres Anchiêta e Maria das Graças Volpe Nunes. 2020. "A Graph-Based Method for Predicting the Helpfulness of Apps Opinions", In Proceedings of the iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 13(4), 06-21.*

## Declaração de Contribuição

R.F. Sousa desenvolveu a pesquisa, contribuiu com a elaboração, execução e avaliação da experimentação e colaborou com a escrita do manuscrito. R.T. Anchiêta contribuiu com a elaboração do processo de experimentação e auxiliou na escrita do manuscrito. M.G.V. Nunes auxiliou na escrita do manuscrito e supervisou o projeto.

Submission date: 14/10/2019
Resubmission date: 16/03/2020
Camera ready submission date: 25/05/2020

1st round notification: 23/12/2019
2nd round notification: 12/05/2020
Available online: 29/05/2020
Publication date: 31/07/2020

Section: extended version
Extended versions of papers presented at the VIII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2019)

# A Graph-Based Method for Predicting the Helpfulness of Apps Opinions

**Rogério Figueredo de Sousa**[1], **Rafael Torres Anchiêta**[1], **Maria das Graças Volpe Nunes**[1]

[1]Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
Av. Trabalhador São Carlense, 400 – 13.566-590 – São Carlos – SP – Brazil

{rogerfig, rta}@usp.br, gracan@icmc.usp.br

***Abstract.*** *This paper presents a new approach to predict the helpfulness of opinions. Usually, researchers in this area use tables of attribute-value to aggregate the features that represent the evaluated texts. Although that representation is common, it considers that the objects are independent. We argue that among the discriminant factors of the helpfulness of opinions, there are dependent factors of the relationship among the opinion-forming elements. Thus, we modeled this task as a network, considering the information of relations among objects in the network (comments, stars, and words). A regularization technique of graphs is used to extract the relevant features of graph structure and, after that, the comments are classified as helpful or unhelpful. We compared our network model with two baselines methods, one based on fuzzy logic and another based on Neural Networks. Our model outperformed the fuzzy logic and Neutal Network methods in 0.17 and 0.19 of F-measure, respectively. The main advantages of our approach are that few data are necessary to helpfulness classification and the relationships may help in the understanding the classification, explaining the reasons for a determinate classification.*

***Keywords.*** *Natural Language Processing; Helpfulness Prediction; Opinion Mining.*

## 1. Introduction

Choosing a product to buy or a movie to watch is today one of the most frequent activities of Internet users. A common practice for these people is to search for information about products or services for which they are interested in specialized websites. In addition to technical information, it is the opinions of other users that interest them the most. If the advertiser is only interested in revealing the qualities of the product, the user who has had a good or bad experience with the product is only interested in contributing to other consumers. According to [Liu 2012], opinions are central to most human activities and are capable of influencing human behavior.

User-generated content (UGC) is a major source of web content, and product and service comments form a large portion of that content. Not every comment (or opinion

Cite as: Sousa, R. F., Anchiêta, R. T. & Nunes, M. G. V. (2020). A Graph-Based Method for Predicting the Helpfulness of Apps Opinions. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 13(4), 06-21. DOI: 10.5753/isys.2020.821

or review) is considered useful or relevant by other users. Indeed, some of this content may be considered unwanted [Kim et al. 2006]. Mentions that unwanted content includes poorly written texts, vague opinions, texts with questionable content, etc. That is, user-generated content varies greatly in quality and such texts do not help readers' decision making.

Another factor to difficult decision making is the huge amount of comments available on the web, making finding relevant content even more complicated. Moreover, it is impossible for users to read all the good quality content available. Deciding whether a comment carries potentially useful information for decision making is the central problem of this paper.

In Natural Language Processing (NLP), the Opinion Helpfulness Prediction task comprises the definition of models for characterizing good quality content and the proposition of methods for classifying opinions on helpfulness. Identifying relevant (helpful) content in user comments can help other users in decision making as well as support other NLP processes, such as the opinion summarization [Anchiêta et al. 2017].

The e-commerce sites' concern with presenting helpfulness content is great, which is why some of them ask for explicit feedback from the user: is this comment helpful or not? Thus, the comments presented are sorted according to the votes they have received, the most helpful first. Some problems arise from this form or manual voting [Kim et al. 2006, Liu et al. 2007, Singh et al. 2017]:

1. Helpful new comments will hardly be at the top of the ranking. It takes some time for several people to vote and the comment to gain proper visibility;
2. Items that have low visitor traffic will not have enough votes to generate a reliable ranking;
3. People may make a false assessment of the helpfulness of comments. Spammers are the ones who can dishonestly evaluate some comments so that they go up or down the ranking.

To avoid this type of problem, it is necessary to learn existing features in rankings of already consolidated comments and thus automatically evaluate the helpfulness of comments generated by users. The vast majority of known works perform this task using the attribute-value representation. But, despite being widely used, it is not able to capture relationship information between objects. The data structures that best represent relationships between objects are networks.

In this work, the task of helpfulness prediction was modeled as a heterogeneous network. To evaluate this approach, we used the UTLCorpus[Sousa et al. 2019], a recently released corpus, specifically the Google Play Store sub-corpus and, then, compared our approach with a well-known baseline based on fuzzy logic [de Sousa et al. 2015] and its evolution based on Neural Networks (NN)[Santos et al. 2016]. The network-based approach exceeded the fuzzy baseline by 0.17 points in F1 measure and 0.19 points in F1 measure on NN baseline, showing that our approach is feasible to predict whether a comment is helpful or not.

It is important to highlight that, as far as we know, this is the first work that models the helpfulness prediction task as a heterogeneous network. In addition, the approaches

are applied and evaluated in comments written in Portuguese in order to foster research in this area for this language.

The rest of the paper is organized as follows. In Section 2, the main related works are briefly described. In Section 3, we present the corpus and the developed modeling, as well as the steps to predict the helpfulness of the review. In Section 4, the performed experiments are detailed. Finally, Section 5 concludes the work, presenting future directions.

## 2. Related Work

[Zeng et al. 2014] attempted to include the sentiment polarity on binary helpfulness classification task. They model the problem in three classes. In the first class are the Helpful positive reviews (star rating $\epsilon$ [4,5] and helpfulness score $h > threshold$). In second class, are the Helpful negative reviews (star rating $\epsilon$ [1,2] and helpfulness score $h > threshold$), and in the last class, are the Unhelpful reviews (helpfulness score $h < threshold$). The purpose of this class division is to assess the impact of sentiment polarity on the helpfulness prediction task. Their dataset contains $8,690$ reviews from Amazon.com. The helpfulness score threshold was set empirically. The best result reported was 72.82% of accuracy on ten-fold cross-validation and specifically, for each class, the results reported were 69% on macro-f1 for the helpful positives; 79.5% on macro-f1 for the helpful negatives and 80% on macro-f1 for the unhelpful ones.

[Krishnamoorthy 2015] builds a model for binary helpfulness classification task. The authors introduced some linguistic features as features of their model. These features was based on a model named Linguistic Category Model (LCM) [Semin 2011] and these features are capable to identify the emotional state of the reviewer. They assume that linguistic categories are perceived by consumers and impact their vote behavior and, hence, the helpfulness of a review. For experimentation, they used an extracted corpus from Amazon.com (MDSD) and, using a threshold $h = 0.60$, the authors build three machine learning methods for helpfulness binary classification. Among the built models, the best result was achieved by an Random Forest model, reaching 84% of F-measure using all proposed features. The LCM features achieved the best result among other features.

In [Malik and Hussain 2017] the authors treated the utility prediction as a sorting task. They devised a method for calculating the emotion score of comments considering some specific feelings such as confidence, surprise, anger, sadness, etc. They used these scores as a feature in addition to more general ones, such as product ranking on Amazon, product price, number of verbs, nouns, adjectives and adverbs, among others. They modeled a Deep Neural Network and also evaluated the method on an Amazon.com subcorpus. The authors reported results on average of 89% F1 using positive emotions and 87% F1 using negative emotions.

In addition to the work described above, it is worth mentioning the work done focusing on the Portuguese language. Four of them will be described bellow.

The work of [Martins and Tacla 2015] presents a methodology focused on identifying features that have the greatest influence on utility votes. Experiments are applied to service domain (hotel) reviews. The authors propose several features that are able to

characterize comments, and they are divided into two categories: textual and semantic. Textual features consist mainly of intelligibility metrics. For their extraction, they use an adapted version of Coh-Metrix-Port [Scarton and Aluísio 2010]. In addition to the intelligibility index, other textual metrics, such as number of sentences, words and syllables, are used. For semantic features, the LSA (Latent Semantic Analysis) [Landauer et al. 1998] is used. The results of this work confirm the positive impact of semantic features in evaluating the helpfulness of opinions also for the Portuguese language. The intelligibility index revealed that longer and more complex opinions are more helpful than shorter and more intelligible opinions.

A different approach to classifying the importance of comments for the Portuguese language was presented in [de Sousa et al. 2015]. The authors proposed a Fuzzy Inference System to classify product domain comments into 4 classes: Insufficient, Sufficient, Good, and Excellent through 3 features: author reputation, number of type pairs (feature, opinion word), and richness of vocabulary. Comments are ranked according to the value expressed by the system. After sorting the list, several cut points were defined successively, and at each cut point a baseline method was applied to define the polarity of the comments. The authors compared the results on the subsets by applying the same method to the complete set. The results showed that a cut-off point considering only 10% of the comments obtained a better result than the complete set analysis. The authors presented a 10% increase in f-measure for positive comments and about 20% f-measure for negative comments.

The work of [Santos et al. 2016] has extended the work of [de Sousa et al. 2015]. They improved the definitions of some characteristics and proposed an experimental study to compare the previous approach of fuzzy systems with the use of Artificial Neural Networks. Two topologies of artificial neural networks were proposed. The authors reported that they could not improve the previous results, but argued that this is due to some factors: the samples obtained scattered expected results, which made the generalization of the network difficult; and none of the candidate topologies reached the minimum accuracy. The minimum accuracy serves as the minimum limit to consider the trained network. 52.48% of f-measure was achieved for positive comments and 62.53% of f-measure for negative comments.

Finally, [Barbosa and Moura 2016] assessed the helpfulness of opinions in the field of games. The authors collected comments from Steam[1] and used the authoring features of the opinions, textual characteristics and metadata existing on the site as input for an artificial neural network of the MLP (Multi-layer Perceptron) type to infer the helpfulness of the reviews. After the experiments, they reported good results and showed that the metrics related to authorship were more relevant along with the size of the text. On the other hand, the date of posting of the comments did not have a strong impact on the evaluation.

Looking at the involved elements in the scenario of users opinions - the opinion text, the reviewer, the object of the opinion, the reader's reaction, the reader, other opinions on the same topic, etc. - one can realize some relationships among them. In order

[1] http://store.steampowered.com

to verify the relevance of these relations for the task of determining the helpfulness of an opinion, in the next sections, we discuss the use of networks to model this task.

## 3. Helpfulness Prediction

### 3.1. Corpus

In our previous work [de Sousa et al. 2019], we used a small corpus collected from Google Play Store with $2,000$ reviews from $10$ apps of the communication category (see Table 1). But, recently a new corpus has been made available. Therefore, to evaluate our modeling, we used the UTLCorpus corpus presented in [Sousa et al. 2019].

Table 1. Number of applications and comments extracted in previous work

| App | # Comments | App | # Comments |
|---|---|---|---|
| Facebook | 200 | Skype | 200 |
| Google Allo | 200 | Snapchat | 200 |
| Hangouts | 200 | Telegram | 200 |
| Mensagens | 200 | Viber | 200 |
| Messenger | 200 | WhatsApp | 200 |
| **Total** | 2,000 | | |

The UTLCorpus contains $2,881,589$ reviews ($1,839,851$ of movies and $1,041,738$ of apps). For this work, we use only the apps domain. The creators of the UTLCorpus collected the apps reviews by crawling the Google Play Store[2]. They gathered reviews from $243$ apps and after the removal of the irrelevant ones, the final corpus was left with $921,257$ reviews. Table 2 presents some detailed information about UTLCorpus and a comparison between our previous corpus[de Sousa et al. 2019] and the UTLCorpus. That table shows the number of reviews for each class with the applied method. First, the reviews are grouped by their categories. After grouping, we consider only the votes of the comments. Next, we sort the votes in descending order. Then, we consider only the number of votes greater than one. Finally, all reviews with more votes than the first percentile of the distribution are considered helpful. On the other hand, all comments with fewer votes than the previous threshold and which were collected at least five days after their publication are considered non-helpful.

It is important to highlight that the UTLCorpus have apps of several categories which is different from the approach of our previous work [de Sousa et al. 2019]. In that past work, one assumption was that using reviews from the same category would help the prediction task because are believed to have similar terms and/or topics [Anchiêta and Moura 2017]. However, it is important to evaluate our approach with reviews from many different categories. This is one reason for the use of UTLCorpus, besides its size.

For the best of our knowledge, there is only more one available corpus in Brazilian Portuguese that contains information about helpfulness [Hartmann et al. 2014], however, the domain is different from that proposed here.

[2]https://play.google.com/store

Table 2. Comparison between the corpora

| | **Previous Corpus** | **UTLCorpus** |
|---|---|---|
| **# documents** | $2,000$ | $921,257$ |
| **# types** | $6,421$ | $419,713$ |
| **# tokens** | $33,749$ | $11,919,636$ |
| **# Helpful Reviews** | $800$ (40%) | $50,166$ (5%) |
| **# Not Helpful Reviews** | $800$ (40%) | $871,091$ (95%) |

A comment on Google Play contains the comment text itself, the comment's author, the number of stars, the comment's date and the number of likes the comment received (see Figure 1). The number of stars and the number of likes are in the range of $[0, 5]$ and $[0, +\infty)$, respectively. The source code and the previous dataset are still available upon request by email.

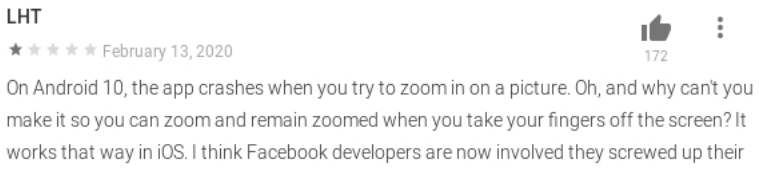



Figure 1. Example of a Google Play comment

The helpfulness votes are often used to define which comments are helpful or not, by calculating the helpfulness score ($h \ \epsilon \ [0, 1]$) using the Equation 1. According to [Diaz and Ng 2018], the task mainly include score regression, binary review classification, and review ranking. For binary classification, a threshold could be applied to helpfulness score to split the reviews on helpful or not.

$$h = \frac{helpful\ votes}{helpful\ votes + unhelpful\ votes} \tag{1}$$

$$utility_{c_i} = \frac{votes_{c_i}}{\sum_{c_j \epsilon C} c_j} \tag{2}$$

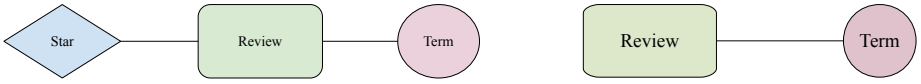
In the Google Play Store, the reviews do not have the "thumbs down" values. For that reason, we had to do an adaptation. Our first idea was applied on a previous paper [de Sousa et al. 2019]. We use the Equation 2 to assign a utility score for each review, where the utility score of the review $i$ is equal to the number of votes of the review $i$ divided by the sum of all votes from reviews ($\sum_{c_j \epsilon C} c_j$ where $C$ is the set of reviews excluding the review $i$). After calculating the score, we sorted the reviews in descending order, considering the first 40% reviews as helpful, and the last 40% reviews as not helpful. The remaining 20% (middle of the list) are disregarded, as they may be noisy, presenting overlap between labels. Therefore, from 2,000 reviews, it has left 1,600

reviews (800 helpful and 800 not helpful). This approach works well for small number of reviews. However, this approach does not work on UTLCorpus, as there are many reviews with no helpfulness votes, and few reviews with many votes. Thus, it is not interesting to split the data set into two equal parts. Fortunately, the authors of the UTLCorpus presented a solution for this issue.

### 3.2. Proposed modeling

The helpfulness prediction task has been seen as a regression, classification or ranking problem, modeled as an attribute-value table. However, in this work, this task was modeled as a heterogeneous network. We propose 4 modeling variations. In Figure 2, the scheme of the modeled networks is shown, where `Star`, `Review`[3], and `Term` are the network nodes. The first model is identical with the depicted in Figure 2a and it has a variation considering weights in edges between `Review` and `Term` nodes. Another model is identical with 2b and it has a weighted variation.

Figure 2. Scheme of heterogeneous network for helpfulness prediction




(a) Star-Reviews-Term network (b) Reviews-Term network

In order to instantiate the heterogeneous network, we developed a methodology organized in 4 steps from the extracted comments. In Figure 3, an overview of the methodology is presented.

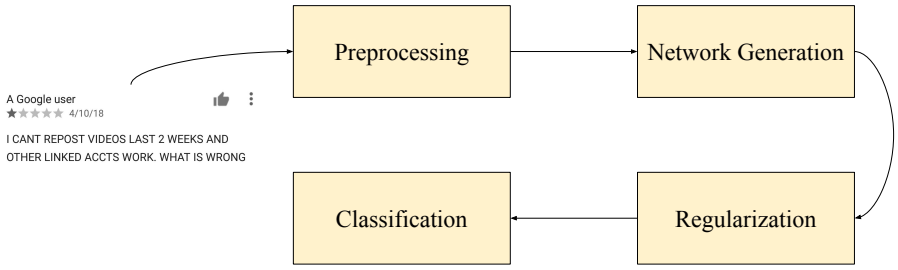



Figure 3. Methodology to instantiate the heterogeneous network

In the first step, the comments were pre-processed (subsection 3.3). Then, the heterogeneous network was generated (subsection 3.4). After that, a regularization algorithm was applied over the network (subsection 3.5). Finally, the comments were classified into helpful or not helpful (subsection 3.6).

It is important to point out that the approach as a whole is language independent. The only step that makes use of linguistic components is pre-processing, but it is possible to replace the linguistic resources so that the whole process is applied to another language.

[3]In this paper, we used the words review and comment on the interchangeable way.

### 3.3. Pre-processing

At this stage, the users' comments were pre-processed in order to be normalized, the most relevant terms of the comments were selected, and the number of stars attributed to each comment were extracted.

Google Play comments are similar to tweets: they are usually short and do not follow grammar and punctuation rules. Thus, a textual normalizer developed for Portuguese [Bertaglia and Nunes 2016] was applied to each comment of the corpus to minimize the amount of textual noise that could hinder further processing steps. These noises include misspellings, abbreviations, internet slangs, repetitions, etc. NLPnet's Part-Of-Speech Tagger [Fonseca and Rosa 2013] was then used to extract open-class words such as noun, verb, adjective, and adverb. Finally, to obtain the stems of each comment word, a stemming algorithm was applied [Orengo and Huyck 2001].

### 3.4. Network generation

The process of generating a network is simple. After the pre-processing step, we created nodes of the types `star`, `review`, and `term`. We present the guidelines to create each network-type below.

- Network Type Star-Review-Term (SRT): The `review` and `star` nodes are linked to each other by an edge according to the number of stars of the review. We adopted this same strategy to link the `review` and `term` nodes, whenever the term is present in the review. The `star` nodes neither can be linked to each other nor to the `term` node. In the same way, the `term` node can not be linked to each other. The network is undirected and in the unweighted variation (USRT) there is no edges with weights. In Figure 4, we illustrated a small instance of our network. But, in the weighted variation (WSRT), there are weights between reviews and terms according to the number of words on that review. We modeled the network in this manner, as we believe that there is a relation between helpfulness and the number of stars of a review. Thus, we also believe that there is a relation between terms, reviews, and stars.
- Network Type Review-Term (RT): In this model, there are no `stars` nodes. The `review` and `term` nodes are linked to each other whenever the term is present in the review. There are no auto-loops. Similarly to the USRT variation, the network is undirected and unweighted (URT). And there is a weighted variation (WRT) where there are weights between revisions and terms according to the number of words in that revision. In Figure 5, we illustrated a small instance of our network. We modeled this variation to evaluate if only `review` and `term` nodes are sufficient to describe the problem satisfactorily.

### 3.5. Regularization

The network generation considers the similarity information among elements of the network, for example, using a Bag-of-Words (BoW) representation and calculating the similarity of documents by the cosine similarity metric. However, it is not always possible to create a network this way. Moreover, modeling a network in this manner is unnatural.

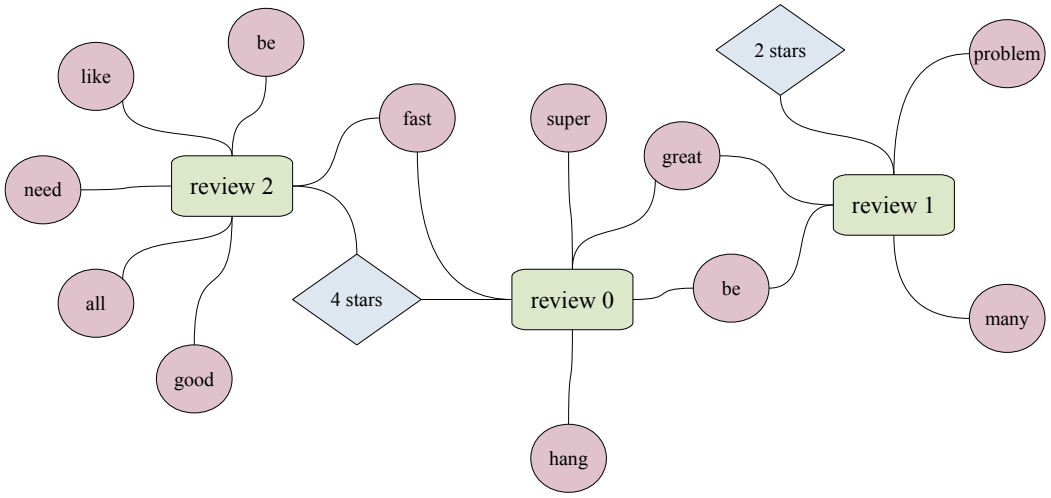



Figure 4. Network SRT instance example

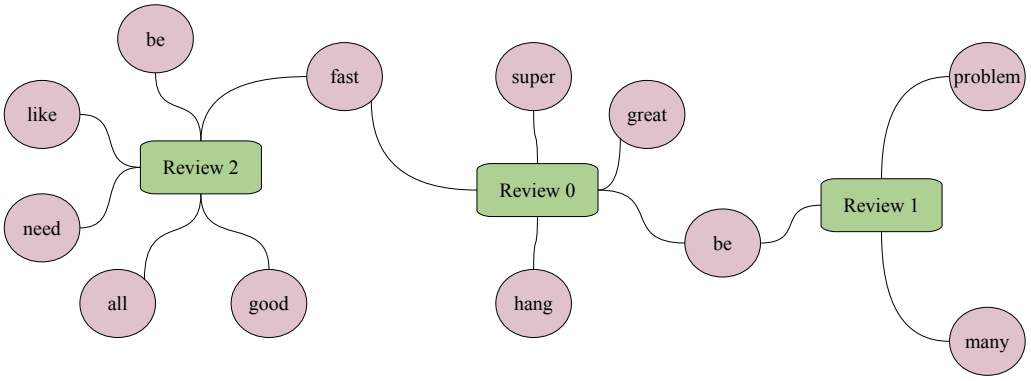



Figure 5. Network RT instance example

More than that, it is difficult to consider and aggregate domain information in the similarity measure between nodes. Hence, in these cases, the methods do not achieve good results.

In this paper, we aim to show an example where that difficulty may be overcome. In previous experiments, the similarity between document vectors did not achieve good results. Thus, we created an heterogeneous network, considering the relationship information among existing elements in the reviews. Using different type nodes increases network informativeness, but it difficulties the use of some learning methods direct in the network.

In this context, we applied regularization methods to perform the features extraction about network objects classes. Regularization is a kind of transductive classification method [Rossi 2016]. It aims to find a set of labels that follows two conditions: (i) the method needs to be consistent with the set of labels manually annotated and; (ii) it needs to be consistent with the network topology, i.e., to consider that nearest neighbors tend to have the same labels.

We adopted three methods for regularization: Gaussian Fields and Harmonic Function (GFHF) [Zhu et al. 2003], Learning with Local and Global Consistence (LLGC) [Zhou et al. 2004], and GnetMine [Ji et al. 2010]. These three methods have slight differences in their equations. For example, GFHF seeks to minimize the function in the Equation 3. For assigning a label into an object, the method computes the weighted average of its neighbors' label information by the weights of the links between objects, as presented in Equation 4, with the following terms:

- $O$ is the set of nodes in the network.
- $O^L$ is the set of pre-annotated nodes in the network.
- $F$ e $f$ is the regularization output. They represent a vector with the relative coordinates of a review in the plane (see Table 3).
- $w$ is the edge weight between nodes $o_i$ and $o_j$.
- $y$ is the information vector for the pre-annotated nodes.

$$Q(F) = \frac{1}{2} \sum_{o_i, o_j \epsilon\, O} w_{o_i,o_j}(f_{o_i} - f_{o_j})^2 + \lim_{\mu \to \infty} \mu \sum_{o_i \epsilon O^L} (f_{o_i} - y_{o_i})^2 \tag{3}$$

$$f_{o_i} = \frac{\sum_{o_j \epsilon O} w_{o_i,o_j} f_{o_j}}{\sum_{o_j \epsilon O} w_{o_i,o_j}} \tag{4}$$

The regularization algorithms requires some nodes to be manually pre-labeled with specific classes. The main difference between GFHF and LLGC is that GFHF does not modify these pre-labeled nodes, unlike the LLGC that performs such modifications. (in Section 4, we detail the labeling process). The GnetMine, in addition to modify the values of pre-labeled nodes, it is an algorithm for heterogeneous networks, so it considers the different types of nodes. The regularization algorithms produce values related to coordinates for each object in the network, as shown in Table 3. These values may be used for several supervised machine learning methods to learn and predict labels [Bui et al. 2018]. In subsection 3.6, we detail the used machine learning algorithms.

Table 3. Example of the regularization algorithm output

| Id | Coordinate 1 | Coordinate 2 |
|---|---|---|
| 0 | 0.13884248 | 0.11291029 |
| 1 | 0.13011554 | 0.12082376 |
| 2 | 0.12334355 | 0.13454545 |
| 3 | 0.12324345 | 0.12324455 |
| ... | ... | ... |

### 3.6. Classification

In this step, we classified each review as helpful or not helpful from the extracted features by regularization algorithms. That is, from the generated coordinates for each review for helpful and not helpful label, we handle the problem of helpfulness prediction as a supervised machine learning problem.

We evaluate several classifies available by the Scikit-Learn library [Pedregosa et al. 2011], such as Support Vector Machine (SVM), Naïve Bayes, C4.5, and Multi-layer Perceptron (MLP).

In what follows, we detail the performed experiments and obtained results. Besides, we compare our approach with a well-known baseline.

## 4. Experiments and Results

We perform an experiment to evaluate the impact of several possible settings of our approach. The variables of the experiment are: amount of pre-annotate nodes to regularization, type of regularization algorithm, and type of the modeling network.

To start the experiment, first, we balanced the data. For that, we applied a downsampling method that randomly assigns to the corpus the same number of helpful and not-helpful reviews, removing samples of the majority class. Next, with the balanced corpus, we explored the regularization methods, aiming to get the features of the corpus. Finally, we used four supervised classifiers with the extracted features to identify if an opinion is helpful or not. Furthermore, we compared our approach against two methods. The first one adopts an approach based on fuzzy logic [de Sousa et al. 2015], while the second one is based on NN [Santos et al. 2016]. This second method is the evolution of first one. Both methods classify reviews into four classes: insufficient, sufficient, good, and excellent. We adapted the baseline methods to consider only two classes: excellent and good as helpful, and insufficient and sufficient as not helpful. This modification was necessary since our method use only two classes: helpful and not helpful.

We adopted the works of [de Sousa et al. 2015] and [Santos et al. 2016] for comparison because they are open source and requires only minor modifications for re-use, and are domain-independent. The main difference between our current approach and previous approaches is that the latter do not consider the relationships among the opinion elements. Moreover, the previous approaches did not focus on binary opinion helpfulness classification but on multi-class classification.

In addition to the works of [de Sousa et al. 2015, Santos et al. 2016], we developed two other baseline methods to compare with our approach. In the first one (Baseline 1), we used the Bag-of-Words as features, while in the second (Baseline 2), we used three well-known features: average sentence length, number of tokens, and number of sentences. We applied both baselines into the Naïve Bayes classifier.

Since the regularization algorithms we used requires a portion of pre-annotated reviews, we followed a strategy for progressively annotate the reviews. For example, we manually annotated from 0.5% to 5.0% of the reviews as helpful and not helpful. We randomly chose these proportions. In this way, we analyzed several supervised algorithms, such as SVM, Naïve Bayes, C4.5, and MLP. To evaluate these algorithms, we applied the $k$-fold cross-validation technique with $k = 10$. Tables 4, 5, 6, and 7 show the results for each pre-annotated portion of reviews and each algorithm. We repeat these steps for each regularization algorithm.

From tables results, we can see that the MLP algorithm reached the better results

Table 4. Results for the classification algorithms for network type Unweighted SRT

| Pre-labeled comment (%) | GFHF | | | | LLGC | | | | GNETMINE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** |
| **0.5** | 0.5017 | 0.5130 | 0.5352 | 0.5500 | 0.5023 | 0.7318 | 0.6593 | 0.5005 | 0.5019 | 0.5089 | 0.5315 | 0.5023 |
| **1.0** | 0.5037 | 0.5087 | 0.5444 | 0.5362 | 0.5049 | 0.7405 | 0.6605 | 0.5003 | 0.5033 | 0.5163 | 0.5403 | 0.5047 |
| **1.5** | 0.5071 | 0.5077 | 0.5530 | 0.5579 | 0.5073 | 0.7462 | 0.6616 | 0.4995 | 0.5059 | 0.5269 | 0.5471 | 0.5070 |
| **2.5** | 0.5090 | 0.5129 | 0.5520 | 0.5831 | 0.5099 | 0.7503 | 0.6715 | 0.4986 | 0.5090 | 0.6162 | 0.5745 | 0.5099 |
| **3.0** | 0.5105 | 0.5139 | 0.5639 | 0.6265 | 0.5123 | 0.7493 | 0.6663 | 0.4997 | 0.5109 | 0.5456 | 0.5963 | 0.5121 |
| **3.5** | 0.5127 | 0.5177 | 0.5708 | 0.6159 | 0.5148 | 0.7515 | 0.6642 | 0.5013 | 0.5134 | 0.5271 | 0.6162 | 0.5147 |
| **4.0** | 0.5187 | 0.5195 | 0.5612 | 0.6238 | 0.5173 | 0.7582 | 0.6767 | 0.7646 | 0.5155 | 0.5525 | 0.6182 | 0.5176 |
| **4.5** | 0.5328 | 0.5221 | 0.5810 | 0.6351 | 0.5198 | 0.7591 | 0.6774 | 0.7664 | 0.5179 | 0.6238 | 0.6525 | 0.5273 |
| **5.0** | 0.5468 | 0.5250 | 0.5841 | 0.6253 | 0.5024 | 0.7602 | 0.6679 | **0.7723** | 0.5204 | 0.5524 | 0.6403 | 0.5245 |

Table 5. Results for the classification algorithms for network type Weighted SRT

| Pre-labeled comment (%) | GFHF | | | | LLGC | | | | GNETMINE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** |
| **0.5** | 0.5015 | 0.5178 | 0.5329 | 0.5061 | 0.5023 | 0.7379 | 0.6539 | 0.5001 | 0.5018 | 0.4843 | 0.5109 | 0.5023 |
| **1.0** | 0.5034 | 0.5114 | 0.5383 | 0.5306 | 0.5049 | 0.7396 | 0.6589 | 0.5010 | 0.5037 | 0.5868 | 0.5392 | 0.5048 |
| **1.5** | 0.5059 | 0.5118 | 0.5499 | 0.5574 | 0.5072 | 0.7474 | 0.6525 | 0.4997 | 0.5059 | 0.5346 | 0.5486 | 0.5105 |
| **2.0** | 0.5078 | 0.5130 | 0.5496 | 0.5929 | 0.5099 | 0.7500 | 0.6602 | 0.5000 | 0.5078 | 0.5276 | 0.5874 | 0.5221 |
| **2.5** | 0.5123 | 0.5153 | 0.5589 | 0.5982 | 0.5123 | 0.7522 | 0.6711 | 0.5010 | 0.5109 | 0.5742 | 0.6153 | 0.5223 |
| **3.0** | 0.5126 | 0.5164 | 0.5539 | 0.6001 | 0.5149 | 0.7585 | 0.6632 | 0.6046 | 0.5122 | 0.5724 | 0.6089 | 0.5149 |
| **3.5** | 0.5156 | 0.5173 | 0.5564 | 0.6102 | 0.5174 | 0.7552 | 0.6683 | 0.7628 | 0.5260 | 0.5703 | 0.6366 | 0.5183 |
| **4.0** | 0.5348 | 0.5211 | 0.5639 | 0.6323 | 0.5198 | 0.7548 | 0.6663 | 0.7592 | 0.5179 | 0.5786 | 0.6274 | 0.5301 |
| **4.5** | 0.5188 | 0.5237 | 0.5582 | 0.6091 | 0.5224 | 0.7587 | 0.6744 | 0.7651 | 0.5209 | 0.5857 | 0.6315 | 0.5354 |
| **5.0** | 0.5381 | 0.5258 | 0.5594 | 0.6251 | 0.5248 | 0.7631 | 0.6683 | **0.7647** | 0.5220 | 0.5628 | 0.6396 | 0.5367 |

Table 6. Results for the classification algorithms for network type Unweighted RT

| Pre-labeled comment (%) | GFHF | | | | LLGC | | | | GNETMINE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** |
| **0.5** | 0.5017 | 0.5110 | 0.5355 | 0.5147 | 0.5023 | 0.7318 | 0.6595 | 0.5004 | 0.5016 | 0.5423 | 0.5222 | 0.5023 |
| **1.0** | 0.5036 | 0.5109 | 0.5483 | 0.5498 | 0.5049 | 0.7411 | 0.6604 | 0.4998 | 0.5039 | 0.5813 | 0.5378 | 0.5046 |
| **1.5** | 0.5055 | 0.5147 | 0.5524 | 0.5461 | 0.5073 | 0.7285 | 0.6557 | 0.5002 | 0.5073 | 0.5866 | 0.5485 | 0.5073 |
| **2.0** | 0.5165 | 0.5179 | 0.5540 | 0.5699 | 0.5008 | 0.7138 | 0.6625 | 0.4989 | 0.5083 | 0.5883 | 0.5607 | 0.5092 |
| **2.5** | 0.5228 | 0.5183 | 0.5662 | 0.5766 | 0.5122 | 0.7180 | 0.6665 | 0.5012 | 0.5109 | 0.6144 | 0.5847 | 0.5153 |
| **3.0** | 0.5177 | 0.5213 | 0.5415 | 0.5822 | 0.5149 | 0.7323 | 0.6730 | 0.5517 | 0.5153 | 0.5837 | 0.6309 | 0.5153 |
| **3.5** | 0.5194 | 0.5213 | 0.5635 | 0.5833 | 0.5174 | 0.7452 | 0.6668 | 0.7632 | 0.5154 | 0.6089 | 0.6246 | 0.5241 |
| **4.0** | 0.5195 | 0.5234 | 0.5565 | 0.5779 | 0.5197 | 0.7370 | 0.6771 | 0.7645 | 0.5174 | 0.6061 | 0.6337 | 0.5313 |
| **4.5** | 0.5254 | 0.5259 | 0.5673 | 0.6119 | 0.5224 | 0.7501 | 0.6705 | 0.7673 | 0.5211 | 0.6149 | 0.6518 | 0.5226 |
| **5.0** | 0.5345 | 0.5293 | 0.5688 | 0.6173 | 0.5247 | 0.7466 | 0.6809 | **0.7685** | 0.5227 | 0.6028 | 0.6366 | 0.5271 |

Table 7. Results for the classification algorithms for network type Weighted RT

| Pre-labeled comment (%) | GFHF | | | | LLGC | | | | GNETMINE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** | **SVM** | **Naïve Bayes** | **C4.5** | **MLP** |
| **0.5** | 0.5010 | 0.5073 | 0.5395 | 0.5132 | 0.5023 | 0.7308 | 0.6600 | 0.5011 | 0.5014 | 0.5331 | 0.5336 | 0.5023 |
| **1.0** | 0.5034 | 0.5117 | 0.5432 | 0.5449 | 0.5048 | 0.6869 | 0.6532 | 0.5022 | 0.5039 | 0.5756 | 0.5399 | 0.5049 |
| **1.5** | 0.5073 | 0.5139 | 0.5451 | 0.5331 | 0.5073 | 0.7461 | 0.6619 | 0.5000 | 0.5057 | 0.5780 | 0.5570 | 0.5078 |
| **2.0** | 0.5082 | 0.5209 | 0.5526 | 0.5406 | 0.5099 | 0.7423 | 0.6604 | 0.5016 | 0.5086 | 0.5836 | 0.5636 | 0.5115 |
| **2.5** | 0.5286 | 0.5210 | 0.5521 | 0.5932 | 0.5122 | 0.7249 | 0.6708 | 0.5008 | 0.5109 | 0.5941 | 0.6089 | 0.5128 |
| **3.0** | 0.5176 | 0.5196 | 0.5574 | 0.5633 | 0.5149 | 0.7540 | 0.6661 | 0.7575 | 0.5125 | 0.5871 | 0.6137 | 0.5230 |
| **3.5** | 0.5205 | 0.5227 | 0.5631 | 0.5951 | 0.5173 | 0.7359 | 0.6737 | 0.7599 | 0.5139 | 0.5988 | 0.6441 | 0.5336 |
| **4.0** | 0.5237 | 0.5253 | 0.5734 | 0.6065 | 0.5196 | 0.7359 | 0.6749 | 0.7636 | 0.5179 | 0.6130 | 0.6381 | 0.5319 |
| **4.5** | 0.5260 | 0.5276 | 0.5783 | 0.6053 | 0.5224 | 0.7578 | 0.6795 | 0.7634 | 0.5208 | 0.6142 | 0.6368 | 0.5234 |
| **5.0** | 0.5285 | 0.5285 | 0.5689 | 0.6025 | 0.5246 | 0.7502 | 0.6755 | **0.7652** | 0.5221 | 0.6053 | 0.6422 | 0.5316 |

in all regularization algorithms and variations of networks, achieving the best result with 5.0% of pre-annotated reviews. The LLGC regularization achieved the best result in all

networks variations. It is important to say that we performed experiments with more pre-annotated reviews, but this did not improve the results, the tables results show that fact, the improving of results after 4.0% of pre-annotated reviews are low. An important observation is that all variations of networks achieved close results, the difference is really small, mainly for the best ones.

We also carried out experimentation using a holdout approach, splitting the corpus in 80% to train and 20% to test. In Table 8, we present the results of the MLP for each class, showing precision, recall, and f-measure. One may see that the MLP produced closely results for each class.

Finally, we compared our approach with the developed baselines, as shown in Table 9. From this table, we compared our network with MLP against the baselines. We can see that our method outperformed the best baseline in 0.5 of f-score.

Table 8. Results of the MLP for LLGC regularization algorithm and USRT network type (Holdout)

| Label | Precision | Recall | F1 |
|---|---|---|---|
| Helpful | 0.81 | 0.62 | 0.70 |
| Not helpful | 0.69 | 0.85 | 0.76 |

Table 9. Comparison among approaches (cross-validation)

| Approach | F1 |
|---|---|
| Baseline 1 | 0.72 |
| Baseline 2 | 0.63 |
| Fuzzy | 0.60 |
| RNA | 0.58 |
| Our network | **0.77** |

These evaluations show the power of the network-based approach to model and predict the helpfulness of opinions. However, it is necessary to investigate other topologies, specific metrics, and adapt other features used in other languages.

Also it is important to highlight that we performed a test of significance in order to verify if our results are statistically significant and different from the baselines. We found out a p-value $< 0.05$, indicating that the results are statistically different and significant with 95% confidence.

## 5. Conclusion and Future Work

Most papers in the literature model the opinions helpfulness prediction task as an attribute-value table. In this way, the relationship information among objects did not consider. Our hypothesis is that, it is possible to improve the results of the helpfulness prediction using relationship information among opinion elements. In this paper, we modeled the opinions helpfulness prediction task as a heterogeneous network. From this network, we applied three regularization algorithms for feature extraction from user reviews. At last, we evaluated several supervised machine learning algorithms on the extracted features. We compared our approach with four baselines methods. For that, we performed an experiment on $100,332$ reviews about apps from Google Play belong to the UTLCorpus. The results showed that our network outperformed the baselines.

Despite the good results, it is important to say that our approach is network-modeling dependent, however, to use neural-graph machines allow to classifier a larger

dataset with few annotated data. The close results among the regularization methods, indicating that it is possible to adopt any regularization method for the classification.

The main contribution of this work is the modeling of a heterogeneous network for the opinions helpfulness prediction task. The informativity power of a heterogeneous network is a great differential in relation to based attribute-value approaches. Although our approach is relatively new, the obtained results showed their benefits and potentialities. However, as this work is a pioneer in considering the helpfulness task for Portuguese, a comparison with other works becomes difficult. However, as future work, we intend to adapt methods from other languages to Portuguese, aiming to compare them with our approach.

For future work, besides to apply the approach on multi-domains, we intend to follow two research lines. In the first, we will explore linguistically motivated approaches, that is, we will evaluate linguistic features that indicate if a review is helpful or not. In the second, we will adapt works of other languages, aiming to investigate language independent-features.

In addition to these two directions, we will explore other network topologies in order to achieve better results. For example, topologies with many layers (a common layer and a layer with semantic information: synonyms, named entities, sentiment words, etc.). Furthermore, these topologies may allow the use of several network metrics, such as hubs, betweenness, and closeness, among others. And, finally, to explore the use of Deep Neural Networks applied to regularization extracted features.

## 6. Acknowledges

The authors are grateful to the IFPI for supporting this work.

## References

Anchiêta, R., Sousa, R. F., Moura, R., and Pardo, T. (2017). Improving opinion summarization by assessing sentence importance in on-line reviews. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 32–36.

Anchiêta, R. T. and Moura, R. S. (2017). Exploring unsupervised learning towards extractive summarization of user reviews. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, pages 217–220. ACM.

Barbosa, J. L. and Moura, R. S. (2016). Avaliaç ao automática da utilidade de reviews usando redes neurais artificiais no corpus do steam. In *Anais do XXVI Congresso da Sociedade Brasileira de Computação: BraSNAM - 5º Brazilian Workshop on Social Network Analysis and Mining*. Brazilian Computer Society.

Bertaglia, T. F. C. and Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120.

Bui, T. D., Ravi, S., and Ramavajjala, V. (2018). Neural graph learning: Training neural networks using graphs. In *Proceedings of 11th ACM International Conference on Web Search and Data Mining (WSDM)*.

de Sousa, R., Anchieta, R., and Nunes, M. (2019). Um método baseado em grafos para predição da utilidade de opiniões sobre produtos. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 95–106, Porto Alegre, RS, Brasil. SBC.

de Sousa, R. F., Rabêlo, R. A., and Moura, R. S. (2015). A fuzzy system-based approach to estimate the importance of online customer reviews. In *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*, pages 1–8. IEEE.

Diaz, G. O. and Ng, V. (2018). Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 698–708.

Fonseca, E. R. and Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian symposium in information and human language technology*, pages 98–107.

Hartmann, N. S., Avanço, L. V., Balage Filho, P. P., Duran, M. S., Nunes, M. D. G. V., Pardo, T. A. S., Aluisio, S. M., et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *International Conference on Language Resources and Evaluation*. European Language Resources Association-ELRA.

Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics.

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Malik, M. and Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, 73:290–302.

Martins, A. C. S. and Tacla, C. A. (2015). Assessement of features influencing the voting for opinions' helpfulness about services in portuguese. In *Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective-Volume 1*, page 21. Brazilian Computer Society.

Orengo, V. and Huyck, C. (2001). A stemming algorithmm for the portuguese language. In *String Processing and Information Retrieval*, pages 186–193.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rossi, R. G. (2016). *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. PhD thesis, Universidade de São Paulo.

Santos, R. L. d. S., de Sousa, R. F., Rabelo, R. A., and Moura, R. S. (2016). An experimental study based on fuzzy systems and artificial neural networks to estimate the importance of reviews about product and services. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 647–653. IEEE.

Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

Semin, G. R. (2011). The linguistic category model. *Handbook of theories of social psychology*, 1:309–326.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., and Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, 70:346–355.

Sousa, R. F., Brum, H. B., and Nunes, M. d. G. V. (2019). A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Proceedings of the 12th Brazilian Symposium in Information and Human Language Technology*, pages 209–218. Sociedade Brasileira de Computação.

Zeng, Y.-C., Ku, T., Wu, S.-H., Chen, L.-P., and Chen, G.-D. (2014). Modeling the helpful opinion mining of online consumer reviews as a classification problem. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014*, 19(2).

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919.

## 6.2 Avaliação de Métodos: Um Benchmark

O trabalho apresentado nesta seção é o seguinte:

*Rogério Figueredo de Sousa e Thiago Alexandre Salgueiro Pardo. 2022. "Evaluating Content Features and Classification Methods for Helpfulness Prediction of Online Reviews: Establishing a Benchmark for Portuguese". In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 204–213*

## Declaração de Contribuição

R.F. Sousa desenvolveu a pesquisa, contribuiu com a elaboração, execução e avaliação da experimentação e colaborou com a escrita do manuscrito. T.A.S. Pardo contribuiu com a elaboração do processo de experimentação, auxiliou na escrita do manuscrito e supervisionou o projeto.

# Evaluating Content Features and Classification Methods for Helpfulness Prediction of Online Reviews: Establishing a Benchmark for Portuguese

**Rogério Figueredo de Sousa, Thiago Alexandre Salgueiro Pardo**
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
Av. Trabalhador São Carlense, 400 – 13.566-590 – São Carlos – SP – Brazil
rogerfig@usp.br, taspardo@icmc.usp.br

## Abstract

Over the years, the review helpfulness prediction task has been the subject of several works, but remains being a challenging issue in Natural Language Processing, as results vary a lot depending on the domain, on the adopted features and on the chosen classification strategy. This paper attempts to evaluate the impact of content features and classification methods for two different domains. In particular, we run our experiments for a low resource language – Portuguese –, trying to establish a benchmark for this language. We show that simple features and classical classification methods are powerful for the task of helpfulness prediction, but are largely outperformed by a convolutional neural network-based solution.

## 1 Introduction

The concern to facilitate users' decision-making is common in most e-commerce platforms. The possibility for customers to publicly provide product reviews is one of the consequences of this concern. This functionality allows future customers to read reviews from other customers and take their buying decision. Despite being useful, the amount of generated data is very large, making it impossible for a human to read them all. Moreover, a large part of this data can be considered unwanted, containing poorly written texts, vague opinions and texts of dubious quality (Kim et al., 2006), making it difficult to find relevant content.

The helpfulness voting functionality that some e-commerce platforms adopt tries to address the above problem, ranking the reviews and showing the most helpful ones to the customers. However, manual voting has some drawbacks, as new helpful reviews take time to get enough votes and gain a visible position. The solution is to automatically predict the helpfulness of reviews.

Despite the usefulness of the task of helpfulness prediction and its practical implications, literature has shown that it is a challenging open issue in Natural Language Processing (NLP). Performance results vary drastically across domains and there are several different features and classification methods in the area, as discussed in (Sousa and Pardo, 2021).

This paper aims to investigate such issues and to identify relevant features and methods for helpfulness prediction. We provide a qualitative and quantitative study of the impact of key content features in two different domains (apps and movies). By content features, we mean those that are related to the information that can be extracted directly from the review, such as the text and the "stars" given by the author. We also perform a comparative study of various classical and deep machine learning classifiers. We show that simple features and classical classification methods may be powerful for the task, but they are largely outperformed by a convolutional neural network-based approach, which reaches a f1-score of $0.90$ for apps and $0.74$ for movies. It is also relevant to cite that we run our experiments for a low resource language – Brazilian Portuguese –, bringing relevant contributions for NLP for Portuguese and establishing a benchmark for the task.

The rest of the paper is organized as follows. Section 2 shows the main related work. In Section 3, we describe the experimental setting adopted in this work. Section 4 reports the achieved results and Section 5 brings some final remarks.

## 2 Related Work

The main research line in review helpfulness prediction aims to predict the helpfulness score for a set of reviews. The helpfulness score is defined as shown in Equation 1 and can be used as the target for regression, binary classification, or ranking. The score regression aims to predict the helpfulness score $h \epsilon [0, 1]$. For binary classification, a threshold is applied in helpfulness score (e.g., $h > 0.5$)



*Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 204 - 213

and all reviews with a helpfulness score above the threshold are classified as helpful; otherwise, they are classified as not helpful. Review ranking seeks to order the reviews by their helpfulness according to a reference ranking.

$$h = \frac{helpful\ votes}{helpful\ votes + unhelpful\ votes} \quad (1)$$

In order to understand the helpfulness of online customer reviews, researches have performed several analyses. It is worth mentioning classical works like the ones of Kim et al. (2006) and Zhang and Varadarajan (2006) that introduce many types of features for helpfulness prediction. Kim et al. (2006) split the features in 5 categories, all considered to be content features: Structural, Lexical, Syntactic, Semantic and Meta-Data Features. They build a model for a regression task and a model for a ranking task using the SVM algorithm. Using a dataset of reviews on two products (MP3 players and Digital Cameras) extracted from Amazon.com, the best results are achieved with the combination of length, unigram and number of stars features. In a similar way, Zhang and Varadarajan (2006) propose three categories of features, also for a dataset extracted from Amazon.com. Their features include Lexical Similarity (Cosine similarity over TF-IDF vectors), Shallow Syntactic Features (Proper nouns, Modal verbs, Interjection, etc.) and Lexical Subjectivity Clues (Subjective adjectives, Subjective nouns, etc.). The authors model two regressors using SVR (Support Vector Regression) and SLR (Simple Linear Regression) techniques, obtaining the best results by combining all the features.

Zeng et al. (2014), in addition to the features already used by Kim et al. (2006), propose the use of Trigrams, Comparison Expressions ("Compare to" or "ADJ + er than"), Degree of detail and Pros and Cons. Using an SVM classifier, the authors address the helpfulness prediction task as a three-class classification: Helpful positive reviews, Helpful negative reviews, and Unhelpful reviews. Furthermore, by running a series of experiments with one less feature each time, they found that the "detail" feature is the most important one, followed by length, number of stars and unigram.

More recently, researchers are using more robust methods for helpfulness prediction. It is the case of Xu et al. (2020), that use BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) along with the features of Star Rating and Product Type. With this combination, the authors model a Neural Network to predict the helpfulness score for reviews extracted from Amazom.com. Wang et al. (2020) also use BERT, but the authors add more features (Number of Words, Number of Sentences, Rating, etc.) than Xu et al. (2020) and compare the BERT-based approach to SVM and CNN models. The neural network-based classifiers achieved similar results to SVM using all features. Wu and Wang (2019) propose the use of syntactic features along with BERT sentence embeddings to helpfulness classification. The work compares some CNN models with BERT and perform an ablation study with all syntactic features. Their results showed high recall but very low precision values. In terms of f1-score, BERT achieved the best results and the main feature was Star Rating.

All these researches have in common the use of content features. The results of methods using handcrafted features were better or very close to state-of-the-art classifiers (using BERT and CNN, for instance). In such setting, this paper aims at further exploring such issues, specially for the context of Portuguese, a low resource language. We present our experiment setting in what follows.

## 3 Experiment Setting

### 3.1 Data Overview

We adopt the dataset of Sousa et al. (2019) that includes reviews written in Portuguese for two very different domains: Movies and Apps. While movie reviews are usually largely subjective and passionate, app reviews tend to be more objective and focus on technical aspects. The dataset (namely UTLCorpus) contains a total of $2,732,538$ reviews ($1,833,691$ for movies and $898,847$ for apps).

Figure 1 presents two examples of reviews extracted from the corpus (from the apps domain). The first is considered not helpful, while the second is helpful. According to the creators of the corpus, the helpfulness status is based on the number of votes the reviews received (0 and 335 helpful votes, respectively) and the posting time (more than 5 days).

As the authors report, each review includes the review text, number of stars given by its author, the number of helpfulness votes, and publication time, among some other information. As shown in Table 1, the UTLCorpus is highly unbalanced. We address the unbalancing problem using an under-

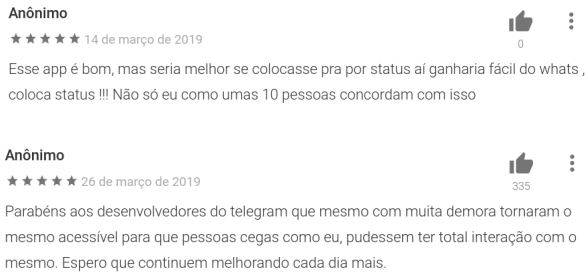



Figure 1: Examples of reviews

sampling approach, randomly removing samples of the majority class. Due to the amount of data, we decided not to carry out the oversampling strategy. Besides the class balancing information, the details of tokens and types in the table show us that the average size of movie reviews is much bigger than that of apps. This difference can make the movies' reviews more challenging than the apps' reviews. Section 4 will further elucidate this assumption.

| | **Movies** | **Apps** |
|---|---|---|
| **# reviews** | 1,833,691 | 898,847 |
| **# movies or apps** | 4,283 | 243 |
| **# types** | 1,828,647 | 419,713 |
| **# tokens** | 60,177,264 | 11,919,636 |
| **Avg. of Tokens p/ doc** | 32.7994 | 12.9384 |
| **Helpfulness Label** | *helpful: 20%* | *helpful: 5%* |

Table 1: UTLCorpus numbers. The helpfulness label refers to the percentage of reviews labeled as helpful.

For our experiments, which we report in the next section, we have randomly split our dataset in three parts: 70% for training, 20% for testing, and 10% for development.

### 3.2 Features

The literature on online review helpfulness explores several features. The researchers often split the features in two big groups: Content and Context features. The content features are related to the information that can be extracted directly from the review, such as the text and the "stars" given by the author. Context features are those extracted from outside the review, such as reviewer information. (Ocampo Diaz and Ng, 2018; Almutairi et al., 2019; Arif et al., 2018). Most of these features are used in domains such as products, books, hotels and so on. We desire to experiment them in apps and movies domains, which are the domains available in the dataset that we adopted in this work and that are remarkably different (which interests us in this paper).

We selected and adapted several content features to the Portuguese language. This process involved finding resources and tools that could support the use of the features in the target language. Table 2 summarizes the implemented features.

We explored the features in machine learning classification solutions. We performed a selection of the best features employing three different strategies. The first method of feature selection is the classical Information Gain (Kozachenko and Leonenko, 1987), which produces values from 0 (no information) to 1 (maximum information) for each feature. The features that contribute with more information are selected to the experiments. The second well-known method for feature selection is using the Random Forest classifier (Breiman, 2001), which is a meta estimator that uses several tree-based classifiers in various subsamples of the dataset to classify the target. Due to its characteristic of using decision trees, it can indicate the importance of features used in the classification process. The third method for feature selection consists in using the correlation values of the features with the helpfulness classes. The previous work of Sousa and Pardo (2021) presents studies of correlation among the feature values and helpfulness status using the correlation coefficients of Pearson and Spearman. Using these correlations, we order the absolute values and select the features with better values.

In addition to the previous features, we also test Term-Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques to generate specific text features and compare the results of the handcrafted features with these two well-known baseline features. It is important to mention that all feature values were normalized for the experimentation process. Table 3 shows an overview of all the features used in this paper.

We comment on the machine learning classifiers and report the achieved results in the next section.

## 4 Results

We explored the following classical classification strategies in this work: Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Neural Network Multilayer Perceptron (NN) and a Dummy Classifier. More sophisticated (deep) strategies that we tested are a BERT-based classifier and a Convolutional Neural

| Feature | Description | Portuguese Resource/Tool |
|---|---|---|
| Average Sentence Length (Avg-SL) | Average sentence size in terms of words (Liu et al., 2007; Lu et al., 2010) | spaCy with portuguese language model |
| Number of Sentences (Num-S) | Total of sentences in the review (Liu et al., 2007; Lu et al., 2010) | |
| Number of Words (Num-W) | Total of words in the review (Kim et al., 2006; Mudambi and Schuff, 2010) | |
| Star Rating (Star-R) | The review-assigned product star rating (Huang et al., 2015) | - |
| Readability Features (READ) | Measure how easy a text is to read and include the following features: Automated Readability Index (ARI), Coleman-Liau Index (CLI), Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning fog index (GFI) and SMOG (Dubay, 2004; Ghose and Ipeirotis, 2011) | Readability features based on (Antunes and Lopes, 2019) |
| Spelling Errors (SPELL) | Number of misspelled words in review (Ghose and Ipeirotis, 2011) | Number of words not found in Wiktionary[1] and Unitex-PB lexicons (Muniz, 2004) |
| Dominant Terms (Dom-Terms) | Presence of important terms in reviews, considering their specificity for the domain (Tsur and Rappoport, 2009) | We use the NILC Corpus (Nunes et al., 1996) to calculate the frequencies of words that do not belong to the domains |
| Product Aspects (Prod-Feat) | Presence of product aspects in the reviews (Kim et al., 2006; Hong et al., 2012; Liu et al., 2007) | We manually extract the features of texts from the corpus development set. |
| Sentiment Words (SENT) | Number of words that express sentiments (Kim et al., 2006) according to the following categories of the LIWC dictionary (Pennebaker et al., 2001): Negate, Swear, Affect, Posemo, Negemo, Anxiety, Anger and Sad | We used a Portuguese version of LIWC dictionary (Balage Filho et al., 2013) |
| Sentiment Divergence (Sent-Div) | Difference between the general sentiment about the movie/app and the sentiment expressed by the author of a review (Hong et al., 2012) | Sentilex sentiment lexicon (Silva et al., 2012) |
| Subjectivity (SUB) | The probability of a review being subjective (Ghose and Ipeirotis, 2011) | |
| Morpho-Syntactic Tokens (SYN) | Number of tokens with the following Part-of-Speech tags: Noun (N), Verb (V), Adverb (ADV) and Adjective (ADJ). It also includes counting for open class words (Open) (Kim et al., 2006) | NLPNet POS-Tagger (Fonseca and Rosa, 2013) |
| Star Deviation (Star-Dev) | Difference between the number of stars in a review and the average star rating for the movie/app (Hong et al., 2012) | - |

Table 2: List of content features

Network (CNN).

### 4.1 Feature Selection

As explained before, we performed feature selection using the techniques of Information Gain and Random Forest. Figures 2a and 2b show the results of feature ranking for the apps domain, while Figure 2c and 2d show the results for movies domain. We performed the classification for the top 8 features of each method of feature selection. As an alternative, we also selected the most correlated features to helpfulness status using the Pearson and Spearman values.

### 4.2 Classification Results

We divided the process of training classifiers in some distinct phases. In the first phase, we trained the classifiers considering the feature selection methods against the TF and TF-IDF techniques. This phase shows us the best sets of features and the best classifiers for both types of features: handcrafted and TF/TF-IDF features. In the second phase, we merged the handcrafted features with the TF/TF-IDF ones. This feature combination process

| Feature category (number of features) | Description |
|---|---|
| Handcrafted Content Features (29) | The content features adapted from previous literature works. |
| Information Gain (8) | The handcrafted content features selected by Information Gain technique. |
| Random Forest (8) | The handcrafted content features selected by Random Forest Classifier. |
| Correlation Coefficients (8) | The handcrafted content features selected by the intersection of correlation coefficients. |
| Baseline TF (500) | The features selected by TF method. |
| Baseline TF-IDF (500) | The features selected by TF-IDF method. |

Table 3: Overview of the features

consists of concatenating the vectors of each text (i.e., TF or TF-IDF vectors) with the vectors of each group of features, both with the same weight. Finally, in the third phase, we decided to use the results of the second phase to model voting-based ensemble classifiers. The classifiers with good results and fewer errors in common were selected to compose the ensembles. The chosen classifiers for the ensembles were Decision Trees and Neural Networks for apps, and Decision Trees and Random Forest for movies. Ensembles with three classifiers obtained similar results (never higher) to those with two classifiers, so we only report the results for ensembles of two classifiers[2]. Finally, in a fourth phase, we used a BERT-based classifier over a pre-trained Portuguese model (Souza et al., 2020) for both domains and a CNN using the GloVe[3] (Hartmann et al., 2017; Pennington et al., 2014) embeddings as input features.

The results referring to the first phase are shown in Figures 3a and 3b, where we show F1 scores (the best ones are written in the chart). Notice that we show in the charts the *F1-Measure* that is the average F1 score for the two classes. One may see that, for apps, the best results were 72%, which may be achieved with simple TF features with SVM and Random Forest; for movies, the best results were 63% for TF-IDF, with the same classifiers. Overall, for both domains, there were no significant performance differences for the two classes.

When we merge the two big groups of features (handcrafted and TF/IDF features), the results are better, as one may see in Figures 3c and 3d. Considering the best situation, apps classification achieved 78% with correlation-based feature selection and TF for SVM (results 8.3% better than before); movies achieved 66% with all the features and TF-IDF for SVM too (4.7% better). Again, SVM showed to be a distinctive technique, with stable classification performances for the two classes.

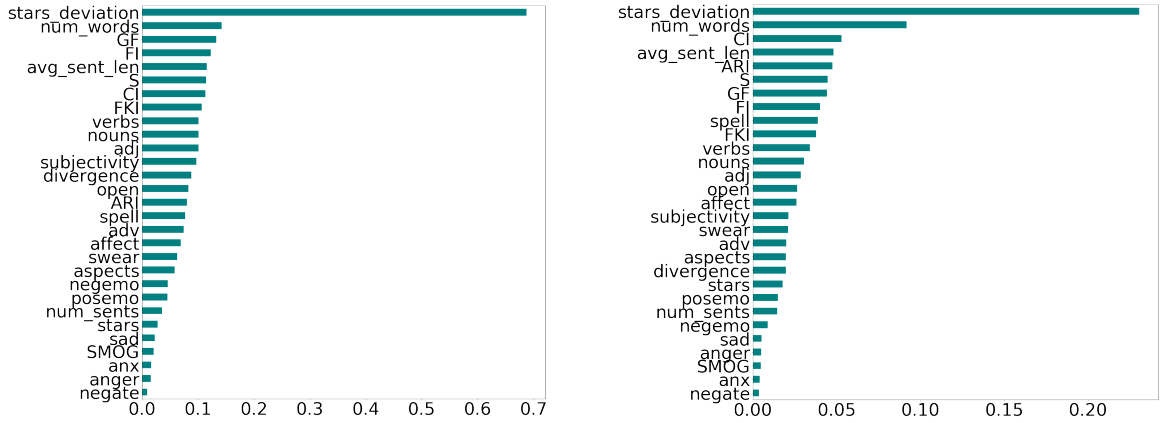
The results for our ensemble, the BERT-based[4] and the CNN classifiers are shown in Figure 3e. For better understanding, the X-axis in Figure 3e mentions the use of the handcrafted features along with BERT (BERT-PT+Hand). For this strategy, we appended all handcrafted features to *CLS* vector ($768 + 29$ $dimensions$), and then the method proceeds normally, using the resulting vector in the next layer to perform the classification. In the same way, for clarification, the strategy BERT-PT+CNN was modeled to merge the BERT architecture to CNN, presented before. We used the four last layers of BERT as features for CNN. The fine-tuning of BERT model was made at the same time as the CNN training. Figure 4 shows the architecture of the CNN.

Despite BERT being a new standard technique in the NLP area, it achieved results very similar to those presented by the ensemble. In the application domain, BERT shows a slight drop in performance. Further investigation is needed to find out why the

[2]We adopted a soft classification, in which the classes are weighted by their probabilities given by the classifiers; if it happens that the two classes end up with the same score, we opt for the not helpful class.

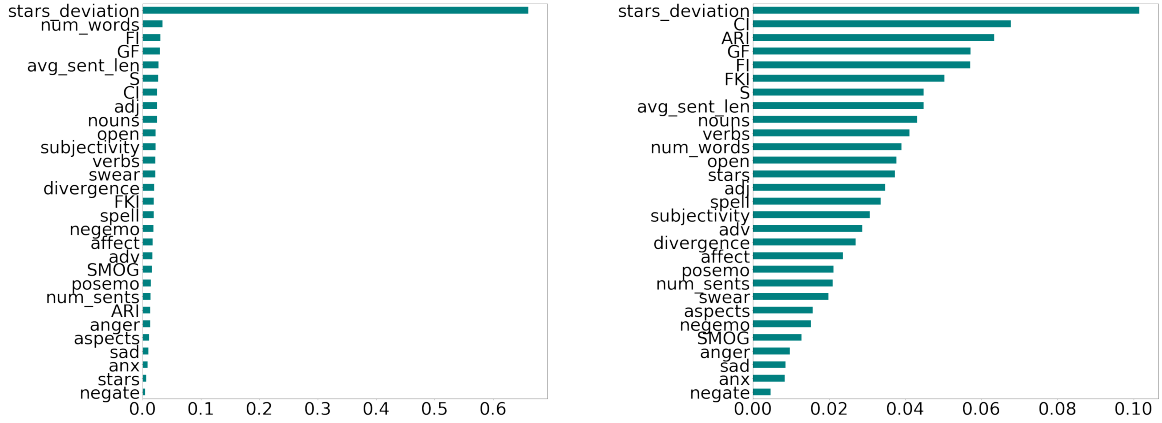
[3]http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

[4]This model was fine-tuned and the pre-trained parameters were not frozen during fine-tuning. The reviews were tokenized using the default tokenizer of Bertimbau model. We applied a single layer feed forward network in CLS output vector (768 dimensions) to classify the instances. The main hyperparameters are as follows: *epochs = 2*, *learning rate = 4e-5*, *optimizer = AdamW*, *train batch size = 8*, *max sequence length = 128*. These hyperparameters were empirically chosen.




(a) Information Gain – Apps

(b) Random Forest – Apps




(c) Information Gain – Movies

(d) Random Forest – Movies

Figure 2: Results of feature importance

results are so low for this case. Possible explanations include the more "passionate" and subjective nature of the movie reviews (while apps' reviews tend to discuss more "technical" aspects). Overall, the ensemble classification could not outperform the previous experiments, while the CNN model outperformed all classifiers.

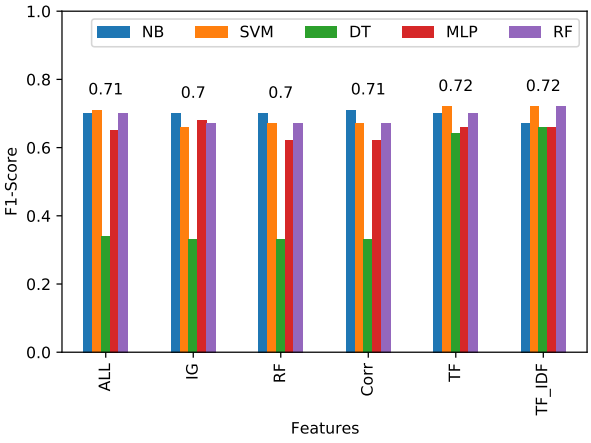
Considering all the experiments, we have some valuable learned lessons. We may see that simple textual features such as TF and TF-IDF may be powerful features for helpfulness prediction. However, merging handcrafted content features with TF-IDF features allows us to achieve better results. Other interesting result is that traditional machine learning techniques may rival more sophisticated strategies as ensemble or BERT-based classifiers. SVM, in special, showed to be an important technique among the classical methods. Anyway, all of them were outperformed by a CNN approach.

Finally, regarding the feature selection processes, the correlation-based one was slightly better than information gain and the Random Forest-based one, but the differences appear to be insignificant. Among the best selected features, although there is some variation depending on the used correlation measure, it is possible to highlight some of them: for apps domain, we highlight average sentence length, star rating and part of speech tags; for movies domain, average sentence length, SMOG readability score, sentiment words and dominant terms.

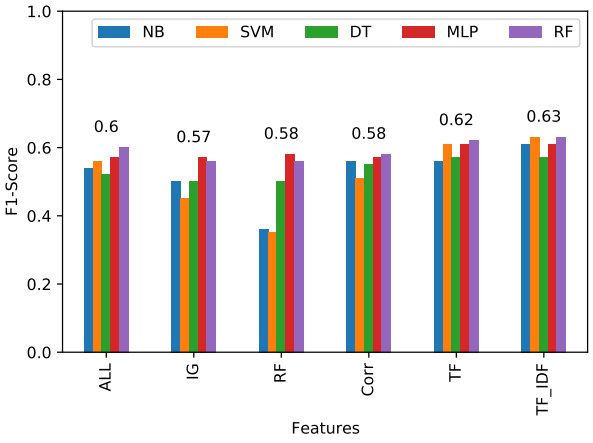
## 5 Final Remarks

This paper synthesized a series of experiments on predicting review helpfulness, showing some relevant learned lessons and contributions (in particular, for Brazilian Portuguese, which is considered a low resource language). However, a lot remains to be investigated. We highlight two issues that concern us the most at this time.

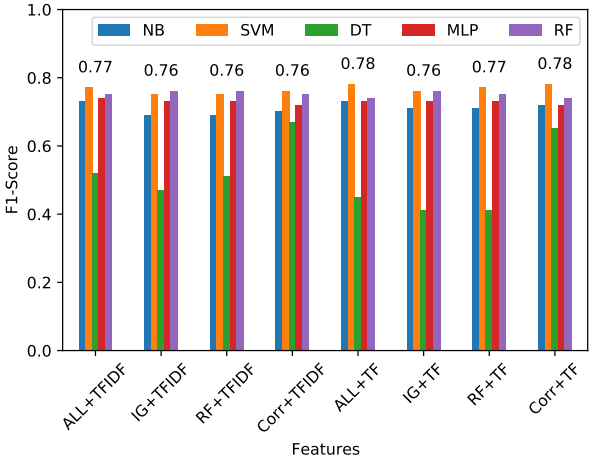
Firstly, the different performances for different domains (across different classification methods) keep intriguing us. This is a known behavior in the sentiment analysis area, and we corroborate it by testing new domains in this paper. We wonder whether new methods or features should be tested,



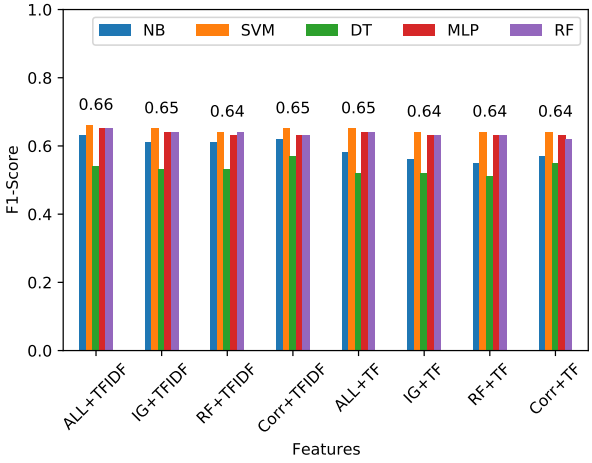

(a) Classification results for apps



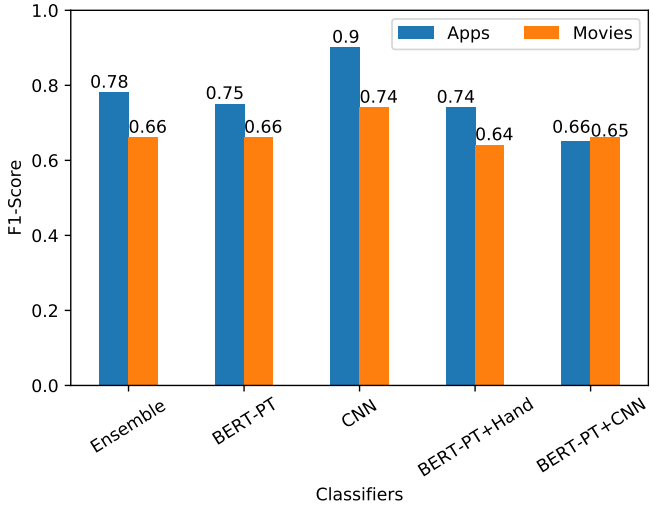

(b) Classification results for movies




(c) Classification results with merged features for apps




(d) Classification results with merged features for movies




(e) Results of ensemble classification and deep models with their combinations

Figure 3: Classification Results

maybe focusing on those that are more domain independent, or whether we should "transform" our data, "eliminating" domain specific traits.

The other issue refers to the helpfulness prediction task itself. Although the literature (including us) have exhaustively tried with this task, it is a highly subjective task that (indirectly) incorporate several other tasks, as subjectivity classifi-

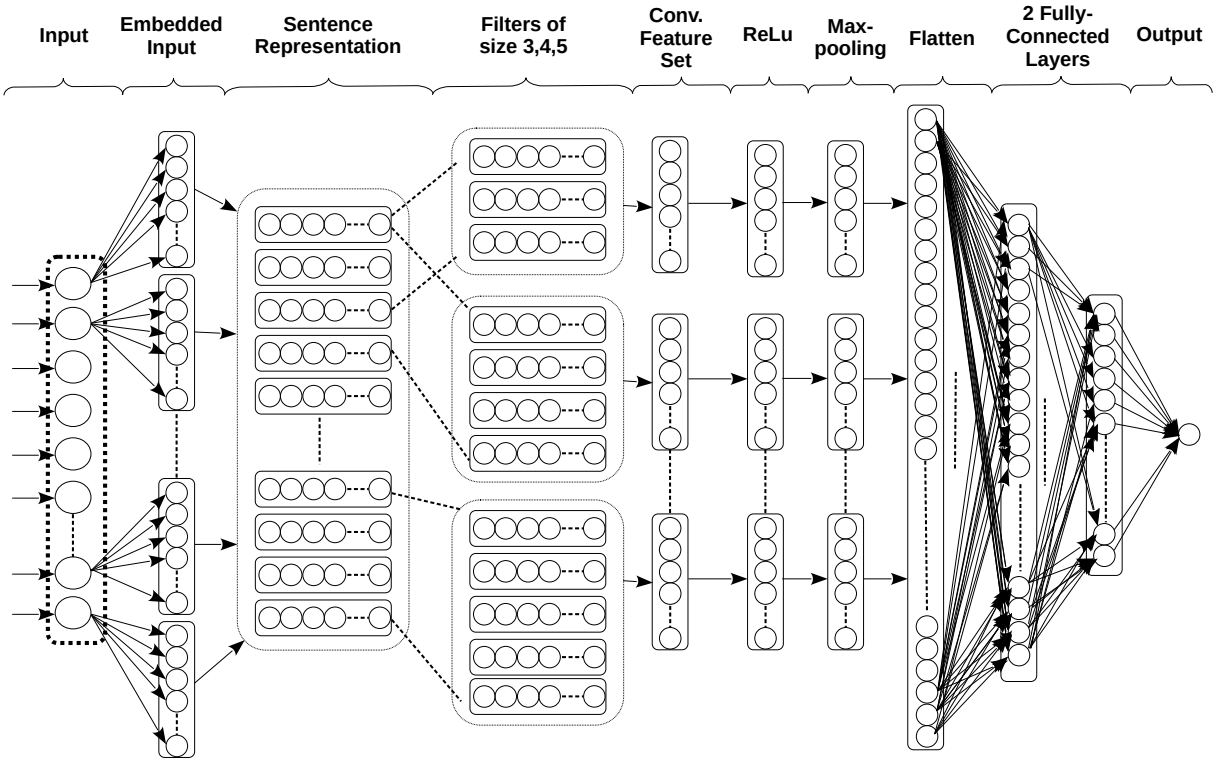



Figure 4: CNN's Architecture. We use 300-dimensional GloVe embeddings as input features. As we can see, we employ three paralels convlayers and set to 100 the size of the output channel for each convlayer. Also, the other parameters are: *epochs* = *5*, *optimizer* = *Adam*, *batch size* = *32*. Fully connected layers: input 1 = 300, output 1 = 32 and Dropout = 0.7

cation (more "personal" reviews look to be more interesting), polarity classification (more "radical" opinions call more attention), aspect identification (as reviews that directly cite some aspects look to be more useful), and detection of user information need (ultimately, a review is helpful only if it attends the information need of the user). Future efforts might explore such supporting tasks for helpfulness prediction.

The complete code for our features and models are available online at https://github.com/RogerFig/deep-helpfulness. The interested reader may also find more information at the POeTiSA project web portal (https://sites.google.com/icmc.usp.br/poetisa).

## Acknowledgments

The authors are grateful to the Center for Artificial Intelligence (C4AI) of the University of São Paulo, sponsored by IBM and FAPESP (grant #2019/07665-4), and to *Instituto Federal do Piauí* (IFPI).

## References

Yasamyian Almutairi, Manal Abdullah, and Dimah Alahmadi. 2019. Review helpfulness prediction: Survey. *Periodicals of Engineering and Natural Sciences*, 7(1):420–432.

Hélder Antunes and Carla Teixeira Lopes. 2019. Analyzing the adequacy of readability indicators to a non-english language. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 149–155. Springer.

Madeha Arif, Usman Qamar, Farhan Hassan Khan, and Saba Bashir. 2018. A survey of customer review helpfulness prediction techniques. In *Proceedings of SAI Intelligent Systems Conference*, pages 215–226. Springer.

Pedro Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Aluísio. 2013. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 215–219.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Dubay. 2004. The principles of readability. *CA*, 92627949:631–3309.

Erick Rocha Fonseca and João Luís G Rosa. 2013. Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian symposium in information and human language technology*, pages 98–107.

Anindya Ghose and Panagiotis G Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Yu Hong, Jun Lu, Jianmin Yao, Qiaoming Zhu, and Guodong Zhou. 2012. 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*, pages 495–504, New York, NY, USA. ACM.

Albert H Huang, Kuanchin Chen, David C Yen, and Trang P Tran. 2015. A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, 48:17–27.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430, Sydney, Australia. Association for Computational Linguistics.

LF Kozachenko and Nikolai N Leonenko. 1987. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16.

Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342, Prague, Czech Republic. Association for Computational Linguistics.

Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700.

Susan M Mudambi and David Schuff. 2010. Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200.

Marcelo Caetano Martins Muniz. 2004. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB*. Ph.D. thesis, Universidade de São Paulo.

Maria das Graças Volpe Nunes, Fabiano M Costa Vieira, Cláudia Zavaglia, Cássia RC Sossolote, and Josélia Hernandez. 1996. A construção de um léxico para o português do brasil: lições aprendidas e perspectivas. In *Anais do II Encontro para o Processamento de Português Escrito e Falado*, pages 61–70.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708, Melbourne, Australia. Association for Computational Linguistics.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Mário J Silva, Paula Carvalho, and Luís Sarmento. 2012. Building a sentiment lexicon for social judgement mining. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 218–228. Springer.

Rogério Figueredo Sousa, Henrico Bertini Brum, and Maria das Graças Volpe Nunes. 2019. A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Proceedings of the 12th Brazilian Symposium in Information and Human Language Technology*, pages 209–218. Sociedade Brasileira de Computação.

Rogério Figueredo Sousa and Thiago Alexandre Salgueiro Pardo. 2021. The challenges of modeling and predicting online review helpfulness. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 727–738. Sociedade Brasileira de Computação.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Proceedings of the 9th*

*Brazilian Conference on Intelligent Systems*, pages 403–417.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, pages 154–161.

Xi Wang, Iadh Ounis, and Craig Macdonald. 2020. Negative confidence-aware weakly supervised binary classification for effective review helpfulness classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1565–1574.

Shih-Hung Wu and Jun-Wei Wang. 2019. Integrating neural and syntactic features on the helpfulness analysis of the online customer reviews. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1013–1017. IEEE.

Shuzhe Xu, Salvador E Barbosa, and Don Hong. 2020. Bert feature based model for predicting the helpfulness scores of online customers reviews. In *Future of Information and Communication Conference*, pages 270–281. Springer.

Yi-Ching Zeng, Tsun Ku, Shih-Hung Wu, Liang-Pu Chen, and Gwo-Dong Chen. 2014. Modeling the helpful opinion mining of online consumer reviews as a classification problem. *International Journal of Computational Linguistics & Chinese Language Processing*, 19(2):17–32.

Zhu Zhang and Balaji Varadarajan. 2006. Utility scoring of product reviews. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 51–57, New York, NY, USA. ACM.

CAPÍTULO

7

# CONCLUSÃO

No presente capítulo, primeiramente, são tecidas considerações acerca deste trabalho de doutorado, iniciando com a retomada do objetivo geral e, em seguida, apresentando as respostas às questões de pesquisa e a avaliação das hipóteses estabelecidas. Em seguida, apresentam-se as contribuições decorrentes da pesquisa, e, por fim, são enumeradas as limitações e algumas sugestões para trabalhos futuros.

## 7.1 Considerações sobre o trabalho

O objetivo geral deste trabalho consistiu em investigar e propor métodos para a tarefa de classificação automática da utilidade de opiniões para a língua portuguesa. Tornando mais viável o tratamento da utilidade de opiniões nessa língua. Com base no que foi apresentado, pode-se afirmar que o objetivo geral desta pesquisa foi atingido após uma série de trabalhos, consumando-se nos trabalhos apresentados no Capítulo 6, nos quais são propostos vários métodos de classificação, usando algoritmos de aprendizado de máquina clássicos e modernos, com o uso de atributos avaliados no Capítulo 5 e, além desses atributos, foram incluídos atributos gerados por meio de métodos de regularização de grafos. Em particular, na Seção 6.1, apresentou-se uma abordagem alternativa para o problema de classificação de utilidade de opiniões. Os comentários são modelados como grafos e diversos métodos de regularização foram avaliados, o que possibilita classificar os comentários em úteis e não úteis com elevada precisão em um cenário com recursos limitados. Na Seção 6.2, dentre diversos experimentos, é proposto um método baseado em Redes Neurais Convolucionais que obteve resultados muito bons em ambos os domínios. Considera-se, também, que tanto a investigação dos métodos quanto as abordagens propostas neste trabalho possam servir de fundamento para futuras pesquisas na área, especialmente para a língua portuguesa.

Os resultados alcançados pelos experimentos comprovam a hipótese de que existem fatores linguísticos e contextuais que podem ser usados para distinguir as opiniões úteis das não

úteis, considerando que os diversos atributos utilizados conseguem representar bem as opiniões textual e contextualmente e dessa forma ajudam a discriminá-las. Os atributos apresentados e estudados no Capítulo 5 conseguem representar a superfície textual dos comentários de diversas maneiras, principalmente por meio da identificação de elementos-chave dentro da sua estrutura, inclusive considerando elementos semânticos, mesmo que superficialmente. Além disso, vale ressaltar que o uso de representações vetoriais, como as apresentadas na Seção 6.2, com as *Word Embeddings* tradicionais e as contextuais (BERT), ajudam a destacar informações presentes nas palavras e nas suas vizinhanças, reforçando a hipótese. Adicionalmente, a representação alternativa proposta na Seção 6.1 reforça a comprovação da hipótese, pois, de acordo com os resultados, a topologia adotada nos grafos permitiu uma boa propagação de rótulos, usando simplesmente a presença de termos em comum entre os comentários.

A segunda hipótese definida para esta pesquisa afirma que as pessoas conseguem decidir a utilidade das opiniões com consistência observando apenas o texto e os metadados explícitos aos quais elas são expostas, a qual foi comprovada no Capítulo 5. O experimento de similaridade lexical mostrou fortes indícios que as pessoas não avaliam a utilidade de comentários aleatoriamente, mas, pelo contrário, são consistentes e sistemáticas. Os resultados mostram que quanto mais semelhantes lexicalmente são os textos expostos aos avaliadores, mais eles tendem a avaliá-los com a mesma utilidade.

Em relação às questões de pesquisa, as abordagens propostas neste trabalho e os experimentos realizados responderam com sucesso a todas:

1. *É possível identificar automaticamente opiniões úteis ou não úteis escritas em português por meio do uso de fatores textuais e contextuais?*

   Sim. Os resultados dos experimentos com atributos extraídos manualmente, *Embeddings*, BERT e Grafos mostraram que é possível diferenciar as opiniões úteis das não úteis com acurácia satisfatória.

2. *As pessoas concordam entre si ao avaliarem a utilidade de opiniões?*

   Sim. O experimento de anotação manual proposto no Capítulo 5 mostrou que, mesmo sem o interesse real no produto sugerido, os anotadores concordaram ao avaliar a utilidade de opiniões. O experimento de similaridade lexical no mesmo capítulo mostrou que o conteúdo textual é relevante para decidir a utilidade de uma opinião, logo, colaborando com a alta concordância.

3. *Os fatores que podem ser identificados, capazes de distinguir as opiniões úteis das não úteis, podem ser extraídos por meio do uso das técnicas, ferramentas e recursos existentes em PLN?*

   Sim. A criação e disponibilização do UTLCorpus foi essencial para a realização do estudo e a extração dos atributos necessários para a presente pesquisa. De fato, foi somente com o

emprego das técnicas, ferramentas e recursos de PLN que foi possível realizar a extração dos atributos.

4. *É possível adaptar os métodos e atributos desenvolvidos para outras línguas para a classificação de utilidade das opiniões escritas em português?*

   Sim. Vários atributos (Seção 2.2) foram adaptados com sucesso de outras línguas para o português, assim como vários métodos apresentados na Seção 6.2. Entretanto, alguns atributos que dependem de informações mais específicas, como conexões entre pessoas, histórico de revisões, histórico de votos de utilidade, e outros, ainda dependem de adequações em sites de domínios específicos. Por exemplo, nos domínios de produtos e aplicativos para Android, não é possível extrair estas informações, mas no domínio de filmes, com o Filmow, já é possível coletá-las. Até o momento, salvo melhor juízo, o Filmow é a única plataforma que disponibiliza essas informações.

## 7.2 Contribuições

Inicialmente, com o objetivo de contribuir com o problema de escassez de dados para a tarefa de classificação da utilidade de opiniões no português brasileiro, foi proposto e disponibilizado publicamente o UTLCorpus, detalhado no Capítulo 4, o qual é um córpus de opiniões escritas em português, multidomínio, contendo informações de utilidade das opiniões. Ele é composto por milhões de comentários dos domínios de aplicativos para *Android* e filmes. Espera-se que ele possa ser usado para apoiar pesquisas futuras, não só para a tarefa de classificação da utilidade de opiniões, mas para quaisquer outras tarefas que tenham as opiniões como objeto de pesquisa.

À época de geração do UTLCorpus, nenhum site que permitia a publicação de opiniões dos usuários disponibilizava os valores de votos de não útil em suas interfaces, portanto, uma das contribuições deste trabalho é a proposição de um modo de anotação automática da utilidade de opiniões usando apenas os votos úteis e a data de publicação da opinião. Esse método de anotação também está detalhado no Capítulo 4.

No Capítulo 5, foram explorados diversos atributos de conteúdo adaptados para a língua portuguesa, o que permitiu a identificação de atributos (*features*) relevantes para a classificação de utilidade das opiniões em português, bem como a especificação dos atributos mais relevantes para a tarefa. Além disso, no mesmo capítulo, foram propostos um processo de anotação manual da utilidade de comentários e um experimento de similaridade lexical, e, por meio deles, foi possível constatar que há concordância entre as pessoas ao avaliarem a utilidade de opiniões, e também que existe um padrão não aleatório nesse processo.

Finalmente, no Capítulo 6, estão descritas as contribuições em termos de métodos para classificação da utilidade das opiniões. Na Seção 6.1, foi proposto um método para classificação

das opiniões em um contexto de escassez de comentários. A estratégia de modelagem dos comentários usando grafos permitiu o uso de métodos de regularização (similar a propagação de rótulos) que necessitam de poucos nós anotados para inferência da utilidade das opiniões. Já a Seção 6.2, por outro lado, considera a aplicação de métodos que funcionam melhor em um contexto de abundância de recursos, e dessa forma os métodos atingiram os melhores resultados apresentados nessa pesquisa. Vale mencionar que todos os recursos produzidos neste trabalho (córpus e códigos) estão publicamente disponíveis em <https://github.com/RogerFig/>.

## 7.3 Limitações e Trabalhos Futuros

Entre as limitações deste trabalho, destaca-se a impossibilidade de trabalhar com regressão e ranking devido à falta de informação sobre a quantidade de votos não úteis que os comentários receberam nos domínios presentes no UTLCorpus. Embora a classificação já possibilite realizar operações de pré-processamento, como a filtragem de comentários em diversas tarefas, seu uso pode apresentar limitações em contextos em que os comentários devem ser apresentados diretamente aos usuários. Isso se deve ao fato de que, nesses casos, espera-se apresentar aos usuários uma quantidade mínima de comentários, o que pode não ser totalmente viável com o uso da classificação.

Como trabalhos futuros da pesquisa, podemos listar as seguintes sugestões:

1. Ampliar os domínios cobertos pelo UTLCorpus. Apesar de conter dois domínios pouco explorados na tarefa, outros domínios podem ter particularidades que podem ser investigadas. Por exemplo, os domínios de análises de livros e análises de hospedagens.

2. Explorar o uso de relações sociais no contexto da utilidade de opiniões. Alguns sites permitem esse tipo de investigação, pois funcionam como redes sociais. É o caso do Filmow, do qual foram extraídos os comentários de filmes do UTLCorpus. Essas relações sociais podem auxiliar a personalização das opiniões apresentadas aos usuários e a entender a subjetividade inerente à tarefa.

3. Ampliar os estudos sobre os fatores determinantes para a utilidade de opiniões para níveis mais altos do PLN, como semântica e discurso. Dada a elevada subjetividade da tarefa, é provável que níveis mais avançados possam revelar fatores ainda mais relevantes.

4. Expandir os estudos apresentados na Seção 6.1 usando elementos mais específicos da área de redes complexas. Os bons resultados apresentados podem indicar um caminho promissor para a tarefa.

# REFERÊNCIAS

ALMUTAIRI, Y.; ABDULLAH, M.; ALAHMADI, D. Review helpfulness prediction: Survey. **Periodicals of Engineering and Natural Sciences**, v. 7, n. 1, p. 420–432, 2019. Citado nas páginas 15, 33 e 34.

ALUÍSIO, S.; PELIZZONI, J.; MARCHI, A. R.; OLIVEIRA, L. de; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: SPRINGER. **International Workshop on Computational Processing of the Portuguese Language**. [S.l.], 2003. p. 110–117. Citado na página 38.

ANCHIÊTA, R.; SOUSA, R. F.; MOURA, R.; PARDO, T. Improving opinion summarization by assessing sentence importance in on-line reviews. In: **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2017. p. 32–36. Citado na página 23.

BAOWALY, M. K.; TU, Y.-P.; CHEN, K.-T. Predicting the helpfulness of game reviews: A case study on the steam store. **Journal of Intelligent & Fuzzy Systems**, IOS Press, v. 36, n. 5, p. 4731–4742, 2019. Citado nas páginas 32, 38, 41, 54 e 61.

BARBOSA, J. L.; MOURA, R. S. Avaliação automática da utilidade de reviews usando redes neurais artificiais no corpus do steam. In: BRAZILIAN COMPUTER SOCIETY. **Anais do XXVI Congresso da Sociedade Brasileira de Computação: BraSNAM - 5º Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2016. Citado na página 51.

BARBOSA, J. L.; MOURA, R. S.; SANTOS, R. L. d. S. Predicting portuguese steam review helpfulness using artificial neural networks. In: ACM. **Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web**. [S.l.], 2016. p. 287–293. Citado nas páginas 37 e 59.

BENESTY, J.; CHEN, J.; HUANG, Y.; COHEN, I. Pearson correlation coefficient. In: **Noise reduction in speech processing**. [S.l.]: Springer, 2009. p. 1–4. Citado nas páginas 33 e 43.

BILAL, M.; ALMAZROI, A. A. Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews. **Electronic Commerce Research**, Springer, p. 1–21, 2022. Citado nas páginas 53, 54 e 61.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S.l.]: "O'Reilly Media, Inc.", 2009. Citado nas páginas 15, 39 e 40.

BLITZER, J.; DREDZE, M.; PEREIRA, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: **Proceedings of the 45th annual meeting of the association of computational linguistics**. [S.l.: s.n.], 2007. p. 440–447. Citado na página 40.

BORDES, A.; USUNIER, N.; GARCIA-DURAN, A.; WESTON, J.; YAKHNENKO, O. Translating embeddings for modeling multi-relational data. **Advances in neural information processing systems**, v. 26, 2013. Citado na página 53.

BOYD, R. L.; ASHOKKUMAR, A.; SERAJ, S.; PENNEBAKER, J. W. The development and psychometric properties of liwc-22. **Austin, TX: University of Texas at Austin**, 2022. Citado na página 38.

CAMBRIA, E.; OLSHER, D.; RAJAGOPAL, D. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: **Twenty-eighth AAAI conference on artificial intelligence**. [S.l.: s.n.], 2014. Citado na página 50.

CAO, C.; YANG, H. Explore how factors that contribute to online review helpfulness in automotive marketing. 2022. Citado na página 57.

CARVALHO, F.; RODRIGUES, R.; SANTOS, G.; CRUZ, P.; FERRARI, L.; GUEDES, G. Avaliação da versão em português do liwc lexicon 2015 com análise de sentimentos em redes sociais. In: **Anais do VIII Brazilian Workshop on Social Network Analysis and Mining**. Porto Alegre, RS, Brasil: SBC, 2019. p. 24–34. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/6545>. Citado na página 39.

CHEN, Y.; CHAI, Y.; LIU, Y.; XU, Y. Analysis of review helpfulness based on consumer perspective. **Tsinghua Science and Technology**, TUP, v. 20, n. 3, p. 293–305, 2015. Citado na página 36.

CHUA, A. Y. K.; BANERJEE, S. Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. **Computers in Human Behavior**, v. 54, p. 547 – 554, 2016. ISSN 0747-5632. Citado nas páginas 34, 35, 53 e 59.

CHUNLI, H.; WENJUN, J. Aspect-based personalized review ranking. In: IEEE. **2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)**. [S.l.], 2018. p. 1329–1334. Citado na página 32.

DANESCU-NICULESCU-MIZIL, C.; KOSSINETS, G.; KLEINBERG, J.; LEE, L. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In: ACM. **Proceedings of the 18th international conference on World wide web**. [S.l.], 2009. p. 141–150. Citado na página 36.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2018. Citado na página 50.

DIAZ, G. O.; NG, V. Modeling and prediction of online product review helpfulness: A survey. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2018. v. 1, p. 698–708. Citado nas páginas 15, 20, 23, 32, 33, 34 e 40.

DU, J.; ZHENG, L.; HE, J.; RONG, J.; WANG, H.; ZHANG, Y. An interactive network for end-to-end review helpfulness modeling. **Data Science and Engineering**, Springer, v. 5, n. 3, p. 261–279, 2020. Citado nas páginas 53 e 61.

DUMAIS, S. T. Latent semantic analysis. **Annual review of information science and technology**, Wiley Online Library, v. 38, n. 1, p. 188–230, 2004. Citado na página 52.

EBERHARD, L.; KASPER, P.; KONCAR, P.; GÜTL, C. Investigating helpfulness of video game reviews on the steam platform. In: IEEE. **2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)**. [S.l.], 2018. p. 43–50. Citado na página 41.

FILHO, P. B.; PARDO, T. A. S.; ALUÍSIO, S. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. Citado na página 39.

FLESCH, R. A new readability yardstick. **Journal of applied psychology**, American Psychological Association, v. 32, n. 3, p. 221, 1948. Citado na página 52.

FONSECA, E. R.; ROSA, J. L. G. Mac-morpho revisited: Towards robust part-of-speech tagging. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [s.n.], 2013. Disponível em: <https://aclanthology.org/W13-4811>. Citado nas páginas 37, 38 e 39.

FONSECA, E. R.; ROSA, J. L. G.; ALUÍSIO, S. M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. **Journal of the Brazilian Computer Society**, SpringerOpen, v. 21, n. 1, p. 2, 2015. Citado na página 38.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001. Citado na página 54.

GAMZU, I.; GONEN, H.; KUTIEL, G.; LEVY, R.; AGICHTEIN, E. Identifying helpful sentences in product reviews. **arXiv preprint arXiv:2104.09792**, 2021. Citado na página 20.

GAO, B.; HU, N.; BOSE, I. Follow the herd or be myself? an analysis of consistency in behavior of reviewers and helpfulness of their reviews. **Decision Support Systems**, Elsevier, v. 95, p. 1–11, 2017. Citado nas páginas 34, 36, 37 e 41.

GHOSE, A.; IPEIROTIS, P. G. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In: ACM. **Proceedings of the ninth international conference on Electronic commerce**. [S.l.], 2007. p. 303–310. Citado nas páginas 31, 52 e 58.

____. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 23, n. 10, p. 1498–1512, 2011. Citado nas páginas 34, 35, 36, 52 e 58.

HAN, W.; CHEN, H.; HAI, Z.; PORIA, S.; BING, L. Sancl: Multimodal review helpfulness prediction with selective attention and natural contrastive learning. **arXiv preprint arXiv:2209.05040**, 2022. Citado na página 57.

HARTMANN, N. S.; AVANÇO, L. V.; FILHO, P. P. B.; DURAN, M. S.; NUNES, M. D. G. V.; PARDO, T. A. S.; ALUISIO, S. M. *et al.* A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION-ELRA. **International Conference on Language Resources and Evaluation**. [S.l.], 2014. Citado na página 41.

HE, R.; MCAULEY, J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: **Proceedings of the 25th International Conference on World Wide Web**. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016. (WWW '16), p. 507–517. ISBN 978-1-4503-4143-1. Disponível em: <https://doi.org/10.1145/2872427.2883037>. Citado nas páginas 40 e 41.

HONG, H.; XU, D.; WANG, G. A.; FAN, W. Understanding the determinants of online review helpfulness: A meta-analytic investigation. **Decision Support Systems**, Elsevier, v. 102, p. 1–11, 2017. Citado na página 36.

HONG, Y.; LU, J.; YAO, J.; ZHU, Q.; ZHOU, G. 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 nas páginas 35, 55, 56 e 60.

HUANG, A. H.; CHEN, K.; YEN, D. C.; TRAN, T. P. A study of factors that contribute to online review helpfulness. **Computers in Human Behavior**, Elsevier, v. 48, p. 17–27, 2015. Citado na página 36.

HUANG, S.; SHEN, D.; FENG, W.; ZHANG, Y.; BAUDIN, C. Discovering clues for review quality from author's behaviors on e-commerce sites. In: ACM. **Proceedings of the 11th International Conference on Electronic Commerce**. [S.l.], 2009. p. 133–141. Citado nas páginas 49 e 58.

JÄRVELIN, K.; JÄRVELIN, K.; KEKÄLÄINEN, J. Ir evaluation methods for retrieving highly relevant documents. In: ACM. **Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 2000. p. 41–48. Citado na página 33.

JORGE, G. A. Z. **Prevendo a utilidade de comentários em Português Brasileiro de jogos no site Steam**. 65 f. Monografia (Especialização) — MBA em Inteligência Artificial, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2022. Citado nas páginas 15 e 42.

KARIMI, S.; WANG, F. Online review helpfulness: Impact of reviewer profile image. **Decision Support Systems**, Elsevier, v. 96, p. 39–48, 2017. Citado nas páginas 34, 37 e 41.

KAUSHIK, K.; MISHRA, R.; RANA, N. P.; DWIVEDI, Y. K. Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on amazon. in. **Journal of Retailing and Consumer Services**, Elsevier, v. 45, p. 21–32, 2018. Citado na página 36.

KIM, S.-M.; PANTEL, P.; CHKLOVSKI, T.; PENNACCHIOTTI, M. Automatically assessing review helpfulness. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2006 Conference on empirical methods in natural language processing**. [S.l.], 2006. p. 423–430. Citado nas páginas 19, 21, 31, 32, 33, 34, 35, 47, 48 e 58.

KINCAID, J. P.; JR, R. P. F.; ROGERS, R. L.; CHISSOM, B. S. **Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel**. [S.l.], 1975. Citado na página 51.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. **Journal of the ACM (JACM)**, ACM, v. 46, n. 5, p. 604–632, 1999. Citado na página 55.

KONG, L.; LI, C.; GE, J.; NG, V.; LUO, B. Predicting product review helpfulness a hybrid method. **IEEE Transactions on Services Computing**, IEEE, 2020. Citado nas páginas 53 e 61.

KORFIATIS, N.; GARCíA-BARIOCANAL, E.; SáNCHEZ-ALONSO, S. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. **Electronic Commerce Research and Applications**, Elsevier, v. 11, n. 3, p. 205–217, 2012. Citado na página 34.

KRESTEL, R.; DOKOOHAKI, N. Diversifying product review rankings: Getting the full picture. In: IEEE COMPUTER SOCIETY. **Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01**. [S.l.], 2011. p. 138–145. Citado na página 35.

KRISHNAMOORTHY, S. Linguistic features for review helpfulness prediction. **Expert Systems with Applications**, Elsevier, v. 42, n. 7, p. 3751–3759, 2015. Citado na página 32.

KWOK, L.; XIE, K. L. Factors contributing to the helpfulness of online hotel reviews: does manager response play a role? **International Journal of Contemporary Hospitality Management**, Emerald Group Publishing Limited, v. 28, n. 10, p. 2156–2177, 2016. Citado na página 34.

LEE, S.; CHOEH, J. Y. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. **Expert Systems with Applications**, Elsevier, v. 41, n. 6, p. 3041–3046, 2014. Citado na página 32.

LI, S.-T.; PHAM, T.-T.; CHUANG, H.-C. Do reviewers' words affect predicting their helpfulness ratings? locating helpful reviewers by linguistics styles. **Information & Management**, Elsevier, v. 56, n. 1, p. 28–38, 2019. Citado na página 34.

LIU, B. Sentiment analysis and subjectivity. **Handbook of natural language processing**, v. 2, p. 627–666, 2010. Citado nas páginas 27, 28 e 30.

____. Sentiment Analysis and Opinion Mining. **Synthesis Lectures on Human Language Technologies**, v. 5, n. 1, p. 1–167, 5 2012. ISSN 1947-4040. Citado nas páginas 19, 20, 27, 28 e 32.

LIU, J.; CAO, Y.; LIN, C.-Y.; HUANG, Y.; ZHOU, M. Low-quality product review detection in opinion summarization. In: **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)**. [S.l.: s.n.], 2007. Citado nas páginas 21, 23, 32, 34 e 35.

LIU, J.; HAI, Z.; YANG, M.; BING, L. Multi-perspective coherent reasoning for helpfulness prediction of multimodal reviews. In: **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. [S.l.: s.n.], 2021. p. 5927–5936. Citado na página 57.

LIU, Y.; HUANG, X.; AN, A.; YU, X. Modeling and predicting the helpfulness of online reviews. In: IEEE. **Data mining, 2008. ICDM'08. Eighth IEEE international conference on**. [S.l.], 2008. p. 443–452. Citado nas páginas 48 e 58.

LU, Y.; TSAPARAS, P.; NTOULAS, A.; POLANYI, L. Exploiting social context for review quality prediction. In: ACM. **Proceedings of the 19th international conference on World wide web**. [S.l.], 2010. p. 691–700. Citado nas páginas 34, 35, 37, 49 e 59.

MADNANI, N. Getting started on natural language processing with python. **XRDS: Crossroads, the ACM Magazine for Students**, ACM New York, NY, USA, v. 13, n. 4, p. 5–5, 2007. Citado na página 39.

MALIK, M.; HUSSAIN, A. Helpfulness of product reviews as a function of discrete positive and negative emotions. **Computers in Human Behavior**, Elsevier, v. 73, p. 290–302, 2017. Citado na página 32.

MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank. **Using Large Corpora**, MIT Press, v. 273, 1994. Citado na página 39.

MARCUS, M.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank. 1993. Citado na página 38.

MARTINS, A. C. S.; TACLA, C. A. Assessement of features influencing the voting for opinions' helpfulness about services in portuguese. In: BRAZILIAN COMPUTER SOCIETY. **Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective-Volume 1**. [S.l.], 2015. p. 21. Citado nas páginas 51 e 59.

MAVERICK, G. V. **Computational analysis of present-day american english**. [S.l.]: JSTOR, 1969. Citado na página 39.

MCAULEY, J.; TARGETT, C.; SHI, Q.; HENGEL, A. van den. Image-based recommendations on styles and substitutes. In: **Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2015. (SIGIR '15), p. 43–52. ISBN 978-1-4503-3621-5. Disponível em: <http://doi.acm.org/10.1145/2766462.2767755>. Citado nas páginas 40 e 41.

MELLENG, A.; JUREK-LOUGHREY, A.; P, D. Ranking online reviews based on their helpfulness: An unsupervised approach. In: **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)**. Held Online: INCOMA Ltd., 2021. p. 959–967. Disponível em: <https://aclanthology.org/2021.ranlp-1.109>. Citado nas páginas 56 e 62.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2013. p. 3111–3119. Citado na página 51.

MORADI, M.; DASS, M.; KUMAR, P. Differential effects of analytical versus emotional rhetorical style on review helpfulness. **Journal of Business Research**, Elsevier, v. 154, p. 113361, 2023. Citado na página 57.

MUKHERJEE, S.; POPAT, K.; WEIKUM, G. Exploring latent semantic factors to find useful product reviews. In: SIAM. **Proceedings of the 2017 SIAM International Conference on Data Mining**. [S.l.], 2017. p. 480–488. Citado na página 32.

NGO-YE, T. L.; SINHA, A. P. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. **Decision Support Systems**, Elsevier, v. 61, p. 47–58, 2014. Citado nas páginas 36 e 41.

NI, J.; LI, J.; MCAULEY, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: **Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)**. [S.l.: s.n.], 2019. p. 188–197. Citado nas páginas 40 e 41.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The PageRank citation ranking: Bringing order to the web.** [S.l.], 1999. Citado na página 55.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Found. Trends Inf. Retr.**, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1–2, p. 1–135, jan 2008. ISSN 1554-0669. Disponível em: <https://doi.org/10.1561/1500000011>. Citado nas páginas 21, 32 e 48.

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. Linguistic inquiry and word count: Liwc 2001. **Mahway: Lawrence Erlbaum Associates**, v. 71, n. 2001, p. 2001, 2001. Citado nas páginas 38 e 54.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 51.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Bioinfo Publications, 2011. Citado na página 43.

QAZI, A.; SYED, K. B. S.; RAJ, R. G.; CAMBRIA, E.; TAHIR, M.; ALGHAZZAWI, D. A concept-level approach to the analysis of online review helpfulness. **Computers in Human Behavior**, Elsevier, v. 58, p. 75–81, 2016. Citado nas páginas 50 e 59.

REN, G.; HONG, T. Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. **Information Processing & Management**, Elsevier, v. 56, n. 4, p. 1425–1438, 2019. Citado na página 35.

RIJSBERGEN, C. van. **Information Retrieval**. 1979. Citado na página 43.

SALEHAN, M.; KIM, D. J. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. **Decision Support Systems**, Elsevier, v. 81, p. 30–40, 2016. Citado na página 34.

SANTOS, R. L. d. S.; SOUSA, R. F. de; RABELO, R. A.; MOURA, R. S. An experimental study based on fuzzy systems and artificial neural networks to estimate the importance of reviews about product and services. In: IEEE. **Neural Networks (IJCNN), 2016 International Joint Conference on**. [S.l.], 2016. p. 647–653. Citado nas páginas 55 e 60.

SAUMYA, S.; SINGH, J. P.; BAABDULLAH, A. M.; RANA, N. P.; DWIVEDI, Y. K. Ranking online consumer reviews. **Electronic Commerce Research and Applications**, Elsevier, v. 29, p. 78–89, 2018. Citado nas páginas 32, 56 e 60.

SAUMYA, S.; SINGH, J. P.; DWIVEDI, Y. K. Predicting the helpfulness score of online reviews using convolutional neural network. **Soft Computing**, Springer, p. 1–17, 2019. Citado na página 32.

____. Predicting the helpfulness score of online reviews using convolutional neural network. **Soft Computing**, Springer, v. 24, n. 15, p. 10989–11005, 2020. Citado nas páginas 50, 51 e 61.

SCARTON, C. E.; ALUíSIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010. Citado na página 52.

SHAH, A. M.; MUHAMMAD, W.; LEE, K. Examining the determinants of patient perception of physician review helpfulness across different disease severities: A machine learning approach. **Computational Intelligence and Neuroscience**, Hindawi, v. 2022, 2022. Citado na página 57.

SHEN, Y.; CHOI, P.; LI, J.; ZHANG, X.; BISER, J. What women and men want in online product reviews: Gender effects on review helpfulness perceptions. **The Journal of Applied Business and Economics**, North American Business Press, v. 24, n. 5, p. 32–40, 2022. Citado na página 57.

SIERING, M.; MUNTERMANN, J.; RAJAGOPALAN, B. Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. **Decision Support Systems**, Elsevier, v. 108, p. 1–12, 2018. Citado na página 36.

SOUSA, R.; PARDO, T. Evaluating content features and classification methods for helpfulness prediction of online reviews: Establishing a benchmark for Portuguese. In: **Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis**. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 204–213. Disponível em: <https://aclanthology.org/2022.wassa-1.19>. Citado na página 26.

SOUSA, R. F. d.; BRUM, H. B.; NUNES, M. d. G. V. A bunch of helpfulness and sentiment corpora in brazilian portuguese. In: **Symposium in Information and Human Language Technology - STIL**. Salvador, BA: SBC, 2019. p. 209–218. Citado na página 26.

SOUSA, R. F. D.; PARDO, T. A. S. The challenges of modeling and predicting online review helpfulness. In: SBC. **Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2021. p. 727–738. Citado na página 26.

SOUSA, R. F. de; ANCHIêTA, R. T.; NUNES, M. d. G. V. A graph-based method for predicting the helpfulness of product opinions. **iSys - Brazilian Journal of Information Systems**, v. 13, n. 4, p. 06–21, Jul. 2020. Disponível em: <https://sol.sbc.org.br/journals/index.php/isys/article/view/821>. Citado na página 26.

SOUSA, R. F. de; RABÊLO, R. A.; MOURA, R. S. A fuzzy system-based approach to estimate the importance of online customer reviews. In: IEEE. **Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on**. [S.l.], 2015. p. 1–8. Citado nas páginas 54 e 60.

SQUARISI, D.; SALVADOR, A. **A arte de escrever bem: um guia para jornalistas e profissionais do texto**. [S.l.]: Editora Contexto, 2008. Citado na página 51.

SUSAN, M. M.; DAVID, S. What makes a helpful online review? a study of customer reviews on amazon. com. **MIS Quarterly**, v. 34, n. 1, p. 185–200, 2010. Citado nas páginas 22, 33, 34 e 36.

TANAKA, Y.; NAKAMURA, N.; HIJIKATA, Y.; NISHIDA, S. Estimating reviewer credibility using review contents and review histories. **IEICE TRANSACTIONS on Information and Systems**, The Institute of Electronics, Information and Communication Engineers, v. 95, n. 11, p. 2624–2633, 2012. Citado na página 34.

TANG, J.; GAO, H.; HU, X.; LIU, H. Context-aware review helpfulness rating prediction. In: **Proceedings of the 7th ACM Conference on Recommender Systems**. New York, NY, USA: ACM, 2013. (RecSys '13), p. 1–8. ISBN 978-1-4503-2409-0. Disponível em: <http://doi.acm.org/10.1145/2507157.2507183>. Citado nas páginas 37 e 41.

TSAPARAS, P.; NTOULAS, A.; TERZI, E. Selecting a comprehensive set of reviews. In: ACM. **Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2011. p. 168–176. Citado na página 32.

TSUR, O.; RAPPOPORT, A. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In: **ICWSM**. [S.l.: s.n.], 2009. Citado nas páginas 20, 32 e 34.

WANG, X.; TANG, L. R.; KIM, E. More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness? **International Journal of Hospitality Management**, Elsevier, v. 77, p. 438–447, 2019. Citado na página 38.

WANG, Y.; WANG, L.; LI, Y.; HE, D.; LIU, T.-Y. A theoretical analysis of ndcg type ranking measures. In: **Conference on Learning Theory**. [S.l.: s.n.], 2013. p. 25–54. Citado na página 33.

WU, J. Review popularity and review helpfulness: a model for user review effectiveness. **Decision Support Systems**, Elsevier, v. 97, p. 92–103, 2017. Citado nas páginas 34, 36 e 37.

WU, J.; XU, B.; LI, S. An unsupervised approach to rank product reviews. In: IEEE. **2011 eighth international conference on fuzzy systems and knowledge discovery (FSKD)**. [S.l.], 2011. v. 3, p. 1769–1772. Citado nas páginas 55 e 60.

WU, R.; WU, H.-H.; WANG, C. L. Why is a picture 'worth a thousand words'? pictures as information in perceived helpfulness of online reviews. **International Journal of Consumer Studies**, Wiley Online Library, v. 45, n. 3, p. 364–378, 2021. Citado na página 57.

XIA, L. The impacts of geographic and social influences on review helpfulness perceptions: A social contagion perspective. **Tourism Management**, Elsevier, v. 95, p. 104687, 2023. Citado na página 57.

XU, C.; ZHENG, X.; YANG, F. Examining the effects of negative emotions on review helpfulness: The moderating role of product price. **Computers in Human Behavior**, Elsevier, v. 139, p. 107501, 2023. Citado na página 57.

XU, S.; BARBOSA, S. E.; HONG, D. Bert feature based model for predicting the helpfulness scores of online customers reviews. In: SPRINGER. **Future of Information and Communication Conference**. [S.l.], 2020. p. 270–281. Citado nas páginas 50, 51 e 61.

YANG, Y.; CHEN, C.; BAO, F. S. Aspect-based helpfulness prediction for online product reviews. In: IEEE. **2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)**. [S.l.], 2016. p. 836–843. Citado na página 35.

YANG, Y.; YAN, Y.; QIU, M.; BAO, F. 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)**. [S.l.: s.n.], 2015. p. 38–44. Citado nas páginas 20, 32, 34, 35 e 38.

ZAR, J. H. Significance testing of the spearman rank correlation coefficient. **Journal of the American Statistical Association**, Taylor & Francis, v. 67, n. 339, p. 578–580, 1972. Citado nas páginas 33 e 43.

ZENG, Y.-C.; KU, T.; WU, S.-H.; CHEN, L.-P.; CHEN, G.-D. Modeling the helpful opinion mining of online consumer reviews as a classification problem. **International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014**, v. 19, n. 2, 2014. Citado na página 32.

ZHANG, Y.; LIN, Z. Predicting the helpfulness of online product reviews: A multilingual approach. **Electronic Commerce Research and Applications**, Elsevier, v. 27, p. 1–10, 2018. Citado nas páginas 34 e 41.

ZHANG, Y.; ZHANG, D. Automatically predicting the helpfulness of online reviews. In: IEEE. **Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)**. [S.l.], 2014. p. 662–668. Citado nas páginas 34 e 36.

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. Citado nas páginas 31, 32, 35, 48 e 58.

ZHENG, J.; WEN, P.; JI, X.; LYU, X.; YANG, Y. Helpfulness prediction of online drug reviews. In: IEEE. **2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)**. [S.l.], 2021. p. 528–537. Citado na página 57.


ICMC
USP
SÃO CARLOS