Estudo de Método de Extração de Aspectos para Português do Brasil Baseado em Regras

Vanessa de Souza Câmara[1], **Tiago de Melo**[1]

[1]Laboratório de Sistemas Inteligentes (LSI)
Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brazil

{vdsc.snf19, tmelo}@uea.edu.br

Abstract. *Aspect extraction is considered to be one of the most important tasks in sentiment analysis. This task focuses on identifying the targets described in online user reviews. Despite its relevance and not being a new task, most of the work is focused on the English language. In this paper, a rule-based aspect extraction approach is proposed and evaluated. We evaluated 42 rules originally proposed for the English language and also a greedy rule selection algorithm to identify the best subset. The approach was evaluated in four different datasets. The results obtained indicate that it is possible to use extraction rules from another language and that the greedy algorithm is an efficient strategy for rule selection.*

Resumo. *Extração de aspectos é considerada como uma das mais importantes tarefas na análise de sentimentos. Essa tarefa foca na identificação dos alvos descritos nos comentários online de usuários. Apesar da sua relevância e de não ser uma tarefa nova, a maioria dos trabalhos é voltada para o idioma inglês. Neste artigo, é proposta e avaliada uma abordagem de extração de aspectos baseada em regras. Foram avaliadas 42 regras propostas originalmente para o idioma inglês e também um algoritmo guloso de seleção de regras para identificar o melhor subconjunto. A abordagem foi avaliada em quatro distintos conjuntos de dados. Os resultados alcançados indicam ser possível o uso de regras de extração a partir de outro idioma e que o algoritmo guloso é uma estratégia eficiente de seleção de regras.*

1. Introdução

Uma grande quantidade de comentários com opiniões de usuários está presente nas publicações de mídias sociais na Web. Esses comentários são importantes na escolha de produtos e serviços por potenciais consumidores, assim como são relevantes para as empresas, pois os fabricantes podem usar as avaliações para analisar e corrigir defeitos de seus produtos [Law et al. 2017].

Análise de sentimentos lida com comentários de usuários e é realizada em diferentes níveis de granularidade. Para uma análise mais refinada das opiniões dos usuários, a polaridade dos principais aspectos de comentários é classificada como positiva ou negativa. Assim, extrair aspectos é considerada como umas das tarefas mais importantes [Shafie et al. 2018], pois esta tem o objetivo de identifcar o termo emitido na sentença juntamente com os sentimentos associados. Um aspecto é o termo na frase sobre o qual

o autor emite uma opinião e eles podem ser atributos, características ou componentes de uma entidade [Pereira 2021]. Por exemplo, na frase "eu adorei o tempero deste restaurante", o aspecto "tempero" seria extraído e classificado como positivo. Os aspectos são empregados em diversas aplicações de redes sociais, tais como na comparação de opiniões de usuários em sistemas de processos judiciais eletrônicos [Nascimento et al. 2020], detecção de notícias falsas [Testoni et al. 2021] ou na tarefa de integração de dados factuais e subjetivos em bancos de dados de comércio eletrônico [da Silva 2021].

Extrair aspectos não é uma tarefa trivial. Ao analisar um produto, por exemplo, as pessoas costumam comentar sobre diversos aspectos e mencionar diferentes sentimentos sobre eles [Shafie et al. 2018]. Além disso, é comum que gírias, abreviações, ironias e sarcasmos apareçam nos comentários, prejudicando as estruturas sintática e semântica da frase e, assim, dificultando a extração. Outro desafio é a precariedade de ferramentas disponíveis para essa tarefa em muitos idiomas. A língua portuguesa, por exemplo, apesar de ser uma das cinco línguas mais utilizadas na Web [Pereira 2021], carece de recursos que processam dados anotados para realização de experimentos e validação de hipóteses, sendo necessário recorrer aos estudos que, majoritariamente, estão disponíveis em inglês [Oliveira and de Melo 2020, Pereira 2021, Rana and Cheah 2016].

Diante deste cenário, este trabalho se propõe a investigar uma abordagem de extração de aspectos baseada em regras e uma estratégia para selecionar o melhor conjunto de regras de extração. Para isso, foram propostas duas hipóteses:

- $\mathcal{H}1$ É possível usar eficientes regras para extração de aspectos em inglês e adaptá-las para português.
- $\mathcal{H}2$ É possível aplicar uma estratégia para selecionar um subconjunto ótimo de regras de extração de aspectos em português.

Para validar a hipótese $\mathcal{H}1$, foram implementadas e adaptadas 42 regras de extração de aspectos em inglês publicadas em diversos trabalhos [Tubishat et al. 2021, Poria et al. 2014] para português. As regras foram avaliadas em quatro diferentes conjuntos de dados e foi possível identificar as melhores regras individuais para cada domínio. Para validar a hipótese $\mathcal{H}2$, foi aplicada uma estratégia gulosa que permite selecionar um subconjunto ótimo do conjunto inicial de regras, conforme proposto recentemente na literatura [Tubishat et al. 2021, Liu et al. 2016]. Todos os subconjuntos de regras alcançaram resultados superiores aos resultados alcançados pelas melhores regras individuais. Assim, o resultado experimental corrobora com a literatura no sentido que a estratégia adotada é eficiente na seleção do conjunto de regras. As implementações das regras e os conjuntos de dados utilizados estão compartilhados[1].

O artigo está organizado da seguinte maneira. Na Seção 2, são descritos os principais trabalhos relacionados. Na Seção 3, são apresentados o conjunto de regras avaliado nesse estudo e o conjunto de dados utilizado nos experimentos. Na Seção 4 são apresentados e discutidos os resultados experimentais. Finalmente, na Seção 5, são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

O trabalho de Cardoso e Pereira [Cardoso and Pereira 2020] tem como objetivo avaliar um método de extração de aspectos usando ferramentas de Procesamento de Lingua-

[1]https://github.com/vanessadcamara/nlp-colab

gem Natural (PLN) para a língua portuguesa. Os autores implementaram um método baseado em CRF (*Conditional Random Field*) com valores de um classificador obtidos através de ferramentas de PLN para português e compararam com os valores traduzidos por ferramentas para o inglês. Os resultados obtidos por Cardoso e Pereira [Cardoso and Pereira 2020] indicam que não vale a pena fazer a tradução de textos para o inglês para se obter maior eficácia.

O trabalho de Costa e Pardo [Costa and Pardo 2020] investigou um método de extração de aspectos em português baseado em léxicos. Porém, neste estudo, o uso de ontologias se mostrou uma melhor opção em relação ao uso de léxicos. Além disso, os autores reconheceram a dificuldade de usar ontologias em domínios variados.

Poria *et al*. [Poria et al. 2014] apresentam uma abordagem baseada em regras para extração de aspectos de comentários feitos sobre produtos. Os autores apresentam 13 regras implementadas para a língua inglesa, sendo que 9 delas são utilizadas para esta pesquisa e adaptadas para a extração de aspectos em português. Quatro regras não foram utilizadas devido às diferenças linguísticas entre os dois idiomas ou porque se tratavam de regras com aspectos implícitos, sendo que, para este trabalho, o foco é na extração de aspectos explícitos.

Tubishat *et al*. [Tubishat et al. 2021] avaliaram um conjunto de 126 regras de extração de aspectos organizadas em diversas categorias, tais como regras baseadas em padrões e regras baseadas em relações de dependência. Das regras baseadas em dependência apresentadas pelos autores, 17 serão adicionadas nesta pesquisa.

Liu *et al*. [Liu et al. 2016] apresentou um método para selecionar automaticamente regras para extração de aspectos. Os autores avaliaram dois algoritmos baseados em uma estratégia gulosa. O algoritmo guloso avaliado pelos autores foi utilizado como referência e adaptado para selecionar o conjunto de regras ótimas para a língua portuguesa.

3. Materias e Métodos

Esta pesquisa lidou com a tarefa de extração de aspectos em português. A hipótese $\mathcal{H}1$ apresentada neste trabalho é que seria possível empregar regras eficientes de extração de aspectos, usadas em outro idioma e adaptá-las para português. Inicialmente, foi selecionado um conjunto de regras de extração de aspectos para a língua inglesa e estas regras foram adaptadas para a extração de aspectos em português do Brasil. A adaptação foi necessária devido à diferença linguística entre os idiomas. O conjunto de regras foi avaliado em quatro conjuntos de dados de domínios distintos. Ademais, para validar a hipótese $\mathcal{H}2$ foi adaptada e analisada uma estratégia gulosa para combinar as regras afim de se obter um subconjunto de regras que fosse mais efetivo do que o uso de regras individuais.

3.1. Conjunto de Dados

Foram utilizados quatro diferentes conjuntos de dados na avaliação do conjunto de regras de extração de aspectos. Estes conjuntos de dados variam em domínio e tamanho. Considerando que o desempenho das regras de extração pode ser afetado por essas características dos dados, avaliar um método em conjuntos de dados com diferentes configurações permite estimar a sua robustez. A Tabela 1 apresenta um sumário quanti-

tativo dos conjuntos de dados empregados neste trabalho. A seguir, é apresentada uma breve descrição dos conjuntos de dados.

Tabela 1. Sumário do conjunto de dados.

	TA-Restaurantes	TV	Reli	ReHol
Comentários	305	1.091	1.600	1.513
Sentenças	316	2.329	2.177	1.540
Aspectos anotados	306	2.388	2.471	2.573
Aspectos únicos	120	350	550	624
Sentenças com aspectos	85.1%	38.2%	100%	100%

TA-Restaurantes. O primeiro conjunto anotado é uma coleta de 350 comentários de usuários sobre restaurantes no Brasil no site TripAdvisor[2]. A coleta ocorreu no período de janeiro a março de 2020 por Oliveira e Melo [Oliveira and de Melo 2020]. Os textos dos comentários foram divididos em 305 sentenças e 306 aspectos foram manualmente identificados pelos autores.

TV[3]. O segundo conjunto de dados é uma coleta de comentários de usuários sobre um modelo de televisão em sites de comércio eletrônico do grupo B2W Digital. Os dados foram manualmente anotados por Cardoso e Pereira [Cardoso and Pereira 2020]. O conjunto de dados contém 1.091 comentários e foram identificados 2.388 aspectos.

ReLi. O dataset ReLi Corpus[4] [Freitas et al. 2012] é um corpus de resenhas de livros composto de 1.600 reviews e que foram identificados 2.471 aspectos. As resenhas que compõem o ReLi foram extraídas do site Skoob.com, uma rede social de livros e leitores, na qual os leitores e colaboradores participam de forma ativa dando opiniões a respeito dos livros.

ReHol[5]. Este conjunto de dados foi produzido e anotado manualmente por Barros e Bona [Barros and Bona 2021] e contém comentários coletados no site TripAdvisor[6] sobre hotéis. ReHol possui 1.513 comentários e foram identificados 2.573 aspectos.

3.2. Conjunto de Regras

Foram avaliadas 42 regras de extração de aspectos. A Tabela 2 apresenta uma descrição de cada regra, com suas relações de dependência e seus respectivos exemplos. Os aspectos estão destacados em negrito para melhor identificá-los tanto na descrição da regra quanto nos exemplos. As regras de 1 a 9 foram inspiradas pelo trabalho de Poria *et al*. [Poria et al. 2014] e as regras de 10 a 42 foram inspiradas pelo trabalho de Tubishat *et al*. [Tubishat et al. 2021].

Algumas regras foram adaptadas para que pudessem ser adequadas para o idioma português. Mais especificamente, as regras 20, 22, 35 e 42 foram adaptadas. Por exemplo, para a Regra 20, o aspecto "*screen displays*" em inglês é traduzido como "exibições na

[2]https://www.tripadvisor.com.br
[3]https://www.kaggle.com/brenexdev/aspect-extraction-portuguese
[4]Dataset disponível em https://dx.doi.org/10.21227/0ej1-br13.
[5]Dataset disponível em https://dx.doi.org/10.21227/0ej1-br13.
[6]https://www.tripadvisor.com.br

Tabela 2. Conjunto de regras. Os termos em negrito na terceira coluna são os aspectos.

#Regra	Descrição da Regra	Exemplo
1	nsubj(**H1**, H2) e (advmod(H2, H3) ou adjmod(H2,H3)) tal que H2 está em SenticNet	O **livro** é muito bom.
2	nsubj(VB1, **NN**) e (advmod(VB1, H1) ou advcl(VB1, H1) ou adjmod(VB1, H1) ou amod(VB1, H1)), tal que a sentença não contém verbo auxiliar	A **bateria** dura pouco.
3	nsubj(H1, VB1) e (obj(VB1, **NN**) ou obj:pass(VB1, **NN**)) tal que NN não está em SenticNet e a frase não contenha verbo auxiliar	Eu gosto das **lentes** da câmera.
4	nsubj(H1, VB1) e (obj(VB1, **NN**) ou obj:pass(VB1, **NN**)), Rel(NN, NN1), em que Rel1 é qualquer relação de dependência NN está em SenticNet e H1 é um sentimento, tal que a frase não contém verbo auxiliar	Eu gosto da **beleza** da tela.
5	nsubj(H1,H2) e xcomp(H2,H3) e Rel1(H3,**NN**) tal que Rel1 é qualquer relação de dependência e a frase não contém verbo auxiliar	Eu gostaria de comentar sobre a **câmera** do celular.
6	nsubj(**NN**,H1) e cop(VB1,H1)	A **câmera** é boa.
7	amod(H1, **NN**) ou nmod(H1, **NN**) ou acl(H1, **NN**) ou acl:relcl(H1, **NN**)	Adorei a **habilidade do jogador**.
8	obj(**NN**,VB1) ou nsubj:pass(**NN**,VB1)	Ana achou o **livro** maravilhoso.
9	nsubj(H2, **NN**) e conj(H1, H2)	A **câmera** é incrível e fácil de usar.
10	nsubj(JJ/OP, **NN**)	O **vídeo** era ruim.
11	ReL1(H1,**NN1**) e ReL2(H1,**NN2**), em que ReL1 e ReL2 podem ser: 'nsubj', 'amod', 'conj', 'prep', 'csubj', 'xsubj', 'dobj', 'iobj'	A **qualidade** e a **lente** da câmera estão aprovadas.
12	nsubj(VB1,H1) e (dobj(VB1,**NN**) ou obj(VB1, **NN**)	Honestamente, eu amo esse **jogador**.
13	nsubj(H1,**NN**) e xcomp(H1,JJ/OP)	Seu **tamanho** também faz dele ideal para viajar.
14	amod(NN1,OP/JJ) e conj(**NN1**,**NN2**)	Toca **dvds** e **cds** originais.
15	nmod(OP/JJ,**NNS**)	O ruim dos **jogos** é que são poucos.
16	amod(**NN**,OP)	O **manual** escasso.
17	ReL1(H1,**NN**) e ReL2(H1, OP/JJ), em que Rel1 e Rel2 podem ser: 'nsubj', 'amod', 'prep', 'csubj', 'xsubj', 'dobj', 'iobj', 'obj'	Esta **câmera** tem uma falha de design.
18	nsubj(**NN**,OP/JJ)	Minha reclamação do hardware são os **botões**.
19	dobj(OP/JJ,**NN**) ou obj(OP/JJ, **NN**)	Eu gostei dos **botões** mais usados.
20	nsubj(H1, H2) e obj(H2,NN1) e nmod(**NN1**, **NN2**) e case (NN2, **H3**)	Eu acho **exibições na tela** irritantes.
21	conj(**NN1**,**NN2**)	**Qualidade** e **lentes** da câmera comprovadas.
22	amod(NN1,OP/JJ) e nmod(**NN1**, **NN2**) e case (NN2, **H3**)	O g3 entrega a melhor **qualidade de imagem**.
23	nsubj(OP/JJ,**NN**) e (advmod(OP/JJ, H1) ou neg(OP/JJ, H1))	As **cores** na tela não são tão nítidas.
24	ReL(**NN**, OP/JJ) tal que ReL é uma relação de dependência de ['nsubj', 'amod', 'prep', 'csubj', 'xsubj', 'dobj', 'iobj']	Definitivamente uma boa **câmera**.
25	nsubj(OP1/JJ1,**NN**) e cop (OP1/JJ1,H1)	Os **cardápios** são fáceis de navegar.
26	**NNS** VBP OP/JJ	Os **controles** são feios
27	**NN** RB JJ/OP	É um **produto** muito incrível
28	JJ/OP **NN**	Baixa **confiabilidade**
29	**NN** JJ/OP	**Software** criativo complexo
30	**NN** IN NN	**Áudio** do vídeo também faltando
31	**NN** IN DT NN	A **construção** do player é a mais cafona que já vi
32	NN IN DT **NN**	Uma boa compra para o **preço**
33	OP + preposição "a"+ **VB**	Ele recusou a **ler** discos secundários
34	JJ + preposição "para"+ JJ + preposição "para"+ VB tal que JJ não está em OP	É extremamente simples de **navegar**
35	**NN JJ1** JJ2, tal que JJ2 não está em op	o g3 tem **deslocamentos brancos** muito mais nítidos
36	**NN** VBZ/VBP DT OP/JJ NN	o **manual** faz um bom trabalho
37	**NN** * PRP/DT OP, tal que * é qualquer padrão encontrado que não seja NN	**apex** é a melhor marca de qualidade barata para leitores de DVD
38	VB OP/JJ **NN**	tem ótima **recepção**
39	**NN** VB OP/JJ	**áudio** é excelente
40	**NN** NN VBZ/VBP RB RB	a tecnologia **mms** é muito bem integrada.
41	**NN NN** * JJ/OP NN	**recurso sunset** bate fotos incríveis
42	amod(NN1,JJ/OP) and amod(**NN1,JJ**) tal que JJ não esteja em OP	a canon g3 é a melhor **câmera digital**

tela" em português. A relação composta (*compound*) entre os termos "*screen*" e "*display*" não faz sentido em português. Assim, os termos "exibições na tela" foram identificados através da relação de modificador nominal (`nmod`). Além disso, há casos em que duas diferentes regras conseguem extrair o mesmo conjunto de aspectos. Por exemplo, as regras 10 e 39 funcionam para frases com estruturas sintáticas idênticas. No entato, decidiu-se por manter as duas regras e investigar o funcinamento de ambas porque a Regra 10 possui extração de aspectos baseada em dependência, enquanto que a Regra 39 realiza a extração baseada na identificação de padrões.

As regras foram implementadas na linguagem Python versão 3.7.13. A biblioteca utilizada para identificar as relações de dependência das sentenças e as classes gramaticais dos textos processados foi o pacote spaCy[7] na versão 3.1.1 para o português. Além disso, utilizou-se o *framework* SenticNet[8] para verificar se determinadas palavras pertenciam a um conjunto de termos subjetivos, conforme descrito nas regras 1, 3 e 4.

A Figura 1 apresenta um exemplo extração de aspectos, em que existe uma relação do tipo `obj` entre o *token* H1 (achou) e o *token* H2 (livro). Conforme descrito na Regra 8 da Tabela 2, se existir esse tipo de relação, o termo "livro" (H2) deve ser extraído como um aspecto.

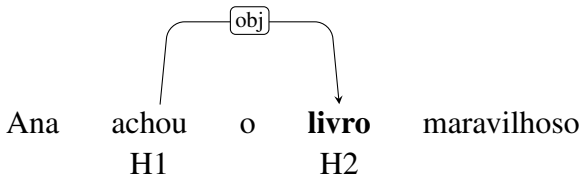


Figura 1. Exemplo de extração de aspectos da Regra 8.

Outro exemplo de extração de aspectos é a sentença apresentada na Figura 2, em que existe uma relação `rel1` do tipo `conj` entre dois substantivos, que são "qualidade" e "lente", e uma relação `rel2` do tipo `nsubj` entre um destes substantivos ("qualidade") e um outro termo "aprovadas" (adjetivo). Essas relações `rel1` e `rel2` estão no conjunto de dependências estabelecidas na Regra 11 da Tabela 2. Nesta regra, os termos "qualidade" e "lente" são extraídos como aspectos.

As regras de extração de aspectos são baseadas em relações de dependência. Por exemplo, na sentença *"a câmera é boa"* existe uma relação nomeada do tipo `nsubj` entre as palavras *câmera* e *boa*. Essa relação `nsubj` representa a ligação entre um adjetivo e um substantivo, conforme ilustrada na Figura 3. Este é o exemplo descrito na Regra 6 da Tabela 2.

Além disso, H é qualquer palavra, OP significa qualquer palavra de opinião a partir de léxico de opinião extraído e traduzido da lista de palavras disponível em [Hu and Liu 2004], NN é qualquer tipo de substantivo, VB é qualquer tipo de verbo, NNS é um substantivo no plural, JJ qualquer tipo de adjetivo. Eles são *tags* do identificador de classe de discurso (POS) de Stanford.

[7]https://spacy.io
[8]https://sentic.net

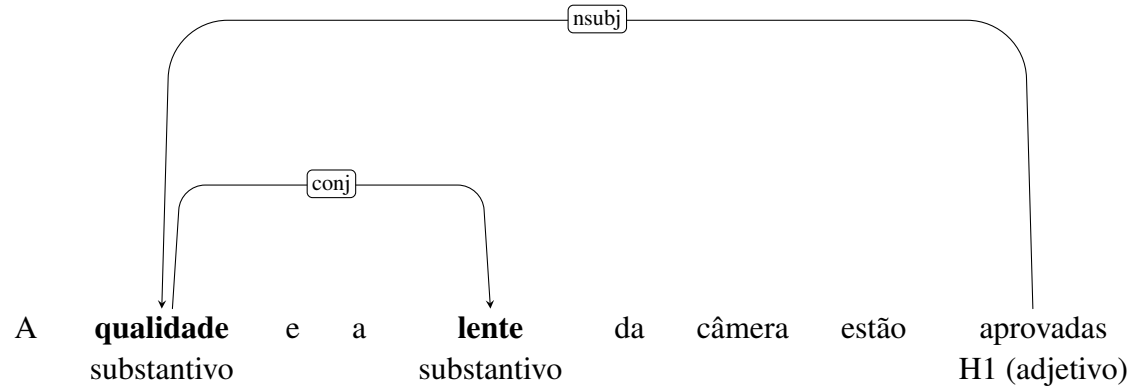


Figura 2. Exemplo de extração de aspectos da Regra 11.

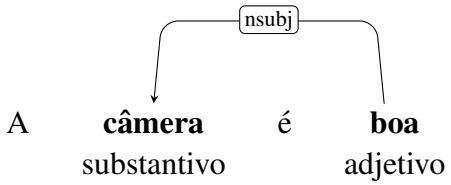


Figura 3. Exemplo de extração de aspectos da Regra 6.

3.3. Estratégia Gulosa

Para selecionar o subconjunto ideal de regras seria necessário executar todas as combinações possíveis das regras de extração. No entanto, isso seria computacionalmente inviável, pois o número de combinações de subconjuntos cresce exponencialmente. Conforme descrito na literatura [Tubishat et al. 2021, Liu et al. 2016], a seleção de um subconjunto ideal de regras é um problema NP-difícil. Por este motivo, uma abordagem viável é a utilização de uma estratégia gulosa para seleção do melhor subconjunto de regras.

Neste trabalho, adotou-se uma estratégia inspirada e adaptada do algoritmo de seleção de regras ótimas proposto por Tubishat *et al*. [Tubishat et al. 2021]. A estratégia é utilizada a fim de que seja computacionalmente viável realizar as combinações das regras para se obter resultados melhores do que os resultados alcançados por regras individuais. O algoritmo funciona da seguinte forma:

Passo 1 Seja o conjunto $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ de regras de extração de aspectos. Inicialmente, calcula-se a precisão, revocação e F1-Score de cada regra individualmente.

Passo 2 Seja o conjunto ótimo $\mathcal{R}_o$ de regras de extração de aspectos. Inicialmente o conjunto ótimo é vazio ($\mathcal{R}_o = \emptyset$).

Passo 3 Ordenam-se as regras de $\mathcal{R}$ de modo decrescente pela precisão. A regra com o maior valor de precisão $r_m \in \mathcal{R}$ será inserida ao conjunto ideal $\mathcal{R}_o$ e a regra r_m será removida de $\mathcal{R}$. Caso haja mais de uma regra com valores iguais de precisão, escolhe-se a regra com maior revocação.

Passo 4 Em seguida, combina-se cada uma das regras restantes de $\mathcal{R}$ com as regras do conjunto ideal $\mathcal{R}_o$ e calcula-se o F1-Score de cada combinação. Caso a combinação do conjunto com uma nova regra r_t testada resulte em um F1-Score maior que o conjunto anterior $\mathcal{R}_o$, a regra r_t é adicionada ao conjunto ótimo $\mathcal{R}_o = \mathcal{R}_o \cup \{r_t\}$

e removida do conjunto $\mathcal{R}$.

Passo 5 O Passo 4 é executado continuamente até que não seja mais possível obter um resultado de F1-Score maior do que o valor atual do conjunto $\mathcal{R}_o$.

O conjunto $\mathcal{R}_o$ representa o conjunto ideal de regras de extração de aspectos. Deve-se observar que é possível termos um conjunto $\mathcal{R}_o$ com apenas uma única regra que apresentou um valor alto de precisão e que depois não foi possível melhorar o valor de F1-Score combinando com as demais regras.

4. Experimentos

4.1. Métricas

Foram utilizadas as bastante conhecidas métricas de precisão (P), revocação (R) e F1-Score (F_1) para avaliar as regras de extração de aspectos. Seja *A* o conjunto de aspectos extraídos corretamente, de acordo com um conjunto de referência, e seja *B* o conjunto de aspectos extraídos pela regra que está sendo avaliada. Precisão (P), revocação (R) e F1-Score (F_1) foram definidos como:

$$P = \frac{|A \cap B|}{|B|} \quad (1) \qquad R = \frac{|A \cap B|}{|A|} \quad (2) \qquad F_1 = \frac{2 \times (P \times R)}{(P + R)} \quad (3)$$

4.2. Resultados e Discussão

A obtenção de resultados está diretamente ligada ao algoritmo guloso descrito na Seção 3.3. Inicialmente, cada regra foi executada individualmente na tarefa de extração de aspectos. A Tabela 3 apresenta os resultados alcançados por cada regra nos quatro domínios. Os melhores resultados foram destacados em negrito. A numeração das regras está de acordo com a numeração apresentada na Tabela 2.

4.3. Geração das regras ótimas

Com a aplicação do algoritmo guloso, foi possível selecionar as regras que pertencem ao conjunto ótimo de regras. Para o *dataset* de TA-Restaurantes, é possível notar que na Tabela 3 há mais de uma regra com precisão de 100% (regras 9, 13, 18 e 27). No entanto, a escolha deve ser feita considerando a maior revocação e, por isso, escolhe-se a regra 23 para compor o conjunto ótimo. Nos demais conjuntos de dados também foram escolhidas as regras com maior precisão: a) regra 18 em TV; b) regra 13 em ReLi; c) regra 23 em ReHol. Seguindo o algoritmo, o próximo passo é combinar este conjunto com cada uma das regras e, aquela cuja combinação resultar em um F1-Score maior pertencerá ao conjunto ótimo. Esse processo se repete até que as combinações resultem em F1-Scores menores ou iguais ao do conjunto.

A Tabela 4 apresenta uma comparação entre a regra com maior F1-Score e o conjunto de regras ótimas para cada conjunto de dados. Os resultados alcançados pelo conjunto de regras ótimas são superiores ao melhor resultado individual das regras em todos os domínios. O resultado obtido demonstra que houve um ganho de 10% de F1-Score em TA-Restaurantes, 6% de F1-score em TV, 4% de F1-score em ReLi e de 8%

Tabela 3. Resultados alcançados pelas regras nos domínios.

	TA-Restaurantes			TV			Reli			ReHol		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
1	0.03	0.00	0.01	0.06	0.02	0.03	0.04	0.02	0.03	0.03	0.01	0.02
2	0.44	0.03	0.06	0.42	0.04	0.08	0.43	0.06	0.11	0.71	0.07	0.12
3	0.00	0.00	0.00	0.16	0.02	0.04	0.20	0.05	0.08	0.21	0.03	0.06
4	0.00	0.00	0.00	0.16	0.02	0.04	0.20	0.05	0.08	0.21	0.03	0.06
5	0.00	0.00	0.00	0.08	0.00	0.01	0.15	0.02	0.04	0.13	0.01	0.02
6	0.72	0.22	0.34	0.49	0.14	0.21	0.37	0.17	0.24	0.67	0.20	0.31
7	0.35	0.31	0.33	0.25	0.25	0.25	0.12	0.22	0.15	0.36	**0.39**	**0.37**
8	0.30	0.15	0.20	0.28	0.19	0.23	0.19	0.20	0.20	0.29	0.18	0.22
9	**1.00**	0.00	0.01	0.14	0.00	0.01	0.06	0.00	0.00	0.05	0.00	0.00
10	0.76	0.17	0.29	0.55	0.12	0.20	0.44	0.12	0.19	0.77	0.18	0.30
11	0.35	**0.32**	0.33	0.27	**0.30**	0.28	0.15	**0.28**	0.20	0.34	0.34	0.34
12	0.33	0.08	0.13	0.36	0.12	0.18	0.20	0.14	0.17	0.36	0.10	0.16
13	**1.00**	0.02	0.05	0.60	0.02	0.04	**0.55**	0.01	0.03	0.83	0.03	0.07
14	0.46	0.11	0.18	0.68	0.12	0.20	0.29	0.08	0.13	0.59	0.12	0.20
15	0.50	0.01	0.03	0.18	0.00	0.01	0.14	0.00	0.01	0.50	0.01	0.02
16	0.69	0.15	0.24	0.59	0.15	0.25	0.23	0.08	0.12	0.58	0.12	0.20
17	0.50	0.02	0.05	0.40	0.03	0.06	0.18	0.03	0.06	0.56	0.04	0.07
18	**1.00**	0.01	0.03	**0.69**	0.02	0.04	0.30	0.01	0.02	0.66	0.01	0.02
19	0.28	0.03	0.06	0.42	0.09	0.15	0.18	0.07	0.10	0.38	0.06	0.11
20	0.05	0.00	0.01	0.10	0.03	0.05	0.00	0.00	0.00	0.09	0.02	0.03
21	0.42	0.20	0.27	0.39	0.20	0.26	0.14	0.13	0.14	0.43	0.24	0.30
22	0.10	0.02	0.04	0.04	0.01	0.02	0.01	0.00	0.00	0.11	0.03	0.05
23	**1.00**	0.10	0.18	0.62	0.08	0.14	0.49	0.05	0.10	**0.84**	0.10	0.18
24	0.57	0.30	**0.39**	0.47	0.22	**0.30**	0.23	0.19	0.21	0.51	0.25	0.34
25	0.68	0.20	0.31	0.49	0.14	0.21	0.36	0.18	**0.24**	0.67	0.19	0.30
26	0.80	0.03	0.07	0.50	0.01	0.03	0.34	0.01	0.03	0.82	0.04	0.08
27	**1.00**	0.07	0.14	0.61	0.06	0.12	0.43	0.06	0.11	0.76	0.09	0.16
28	0.54	0.16	0.25	0.51	0.15	0.23	0.15	0.08	0.10	0.31	0.09	0.14
29	0.50	0.22	0.31	0.44	0.18	0.26	0.24	0.17	0.20	0.52	0.21	0.30
30	0.27	0.12	0.16	0.29	0.16	0.21	0.16	0.17	0.10	0.36	0.17	0.23
31	0.26	0.12	0.16	0.30	0.17	0.21	0.15	0.08	0.10	0.35	0.17	0.23
32	0.16	0.01	0.03	0.42	0.06	0.11	0.18	0.19	0.12	0.40	0.05	0.09
33	0.00	0.00	0.00	0.28	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00
34	0.00	0.00	0.00	0.06	0.02	0.03	0.00	0.00	0.00	0.02	0.00	0.00
35	0.75	0.02	0.05	0.38	0.01	0.02	0.00	0.00	0.00	0.66	0.00	0.01
36	0.00	0.00	0.00	0.60	0.00	0.01	0.5	0.00	0.00	0.57	0.00	0.01
37	0.28	0.01	0.03	0.35	0.05	0.09	0.22	0.08	0.12	0.45	0.04	0.08
38	0.25	0.00	0.01	0.60	0.02	0.04	0.03	0.00	0.00	0.45	0.01	0.03
39	0.77	0.19	0.31	0.61	0.10	0.17	0.44	0.08	0.13	0.82	0.16	0.28
40	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
41	0.33	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00
42	0.05	0.01	0.02	0.05	0.03	0.04	0.00	0.00	0.00	0.02	0.01	0.01

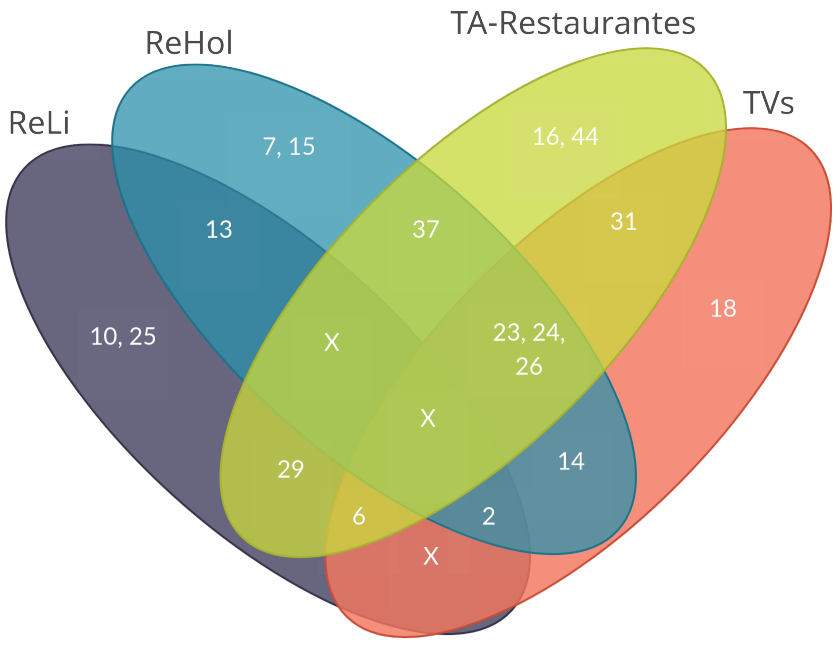
de F1-Score na ReHol. Assim, o resultado experimental corrobora com a literatura no sentido que o algoritmo guloso é uma estratégia eficiente de seleção de regras. Interessantemente, observa-se que as melhores regras individuais possuem uma relação de dependência (*amod* ou *nsubj*) entre substantivos e adjetivos.

Tabela 4. Resultados alcançados pela melhor regra individual versus o conjunto de regras ótimas.

Dataset	Regras individuais				Regras ótimas			
	Regra	Precisão	Revocação	F1-Score	Regras	Precisão	Revocação	F1-Score
TA-Restaurantes	24	0.57	0.30	0.39	6, 16, 23, 24 26, 29, 31, 37	0.53	0.46	0.49
TVs	24	0.47	0.22	0.30	2, 6, 14, 18 23, 24, 31, 26	0.39	0.33	0.36
ReLi	25	0.36	0.18	0.24	2, 6, 10 13, 25, 29	0.30	0.26	0.28
ReHol	07	0.36	0.39	0.37	2, 7, 13, 15, 14 23, 24, 26, 37	0.40	0.55	0.45

A distribuição do conjunto de regras ótimas é representado através do diagrama de Venn da Figura 4. Os números que aparecem são as regras descritas na Tabela 2. Os valores X indicam que não houve regras ótimas. Inicialmente, nota-se que os domínios ReHol e TA-Restaurantes apresentam o maior conjunto de regras quando comparado aos demais domínios. Uma possível razão para isso acontecer é devido ao fato que ReHol e TA-Restaurantes são domínios relacionados a avaliações turísticas e estas possuem um conjunto mais diversificado de aspectos. Consequentemente, surge a necessidade de mais regras de extração de aspectos para lidar com diferentes tipos de aspectos. É possível observar ainda que não existe uma regra ótima que pertença concomitantemente aos quatro domínios. Uma razão para isso é que os domínios possuem comentários bastante diferentes e, consequentemente, as regras aplicáveis têm uma utilidade diferente.

Figura 4. A distribuição do conjunto de regras ótimas entre os domínios.



A fim de se fazer uma generalização do método proposto, foi realizada uma análise com as regras que estão no conjunto ótimo de mais de um domínio. O critério de escolha foi selecionar as regras que participam como regras ótimas de três dos quatro conjuntos de dados. Assim, foram selecionadas as regras 2, 6, 23, 24 e 26. A Tabela 5 apresenta os resultados alcançados pelo conjunto geral de regras. Os resultados desse conjunto de regras são superiores aos resultados das regras individuais, mas inferiores aos resultados do conjunto de regras ótimas em cada domínio (ver Tabela 4). Apesar de não ser o melhor resultado possível em cada domínio, pode-se indicar essas regras como o conjunto geral de regras para diferentes domínios.

Tabela 5. Generalização do método.

Dataset	**Precisão**	**Revocação**	**F1-Score**
TA-Restaurantes	0.51	0.34	0.41
TVs	0.39	0.27	0.31
ReLi	0.22	0.26	0.24
ReHol	0.49	0.32	0.39

5. Conclusão

Este trabalho investigou uma abordagem de extração de aspectos para a língua portuguesa baseada em um vasto conjunto de regras. A hipótese proposta neste estudo é que seria possível adaptar regras que reconhecidamente funcionam bem em inglês para português. Os resultados alcançados indicam ser possível o uso de regras a partir de outro idioma. O estudo ainda contribui na avaliação de quais regras funcionaram melhor e pior no conjunto de dados. Foi ainda proposta uma segunda hipótese: a de que seria possível adotar uma estratégia gulosa com o objetivo de selecionar um conjunto ótimo de regras. Os resultados alcançados corroboram com a hipótese proposta.

Nos próximos passos da pesquisa, pretende-se utilizar o conjunto de regras ótimas que foram selecionadas nesse trabalho para treinar classificadores para extração de aspectos. O objetivo será desenvolver um método supervisionado à distância. Além disso, deseja-se testar o conjunto de regras com outras bases de dados em português, a fim de se obter mais resultados e garantir a robustez da abordagem.

Agradecimentos

Os autores ainda agradecem pelo apoio fornecido pela Universidade do Estado do Amazonas (UEA) através da Gratificação de Produtividade Acadêmica (GPA) (Portaria 086/2021).

Referências

Barros, J. M. and Bona, G. D. (2021). A deep learning approach for aspect sentiment triplet extraction in portuguese. In *Brazilian Conference on Intelligent Systems*, pages 343–358. Springer.

Cardoso, B. and Pereira, D. (2020). Evaluating an aspect extraction method for opinion mining in the portuguese language. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 137–144, Porto Alegre, RS, Brasil. SBC.

Costa, R. W. M. and Pardo, T. A. S. (2020). Métodos baseados em léxico para extração de aspectos de opiniões em português. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72. SBC.

da Silva, A. S. (2021). Integração de dados factuais e subjetivos: Um estudo de caso em comércio eletrônico. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 241–252. SBC.

Freitas, C., Motta, E., Milidiú, R., and César, J. (2012). Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. *Encontro de Linguística de Corpus*, 11:22.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Law, D., Gruss, R., and Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67:84–94.

Liu, Q., Gao, Z., Liu, B., and Zhang, Y. (2016). Automated rule selection for opinion target extraction. *Knowledge-Based Systems*, 104:74–88.

Nascimento, R. S., Nascimento, G., Carvalho, F., and Guedes, G. (2020). Mineração de opiniões com liwc: abordagem prática sobre sistemas judiciais eletrônicos brasileiros. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 132–141. SBC.

Oliveira, M. V. and de Melo, T. (2020). Investigating sets of linguistic features for two sentiment analysis tasks in brazilian portuguese web reviews. In *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 45–48. SBC.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Poria, S., Cambria, E., Ku, L.-W., Gui, C., and Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37.

Rana, T. A. and Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review*, 46(4):459–483.

Shafie, A. S., Sharef, N. M., Murad, M. A. A., and Azman, A. (2018). Aspect extraction performance with pos tag pattern of dependency relation in aspect-based sentiment analysis. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–6. IEEE.

Testoni, G. A., Souza, M. P., Freire, P. M. S., and Goldschimidt, R. R. (2021). Um método linguístico que combina polaridade, emoção e aspectos gramaticais para detecção de fake news em inglês. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 151–162. SBC.

Tubishat, M., Idris, N., and Abushariah, M. (2021). Explicit aspects extraction in sentiment analysis using optimal rules combination. *Future Generation Computer Systems*, 114:448–480.