UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PPG EM INFORMÁTICA APLICADA

AVALIAÇÃO DE CARACTERÍSTICAS PARA EXTRAÇÃO AUTOMÁTICA DE ASPECTOS

Roberto Márcio Mota de Lima

Trabalho apresentado ao Programa de Pós-graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco como requisito parcial para obtenção do grau de Mestre em Informática Aplicada

ORIENTADOR: PROF. DR. RAFAEL FERREIRA LEITE DE MELLO

CO-ORIENTADOR: PROF. DR. RINALDO JOSÉ DE LIMA

RECIFE-PE – AGOSTO/2018

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

M827a Morais, Renê Douglas Nobre de.
Avaliação de características para extração automática de aspectos / Roberto Márcio Mota de Lima. – Recife, 2018.
51 f.: il.

Orientador: Rafael Ferreira Leite de Mello.
Coorientador: Rinaldo José de Lima.
Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2018.
Inclui referências e apêndice(s).

1. Banco de dados - Pesquisa 2. Inteligência artificial 3. Processamento eletrônico de dados 5. Sistemas especialistas (Computação) 6. Mineração de dados (Computação) I. Mello, Rafael Ferreira Leite de, orient. II. Lima, Rinaldo José de, coorient. III. Título

CDD 004

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PPG EM INFORMÁTICA APLICADA

AVALIAÇÃO DE CARACTERÍSTICAS PARA EXTRAÇÃO AUTOMÁTICA DE ASPECTOS

Roberto Márcio Mota de Lima

Dissertação julgada adequada para obtenção do título de Mestre em Informática Aplicada, defendida e aprovada em 31/08/2018 pela Comissão examinadora

Orientador:

_____

Prof. Dr. Rafael Ferreira Leite de Mello - UFRPE

Banca Examinadora:

_____

Prof. Dr. Rinaldo José de Lima – UFRPE

_____

Prof. Dr. Renato Fernandes Corrêa - UFPE

_____

Prof. Dr. Rafael Dueire Lins – UFRPE

RECIFE-PE – AGOSTO/2018

A meu avô, Antônio, que no decorrer da escrita desta dissertação acometeu-se de mal grave, indo para os braços de Deus, mas que permanecerá em meu coração e de minha família, DEDICO.

Agradecimentos

Agradeço, primeiramente, e acima de tudo, a Deus, por ter me dado serenidade, sabedoria e confiança nos momentos que mais Lhe pedi;

A meus pais, Dulce e Roberto, por terem investido no meu futuro, desde sempre, acreditando nas minhas capacidades e ideais;

Aos meus orientadores, professor Rafael Ferreira e professor Rinaldo Lima, pelo apoio incondicional que me deram durante toda a pesquisa, por estarem sempre presentes, ensinando-me, aconselhando-me e incentivando-me a sempre prosseguir;

A todos os meus familiares e amigos, que sempre me apoiaram em todos os meus projetos de vida;

Ao professor Tiago Ferreira, pelo grande apoio.

Aos professores, colegas de pesquisa e demais funcionários do PPGIA.

À Defensoria Pública e ao Tribunal de Justiça de Pernambuco por terem feito cumprir-se a Lei.

À FACEPE pelo apoio financeiro.

O que sabemos é uma gota; o que ignoramos é um oceano.

Isaac Newton

RESUMO

A utilização cada vez mais crescente da Internet e demais interações *online* entre pessoas, tais como *chats*, participação em fóruns, transações em comércio eletrônico, revisões de produtos e serviços, etc., tem levado à necessidade cada vez maior de extrair, transformar e analisar uma quantidade enorme de dados, utilizando-se de uma combinação de processos de mineração de textos e demais conteúdos obtidos diretamente da Web. Dentre as principais demandas para a área de mineração de dados, sobretudo de opinião, estão as grandes empresas dos mais diversos ramos. Essas instituições estão cada vez mais interessadas em saber o que seus clientes reais, ou em potencial, comentam sobre elas. De restaurantes a hotéis, de celulares a câmeras fotográficas, os fóruns de revisões espalhados na internet são os cartões de visita mais importantes para as empresas pelo simples fato de não serem criados por elas, mas por pessoas que de alguma forma fazem ou farão uso de seus serviços. As pessoas são mais propensas a expressar suas opiniões e experiências práticas em produtos ou serviços que eles utilizaram. Essas revisões são importantes para organizações empresariais e consumidores. Contudo, analisar todas as críticas de clientes é difícil, já que tal número de comentários pode ser de centenas ou até milhares. Portanto, é necessário fornecer informações coerentes e sumários concisos para essas revisões. A Análise de Sentimento Baseada em Aspecto é uma tendência recente e uma abordagem que tem muito a ser explorada, uma vez que tem demonstrado bons resultados na literatura como uma técnica de extração de opiniões mais refinada e pontual do que as baseadas puramente em léxica e regras. Esta dissertação de Mestrado teve por objetivo pesquisar, implementar e avaliar um novo método para extração de termos de opinião em revisões de restaurantes levando em conta e escolha da algumas das melhores características descritas no estado da arte. Como modelo de classificação, foi utilizado o CRF, dada sua alta eficiência como classificador condicional. Os resultados obtidos demonstraram um bom desempenho quando comparado aos principais trabalhos da área, tendo como destaque a alta cobertura alcançada pelo método desenvolvido.

Palavras-chave: Análise de Sentimento Baseada em Aspectos, CRF, extração de características, termo de opinião.

ABSTRACT

The increasing use of the Internet and other online interactions between people, such as chats, forum participation, e-commerce transactions, reviews of products and services, among others has led to the increasing need to extract, transform and analyze a vast amount of data, using a combination of text mining processes and others directly from the Web. Companies from several different sectors demand for customer feedback. Such institutions are increasingly interested in knowing what their real, or potential, customers say about them. From restaurants to hotels, from smartphones to cameras, the reviews are spread out over the internet, and they are essential to companies because they are created by people who somehow make, or use their services. People are more likely to express their opinions and practical experiences in products or services that they have used. Such feedback is important to business organizations and consumers. However, analyzing the hundreds or even thousands of customer input is difficult to be handled by humans. Therefore, it is necessary to provide concise information and concise summaries of such reviews. The Aspect-Based Sentiment Analysis is a recent trend and an approach that has much to explore since it has demonstrated relevant results in the literature as a more refined and punctual opinion extraction technique than those based purely on lexicon and rules. This MSc thesis aimed to research, implement and evaluate a new method for extraction of opinion terms in restaurant reviews taking into account and choosing some of the best features described in the state of the art. As the classification model, CRF was used, given its high efficiency as a conditional classifier. The results obtained showed good performance when compared to the principal related works of the area, highlighting the high coverage achieved by the developed method.

Keywords: *Aspect Based Sentiment Analysis, Conditional Random Fields, feature extraction, Opinion Term Expression*

Lista de Figuras

Figura 1: Mineração de textos 19
Figura 2: Principais abordagens em Análise de Sentimentos 22
Figura 3: A Análise de Sentimento Baseada em Aspectos e suas abordagens 30
Figura 4: *Content* e *gramatical words* 44
Figura 5: Uma sentença na ferramenta BRAT, anotada com quatro termos de aspecto 59
Figura 7: Modelo de Análise de dependência 68
Figura 8: Trecho do arquivo de treinamento gerado pela aplicação *Python* 73
Figura 9: Ferramentas destinadas à implementação do modelo *Conditional Random Fields* 74
Figura 10: Trecho de um dos *templates* utilizados na fase de treinamento do classificador CRF++ 76
Figura 11: Ilustração representativa do *token* corrente e condicionais 76
Figura 12: Trecho ilustrativo do arquivo de *template* 77

Lista de Tabelas

Tabela 1: Resumo dos principais trabalhos na área de extração de aspectos com classificador CRF 56
Tabela 2: *PoS Tag Table* 61
Tabela 3: Tabela parcial tratada de arquivo de saída do classificador CRF 80
Tabela 4: Resultados da aplicação de *Information Gain* sobre as características utilizadas 85
Tabela 5: Chi quadrado. Classificação pelo método Chi quadrado das características mais relevantes 86
Tabela 6: Comparação das métricas de avaliação em 3 cenários diferentes 88
Tabela 7: Comparação dos resultados de precisão, cobertura e medida-F com os obtidos na SemEval2014 89

Lista de Quadros

Quadro 1: Descrição dos códigos das características utilizadas 84

Lista de Acrônimos

ABSA – *Aspect Based Sentiment Analysis*

AS – Análise de Sentimento

ASBA – Análise de Sentimento Baseada em Aspectos

OTE – *Opinion Target Expression*

PoS – *Part of Speech*

CRF – *Conditional Random Fields*

NLP – *Natural Language Processing*

MT – Mineração de Textos

NER – *Named Entity Recognition*

PLN – Processamento de Linguagem Natural

SVM – *Support Vector Machine*

TF-IDF– *Term Frequency-Inverse Document Frequency*

AM – Aprendizagem de Máquina

IA – Inteligência Artificial

DL – *Deep Learning*

pLSA – *Probabilistic latent semantic analysis*

LDA – *Latent Dirichlet allocation*

EM – Máxima Expectativa

NLTK – *Natural Language Toolkit*

Sumário

1. Introdução 14
1.1. Objetivos 16
1.1.1. Objetivo Geral 16
1.1.2. Objetivos Específicos 16
1.2. Organização do trabalho 16
2. Fundamentação Teórica 18
2.1. Extração de Informação e Mineração de Textos 18
2.2. Conjunto de dados (*dataset*) 20
2.3. Análise de Sentimento 20
2.3.1. Principais abordagens utilizadas em Análise de Sentimento 21
2.3.1.1. Abordagem Baseada em Dicionário (Léxica) 22
2.3.1.2. Abordagem Baseada em Aprendizagem de Máquina 24
2.3.1.3. O modelo *Conditional Random Fields* (CRF) 26
2.3.2. Níveis de classificação em Análise de Sentimento 28
2.3.3. Análise de Sentimento Baseada em Aspectos (ASBA) 30
2.3.3.1. Tarefa de Extração de Aspectos 31
2.3.3.2. Principais características (*features*) utilizadas em ABSA com CRF 32
3. Revisão da Literatura 49
3.1. Trabalhos Relacionados 49
4. Método 58
4.1. Coleta e estrutura do banco de dados (*dataset*) 58
4.2. Entrada de dados 59
4.2.1. Preparação dos *datasets* 59
4.2.1.1. Processo de anotação 59
4.2.2. Formato do *dataset* 61
4.3. Extração de Características 62
4.4. Geração do arquivo de saída 72
4.5. Métodos para avaliar importância das características 73
4.6. Classificação 74
4.6.1. O CRF++ 75

4.6.1.1. Etapa de treinamento 78
4.6.1.2. Etapa de classificação 79
5. Resultados 81
5.1. Medidas de avaliação de desempenho 81
5.2. Avaliação da importância das características 83
5.3. Resultados da classificação de aspectos 87
6. Conclusão 91
Bibliografia 93
Anexo 99

1

Introdução

A utilização cada vez mais crescente da Internet e demais interações online entre pessoas, tais como *Facebook*, chats, participação em fóruns, transações em e-commerce, revisões de produtos e serviços, etc., tem levado à necessidade cada vez maior de extrair, transformar e analisar uma quantidade enorme de dados estruturados e não estruturados [1].

Um exemplo desse crescimento foi a rápida expansão do comércio eletrônico, onde além de compras, as pessoas estão mais propensas a expressar suas opiniões e experiências práticas em produtos ou serviços dos quais fazem uso. Essas revisões são importantes para organizações empresariais e consumidores. As empresas podem decidir sobre suas estratégias de marketing e melhoria para seus próximos lançamentos; os clientes podem tomar uma decisão melhor ao comprar produtos ou utilizar serviços.

Contudo, para grande parte das empresas, ler todas as críticas de clientes é uma tarefa difícil, especialmente para itens populares, onde o número de comentários pode ser de centenas ou até milhares. Portanto, é necessário fornecer informações coerentes e sumários concisos para essas revisões [2], uma vez que tais instituições estão cada vez mais interessadas em saber o que seus clientes reais, ou em potencial, comentam sobre elas.

No tocante à extração de informações, a mineração de dados é uma das tecnologias mais promissoras da atualidade [3]. Um dos fatores desse sucesso é o fato de que milhões de dólares são gastos por empresas na coleta de dados, mas sem que por algumas vezes nenhuma informação útil consiga ser identificada. De restaurantes a hotéis, e de celulares a câmeras fotográficas, os fóruns de revisões espalhados pela internet são os cartões de visita mais importantes para as empresas pelo simples fato de não serem criados por elas, mas por pessoas que de alguma forma fazem ou farão uso de seus serviços. Uma área que vem trabalhando nesse nicho é a mineração de opinião, também, e mais conhecida, como

Análise de Sentimento, um ramo da Inteligência Artificial que trata da identificação e extração de opiniões ou emoções dos usuários, expressas em diferentes blogs, sites sociais, fóruns de discussão, websites de empresas, etc... [4]

Como uma subárea da Mineração de Opiniões, a Análise de Sentimento Baseada em Aspectos (ASBA) é uma tendência recente e uma abordagem que tem muito a ser explorada, uma vez que tem demonstrado bons resultados na literatura como uma técnica de extração mais refinada e pontual do que as técnicas baseadas puramente em léxica e regras [4].

Apesar dos avanços na área, a Análise de Sentimentos traz muitos desafios ainda não solucionados, ou apenas parcialmente resolvidos. É laborioso estabelecer com clareza qual o estado da arte em classificação de opiniões, sobretudo por conta de fatores como domínio (revisões sobre filmes, hotéis, restaurantes, produtos, serviços, política, etc), tipo de texto (*tweets*, revisões, postagens em fóruns, etc), nível textual (documento, sentença ou aspectos), dentre outros, o que é demonstrado por algumas revisões na área, como os trabalhos de Ravi & Ravi [1]. No entanto, a demanda por métodos mais apurados e precisos para extração de opinião exige que abordagens mais "inteligentes" como as baseadas em Aspectos sejam aprimoradas. Diante deste contexto, este trabalho busca analisar especificamente uma subtarefa da área de ASBA: **a extração de aspectos**.

Em Análise de Sentimento Baseada em Aspecto, um aspecto (também chamado de termo de opinião - OTE) é a **palavra alvo** à qual a opinião (positiva, negativa, ou neutra) se refere. Por exemplo, na sentença: "A comida deste restaurante é excelente, mas o atendimento não é bom", os aspectos são "restaurante" e "atendimento".

Diante deste contexto, este trabalho apresenta a pesquisa e aplicação de diferentes características para extração de aspectos no domínio de restaurantes com a elaboração de um método baseado no modelo de classificação CRF (Campos Aleatórios Condicionais).

1.1. Objetivos

1.1.1 – Geral

Esta pesquisa objetiva aplicar um método supervisionado baseado no modelo CRF para extração automática de aspectos em revisões de restaurantes, fazendo uso das principais características descritas na literatura na área de Análise de Sentimento Baseada em Aspectos (ASBA).

1.1.2 - Específicos

- Levantar o estado da arte em ASBA, com foco direcionado para as características e classificadores mais utilizados;
- Fazer um levantamento de recursos semânticos e métodos de extração de características relacionadas aos objetivos da proposta;
- Implementar um método computacional para extração de características em ambiente Python;
- Investigar a influência de diferentes características sobre o desempenho;
- Conduzir experimentos para avaliar a efetividade do método proposto comparando com outros já publicados na literatura, em particular com os obtidos no SemEval 2014;

1.2. Organização do trabalho

Este trabalho se divide da seguinte forma: o Capítulo 1 apresenta uma breve introdução sobre os assuntos abordados e os principais objetivos da pesquisa. Em seguida, o Capítulo 2 aborda os conceitos necessários para um melhor entendimento deste documento por parte do leitor. Já o Capítulo 3 apresenta uma breve revisão do estado da arte, citando os principais trabalhos que se relacionam ao tema desta pesquisa.

Contextualizados os temas introdutórios, no Capítulo 4 serão abordados com detalhes os métodos aplicados no desenvolvimento da presente pesquisa. O Capítulo 5, por sua vez, apresenta os resultados encontrados e uma discussão dos mesmos. O Capítulo 6 elenca as conclusões consequentes da pesquisa, assim como as considerações finais e trabalhos futuros.

2

Fundamentação Teórica

2.1. Extração de Informação e Mineração de Textos

Como já citado, a mineração de dados é uma das tecnologias mais promissoras da atualidade. O principal objetivo da Mineração de Textos (MT) consiste na extração de características em uma grande quantidade de dados não estruturados.

As técnicas utilizadas na mineração de textos são semelhantes às aplicadas em mineração de dados, ou seja, fazem o uso dos mesmos métodos de aprendizagem, independente se uma técnica utiliza-se de dados textuais (MT) e a outra de dados numéricos (MD) [5].

Podemos diferenciar as duas técnicas a partir de dois esclarecimentos básicos: enquanto a Mineração de Dados é caracterizada por extrair informações implícitas, anteriormente desconhecidas, contudo potencialmente úteis, na Mineração de Textos, a informação que se deseja extrair é clara, sendo apresentada em forma de textos. Porém, essa informação não é exprimida de uma maneira que seja passível de processamento automático [6].

De fato, estamos vivenciando o crescimento acelerado de informações não estruturadas (*textos, tweets,* revisões, *posts, etc*). Com isso, a Mineração de Textos ganha espaço não somente no meio acadêmico, mas também no mundo dos negócios e até político, o que já é percebido em debates recentes, com estatísticas sobre popularidade de um candidato nas redes sociais, e os temas mais abordados pelos internautas.

Por vezes, muitas das informações extraídas no processo de mineração são inúteis às empresas. Em seu trabalho, Han [7] refere-se a isso como "dados ricos, pobres em informação." Além do setor privado, o setor público e o terceiro setor também podem se beneficiar da Mineração de Dados [8]. Um dos campos responsáveis pela evolução de

técnicas de Mineração de Dados/Textos é o de Processamento de Linguagem Natural (PLN).

Resumidamente, o processo de Mineração de Textos é dividido em quatro etapas bem definidas: **seleção, pré-processamento, mineração e assimilação** (Figura 1).

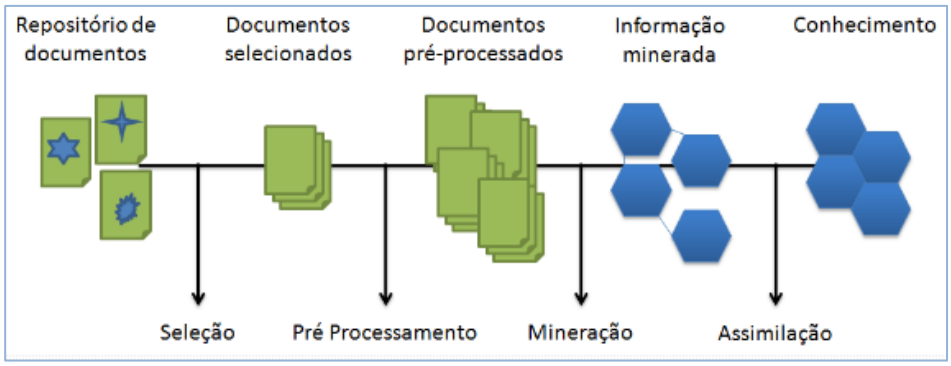


Figura 1: Mineração de textos. Fonte: [9]

Tecnologias baseadas em PLN estão se tornando cada vez mais generalizadas. Ao fornecer interfaces homem-máquina mais naturais, e acesso mais sofisticado a informações armazenadas, o processamento de linguagem natural desempenha um papel central na sociedade da informação multilíngue [10], sendo o principal objetivo da Mineração de Textos (MT) a extração de características em uma grande quantidade de dados não estruturados.

2.1.1. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área da Ciência da Computação e Inteligência Artificial focada nas interações entre os humanos, que se comunicam através de linguagem natural (idioma falado ou escrito), e o computador. Mais especificamente, é a área destinada a inferir dados e informação a partir de um grande conjunto de dados disponibilizados em linguagem natural, seja ela falada ou escrita.

Devido à escalabilidade dos gerenciadores de bancos de dados em armazenar informações, diversos sistemas conseguiram manter disponíveis textos em formato de documentos sem problemas de demanda, acesso e disponibilidade dos dados. Todavia, com o aumento exponencial de documentos circulando em diversos tipos de sistemas, mesmo os computadores modernos podem não comportar essa massa de dados, tendo que restringir a

representação a um conjunto limitado de termos [11]. Ao conjunto limitado de textos estruturados sobre um determinado tópico dá-se o nome de conjunto de dados, ou de forma mais comumente utilizada, do Inglês, *dataset.*

2.2. Conjunto de dados (*dataset*)

Um dos motivos que tem tornado possível o aumento de pesquisas na área de PLN é a disponibilidade cada vez maior de bibliotecas de dados estruturados (*datasets*).

Os *datasets* formam a base de qualquer análise de dados de alto nível [12]. Atualmente há diferentes variedades de *datasets* disponíveis e acessíveis a pesquisadores de todo o mundo. Esse conjunto de dados possui características que diferem entre si, e vão desde a natureza do domínio a que se aplicam à formatação do arquivo. A especificidade de cada *dataset* é requerida ao se trabalhar com sistemas de abordagens diferentes (aprendizado de máquina, supervisionado e não supervisionado, uso de léxica, regras, etc), em diferentes idiomas, ou para uma necessidade específica. Algumas técnicas de aprendizagem de máquina, como as baseadas em redes bayesianas requerem o uso de *corpora* anotados (conjunto de dados previamente classificados em categorias específicas) para o treinamento do classificador.

O papel de marcação das categorias presentes em um *corpus* é feito muitas vezes de forma manual por especialistas em determinado domínio, o que demanda várias horas de trabalho e um custo inerente. Dependendo do tamanho do *dataset*, a tarefa de classificação pode demorar meses, e o custo ser bastante elevado. Por esse motivo, a variedade de *datasets* gratuitos disponíveis é modesta, o que justifica em partes a presença de um mesmo conjunto de *datasets* em centenas de artigos já publicados. A outra parte que justifica o uso de um mesmo *corpus* por diversos pesquisadores é o objetivo de manutenção dessa variável para comparação entre técnicas e avanço nos *baselines* – em sistemas de classificação multidomínio, entretanto, a aplicação de mais de um *corpus* pode ser requerida a fim de contemplar os objetivos de pesquisa. Um dos segmentos da área de Inteligência Artificial que mais fazem uso de *datasets* é a Análise de Sentimento.

2.3. Análise de Sentimento

A Análise do Sentimento (AS) é a tarefa de detectar, extrair e classificar opiniões, sentimentos e atitudes sobre diferentes tópicos. A AS contribui para vários objetivos, como observar o humor do público em relação à política, na inteligência de mercado, na avaliação de satisfação dos clientes de determinada empresa, e muito mais [1]. Assim, a AS é o processo inteligente de determinar quando uma sentença, frase ou palavra é considerada positiva, negativa ou neutra [13], ou se faz parte de uma determinada categoria, dentro de um certo escopo.

Trabalhos mais recentes em Análise de Sentimento podem ser classificados de diferentes pontos de vista: técnica usada, vista do texto, nível de detalhe de análise de texto, nível de classificação, etc. De um ponto de vista técnico, identificam-se como uns dos principais exemplos a aprendizagem de máquina, baseada em léxica, e abordagens baseadas em regras.

Pang e Lee [14] realizaram uma extensa pesquisa de mais de trezentos *papers* cobrindo aplicações, desafios comuns para Análise de Sentimento, tarefas principais de mineração de opinião, extração de opinião, classificação de sentimento, determinação de polaridade e sumarização. Esses trabalhos serviram como base para diversos artigos e revisões, dentre eles os trabalhos de Kumar e Vadlamani Ravi (2015) [1], uns dos mais completos da área.

2.3.1. Principais abordagens utilizadas em Análise de Sentimento

A Análise de Sentimento apresenta duas abordagens principais: a léxica e a de aprendizagem de máquina (Fig.1).

O método de Aprendizagem de Máquina (AM) usa vários algoritmos de aprendizagem para determinar o sentimento por uma formação sobre um conjunto de dados conhecido.

É importante destacar que a AM proporciona uma precisão máxima enquanto a abordagem por Orientação Semântica (léxica) fornece uma melhor generalidade.

A abordagem baseada em léxica envolve o cálculo de polaridade de sentimento a partir de uma revisão usando a orientação semântica de palavras ou frases na avaliação. O método baseado em dicionários usa um dicionário existente, que é uma coleção de palavras

de opinião, juntamente com a marcação positiva (+) ou negativa (-) de acordo com o sentimento. Os dicionários podem ser criados com ou sem o uso de ontologia (relacionamentos). A abordagem baseada em regras, por sua vez, procura palavras de opinião em um texto e, em seguida, classifica-o com base no número de palavras positivas e negativas [15].

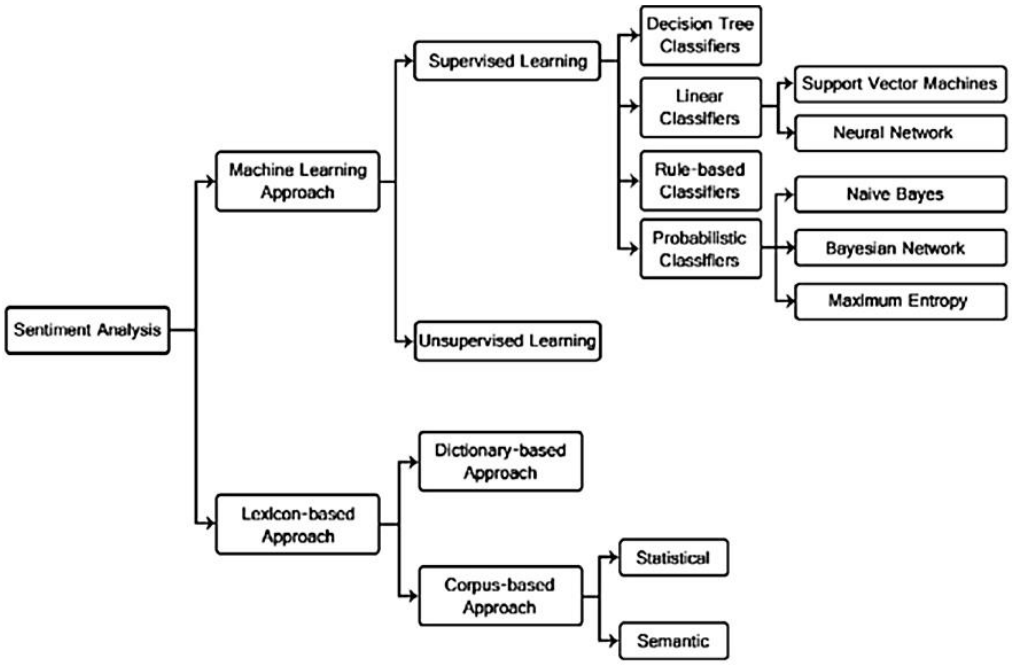


Figura 2: Principais abordagens em Análise de Sentimentos. Fonte: [1]

2.3.1.1. Abordagem Baseada em Dicionário (Léxica)

A abordagem baseada em *dicionário* é também denominada *léxica* ou *linguística*. O princípio central desta abordagem é o uso de léxicos, ou seja, dicionários de sentimentos, que são compilações de palavras ou expressões de sentimento associadas à respectiva polaridade [16].

Um dos métodos mais utilizados na abordagem linguística é o da ocorrência conjunta entre alvo e sentimento, onde não são levadas em consideração nem a ordem dos termos dentro de um documento (*bag-of-words*), nem suas relações sintáticas. Para a classificação do sentimento em um texto, basta que exista uma palavra de sentimento, onde

sua polaridade é dada a partir de um dicionário de sentimentos. Na literatura da área, as palavras de opinião também são conhecidas como palavras de sentimento. Palavras de opinião positiva são usadas para expressar alguns estados desejados, enquanto as de opinião negativa são usadas para expressar alguns estados indesejáveis. Quando ocorrem de forma conjunta e coletiva, as palavras de opinião são chamadas de léxico da opinião [17].

Esse método de coocorrência é extensamente empregado para a correlação de um sentimento a uma entidade em uma sentença. Por exemplo, na sentença "o restaurante é muito bom", a polaridade positiva da palavra "bom" é associada à entidade restaurante. O método por coocorrência apresenta bons resultados quando o nível de análise textual é de granularidade pequena, ou seja, frases curtas (um *tweet*, por exemplo), pois a palavra detentora do sentimento está **próxima** à entidade que qualifica. Quando aplicada em nível de maior granularidade (um *review* extenso, ou um documento, etc.), estabelece-se algum tipo de **média** sobre as palavras de sentimento encontradas [16].

A Equação 1 [16] apresenta um exemplo de função para determinação de polaridade de um documento "D", sendo "S_w" a polaridade de uma palavra "w" em um léxico. Estes cálculos podem levar em conta funções de peso e de modificação. A função *peso* () pode ser, por exemplo, alguma medida de distância entre a palavra de sentimento e o alvo, ou de importância da palavra no texto (ex.: frequência). A função *modificador* () pode ser usada para tratar negações, palavras de intensidade (e.g. muito), etc. Esta função de agregação também pode ser estendida a sentenças, cujas cláusulas podem combinar diferentes palavras de sentimento. A função *modificador* () em especial tem caráter singular, sobretudo quando da ocorrência te palavras negativas, dada sua capacidade de alterar completamente a polaridade de uma sentença. As palavras negativas serão estudadas mais detalhadamente em seção posterior deste trabalho.

$$S(D) = \frac{\sum_{w \in D} S_w \cdot peso(w) \cdot modificador(w)}{\sum peso(w)} \quad (1)$$

Existem métodos linguísticos mais complexos, como a utilização de *parsers* linguísticos, que têm como propósito analisar o texto e aumentar a qualidade da classificação com base em informações morfossintáticas ali presentes (ex.: sujeito, predicado, dependências, funções sintáticas, etc.). No entanto, ferramentas de processamento de linguagem natural são em sua maioria restritas a determinado idioma, a maioria em Inglês. A maioria dos léxicos existentes também são dependentes de idioma, e foram feitos estritamente também para a língua inglesa. Alguns exemplos de léxicos são o *General Inquirer* [18], o *Opinion Finder* [19], o *WordNetAffect* [20] e o *SentiWordNet* [21], este último utilizado no desenvolvimento do presente trabalho.

A composição básica de um léxico de sentimento é a palavra de sentimento e sua respectiva polaridade, expressa como uma categoria, ou como um valor em uma escala. Muitos dicionários possuem adicionalmente: o *lemma* e o *stemming* de cada entrada; a categoria gramatical (*Part-of-Speech - POS*); e o alvo do sentimento (predicado ou sujeito).

A maioria dos dicionários disponíveis são genéricos, ou seja, auxiliam na tarefa de classificação, independentemente do domínio dos textos sendo considerados. Entretanto, os melhores resultados obtidos na tarefa de classificação foram baseados em dicionários dependentes de contextos [22], criados a partir de palavras semente e expandidos utilizando o WordNet [23] ou tesauros [16].

2.3.1.2. Abordagem Baseada em Aprendizagem de Máquina

As técnicas computacionais baseadas em aprendizado de máquina têm como objetivo principal descobrir automaticamente regras gerais, em grandes conjuntos de dados, que permitam extrair informações implicitamente representadas, e aplicar esse aprendizado na predição de novas informações a partir de um novo conjunto de dados.

Os métodos baseados em Aprendizagem de Máquina podem ser divididos basicamente em 2 abordagens: **aprendizagens não supervisionadas** (para entradas não rotuladas) e **aprendizagens supervisionadas** (que "aprendem" a partir de entradas rotuladas).

- ***Aprendizagem Não-Supervisionada:*** *envolvem padrões de aprendizado na entrada quando não são fornecidos valores de saída específicos* [24], *isto significa que o classificador recebe apenas um conjunto de exemplos. Métodos não supervisionados também podem ser usados para rotular um corpus que pode ser usado posteriormente por um classificador de aprendizagem supervisionada* [25]. *Um agente puramente baseado em aprendizado não supervisionado não pode se direcionar ao que fazer, porque não tem informação sobre o que constitui uma ação correta ou um estado desejável* [24]. *Exemplos de métodos de aprendizado não supervisionados são clustering (k-means) ou cluster análise, que discerne várias categorias em uma coleção de objetos* [24] *e o algoritmo de maximização da expectativa, um algoritmo para encontrar a máxima probabilidade de exemplos* [26].

- ***Aprendizagem Supervisionada:*** *técnicas supervisionadas de aprendizado de máquina implicam o uso de um corpus de treinamento rotulado para função de classificação* [25] *e envolvem aprender uma função a partir de exemplos de entradas e saídas* [24]. *A saída desta função é um valor contínuo ('Regressão') ou pode prever uma categoria ou rótulo do objeto de entrada ('classificação'). Exemplos de classificadores supervisionados são o Naïve Bayes (chamado ingênuo porque assume que as probabilidades sendo combinados são independentes umas dos outras); Máxima Entropia (parametrizado por um conjunto de pesos que são usados para combinar os recursos conjuntos que são gerados a partir de um conjunto de recursos por um codificador); Árvores de Decisão (uma árvore na qual os nós internos são rotulados pelas características, o as bordas que saem de um nó são rotuladas por testes no peso da característica, e as folhas são rotuladas por categorias); SVM (muitas vezes considerado como o classificador que produz os mais altos resultados de precisão em problemas de classificação de texto. Eles operam construindo um hiperplano com a máxima distância euclidiana aos exemplos de treinamento mais próximos)* [26] *e o CRF, um*

tipo de modelo gráfico probabilístico não direcionado e discriminativo, que é usado para codificar relações conhecidas entre observações e construir interpretações consistentes. O modelo CRF será estudado mais detalhadamente no tópico seguinte.

2.3.1.3. O modelo *Conditional Random Fields* (CRF)

Algo fundamental para muitas aplicações é a capacidade de prever múltiplas variáveis que dependem umas das outras. Tais aplicações são tão diversas como pontos de classificação de uma imagem [27], como a tarefa de segmentar genes em uma cadeia de DNA [28] ou realizar a análise sintática de um texto em linguagem natural [29].

Em tais aplicações, nós desejamos predizer um vetor de saída y = {y0, y1, ..., yT} de variáveis aleatórias a partir de um vetor de característica x observado. Um exemplo relativamente simples do processamento de linguagem natural (PLN) é a marcação (*tag*) do PoS (*Parto of Speech*), ou classe gramatical da palavra, na qual cada variável y_s é o PoS-tag da palavra na posição s (ex.: NN – substantivo, ADV-adjetivo, etc.), e a entrada x é dividida em vetores de características {x0, x1,..., xT}. Cada xs contém várias informações sobre a palavra na posição s, como identidade, recursos ortográficos como prefixos e sufixos, associação em léxicos específicos de domínio e informações em bancos de dados semânticos como o *WordNet*.

Uma abordagem para este problema de previsão multivariada, especialmente se nosso objetivo é maximizar o número de etiquetas que são corretamente classificadas, é aprender um classificador independente por posição que mapeia $x \rightarrow y_s$ para cada s. A dificuldade, no entanto, é que as variáveis de saída têm **dependências complexas**. Por exemplo, em inglês, os adjetivos em muitos casos, não seguem substantivos (ex: "*She is nice*", "*The hospital was full of people when we arrived*"), e em visão computacional, regiões vizinhas em uma imagem, por exemplo, tendem a ter rótulos semelhantes. Outra dificuldade é que as variáveis de saída podem representar uma estrutura complexa, como

uma *parse tree*, na qual a escolha de qual regra gramatical usar próximo ao topo da árvore pode ter um grande efeito no resto da árvore.

Uma maneira natural de representar o modo pelo qual variáveis de saída dependem uma da outra é fornecido por **modelos gráficos**. Modelos gráficos - que incluem uma variedade de modelos de famílias como as redes Bayesianas, redes neurais, gráficos de fatores, campos aleatórios de Markov, e outros - representam uma distribuição complexa sobre muitas variáveis como um **produto de fatores locais em subconjuntos menores de variáveis.** Então é possível descrever como uma dada fatoração da densidade de probabilidade corresponde a um conjunto particular de relações de independência condicional satisfeitas pela distribuição. Esta correspondência faz a modelagem muito mais conveniente, porque muitas vezes o nosso conhecimento de o domínio sugere suposições razoáveis de independência condicional, que então determinam nossa escolha de fatores [30].

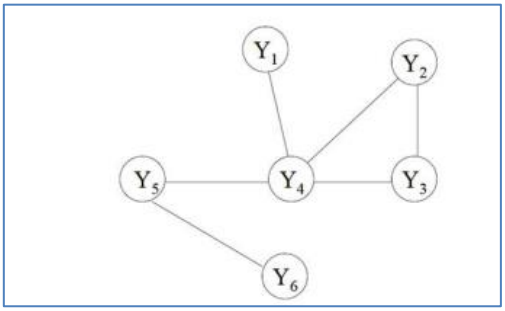
Muito trabalho na aprendizagem com modelos gráficos, especialmente no processamento estatístico de linguagem natural, tem se focado em modelos generativos que explicitamente tentam modelar uma distribuição de probabilidade conjunta p (y, x) sobre entradas e saídas. Embora esta abordagem tenha vantagens, ela também tem limitações importantes. Não só a dimensionalidade de x pode ser muito grande, mas as características (*características*) podem ter dependências complexas, portanto é difícil construir uma distribuição de probabilidade sobre eles. Modelar as dependências entre as entradas pode levar a modelos intratáveis, mas ignorá-las pode levar a um desempenho reduzido.

Uma solução para este problema é uma abordagem discriminativa, similar à tomada em classificadores como o de regressão logística. Em seu trabalho Sutton *et al* [30] modelaram a distribuição condicional p (y | x) diretamente, tudo o que é necessário para classificação. Esta é a abordagem adotada pelo modelo CRF (*Conditional Random Fields*), ou Modelo de Campos Aleatórios Condicionais. Os CRFs são essencialmente uma maneira de combinar as vantagens da classificação discriminativa e da modelagem gráfica, combinando a habilidade de compactamente modelar saídas multivariadas **y** com a capacidade de alavancar um grande número de recursos de entrada **x** para previsão. A vantagem de um modelo condicional é que as dependências que envolvem apenas variáveis

em x não desempenham nenhum papel no modelo condicional, de modo que um modelo condicional preciso pode ter estrutura muito mais simples que um modelo comum. A diferença entre modelos generativos e CRFs é, portanto, exatamente análoga à diferença entre os classificadores Naive Bayes e os classificadores de regressão logística. De fato, o modelo de regressão logística multinomial pode ser visto como o tipo mais simples de CRF, no qual existe apenas uma variável de saída.

Muitos problemas diferentes têm sido trabalhados com o modelo CRF. Aplicações bem sucedidas incluíram processamento de texto [31], [32], [33], visão computacional [34] e bioinformática [35]. Embora as aplicações precoces dos CRF usassem cadeias lineares, algumas aplicações de CRFs também usaram estruturas gráficas mais gerais. Estruturas gráficas gerais são úteis para prever estruturas complexas, como gráficos e árvores, e para relaxar a independência pressuposto entre entidades, como na aprendizagem relacional [36].

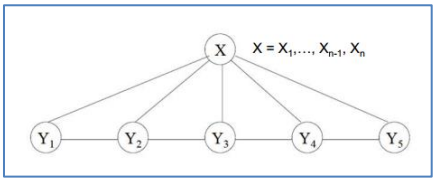
Seja G = (Y, E) um grafo onde cada vértice Y_v é uma variável aleatória. Supondo P(Yv | todos os outros Y) = P(Yv | vizinhos (Y_v)) então Y é um campo aleatório.



P(Y5 | todos outros Y) = P(Y5 | Y4, Y6)

Supondo P(Y_v | X, todos outros Y) = P(Y_v |X, vizinhos(Y_v))
então X com Y é um campo aleatório condicional.

X: observações, Y: etiquetas
P(Y3 | X, todos outros Y) = P(Y3 |X, Y2, Y4)



2.3.2. Níveis de classificação em Análise de Sentimento

Em Análise de sentimento, há três níveis de classificação: Classificação a nível de **documento**; classificação a nível de **sentença,** e classificação a nível de **aspecto.**

Considerem-se os seguintes exemplos de revisões de restaurantes:

1. *Great for large groups and celebrations - our SUPER HAPPY waiter was the entertainment of the evening. The place has proportioned good moments every time we go there. Last night I took some friends of mine to there, after leaving work. My girlfriend came after, and easily parked her car in the parking lot, however we did not. The maître was attentive as always. The food was perfect! Patty´s Restaurant has proved it is a good place to go with your family and friends in London.*
2. *Richard´s Pub and Meals is an excellent place to go with friends!*
3. *I loved eating at Salads House. The dishes are perfect, the staff is exceptional, but, unfortunately, the drinks are not so good.*

Analisando-se as revisões de restaurantes acima, percebe-se que no exemplo 1 temos um texto que contém várias sentenças sobre um mesmo tópico (*Patty´s Restaurant*). Apesar de o número de frases não ser tão alto, não haveria problema se considerássemos este *review* como um **documento**. Uma avaliação do sentimento médio expresso pelo autor deste texto, pode nos levar a uma ideia de algo **positivo,** uma vez que ele teceu mais elogios do que críticas ao estabelecimento**.** De forma geral, o autor aprecia o *Patty's Restaurant*.

No exemplo 2, temos uma revisão **bem menor** sobre o mesmo tópico (restaurante). Neste caso, a Análise de Sentimento não ocorre sob um cálculo médio entre as frases de um documento, mas apenas sobre uma única **sentença**. A polaridade é definida por uma análise direta da frase, que apresenta uma polaridade definida, no caso positiva.

O exemplo 3 é um pouco mais complexo, apesar de ter semelhanças com o exemplo 1, já que apresenta mais de uma sentença, há uma particularidade: apesar de o sentimento geral do *review* levar o leitor a interpretar este como **positivo**, o autor faz uma ressalva (*"the drinks are not so good"*). Ou seja, mesmo amando

comer no *Salads House*, as bebidas não são boas; um **aspecto** dentre todos os avaliados não o agradou – e talvez essa característica seja a mais importante para um leitor. É a análise minuciosa sobre um produto ou serviço, descritiva em cada aspecto individual, que pode agregar valor e informação na avaliação dos sentimentos presentes em um documento ou sentença.

2.3.3. Análise de Sentimento Baseada em Aspectos (ASBA)

A maioria das abordagens atuais utilizadas em mineração de opinião tenta detectar a polaridade geral de uma sentença (ou documento) independentemente das entidades alvo (por exemplo, restaurantes, laptops, hotéis) e de seus aspectos (por exemplo, comida, preço, bateria, tela) [37]. Em contrapartida, a ASBA identifica os aspectos de uma determinada entidade a ser analisada e estima a polaridade do sentimento para cada aspecto mencionado de forma individual. Isto abre possibilidades completamente novas na forma de analisar os dados (Fig.2).

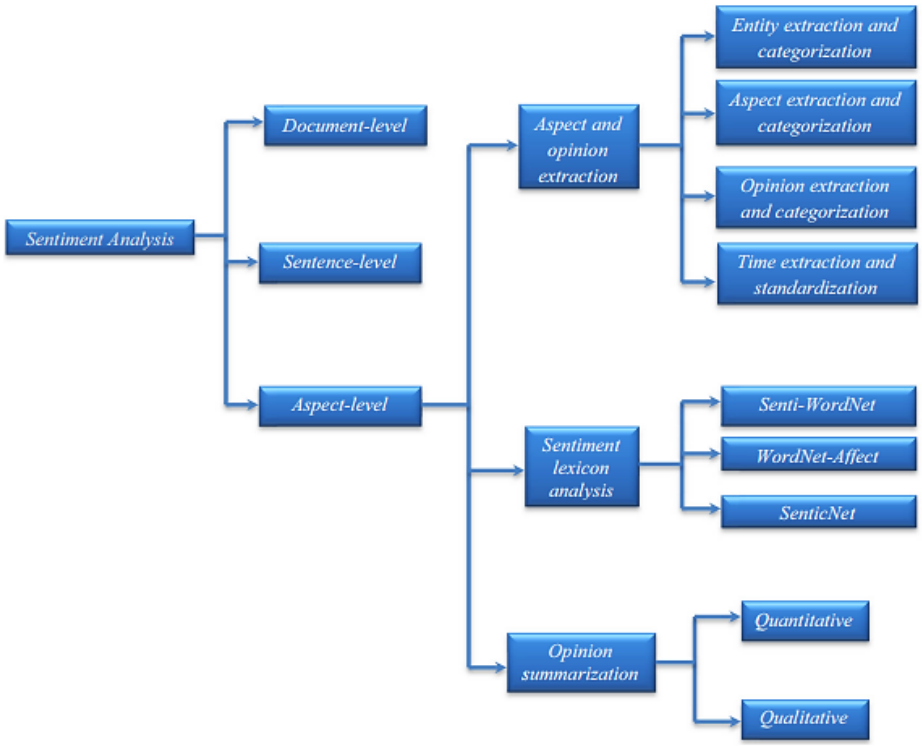


Figura 3: A Análise de sentimento Baseada em Aspectos e suas abordagens. Fonte: [4]

A Análise de Sentimento Baseada em Aspecto é uma tendência recente e uma abordagem que tem muito a ser explorada, uma vez que tem demonstrado bons resultados na literatura, estando presente, inclusive, como tarefa nos anos de 2014, 2015 e 2016 na SemEval, uma série de avaliações em análise semântica computacional que ocorre anualmente em Workshops e conferências sobre avaliação semântica. Na edição de 2016, Falk *et al* [38] propuseram um trabalho para detecção de palavras alvo utilizando vetores de palavras e seus relacionamentos de dependência gramatical. Yanase *et al*. [39], no mesmo ano, apresentaram uma abordagem para estimar a relação entre entidades alvo e expressões de sentimento, usando para tal um modelo de atenção neural e um sistema baseado em regras. A execução das tarefas propostas na SemEval é uma atividade de grande valia, e a superação dos *baselines* propostos funciona como indicador do estado da arte nas pesquisas em ASBA.

2.3.3.1 Tarefa de Extração de Aspectos

Para a tarefa de extração de aspectos explícitos, existem 4 (quatro) abordagens principais [40]**:** extração baseada em substantivos frequentes; extração através da relação entre o alvo e o sentimento; extração usando aprendizagem supervisionada, e extração usando modelos de tópicos.

1. **Extração baseada em substantivos frequentes**
   Utiliza analisador gramatical para identificar a substantivos mais frequentes. Esta abordagem foi desenvolvida inicialmente por Hu e Liu [41]

2. **Extração através da relação entre alvo e sentimento**
   Utiliza analisador gramatical e relações de dependência para obter evidências de relação entre o alvo e as palavras de sentimento. Um trabalho bastante detalhado foi proposto por [42];

3. **Extração via aprendizagem supervisionada**

Utiliza modelos de aprendizado supervisionado para determinar se uma opinião refere-se a uma entidade ou aspecto da entidade. Diversos trabalhos utilizam esta abordagem sendo mais comum para identificação de entidades implícitas [43];

4. **Extração usando modelos de tópicos**

Utiliza métodos baseados em agrupamentos (ou *cluster*) de tópicos buscando obter distribuições que representem aspectos. Diversas abordagens têm sido apresentadas na literatura tais como utilização de pLSA (Análise Semântica Latente Probabilística), LDA (Alocação Latente de Dirichlet), EM (Máxima Expectativa), entre outros.

2.3.3.2. Principais características (*features*) utilizadas em ABSA com CRF

A fase de extração de características é uma das principais etapas no processo de extração de informação de um texto. As características são os elementos que os algoritmos de classificação por aprendizagem de máquina vão utilizar como dados de entrada para a fase de treinamento e posterior classificação.

No entanto, tomando-se como dados brutos textos escritos em linguagem natural, os modelos de classificação não são capazes de interpretar o que cada termo de uma sentença significa – são apenas caracteres, apresentados em combinações aleatórias em milhões de possibilidades. Além disso, muitas informações contidas em um texto ou sentença podem ser irrelevantes para o que se deseja inferir.

Por exemplo, dada uma situação cujo objetivo fosse identificar se o conteúdo de um determinado *review* tem caráter positivo (elogio) ou negativo (crítica), e tomando-se para isso duas sentenças-exemplo: 1: "*I and my friend Thomas, an old man with brown hair and a silver wrist-watch, appreciate your snack bar very much*" e 2: "*We appreciate your restaurant a lot*" percebe-se que ambas as sentenças têm caráter positivo (elogio), mas que na sentença 1 há uma quantidade de palavras muito maior que na 2, e a maioria delas

irrelevantes para a classificação do caráter positivo da frase. Ainda mais, comparando-se as duas sentenças, percebe-se que o termo que está sendo avaliado na frase 1 é "*snack-bar*" e na 2, "*restaurant*". O tipo de entidade que está sendo avaliada não interfere no caráter positivo ou negativo do que está escrito, assim, o termo "*restaurant*" poderia ser substituído por "*pub*", "*market*", "car" ou até "*country*". Enfim, por um substantivo qualquer.

Assim, percebe-se que é necessário que haja um pé-tratamento desses dados, uma extração das características mais relevantes para o classificador e uma categorização de termos cambiáveis, semelhantes, que não alterem a estrutura lógica da sentença se forem substituídos por outros de mesma categoria.

O exemplo apresentado apenas ilustra a importância da seleção das características de uma sentença para análise. No decorrer desta seção serão apresentadas as principais características escolhidas para o desenvolvimento dos ensaios experimentais deste trabalho, baseadas em revisão da literatura e adequação ao método proposto.

Stemming

Em inglês, *stem* é o mesmo que radical ou tema. Segundo Santana [44], radical é o elemento mórfico que funciona como base do significado. É o elemento comum a palavras de mesma família. Raiz seria "root". Assim, para o *The Oxford Dictionary* and *Thesaurus*, o stem é "*The root or main part of a noun, verb, etc. to which inflections are added; the part that appears unchanged throughout the cases and derivatives of a noun, persons of a tense, etc*." Ainda de acordo com ambos, *stemmer* é a remoção de *stems* ou radicais. *The Cambridge International Dictionary of English* define e exemplifica o que é stem: *"The stem of a word is what is left when you take off the part which changes in order to show a different tense or a plural form etc.: From the stem 'sav-' you get 'saves', 'saved', 'saving' and 'saver."* Para PORTER, *"stemmer or 'Porter stemmer' is a process for removing the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems"*.

Lema

A lematização consiste em encontrar um item, isto é, uma forma gráfica representativa de todas as formas que uma unidade de significação lexicográfica (tradicionalmente palavra ou palavras compostas) pode tomar [45].

Na linguística computacional, a lematisação é o processo algorítmico de determinar o lema para uma dada palavra. Como o processo pode envolver tarefas complexas, como compreender o contexto e determinar POS tag de uma palavra em uma frase (exigindo, por exemplo, conhecimento da gramática de uma língua), pode ser uma tarefa difícil implementar um *lemmatiser* para uma nova língua. Em muitos idiomas, as palavras aparecem em várias formas flexionadas. Por exemplo, em inglês, o verbo "*walk*" pode aparecer como "*walk*", "*walked*", "*walks*", "*walking*". A forma básica, "*walk*", que pode ser encontrada em um dicionário, é chamada de lema da palavra. A combinação da forma básica com a parte da fala é freqüentemente chamada de lexema da palavra [46].

Ainda segundo Lucca [45], apesar de muito parecidos, lematização difere fundamentalmente de *stemming*. Enquanto lematização existe puramente no contexto lexicográfico, *stemming* não. Lematização é, pois, a representação da palavra através de seu masculino singular, adjetivos e substantivos e infinitivo (verbos), apenas no contexto da lexicologia. *Stemming* é a retirada de sufixos do radical, enquanto *stem* é o radical. Assim, as estruturas são distintas, embora eventualmente possam ser graficamente semelhantes.

Diferenças entre Lema e *Stemming*

O algoritmo de Porter é conhecido como “um algoritmo para remoção de sufixos”. Portanto, o algoritmo de Porter, na verdade, retira os sufixos das palavras, ao contrário da lematização, que representa as palavras, no caso dos verbos, por meio de seu infinitivo e, no caso dos substantivos e adjetivos, por meio de seu masculino singular. A segunda

diferença é que *stemming* não é necessariamente empregado em lexicografia, ao passo que lematização é [45].

Adjetivos Comparativos e Superlativos

De acordo com o *Education First* [47]: *Comparative adjectives* (Inglês) são usados para comparar diferenças entre os dois objetos que eles modificam (maiores, menores, mais rápidos, mais altos). Eles são usados em frases onde dois substantivos são comparados, neste padrão:

Noun (subject) + verb + comparative adjective + *than* + noun (object).

O segundo item de comparação pode ser omitido se estiver claro no contexto. Exemplo: "*My car is **faster** than yours"*; "*This hot dog is **bigger** than the other one we ate last night"*

Superlative adjectives (Inglês) são usados para descrever um objeto que está no limite superior ou inferior de uma qualidade (o mais alto, o menor, o mais rápido, o mais alto). Eles são usados em frases em que um assunto é comparado a um grupo de objetos.

Noun (subject) + verb + the + superlative adjective + noun (object).

O grupo que está sendo comparado pode ser omitido se estiver claro no contexto. Exemplo:

"This pasta is the hottest we´ve tried". "Tom´s hamburger is the best!"

Negative Words

Negative words são quaisquer palavras que têm o sentido de objeção, negação ou contrariedade a outros termos tidos como afirmativos. Por exemplo, *somebody* / ***nobody***; *something* / ***nothing***; *yes* / **not.** Uma palavra negativa pode ter seu sentido de negação

completo em si ("*Nobody asked* for the *bill*"), ou serem **modificadores** de termos ou expressões subsequentes e/ou antecedentes. Ou seja, modificadores negativos podem ser considerados termos que ao negar uma palavra ao qual estão relacionados, mudam o seu sentido de verdadeiro para falso, ou de falso para verdadeiro. Por exemplo, na expressão 1: "*The service at this restaurant is excellent*". O termo "*excellent*" é um adjetivo diretamente relacionado a "*service*". A frase, que está em seu sentido afirmativo indica que o atendimento do restaurante em questão é excelente. Na expressão 2: "*The service at this restaurant is NOT excellent*", o termo NOT nega a palavra excelente, falseando a ideia passada na expressão 1, ou seja, uma palavra negativa tem o poder de modificar o termo ao qual ela está relacionada. Percebe-se, entretanto, que de uma forma contextual, a influência se estende à frase como um todo, modificando todo o sentido / sentimento transmitido. Assim, para ABSA, o reconhecimento de palavras negativas é imprescindível no processo de classificação semântico.

Algumas das principais palavras negativas no idioma Inglês são: *no, not, nothing, negative, bad, old, cannot, zero, deplorable, ugly, never, annoy, don´t, dirty, sad, nonsense...*

Certos termos classificados considerados negativos como "*cry*", "*anxious*" e "*hard*" podem expressar uma ideia positiva, ou intensificar um sentido positivo em uma frase ("*She cried of happiness when she won a new car"; "Paul and Jessica worked very hard to earn their salaries*"). Deste modo, a aplicação de palavras negativas para classificação textual deve ser realizada de forma criteriosa.

Tendo como um de seus artifícios a utilização de extração de palavras negativas, e a presença ou não de termos negativos margeando em 4 posições cada token, Li [2] propôs um framework de aprendizagem de máquina baseado em CRF (*Conditional Random Fields*) capaz de empregar recursos avançados para extrair conjuntamente opiniões positivas, negativas e neutras em revisões de filmes e produtos.

Palavras Positivas e Negativas

Uma abordagem em Análise de Sentimento é coletar palavra efetivas em um léxico de sentimentos. No entanto, muitas vezes, tal conjunto de dados é dependente de domínio e não levam em conta as relações entre as palavras [48], limitando-se à semântica isolada dos termos.

É sabido que em qualquer idioma, a comunicação pode expressar sentimentos positivos, negativos e neutros. Essa classificação pode ser subjetiva (o que um leitor, ou interlocutor, entende como algo positivo, pode não o ser para outro); o conceito plural de muitas palavras e a individualidade de cada um na sua forma de entender o mundo em sua forma escrita ou falada leva a essa subjetividade. Entretanto, tomando-se como média geral, alguns termos tendem a apresentar-se muito mais com caráter positivo do que negativo, enquanto outros são mais tomados como negativos do que positivos. Ainda, há termos que não apresentam nem tendência positiva, nem negativa – são os neutros.

A quantidade de termos positivos ou negativos em uma sentença pode sugerir se tal sentença tem valor positivo, ou negativo. A aplicação de algumas técnicas como TF-IDF (*term frequency–inverse document frequency*), uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico [49], leva em conta a frequência de uma determinada palavra em uma sentença ou texto, equilibrado à frequência desse termo no corpus linguístico. Assim, um texto poderia ser classificado como positivo, negativo ou neutro de acordo com a frequência de palavras dessas categorias dentro de uma sentença ou texto. Desse modo, percebe-se que a polaridade (positiva, negativa ou neutra) de uma expressão pode ser diretamente influenciada pelo conjunto das polaridades específicas das palavras que a compõem.

Ex: "*That restaurant is excellent! The food is delicious, the service is fast and the drinks are wonderful*"

No exemplo acima, os termos *excellent, delicious, fast* e *wonderful* são classificados como palavras positivas [50] – a sentença acima passa uma ideia positiva acerca do restaurante.

Morfologia – as unidades do enunciado

Segundo Bechara [51], quase sempre a gramática engloba numa mesma relação palavras que pertencem a grupos bem diferentes: *substantivo*, *adjetivo*, *artigo*, *numeral*, *pronome*, *verbo*, *advérbio*, *preposição*, *conjunção* e *interjeição*. Um exame atento facilmente nos mostrará que a relação junta palavras de natureza e funcionalidade bem diferentes com base em critérios categoriais, morfológicos e sintáticos misturados. E o elemento que as diferencia são os diversos significados que lhes são próprios. Para tanto, devem-se distinguir os seguintes significados:

Significado Lexical - é o significado que corresponde ao *quê* da apreensão do mundo extralinguístico, isto é, é o que corresponde à organização do mundo extralinguístico mediante as línguas. A linguagem *classifica* a realidade segundo interesses e atitudes humanas; por isso suas distinções *podem* coincidir com realidades e delimitações objetivas, mas isso não é necessário. A língua é um saber acerca de modelos e esquemas linguísticos, e não sobre os objetos, a respeito dos quais informam nossa **experiência** (o nosso saber do mundo) e as ciências não linguísticas; assim, que um furacão se chame *Katrina*, e só haja um com este nome, é um fato da geografia, e não um *fato de língua*, tal como *Windows* e *Apple* também ilustram essa relação. É o significado que é comum a cada uma das séries de palavras: *comer, comida, comestível*, ou no Inglês: *love*, *lover*, *lovely* e *loveless.*

Significado Categorial - é o que corresponde ao *como* da apreensão do mundo extralinguístico, a forma da intuição da realidade ou, ainda, o modo de ser das palavras no discurso, e não classes léxicas fixas: *love* (quando empregado como substantivo), *lover* (quando empregado como adjetivo), *love* (quando empregado como verbo), *lovely* (quando advérbio). Cabe à gramática geral, definir a categoria linguística "substantivo"; à gramática descritiva cabe tão somente comprovar se a língua objeto da descrição tem ou não substantivos, e, em caso afirmativo, quais são os meios materiais, isto é, os esquemas formais para expressar a categoria "substantivo" [51]. Assim, a categoria "substantivo" expressa substantivos por meio de nomes como *car, food, room*, por meio de pronomes

como *this*, *that*, *mine*, por sintagmas como *New York City* ou orações (como as *Noun phrases*, *em Inglês*). As palavras lexemáticas e categoremáticas só estão categorialmente determinadas como substantivo, adjetivo, verbo e advérbio quando **integradas na oração**, atualizadas no discurso. No momento em que a gramática geral define o que é "adjetivo", essa definição, caso correta, deverá servir a **todas as línguas que tenham adjetivos**. É **equivocado** o querer definir "o adjetivo em português". À gramática descritiva cabe descrevê-lo *numa* língua. Constituem o substantivo, o adjetivo, o verbo e o advérbio as quatro únicas reais "categorias gramaticais" da língua, confusamente misturadas na gramática tradicional, e que os estudiosos chamam "categorias verbais", porque são as únicas dotadas do **significado categorial**. Admitem, como já ensina a gramática tradicional, subdivisões. O significado categorial não caracteriza apenas os lexemas, mas ainda sintagmas e orações inteiras. Também o significado categorial está sempre implicado com certas funções específicas na estruturação gramatical; por isso, só o "substantivo" (representado por nome, pronome, sintagma nominal, oração nominalizada) pode ser o sujeito da oração, assim como o verbo exerce a função de predicado [52].

Significado instrumental - este é o significado dos morfemas, isto é, dos elementos pertencentes ao universo da gramática. Podem apresentar-se como palavras morfemáticas (como os *artigos* e as *preposições*, por exemplo, ou como elementos de palavras: o *-s* de *car-s* ou de *pizza-s*, etc.); são os chamados *instrumentos gramaticais*. O artigo *The* em "*The books are about Science"* tem o significado "atualizador" e o *-s* em *book-s* tem o significado "pluralizador". Incluem-se como dotados desse significado instrumental, ou seja, "morfemas", nas combinações gramaticais, os prefixos, os sufixos, as desinências, o ritmo, a entoação, a ordem das palavras, etc. Os significados instrumentais são modos da **expressão** material [51].

Significado estrutural – é o que resulta das combinações de unidades lexemáticas com os morfemas (menor unidade linguística que possui significado), dentro da oração. A exemplo têm-se "singular", "plural", "atual", "virtual", "ativo", "passivo", "presente", "passado", "futuro", etc. Assim o *-s* de *book-s* acima tem o significado instrumental" pluralizador" (e não "plural") ao lado do termo que foi pluralizado *book*; da combinação resulta o significado estrutural ou sintático "plural".

As classes de palavras

Os significados léxico, categorial e instrumental nos permitem dividir as palavras em lexemáticas (substantivo, adjetivo, verbo e advérbio), categoremáticas (pronome e numeral) e morfemáticas (artigo, preposição e conjunção). Isto não impede que uma palavra categoremática possa também aparecer com significado instrumental, como é o caso de *your computer*, em que *your* tem o significado categorial "adjetivo" e o significado instrumental em relação ao substantivo *computer*.

Apresentadas as classes gramaticais, pode-se agora discorrer um pouco sobre algumas das principais:

Substantivo

O substantivo (no Inglês, *noun*) é uma classe de palavras variável com que se designam ou se nomeiam os seres em geral, ou são as palavras variáveis com que se designam os seres (pessoas, animais e coisas) [53]. É a classe de lexema que se caracteriza por significar o que convencionalmente chamamos *objetos substantivos*, isto é, em primeiro lugar, substâncias (*hotel, city, girl*) e, em segundo lugar, quaisquer outros objetos mentalmente apreendidos como substâncias, quais sejam qualidades (*happiness, love)*, estados (*health, illness*), etc.. [51]. O substantivo é a palavra que serve, de modo primário, de núcleo de sujeito, do objeto direto, do objeto indireto e do agente da passiva. Qualquer palavra de outra classe que desempenhe uma dessas funções equivalerá, forçosamente, a um substantivo [53].

De acordo com Bechara (2015), os substantivos podem ser classificados em:

- Concretos e abstratos – concretos: é o que designa ser de existência independente (*house, stone, cloud*) – abstratos: são aqueles que são tidos como de existência independente (*thinking*, *sadness, dream*).

- Comuns e próprios – comum: é o que se aplica a um ou mais objetos particulares que reúnem características inerentes a dada classe (*bus, pen, egg*) – próprio: é o que se aplica a um objeto ou a um conjunto de objetos, mas sempre individualmente, nomeando e individualizando-o (*London, Peter, Brazil).*
- Primitivos e derivados – primitivos são aqueles de que não derivam de outros vocábulos (*life, car, play*) – derivados: procedem de outras palavras (*player, neighbourhood*)
- Simples e compostos – simples: são constituídos de um só radical (*house, dish*) – compostos: são formados da união de dois ou mais radicais (*stepfathe, policeman)*
- Coletivos: São substantivos comuns que, no singular, designam um conjunto de seres ou coisas da mesma espécie (*staff, bunch*)

Adjetivo

O adjetivo é a classe de lexema que se caracteriza por constituir a *delimitação*, isto é, por caracterizar as possibilidades designativas do substantivo, orientando delimitativamente a referência a uma *parte* ou a um *aspecto* do denotado [51]. Apesar de no Português flexionarem-se em gênero, número e grau, no idioma Inglês são **invariáveis**. (*That **beautiful girl** over there is my sister, and those **beautiful women** are my aunts*)

Artigo

Artigos são palavras que definem um substantivo como específico ou inespecífico. Exemplo: *The president has visited other countries a lot since last year*

Os artigos podem ser definidos ou indefinidos

- Definidos: Em Inglês, o artigo definido é a palavra "*The*". Ele limita o significado de um nome a uma coisa em particular. Por exemplo: "*Are you going to **the** party this weekend*?" (uma festa específica).

- Indefinidos: Há duas formas para o artigo indefinido em Inglês (*a* e *an*), que assumem o mesmo papel: indicar que um nome se refere a uma ideia geral, em vez de algo em específico. Exemplo: "*Should we bring a bottle of wine to the meet*" (qualquer garrafa de vinho) [54].

Pronome

Pronome é a classe de palavras categoremáticas que reúne unidades em número limitado e que se refere a um significado léxico pela situação ou por outras palavras do contexto. De modo geral, esta referência é feita a um objeto substantivo considerando-o apenas como pessoa localizada do discurso [51]. Os pronomes podem então ser entendidos como palavras que substituem nomes: *I, me, they, we, his, her, us, etc...* Exemplo: "*I like Angela, she is one of the best persons I have ever known.*"

Os pronomes precisam de antecedentes. Isso significa que a coisa (ou pessoa) a que o pronome se refere precisa já ter sido mencionada pelo nome em alguma parte da frase ou do parágrafo. Se não estiver claro a que o pronome se refere, o leitor pode ficar bastante confuso [54]. Tal ambiguidade pode levar à incompreensão, ou compreensão parcial, da mensagem que está sendo transmitida. Essa dificuldade de compreensão por um interlocutor/leitor é mimetizada pelos métodos de Processamento de Linguagem Natural.

Em Inglês, os pronomes podem ser: pessoais (*I, he, she, us, them, etc.*), possessivos (*my, his, her, hers,* etc.), indefinidos (*somebody, someone, anybody*, etc.), relativos (*who, which,* etc.), demonstrativos (*this, that, those,* etc.), interrogativos (*what, when, where, how*, etc.), reflexivos / intensivos (*myself, yourself,* etc.) e recíprocos (*each other, one another*) [55].

Verbo

Os verbos são palavras usadas para indicar uma ação praticada (*walk, study, drink,...*), um estado em que alguém ou alguma coisa se encontra (*is, was, stays,...*) ou uma mudança de estado (*turn, become,...*) [55].

Os verbos são flexionados de acordo com a pessoa ou tempo verbal. Podem variar sua estrutura escrita (*play, played*), ou tornarem-se outra palavra sem nenhum radical

comum (*is, was, go, went*). Os verbos ainda podem requerer palavras auxiliares para assumirem um determinado tempo verbal (*She* ***will*** *travel tomorrow*), ou alguns ainda podem participar como verbos auxiliares ou modais na composição de novas estruturas frasais.

Advérbios

Assim como os adjetivos têm o poder de acompanhar um substantivo, modificando, ou a ele dando uma determinada característica, os advérbios são palavras que dão qualidades de modo, tempo e lugar para os verbos, adjetivos e também para os próprios advérbios. Portanto, os advérbios funcionam como modificador de verbos, adjetivos e de outros advérbios, usados para dizer como, quando ou onde uma coisa ocorreu, vindo geralmente depois dos verbos principais (*Becky drives* ***carefully****; Thomas speaks Spanish* ***perfectly****; He drives* ***very fast)*** [55].

Numeral

Numeral é a palavra que indica os seres em termos numéricos, ou seja, que atribui quantidade aos seres ou os situa em determinada sequência. (*There are two seats available*). Podem apresentar-se como forma cardinal (*one, two, three, four,...*), ou ordinal (*first, second, third, ...*). No Inglês, os numerais não sofrem variação de gênero, como no Português, por exemplo.

Preposição

Chama-se preposição a uma unidade linguística desprovida de independência – isto é, não aparece sozinha no discurso, salvo por hipertaxe – e, em geral, átona, que se junta a substantivos, adjetivos, verbos e advérbios para marcar as relações gramaticais que elas desempenham no discurso, quer nos grupos unitários nominais, quer nas orações.

A preposição não exerce nenhum outro papel que não seja ser índice da função gramatical de termo que ela introduz [51].

No Inglês, as palavras são divididas em 2 grandes grupos: as *content words* e as *gramatical word.* Como o próprio nome diz, as *content words* são palavras com **conteúdos**, ou seja, aquelas que apresentam algum sentido relevante para a frase, onde a sua ausência em uma frase mudará a ideia daquela sentença.

Por exemplo, considere-se a frase: "*I have a dog, its name is Rex"*. Se esta fosse dividida em 2 linhas, como demonstrado na Figura 3, ter-se-iam na primeira linha os termos mais importantes para o entendimento da mensagem, e na segunda, termos acessórios.

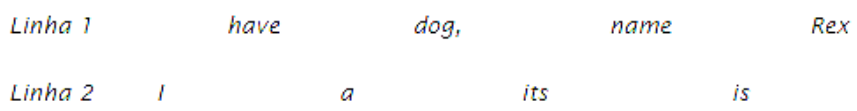


Figura 4: *Content* e *gramatical words*. Fonte: https://www.englishexperts.com.br/preposicoes-em-ingles/

Essas palavras da primeira linha são as *content words*, pois possuem realmente conteúdo e sentido na frase e, mesmo na ausência de alguns termos, ainda é possível compreender a ideia da frase, uma vez que elas têm sentido completo por si próprias. Já na segunda linha, pouco pode ser entendido. Na verdade, não faz qualquer sentido, pois são apenas palavras soltas que precisam de "algo mais" para serem compreendidas, pois elas não possuem sentido completo por si próprias. Essas são as *grammatical words*, ou seja, palavras que servem mais para formar uma estrutura de uma frase do que para dar conteúdo a ela. A função das *grammatical words* é de ligar uma *content word* à outra e assim tornar uma frase harmônica e de fácil entendimento [55].

A preposição, por fazer parte do grupo de *grammatical words*, não possui um sentido completo quando isolada. Sua função é de ligar um substantivo, pronome ou *noun phrase* (ou seja uma frase constituída de substantivo / nome / pronome + todos os agentes que podem modificar o sentido das mesmas, como os adjetivos, advérbios, infinitivos e particípios) às outras partes da oração e assim explicar qual tipo de relação esse substantivo / pronome / *noun phrase* tem com essas outras partes.

Como a função da preposição é constituir uma relação de sentido entre dois ou mais termos de uma oração, há quatro tipos de sentido que um pronome pode estabelecer: lugar, direção, tempo e relação lógica geral. Exemplo: ("*Kate will be arriving* ***at*** *6 o´clock"," I´ve lived here* ***for*** *5 years"*)

As principais (mais comuns) preposições no idioma Inglês são: *of, in, to, for, with, on, at, from, by, about, like, through, after, between, out, against, during, without, before, under, among* [56].

Conjunção

As conjunções são palavras invariáveis que servem para conectar orações ou dois termos de mesma função sintática [57], e estabelece entre eles uma relação de dependência ou coordenação. As conjunções ligam outras palavras, frases ou cláusulas juntas.

As conjunções podem ser:

- *Coordenadas***:** permitem que seja possível a junção de palavras, frases e cláusulas de igual classificação gramatical em uma frase. As conjunções coordenadas mais comuns em Inglês são *from and, nor, but, or, yet* e *so*. Exemplo*: I like cooking and eating, but I don't like washing dishes afterward.* [54]

- *Subordinadas***:** as conjunções subordinadas unem cláusulas independentes e dependentes. Uma conjunção subordinada pode sinalizar uma relação de causa e efeito, um contraste ou algum outro tipo de relação entre as cláusulas. Em Inglês, as conjunções subordinadas mais comuns são *because, since, as although, while* e *whereas* [58]. Exemplo: *Since you passed the test, the job is yours!*

A estruturação do enunciado: frase e oração

Segundo Frances Peck [59], as partes da sentença são um conjunto de termos para descrever como as pessoas constroem sentenças a partir de partes menores. Não há uma correspondência direta entre as partes da sentença e as partes do discurso - o sujeito de uma frase, por exemplo, poderia ser um substantivo, um pronome ou até mesmo uma frase ou cláusula inteira. Como as partes do discurso, no entanto, as partes da frase formam parte do vocabulário básico da gramática. Abaixo elas são conceituadas de forma resumida:

- *Sujeito e predicado*

  Cada frase completa contém duas partes: um sujeito e um predicado. O sujeito é sobre o que (ou quem) a sentença se refere, é aquele que define a ação do verbo enquanto o predicado diz algo sobre o assunto. O predicado é o assunto em si, a ação executada, que pode ocorrer inclusive sem um sujeito agente (no Português). No Inglês, entretanto, todo predicado tem um sujeito correspondente, mesmo que este não execute propriamente a ação do verbo (***It*** *rainned again last night).* A presença de verbo define o predicado de uma oração

- *Objetos direto e indireto*

  Objetos nada mais são do que complementos verbais dentro de uma ou mais orações. Eles são classificados basicamente em dois tipos: objeto direto e objeto indireto.

  O objeto direto é o complemento dos verbos transitivos diretos, ou seja, aqueles verbos que não precisam de preposição para se ligar a seus objetos (ex: I love you. Verbo: love; objeto direto: you).

  O objeto indireto por sua vez também complementa os verbos transitivos, mas faz isso de forma indireta, por meio de uma preposição, ou indireta apenas no sentido. Em Inglês, os objetos indiretos apenas ocorrem se na sentença também houver um objeto direto. Ex: Thomas gave Jane a *piece* of cake. Neste caso temos sujeito: Thomas, Obj. Direto: *a piece of cake* e Objeto indireto: *Jane*. É interessante perceber que a presença de objetos

sempre indica a presença de um verbo e consequentemente de um sujeito, mas a lógica inversa não é sustentada em todos os casos – verbos intransitivos, como “rain”, “sleep” e “dance” têm seu sentido completo sem a necessidade de um complemento verbal.

- *Verbos de ligação*

  Verbos de ligação (ou copula, em Inglês) são verbos que indicam um estado direto do sujeito ou do objeto. Esses verbos têm como finalidade unir o sujeito ou objeto a um termo que o qualifica, adjetiva, ou lhe indica uma característica de estado. Ou seja, o verbo de ligação não assume um papel protagonista como os demais verbos (ação, fenômeno natural, emoção, etc.), mas sim possibilitam que um nome seja conectado ao termo que o toma como objetivo referencial. Ex: “*The president is worried and upset about the last meeting*”. Percebe-se que a ideia principal da sentença não é o “estar”, mas sim “*worried*” e “*upset*”. Verbos de ligação fazem uma ponte direta entre substantivos e adjetivos, entre a qualidade e o qualificado.

Sinônimos

Sinônimos são palavras que mantêm relações de sinonímia e que representam, basicamente, uma mesma ideia. Ex.: *effort* (*achievement, attempt, battle, creation, etc.*). Um sinônimo é uma palavra ou frase que significa exatamente ou quase o mesmo que outro lexema (palavra ou frase) no mesmo idioma. Palavras que são sinônimos são consideradas sinônimas, e o estado de ser sinônimo é chamado de sinonímia [60].

Antônimos

Os antônimos, por sua vez, são palavras com sentido exatamente contrário, inverso a outra. É o oposto direto da palavra. Ex.: (*bad, good*), (*interesting, unexciting*)

Influência de adjetivos e advérbios

O conceito de adjetivos e advérbios já foi discutido neste trabalho. Sabe-se, portanto, que adjetivos e advérbios são modificadores de nomes, no caso dos primeiros, ou modificadores de verbos, adjetivos, e até outros advérbios, senos referimos ao último.

Na área de Mineração de textos e extração de Aspectos, sobretudo em revisões, a identificação desses modificadores é de suma importância, dado o fato de que esses termos têm o poder de qualificar, caracterizar e / ou modificar uma entidade. São por meio dos adjetivos e advérbios que os sentimentos e opiniões são "materializados" em forma de texto, e, em revisões de produtos e / ou serviços, a intenção do autor é focada na expressão de uma experiência, crítica, sugestão ou simples avaliação sobre uma vivência deste, tornando essas classes gramaticais os personagens principais desse tipo de documento.

3

Revisão da Literatura

Este capítulo descreve os principais trabalhos tomados da literatura que serviram de base para a escolha das principais características utilizadas no processo de extração de aspectos

3.1. Trabalhos Relacionados

Em seu artigo, Li *et* al [61] focaram na sumarização de revisão baseada em características de objeto. Diferentemente de muitos outros trabalhos com regras linguísticas ou métodos estatísticos, eles formularam a tarefa de mineração de revisão como um problema de marcação de estrutura conjunta. Os autores propuseram uma nova estrutura de aprendizado de máquina baseada em Campos Aleatórios Condicionais (CRFs – *Conditional Random Fields*) e sugeriram que ela poderia empregar recursos avançados para extrair conjuntamente opiniões positivas, opiniões negativas e recursos de objetos para sentenças comentários (revisões). A estrutura linguística pode ser naturalmente integrada à representação do modelo. Além da estrutura de cadeia linear, eles também investigaram estrutura de conjunção e estrutura de árvore sintática no *framework* proposto. Por meio de extensos experimentos em revisão de filmes e conjuntos de dados de análise de produtos, eles mostraram que os modelos com suporte a estrutura superam as muitas abordagens de última geração da área de mineração.

Depois de aplicar muitos recursos como token de palavra, lema, POS, palavras negativas, sinonímia, antonímia e elações sintáticas, eles usaram CRFs de cadeia linear para identificar as dependências sequenciais entre palavras contínuas. Eles aprenderam que se duas palavras ou frases são conectadas por conjunção “e”, então ambas as palavras têm a mesma polaridade e se elas estão conectadas por “mas”, então ambas têm polaridades opostas.

Portanto, para superar a dependência de longa distância, eles usaram modelos *CRFs Skip-chain* para encontrar aspectos e opiniões, os CRFs de árvores, para aprender a estrutura sintética das sentenças nas revisões, os CRFs de cadeias de salto, para fornecerem as relações semânticas com respeito a conjunções e os CRFs de Árvore, para fornecerem relações de dependência entre palavras diferentes dentro de uma mesma sentença.
Eles propuseram CRFs *Skip-Tree* para combinar os métodos acima explicados e usar essas árvores para extrair aspectos e opiniões de resenhas de filmes, dando uma lista de aspectos como sementes de entrada.

Por sua vez, Jakob e Gurevych [62] focaram na extração do alvo de opinião (OTE) como parte da tarefa de mineração de opinião. Aplicando recursos como Token, POS, dependência curta, distância de palavras e sentença de opinião, eles modelaram o problema como uma tarefa de extração de informações, que eles abordam com base em campos aleatórios condicionais (CRFs). Como *baseline*, eles empregaram o algoritmo supervisionado de acordo com Zhuang *et* al. [63], que representou o estado-da-arte dos dados empregados. Os autores avaliaram os algoritmos de forma abrangente em conjuntos de dados de quatro domínios diferentes, anotados com instâncias de alvo de opinião individuais em um nível de sentença. Além disso, eles investigaram o desempenho de sua abordagem baseada em CRF e a *baseline* em um cenário de extração de alvo de opinião de domínio único e cruzado. Eles mostraram que sua abordagem baseada em CRF melhorou o desempenho em 0,077, 0,126, 0,071 e 0,178 em relação à F-*Measure* na extração de domínio único nos quatro domínios. No cenário de domínio cruzado, a abordagem melhorou o desempenho em 0,409, 0,242, 0,294 e 0,343 em relação ao F-*Measure* sobre a *baseline*. Eles observaram que, a mesma palavra pode ter representação diferente em domínios diferentes. Por exemplo, "imprevisível" pode ter sentido positivo em críticas de filmes, mas negativo em avaliações de carros. Portanto, os autores usaram a abordagem em diferentes domínios, tendo em mente essas palavras com característica de **portabilidade de domínio**.

Segundo Choi e Cardie [64], o reconhecimento automático de opiniões envolve várias tarefas relacionadas, como identificar os limites da expressão de opinião (OTE), determinar sua polaridade, e sua intensidade. Embora muito progresso tenha sido feito nessa área, a pesquisa existente normalmente trata cada uma das tarefas acima de forma

isolada. Em seu artigo, os autores foram um pouco além, e aplicaram uma técnica hierárquica de compartilhamento de parâmetros usando Campos Aleatórios Condicionais (CRF) para análise de opinião detalhada, detectando conjuntamente os limites das OTEs e determinando dois de seus principais atributos - polaridade e intensidade. Seus resultados experimentais mostraram que a abordagem proposta foi capaz de melhorar o desempenho sobre uma *baseline* que não explora a estrutura hierárquica entre as classes. Além disso, eles descobriram que a abordagem conjunta supera uma *baseline* que é baseada apenas na cascata de dois componentes separados.

Para a extração das expressões de opinião e de seus atributos, utilizou-se uma técnica de compartilhamento de parâmetros hierárquicos usando CRFs. Eles definiram o problema como uma tarefa de marcação de sequência para identificar opiniões e aspectos. Assim, sua abordagem não apenas extraia opiniões e aspectos, mas também classificava os aspectos de acordo com a polaridade de suas palavras de opinião.

Em seu trabalho, Choi e Cardie usaram muitos recursos para extração de características, como Token, PoS-tag, tokens anteriores e próximos, hiperônimos (WordNet), léxico de opinião, polaridade, palavras de intensidade, dentre outros.

Conforme já mencionado neste trabalho, com o crescimento de conteúdos gerados por usuários na Web, a revisão de opiniões de produtos se torna cada vez mais uma prática de pesquisa de grande valor para o comércio eletrônico, busca e recomendação. Infelizmente, o número de resenhas está subindo para centenas ou até milhares, especialmente para alguns itens populares (como revisões de produtos e serviços), o que se torna um tarefa custosa, para os compradores em potencial e para os fabricantes, lerem inúmeros documentos para terem em si a segurança de que tomaram a decisão mais sábia. Além disso, o formato livre e a incerteza das expressões de comentários tornam a extração e a categorização de recursos mais refinados do produto uma tarefa mais difícil do que as técnicas tradicionais de extração de informações. Em seu trabalho, Huang *et* al [65] propuseram tratar a extração de características do produto como uma tarefa de rotulagem de sequências e empregar um modelo de aprendizagem discriminativo usando Campos Aleatórios Condicionais (CRFs) para lidar com isso. Naquela época, eles incorporam de maneira inovadora os recursos PoS e os recursos da estrutura da sentença no processo de aprendizado dos CRFs. Para a categorização das características do produto, os autores

introduziram a medida de similaridade baseada em conhecimento e distribuição de contexto para calcular as semelhanças entre expressões de características do produto. Assim, foi proposto um algoritmo de categorização efetivo baseado na poda do gráfico para classificar a coleção de expressões de características em diferentes grupos semânticos. Os estudos empíricos provaram a eficácia e eficiência de suas abordagens em comparação com outros métodos de contrapartida.

Deste modo, eles propuseram um modelo probabilístico de aprendizagem baseado em CRF para **extrair aspectos** de produto. A tarefa de extração de aspectos foi adotada como tarefa de rotulagem de sequências, e o CRF foi usado para lidar com essa tarefa.
Os aspectos do produto foram divididos em três subcategorias para o processo de aprendizagem, e determinadas regras de marcação foram definidas para identificar esses aspectos.

Após a extração dos aspectos, eles definiram duas categorias para agrupar aspectos semelhantes. A primeira abordagem que eles usaram foi baseada no *WordNet*, que encontrou os cálculos de similaridade usando a abordagem do dicionário *WordNet*.
Já a segunda abordagem foi baseada em dependências sintáticas para o contexto sintático de longa distância. A combinação das duas abordagens foi usada para colocar todos os aspectos do produto em seus respectivos grupos.

Para o trabalho que propuseram, Huang *et* al, aplicaram muitos recursos. Para evitar a influência de diferentes variantes de uma palavra, eles usaram o radical da palavra em vez das diferentes variantes. E usou as hastes em uma janela [-1, +1] como recursos de palavras. Com o objetivo de gerar PoS-tags para as palavras de cada frase, eles também usaram *Stanford PoS-tagger*. Especificamente, eles usaram PoS-tags na janela [-1, +1] como características de PoS.

Como os recursos de palavras e os recursos de PoS envolvem apenas palavras isoladas, e as janelas de partição capturam apenas palavras vizinhas, os autores precisavam aprender as relações sintáticas entre as palavras que abrangem longa distância. Os recursos da estrutura da frase codificavam a informação da estrutura sintática entre as palavras nas sentenças. Finalmente, o analisador de dependência sintática foi usado para extrair as relações de dependência sintáticas para cada sentença.

A maioria das abordagens existentes aborda a extração de entidades de opinião e relações de opinião de uma maneira detalhada, em que as interdependências entre diferentes estágios de extração não são capturadas. Com o objetivo de aplicar uma solução melhor para esse problema, Yang e Cardie [66] abordaram a tarefa de extração de opinião com refinanciamento - a identificação de entidades relacionadas à opinião: as expressões de opinião, os detentores de opinião e as metas das opiniões e relações entre as expressões de opinião e seus alvos e titulares. Eles propuseram um modelo de inferência conjunta que aproveita o conhecimento de preditores que otimizam as subtarefas da extração de opinião e busca uma solução globalmente ideal. Seus resultados experimentais demonstraram que sua abordagem de inferência conjunta superou significativamente os métodos de *pipeline* tradicionais e as *baselines* que lidam com as subtarefas isoladamente para o problema da extração de opinião.

Seu trabalho foi focado em identificar conjuntamente as entidades relacionadas à opinião, ou seja, a expressão da opinião, as opiniões e os titulares de opinião, juntamente com as relações que ligam as opiniões com entidades e essas relações eram IS-ABOUT e IS-FROM. Eles dividiram a tarefa inteira em várias subtarefas. A primeira foi identificar palavras de opinião e entidade e, para isso, aplicaram o CRF para encontrar as sequências entre palavras diferentes. Como os aspectos de opinião foram identificados, o próximo passo foi identificar relações entre diferentes entidades e opiniões. Para identificar as relações, propuseram-se dois classificadores: as relações *Opinion-Arg* e as relações *Opinion-Implicit-Arg*. O primeiro classificador identifica as relações em que os argumentos são explícitos. Para os argumentos implícitos, eles usaram o segundo classificador. Na última fase, eles se uniram a esses métodos com uma série de restrições para encontrar as entidades de opinião e as relações de opinião. Eles indicaram que o conhecimento de diferentes preditores pode ser integrado para alcançar uma melhora significativa no desempenho geral.

Como principais características, eles usaram palavras e PoS-tag: as palavras contidas no candidato e suas tags PoS, *Lexicon* (para cada palavra no candidato, eles incluíram seus hiperônimos do *WordNet* e sua força de subjetividade no *Subjectivity Lexicon3* (por exemplo, *weaksubj*, *strongsubj*), tipo de frase (a categoria sintática do componente mais profundo que cobre o candidato na árvore de análise), a distância relativa

entre a opinião e argumentos candidatos e o caminho de dependência (o caminho mais curto na árvore de dependência entre o candidato de opinião e o alvo candidato).

As tarefas de mineração de opinião a nível de características geralmente incluem a extração de entidades de avaliações em revisões de consumidores, a identificação de palavras de opinião que estão associadas às entidades e a determinação das polaridades dessas opiniões (por exemplo, positivo, negativo ou neutro). Duas abordagens principais foram propostas para determinar opiniões no nível de características: métodos baseados em modelos como o baseado no modelo de Markov oculto lexicalizado (L-HMMs), e métodos estatísticos como a técnica baseada em mineração de regras de associação. No entanto, poucos trabalhos comparam esses algoritmos em relação às suas habilidades práticas na identificação de vários tipos de elementos de revisão, tais como características, opiniões, intensificadores, frases de entidades e entidades pouco frequentes. Por outro lado, poucas atenções foram dadas para aplicar modelos de aprendizagem mais discriminativos para realizar essas tarefas de mineração de opinião. Diante desta necessidade, Chen *et* al [67] não apenas compararam experimentalmente esses métodos com base em um conjunto de dados de revisão do mundo real, mas também adotaram em particular o modelo CRF e avaliaram seu desempenho em comparação com algoritmos relacionados. Além disso, para o algoritmo de mineração baseado em CRFs, os autores testaram o papel de um processo de autotipagem em duas condições de treinamento automáticas e identificaram a combinação ideal de funções de aprendizado para otimizar seu desempenho de aprendizado. O experimento comparativo acabou revelando a precisão de desempenho superior do método baseado em CRFs em termos de mineração de vários elementos de revisão, em relação a outros métodos.

Os autores ainda melhoraram a técnica, integrando o processo de autocomposição com o objetivo de minimizar o esforço manual para rotular os dados e para obter o equilíbrio desejado entre a complexidade e a precisão do algoritmo, identificando o conjunto ideal de funções de aprendizado.

Eles aplicaram algumas técnicas de extração de aspectos para comparar a precisão de diferentes sistemas em diferentes níveis e no mesmo conjunto de dados sobre revisões de produtos. Um dos exemplos foi a criação de regras em relação à distância de adjetivos e advérbios das palavras-alvo.

Em 2015, [68] adotaram um modelo CRF com notação BIO para OTE (*Opinion Term Expression*) com vários grupos de características, incluindo sintaxe, léxico, semântico e características do léxico do sentimento. A submissão para a tarefa de extração de termos de opinião (OTE) foi classificada em quinto lugar dentre mais de vinte submissões. Um modelo de regressão logística com esquema de ponderação de rótulos positivos e negativos foi usado para a polaridade do sentimento; vários grupos de características (léxico, sintático, semântico, léxico e escore Z) foram extraídos. A submissão para detecção de polaridade ficou em terceiro lugar em relação a dez submissões no conjunto de dados do restaurante.

Um dos trabalhos mais recentes na área de extração de aspectos, muito similar ao aqui desenvolvido, foi o de Xiang *et* al [69]. Os autores propõem o MFE-CRF que introduz o armazenamento em cluster *Multi-Feature Embedding* (MFE) baseado no modelo de Campo Aleatório Condicional (CRF) para melhorar o efeito da extração do termo de aspecto em Análise de Sentimento Baseada em Aspectos. Primeiro, a incorporação de múltiplos recursos (MFE) foi proposta para melhorar a representação do texto e capturar mais informações semânticas deste. Em seguida, os autores usam o algoritmo k-means ++ para obter MFE, e agrupamento de palavras para enriquecer as *características* de posição do CRF. Finalmente, as classes de clusterização de MFE e a incorporação de palavras são definidas como características adicionais de posição para treinar o modelo de CRF para extração de termos de aspecto. Os experimentos nos conjuntos de dados do SemEval validaram a eficácia desse modelo. Os resultados de diferentes modelos indicaram que o MFE-CRF pode melhorar muito a taxa de Recall do modelo CRF. Além disso, a taxa de precisão também aumenta obviamente quando a semântica do texto é complexa.

A tabela 1 apresenta um resumo dos trabalhos detalhados até aqui. Ela é dividida em: i) características aplicadas[1], ii) classificadores utilizados, iii) domínio da aplicação e ano.

[1] Mais detalhes sobre os grupos de características se encontram no anexo 1.

Tabela 1: resumo dos principais trabalhos na área de extração de aspectos com classificador CRF.

Trabalhos	(Características usadas)	Classificador	Domínio	Ano
[61]	1, 2, 12, 13, 15, 16, 17, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37	CRF	Produto / Filme	2010
[70]	1, 10, 12, 40, 49, 50, 51	L-HMM / CRF	Câmeras digitais	2012
[62]	1, 12, 38, 39, 41	CRF	Multidomínio	2010
[64]	1, 10, 12, 13, 14, 18, 19, 20, 25, 42, 43, 44, 45, 46, 47, 48	CRF	MPQA (Respostas a Questionários Multi-perspectivos)	2010
[65]	1, 10, 12, 13, 18, 22, 38, 39, 40, 49	CRF	Produtos	2012
[66]	1, 12, 21, 25, 26, 31, 38, 39	CRF	MPQA (Respostas a Questionários Multi-perspectivos)	2013
[68]	1,2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15	CRF / Regressão Logística	Restaurantes / Laptops / Hotéis	2015
[69]	1, 3, 8, 10, 12, 39, 52	MFE-CRF	Restaurantes / Laptops	2018

O campo da Análise de Sentimento (AS) por vezes vai um pouco além da identificação de componentes em um conjunto de dados. De forma trivial, os métodos de AS utilizam regras estatísticas, classificando um segmento textual muitas vezes pelo percentual de determinada característica em detrimento de outra. Abordagens um pouco

melhores conseguem ir além, e podem mensurar a polaridade de uma sentença inteira a partir de aprendizagem de máquina, ou ainda identificar características mais particulares e pontuais (aspectos) em uma sentença ou documento.

Abordagens mais sofisticadas sugerem a aplicação do método de ASBA com o uso de dependências, sendo focadas na inter-relação entre os componentes textuais e na análise de aspectos de forma particular, não apenas identificando, mas também classificando segmentos de uma sentença em categorias, além da inter-relação de dependência entre eles, inclusive com classificadores que trabalham com essas relações contextuais, como é o caso do CRF, o que dá a essa abordagem um caráter bem mais aprofundado e subjetivo, de características qualitativas.

A literatura tem demonstrado que uma das tarefas mais importantes na Análise de sentimento Baseada em Aspectos é a extração do termo de opinião (OTE, ou simplesmente Aspecto). A execução dessa tarefa é um dos objetivos de encontros anuais da comunidade científica, a saber, o *Semantic Evaluation* (SemEval), evento já citado, e que há quase uma década reúne equipes de vários centros de pesquisa ao redor do mundo para a execução de desafios em análise de computação semântica.

No decorrer do desenvolvimento deste trabalho, percebemos que diante das revisões da literatura acerca de onde esta pesquisa poderia contribuir com a comunidade científica, pontuamos o estudo e seleção das principais características para extração de termos de opinião e o desenvolvimento de um framework para esta tarefa.

Os próximos capítulos detalham os métodos aplicados nessas atividades, assim como os resultados obtidos.

4

Método

O presente capítulo trata do caráter científico deste trabalho que, apesar de sugerir resultados discretos, apresenta também características de uma abordagem qualitativa, e descreve o uso de técnicas de extração e avaliação de características destinada à identificação de Aspectos (*Opinion Target Expression* - OTE) em Análise de Sentimento, para aplicação direta em um conjunto de revisões de restaurantes.

Este capítulo foi dividido em quatro seções:

- **Coleta e estrutura do banco de dados**: formatação dos dados para aplicação no sistema;
- **Extração de características:** desenvolvimento de um sistema para extração das características mais relevantes;
- **Análise de características:** selecionar as características com maior ganho de informação;
- **Classificação:** aplicar o modelo classificador para extração dos aspectos

4.1. Coleta e estrutura do banco de dados (*dataset*)

Para avaliação da ferramenta proposta, foram utilizados como dados de entrada um conjunto de dados contendo postagens de revisões de restaurantes. O *dataset* disponível foi obtido de diversas revisões de restaurantes ao longo de vários meses. Este conjunto de dados é distribuído para a Tarefa 4 - Análise de Sentimento Baseado em Aspectos (ABSA), do SemEval-2014[2].

Os datasets oficiais do SemEval 2014 foram apresentados no formato xml.

[2] http://alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools

4.2.: Entrada de dados

Aqui serão explicitados os passos executados para tratamento do conjunto de dados adquiridos para experimentação.

4.2.1 Preparação dos datasets

Os dados de treinamento de restaurantes, com 3044 frases em inglês, é um subconjunto do conjunto de dados do repositório do SemEval2014, que incluiu anotações para categorias de aspecto e polaridades gerais das frases. Foram adicionadas anotações para termos de aspecto que ocorrem nas sentenças (SB1), polaridades de termos de aspecto (SB2) e polaridade da categoria do aspecto (SB4). O conjunto de testes foi composto por 800 sentenças organizadas e estruturadas da mesma forma. Não havia no dataset de teste nenhuma sentença comum ao de treinamento.

4.2.1.1. Processo de Anotação

Para uma determinada entidade-alvo (no caso deste trabalho, restaurante) ser revisada, os anotadores foram solicitados a fornecer dois tipos de informação: termos de aspecto (SB1) e polaridades de termos de aspecto (SB2). Além disso, duas camadas adicionais de anotação foram adicionadas: categoria de aspecto (SB3) e polaridade da categoria do aspecto (SB4). Os anotadores utilizaram o BRAT [71], uma ferramenta de anotação baseada na Web, configurada adequadamente para as necessidades da Tarefa 1 (ABSA) do SemEval. A Figura 4 mostra uma sentença anotada no BRAT, conforme visualizada pelos anotadores.

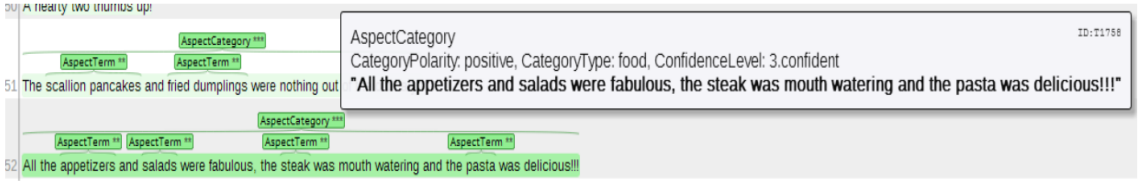


Figura 4: Uma sentença na ferramenta BRAT, anotada com quatro termos de aspecto ("*appetizers*", "*salads*", "*beef*", "*pasta*") e uma categoria de aspecto (FOOD). Para as categorias de aspecto, toda a sentença é marcada. [72].

Estágio 1: termos e polaridades do aspecto. Durante uma primeira fase de anotação, os anotadores marcaram todos os termos de uma ou várias palavras que nomearam aspectos particulares da entidade de destino (por exemplo, "*I liked the service and the staff, but not the food*" → {*'service', 'staff', 'food'*}). Eles foram convidados a marcar somente os termos de aspecto que nomeiam explicitamente aspectos específicos (por exemplo, "tudo a respeito" ou "é caro" não mencionam aspectos específicos). Os termos do aspecto foram anotados como apareceram, mesmo se incorreto (por exemplo, *"warrenty"* em vez de *"warranty").* Cada termo de aspecto identificado também teve que ser atribuído um rótulo de polaridade (positivo, negativo, neutro, conflito). Por exemplo, *" I hated their fajitas, but their salads were great" → {'fajitas': negative, 'salads': positive}*.

Cada sentença do conjunto de dados foi anotada por dois anotadores, um estudante de pós-graduação (anotador A) e um linguista especialista (anotador B). Inicialmente, um subconjunto de sentenças (300 do conjunto de dados total) foi anotado pelo anotador A e as anotações foram inspecionadas e validadas pelo anotador B. As divergências entre os dois anotadores foram confinados a casos limítrofes. Levando em consideração os tipos dessas discordâncias (discutidas abaixo), o anotador A foi fornecido com diretrizes e marcou o restante das sentenças em ambos os conjuntos de dados. 2 Quando A não estava confiante, uma decisão foi tomada em colaboração com B. Quando A e B discordaram, foi tomada uma decisão colaborativamente por eles e mais um terceiro anotador especialista. A maioria das divergências se enquadra em um dos três tipos a seguir:

- **Ambiguidade de polaridade:** em várias frases, não ficou claro se o revisor expressou opinião positiva ou negativa, ou nenhuma opinião (apenas um fato), devido à falta de contexto. Por exemplo, em "*we had a 15 minute wait time for the food"* não está claro se o revisor expressa uma opinião positiva, negativa ou sem opinião sobre o termo "*wait time*".

- **Limites na detecção de aspectos formados por mais de uma palavra:** em vários casos, os anotadores discordaram sobre os limites de termos de aspecto com várias palavras quando apareceram em conjunções ou disjunções (por exemplo, "Seleção de carnes e frutos do mar", "macarrão e pratos de arroz"). Em tais casos, foi solicitado aos anotadores que anotassem todos os termos que descrevem um aspecto como um aspecto único. Outros

desacordos dizem respeito à extensão dos termos do aspecto quando adjetivos que podem ou não ter um significado subjetivo também estavam presentes. Por exemplo, se "grande" em "grande camarão inteiro" é parte do nome do prato, então as diretrizes exigem que o adjetivo seja incluído no termo de aspecto; de outra forma (por exemplo, em "grandes partes") "grande" é um adjetivo modificador genérico, e não deve ser incluído no termo de aspecto. Apesar das orientações, em alguns casos foi difícil isolar e marcar o termo exato, por causa de palavras intervenientes, pontuação ou relações de dependência entre palavras distantes umas das outras.

- **Termo de Aspecto vs. referência à entidade-alvo:** em alguns casos, não ficou claro se um substantivo ou uma frase substantiva foi usado como o termo de aspecto ou se referiu à entidade sendo revista como um todo. Por exemplo, na sentença “Este lugar é incrível”, o termo “lugar" mais provavelmente se refere ao restaurante como um todo (portanto, não deve ser marcado como um termo de aspecto). Por outro lado em "*Cozy place and good pizza*” provavelmente se refere ao ambiente do restaurante. Um contexto mais amplo novamente ajuda em alguns desses casos.

4.2.2 Formato do dataset

Os conjuntos de dados da tarefa do ABSA foram fornecidos em formato XML. Eles estão disponíveis com uma licença não comercial, sem redistribuição através do META-SHARE, um repositório dedicado ao compartilhamento e disseminação de linguagem de recursos humanos [73]. Os datasets estão disponíveis livremente para a comunidade científica, sob única exigência de se possuir uma conta do repositório para download.

Como já exposto, este trabalho fez uso de um dataset de revisões de restaurantes de acordo com as diretrizes do SemEval 2014. O arquivo de treinamento em formato .xml continha 3044 sentenças de revisões previamente anotadas por uma tag específica. A tarefa inicial então foi extrair as informações do arquivo .xml, dentro do ambiente de programação Python. Para isso foi utilizada a biblioteca *Element Tree*. O *Element Tree* é um objeto contêiner simples, mas flexível, projetado para armazenar estruturas de dados hierárquicas, como infosets XML simplificados, na memória. O tipo de elemento pode ser descrito como um cruzamento entre uma lista do Python e um dicionário do Python. Deste

modo, um vetor contendo todas as sentenças do XML foi gerado, podendo cada sentença ser acessada e manipulada individualmente.

4.3. Extração de Características

Em Análise de sentimento nenhuma atividade é tão importante quanto a seleção de características. É nesta etapa que serão extraídas das sentenças todas as informações úteis do texto, e a partir das quais o classificador irá realizar treinamento para posterior inferência e classificação das sentenças. Deste modo, grande parte deste trabalho foi destinada ao estudo e pesquisa das principais características utilizadas em ABSA, as mais aplicadas em CRF e as que poderiam contribuir de forma mais pontual para o domínio utilizado neste trabalho.

Característica 1: Divisão de sentenças e obtenção do Token

Para a etapa de divisão de sentenças, a biblioteca TextBlob foi utilizada. O TextBlob é uma biblioteca para processamento de dados textuais, fornecendo uma API simples para tarefas comuns de processamento de linguagem natural (NLP), como marcação de POS , tokenização, extração de frases nominais, análise de sentimento, classificação, tradução etc.

Assim, aplicando a função de tokenização em cada sentença, no vetor geral de sentenças, foram gerados subvetores com cada palavra de cada frase separadas por vírgulas. Caracteres especiais como: “.”, “;”, “?”, “!”, etc. foram removidos .

A primeira característica de cada Token é o nome do token em si. Assim, para a expressão: *The drinks are really cheap*, a Característica 1 do primeiro token é o termo “*The*”

Característica 2: PoS-tag

A característica Part of Speech (Tabela 2) é de uso quase padrão em Análise de Sentimento, isso porque a classificação morfológica de cada termo de uma sentença é base fundamental para que se entendam as relações morfossintáticas que estruturam a linguagem, em qualquer idioma. No caso do Inglês isto não é diferente. Para se aplicar o

POS Tag de cada termo, mais uma vez foi utilizada a biblioteca TextBlob, já descrita no tópico anterior.

Tabela 2: POS Tag Table. Fonte: o autor.

Part-Of-Speech Tags (POS-TAG)

THE PENN TREEBANK PROJECT |

N.	Tag	Description
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Pre-determiner
17	POS	Possessive ending
18	PRP	Personal pronoun
N.	**Tag**	**Description**
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	to
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

Geração do Vetor de Características

De posse da primeira característica, um arquivo em formato.txt é criado pelo sistema na pasta local da aplicação. A partir deste momento, este arquivo vai receber dados da aplicação a cada ciclo de extração das características (*Features*). O primeiro ciclo constrói um vetor para cada token do dataset e acrescenta a este o valor da Característica 1 (F-1, "token") . Todo o processo de extração ocorre em 23 ciclos. Esta estratégia de criação de um arquivo na memória física local possibilitou uma economia significativa da memória de acesso randômico. Cada ciclo executado pelo sistema representa a extração de uma característica, e a anotação do respectivo valor na F-nésima posição de cada um dos vetores gerados (um para cada token).

Como exemplo, seja a primeira sentença do dataset: "*The food is excellent*", em seu segundo ciclo, os vetores representativos dessa sentença estarão com os elementos relativos às Características 1 e 2 (token e PoS-tag, respectivamente):

FeaturesVector1 = {((*The*), (DT))}
FeaturesVector2 = {((*food*), (NN))}
FeaturesVector3 = {((*is*), (VBZ))}
FeaturesVector4 = {((*excellent*), (JJ))}

Este processo é feito em cada sentença do vetor de sentenças. A cada nova característica aplicada, os vetores de características são incrementados com o novo valor. Cada ciclo de extração de características é aplicado a cada token um a um. Ao final de cada ciclo, uma nova função de extração de características é aplicada e assim sucessivamente até que todos os vetores tenham como elementos os dados relativos às 23 características aplicadas.

Característica 3: PoS-tag do termo subsequente

Como já comentado, em todas as línguas, as palavras se relacionam para a construção da informação. Certas classes de palavras têm uma relação íntima umas com as outras, sobretudo quando uma delas caracteriza-se por ser um termo modificador, ou determinante de uma outra. Como exemplo temos as relações entre determinantes e

substantivos (“The food”), ou de advérbios e adjetivos (“*very tasteful*”). A partir de tais premissas, é mister perceber que um termo subsequente pode ajudar a inferir-se o anterior.

Característica 4: Lemma

Lemma é uma das características mais presentes na literatura para extração de informações em documentos de texto. Neste trabalho, a definição do lema foi feita utilizando-se o módulo *WordLemmatizer* do NLTK (*Natural Language Toolkit*). O NLTK é uma plataforma para criar programas em Python voltado ao trabalho com dados em linguagem humana. Ele fornece interfaces fáceis de usar para mais de 50 recursos corpora e lexicais como o *WordNet*, juntamente com um conjunto de bibliotecas de processamento de texto para classificação, tokenização, *stemming*, *tagging*, análise e raciocínio semântico [74]

Característica 5: Stemming

Assim como o *lemma*, o *stemming* dos tokens também foi realizado a partir da plataforma NLTK em Python. Neste caso, utilizou-se o módulo *SnowballStemmer*, ignorando-se as *stopwords*.

Característica 6: Superlativo

A verificação dos tokens quanto ao caso de serem superlativos ou não, foi realizada gerando-se um vetor de termos contendo os POS-tags referentes à classificação de um termo como superlativo. De acordo com a tabela de códigos PoS-tags (tabela do projeto *Penn Tree Bank*, já demonstrada em seção anterior), o PoS-tag de termos no superlativo é “JJS”, no caso de Adjetivos Superlativos ou “RBS”, para o caso de Advérbios Superlativos. Assim, criado o vetor, a tarefa foi apenas comparar a Característica 2 (PoS-tag) de cada token com o conteúdo do vetor de superlativos. Caso o token em questão tivesse o PoS-tag igual a JJS ou RBS, a Característica 6 recebia o valor binário 1, caso contrário, 0.

Característica 7: Comparativo

Analogamente à classificação dos tokens como superlativos ou não, foi a verificação de termos comparativos. A única diferença limitou-se aos 2 elementos do vetor

de comparativos serem “JJR” e “RBR”, respectivamente, Adjetivo Comparativo e Advérbio Comparativo. Assim, a Característica 7 também recebia valores binários: 0 ou 1.

Característica 8: Termos Negativos nos 4 últimos

Como já explicitado, a presença de termos negativos pode alterar o sentido de uma expressão ou negar um termo ao qual esta palavra esteja relacionada. De acordo com algumas pesquisas (algumas já mencionadas na seção de trabalhos relacionados), a negação de uma ideia não ocorre apenas com palavras de negação imediatamente sucessoras ou antecessoras a um termo em questão. Expressões do tipo “*Nobody at that restaurant liked the food*” trazem o “modificador” do termo “*liked*” (*Nobody*) distante em 4 posições do verbo, que, apesar de, isoladamente ter conotação positiva, dentro do contexto teve o sentido anulado diante da presença do pronome “*nobody*”. Interessante observar o fato de que pronomes **não** são semanticamente estruturas modificadoras, mas dentro de um contexto frasal, o valor positivo de um verbo torna-se nulo ao indicarmos que “ninguém” executa a ação do mesmo. Diante desses fatos, optou-se por considerar a busca por termos negativos em 4 posições anteriores a cada token. As exceções dos termos nas primeiras posições das sentenças foram tratadas pelo código. A verificação dos termos negativos ocorreu com a criação de um vetor com os principais termos negativas do idioma Inglês, alguns descritos abaixo, em estrutura semelhante ao vetor:

negative_vector = {***"no"***, ***"none"***, ***"nobody"***, …… ***"not"***, ***"hardly"***, ***"scarcely"***, ***"cannot"***, ***"no"***, ***"never"***}

Foi então verificada a ocorrência de cada token da sentença de entrada no vetor de negação (*negative_vector*). Em caso positivo, a Característica 8 no vetor de características recebia o valor binário “1”. Do contrário, “0”.

Característica 9: Pontuação Positiva

O *SentiWordNet*[3] é um recurso léxico para a mineração de opinião. Ele atribui a cada sincronia do *WordNet* (uma grande base de dados léxicos em Inglês) três pontuações

[3]http://sentiwordnet.isti.cnr.it

de sentimento: positividade, negatividade, objetividade. O módulo *senti_synset* do *SentiWordNet* consegue gerar valores decimais (entre 0 e 1) para positividade. Uma vez que o sistema classificador empregado neste trabalho não aceita valores decimais como entrada, foi necessário discretizar tais valores. Para isso foi gerada uma função que categoriza cada termo em seis níveis de positividade diferentes (0 a 5). A Característica 9 recebia então o valor de Positividade de cada token da sentença: "0" para um token pouco positivo e "5", para um muito positivo.

Característica 10: Pontuação Negativa

Exatamente os mesmos procedimentos empregados para a geração da Característica 9 o foram para o Score Negativo. Assim, o nível de negatividade de cada token era gerado em níveis de categoria variando de 0 a 5. A Característica 10 então recebia valor "0" para um token pouco negativo e "5" para um muito negativo, por exemplo.

As características de Análise Sintática e dependência (11 a 16)

As classes sintáticas das palavras já foram abordadas na seção de Referencial Teórico. A interdependência entre uma sequência de termos de uma sentença é o que fundamenta a comunicação. Sem relações sintáticas não há significado no contexto linguístico. A identificação de cada ente que compõe uma frase é de suma importância na área de Extração de Aspectos, e métodos automatizados de detecção das interconexões frasais agregam um grande valor e utilidade à área.

O *Spacy*[4] é uma biblioteca de software de código aberto para processamento avançado de linguagem natural, escrita nas linguagens de programação *Python* e *Cython*. A biblioteca é publicada sob a licença do MIT e atualmente oferece modelos estatísticos de redes neurais para vários idiomas. Os sistemas capazes de efetuar análise de dependência conseguem identificar relações sintáticas na estrutura frasal (Fig.4)

[4] https://spacy.io

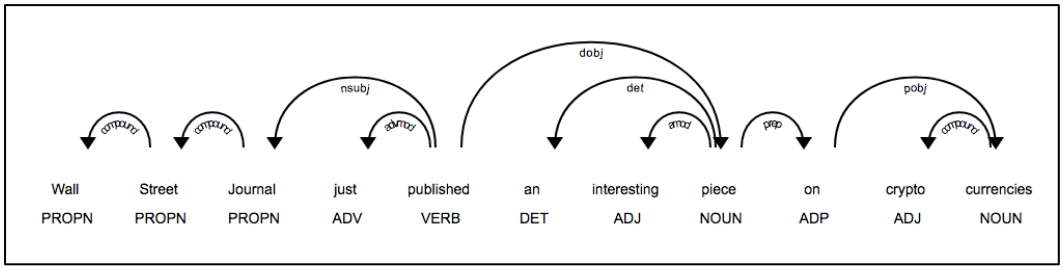


Figura 7: modelo de Análise de dependência. Fonte: Spacy

Neste projeto foram aplicadas análises de dependência para a identificação de:

- **Sujeito** (Característica 11)
- **Objeto direto** (Característica 12)
- **Objeto indireto** (Característica 13)
- **Verbo de Ligação** (Característica 14)
- **Conjunção** (Característica 15)
- **Conjunção Coordenada** (Característica 16)

Para essa tarefa, foi utilizada a biblioteca Spacy. Uma vez que as relações de dependência ocorrem a nível de sentença, um novo vetor com divisão em sentenças foi gerado. Neste novo vetor, as palavras não estavam mais tokenizadas, o módulo do Spacy responsável pela análise de dependência requer sentenças completas, inclusive com pontuações e *stopwords*, uma vez que nessa classe de termos encontram-se os conectores frasais como conjunções, por exemplo, importantes para análise do discurso.

Foram gerados 6 novos vetores (4 para cada um dos componentes do enunciado – sujeito, obj. direto, obj. indireto, Verbo de Ligação – e 2 para as conjunções). A cada sentença analisada, cada um dos vetores recebia como elemento o termo classificado em sua categoria.

Tomando-se como exemplo a sentença:

We enjoyed ourselves thoroughly and will be going back for the desserts

Temos:

subject_vector = [“we”]

direct_object_vector [“ourselves”]

indirect_object_vector [ ]

copula vector [ ]

conjunction_vector [ ]

coord_conj_vector [“and”]

Em seguida, cada token da sentença era comparado com o conteúdo dos vetores. Nos casos de *match*, a característica correspondente àquele vetor recebia valor “1”, do contrário, “0”. No exemplo acima, o token “WE” teria ao final do *parser*:

- Característica 11: 1
- Característica 12: 0
- Característica 13: 0
- Característica 14: 0
- Característica 15: 0
- Característica 16: 0

Por sua vez, o token “AND” receberia:

- Característica 11: 0
- Característica 12: 0
- Característica 13: 0
- Característica 14: 0
- Característica 15: 0
- Característica 16: 1

Característica 17: Sinonímia

Para a geração das características de sinonímia, foi feito uso do componente *syn.lemmas* do WordNet no ambiente Python. A principal relação entre as palavras no *WordNet* é sinonímia. Sinônimos são palavras que denotam o mesmo conceito e são intercambiáveis em muitos contextos. No *WordNet* eles são agrupados em conjuntos não

ordenados (*synsets*). Cada um dos 117.000 *synsets* do *WordNet* é vinculado a outros *synsets* por meio de um pequeno número de "relações conceituais".

Assim, a Característica 17 recebeu como valor o termo sinônimo do token corrente, baseado no banco de dados do *WordNet*, que em alguns casos pode retornar como valor o próprio termo.

Exemplo:

***Food** is always fresh and hot – ready to eat*

Característica 17: food

Característica 18: Antonímia

De forma análoga à última característica descrita, o processo para detecção de antônimos (termo com significado oposto) também fez uso do WordNet. Assim, a Característica 18 de cada token recebia como valor o antônimo deste token.

Para ambas as Características, no caso de o token não ter um sinônimo ou antônimo definido, as respectivas características recebiam o valor "NULL".

Exemplo:

***Food** is always fresh and hot – ready to eat*

Característica 18: NULL

Característica 19: Caractere especial

Para a detecção de caracteres especiais, foi utilizada uma expressão regular comparando-se cada token com uma lista de termos representantes de caracteres especiais. Em caso de match, a Característica 19 recebia valor 1, caso contrário, 0.

Característica 20: Distância a adjetivos e/ou advérbios

Como o próprio nome sugere, a Característica 20 armazenou a distância do token ao adjetivo mais próximo. O procedimento consistiu em verificar o PoS-tag dos tokens anteriores e posteriores ao atual até se encontrar o 1º adjetivo ou advérbio. Em casos que na sentença não houvesse tais classes de palavras, era retornado o valor "NULL" para a característica.

Característica 21: É ou não *StopWord*

Utilizando o Word2Vec, foi criada uma função que verifica cada token e o compara com um vetor contendo uma lista de stopwords em Inglês. Caso o token esteja nessa lista, a Característica 21 recebia o valor 1, e em caso negativo, 0.

Característica 22: Similaridade do token com as categorias.

Para a detecção de similaridade do token com cada categoria de Aspecto definida pelo SemEval 2014 para o domínio de restaurantes, a saber: ***food, service, price, ambience, anecdotes/miscellaneous,*** foi utilizado um método baseado em *Word Embedding* e Word2Vec. O Word2vec é um grupo de modelos relacionados que são usados para produzir encartes de palavras (*word embeddings*). Esses modelos são redes neurais superficiais de duas camadas que são treinadas para reconstruir contextos linguísticos de palavras. O Word2vec leva sua entrada para um grande corpus de texto e produz um espaço vetorial, normalmente de várias centenas de dimensões, com cada palavra exclusiva no corpus sendo atribuída a um vetor correspondente no espaço. Os vetores de palavras são posicionados no espaço de modo que eles compartilham contextos comuns no corpus, e estão localizados em estreita proximidade um do outro no espaço, sendo a similaridade entre elas definida pelo método do cosseno [75].

Coluna 23 – CLASSE: Notação BIO - Tag de marcação (OTE)

A 23ª coluna da matriz de características é a **tag de marcação** (classe) no conjunto de treinamento. Esta coluna informa se o token correspondente é uma *Opinion Target Expression* (OTE) e auxilia para que o classificador "aprenda" quais os termos-alvo e quais não. Para marcação da OTE, foi utilizada a notação BIO.

A notação BIO, ou IOB (abreviatura de *inside, outside, beginning*) é um formato de marcação comum para etiquetagem de tokens numa tarefa de agrupamento em linguística computacional [76]. O prefixo B em uma tag indica que a tag é o começo de uma sentença de termos considerados alvo. O prefixo I indica que a tag está dentro do alvo. A tag B é usada somente quando uma tag é seguida por uma tag do mesmo tipo sem O entre elas. Uma tag O indica que um token não é alvo.

Exemplo: Considere-se a frase: *The hot french fries are delicious*

Neste caso, temos o termo “hot french fries” como o aspecto a ser marcado. Usando-se a notação BIO, a feature 23 para este trigrama seria assim representada:

Tokens	**Feature 23 (BIO)**
The	O
hot	B
french	I
fries	I
are	O
delicious	O

4.4. Geração do arquivo de saída

Uma vez aplicada a extração de todas as características pelo sistema desenvolvido, uma grande matriz em formato.txt foi gerada. Esse documento é composto por todas as sentenças do corpus de entrada, segmentadas token a token e dispostas em n linhas (onde n é o número de tokens de cada sentença) e 23 colunas, onde cada coluna representa uma das 22 características aplicadas e a última, a *tag* de marcação. A separação entre as sentenças é dada por uma linha vazia. Como já descrito, a primeira coluna da matriz representa o token em si, e a última, a tag que define se o termo é uma OTE (aspecto), ou não. Essa é a estrutura tanto do documento de treinamento, como de teste. Esses documentos seguem o formato de entrada exigido pelo classificador utilizado. A figura 5 ilustra esse formato, com representação das primeiras 2 sentenças. O arquivo com todas as sentenças analisadas servirá de entrada para o classificador. Na imagem, a linha em branco divide as duas sentenças tokenizadas. Cada coluna representa uma das 22 características, extraídas após tratamento no sistema desenvolvido em Python, e a *tag* de marcação na última coluna.

F-1	F-2	F-3	F-4	F-5	F-6	F-7	F-8	F-9	F-10	F-11	F-12	F-13	F-14	F-15	F-16	F-17	F-18	F-19	F-20	F-21	F-22	F-23
Food	NN	VBZ	Food	food	0	0	0	0	0	1	0	0	0	0	0	food	NULL	0	2	0	3	B
is	VBZ	RB	is	is	0	0	0	2	1	0	0	0	0	0	0	be	differ	0	1	1	0	O
always	RB	JJ	always	alway	0	0	0	0	0	0	0	0	0	0	0	always	never	0	1	0	0	O
fresh	JJ	CC	fresh	fresh	0	0	0	2	4	0	0	0	0	0	0	fresh	stale	0	1	0	1	O
and	CC	JJ	and	and	0	0	0	0	0	0	0	0	0	0	1	NULL	NULL	0	1	1	0	O
hot-	JJ	NN	hot	hot	0	0	0	0	0	0	0	0	0	1	0	hot	cold	0	2	0	0	O
ready	NN	TO	ready	readi	0	0	0	0	2	0	0	0	0	0	0	ready	unready	0	1	0	0	O
to	TO	VB	to	to	0	0	0	0	0	0	0	0	0	0	0	NULL	NULL	0	2	1	0	O
eat	VB	NULL	eat	eat	0	0	0	0	0	0	0	0	0	0	0	eat	NULL	0	0	0	2	O
Did	NNP	PRP	Did	did	0	0	0	0	0	0	0	0	0	0	0	make	unmake	0	NULL	0	0	O
I	PRP	NN	I	i	0	0	0	0	0	1	0	0	0	0	0	iodine	NULL	0	NULL	0	0	O
mention	NN	IN	mention	mention	0	0	0	0	0	0	0	0	0	0	0	mention	NULL	0	NULL	0	0	O
that	IN	DT	that	that	0	0	0	0	0	0	0	0	0	0	0	NULL	NULL	0	NULL	1	0	O
the	DT	NN	the	the	0	0	0	0	0	0	0	0	0	0	0	NULL	NULL	0	0	1	0	O
coffee	NN	VBZ	coffee	coffe	0	0	0	0	0	1	0	0	0	0	0	coffee	NULL	0	0	0	1	B
is	VBZ	VBN	is	is	0	0	0	2	1	0	0	0	0	0	0	be	differ	0	0	1	0	O
OUTSTANDING	VBN	NULL	)UTSTANDIN(	outstand	0	0	0	4	0	0	0	0	0	0	0	outstanding	NULL	0	0	0	0	O

Figura 8: trecho do arquivo de treinamento gerado pela aplicação Python, após extração das 22 características do dataset. A última coluna representa a tag de marcação na notação BIO.

Após realização do pré-processamento das sentenças de revisões de restaurantes e extração das características relevantes no sistema desenvolvido em Python, foram gerados dois arquivos de saída (treinamento e teste) compostos, cada um, de uma grande matriz de características contendo 23 colunas e 32.949 linhas no arquivo de treinamento e 23 colunas e 9.282 linhas no de teste. Esses dois arquivos foram aplicados a um sistema de análise de ganho de informação (seção 4.3) e utilizados como dados de entrada para o classificador CRF, cuja utilização será discutida em detalhes na seção 4.4.

4.5. Métodos para avaliar importância das características

Após o processo de extração de características, foi realizada a análise de importância das mesmas. Esta etapa é de suma importância, dado que um dos principais problemas de categorização de texto é a alta dimensionalidade dos vetores / matrizes de características [77]. O espaço de características nativo consiste nos termos exclusivos (palavras ou frases) que ocorrem em documentos, que podem ter dezenas ou centenas de milhares de termos, mesmo para uma coleção de textos de tamanho moderado. Isso é proibitivamente alto para muitos algoritmos de aprendizado de máquina. Se reduzirmos o conjunto de características considerado pelo algoritmo, podemos servir dois propósitos: diminuir consideravelmente o tempo de execução do algoritmo de aprendizado e / ou aumentar a precisão do modelo resultante. Nesta linha, um número de pesquisas já abordou a questão da seleção de subconjuntos de características [78], [79]. Yang e Pederson [80]

encontraram o ganho de informação (IG) e o teste do CHI-quadrado (CHI) mais efetivos na remoção de vários termos sem perder a precisão da categorização em seus experimentos. Assim, os métodos de Ganho de Informação (*IG*) e CHI-quadrado foram empregados na matriz de características gerada pelo Python a fim de se otimizarem os resultados, com uma melhor seleção das características mais relevantes.

Após a aplicação dos algoritmos, foram definidos três ensaios diferentes para serem rodados no classificador:

- Treinamento e teste com todas as características;
- Treinamento e teste com todas as características com exceção das 3 pior ranqueadas pelos algoritmos IG e CHI quadrado;
- Treinamento e teste apenas com as 8 características mais bem ranqueadas nos testes de IG e CHI quadrado.

4.6. Classificação

Diversas são as ferramentas descritas na literatura para implementação do modelo CRF. Em seu trabalho, Sutton *et* al ([30]) elenca algumas das principais e mais populares (Fig.6):

CRF++	http://crfpp.sourceforge.net/
MALLET	http://mallet.cs.umass.edu/
GRMM	http://mallet.cs.umass.edu/grmm/
CRFSuite	http://www.chokkan.org/software/crfsuite/
FACTORIE	http://www.factorie.cc

Figura 9: Ferramentas destinadas à implementação do modelo *Conditional Random Fields.* Fonte: [30]

Dentre as aplicações descritas acima, o CRF++[5] foi escolhido por ser gratuito, apresentar bons resultados na literatura, além de bem documentado.

[5] http://wing.comp.nus.edu.sg/~forecite/services/parscit-100401/crfpp/CRF++-0.51/doc/

4.6.1. O CRF++

O CRF ++ é uma implementação personalizável e de código aberto destinada à aplicação do modelo CRF (*Conditional Random Fields*) para segmentar / rotular dados sequenciais. O CRF ++ foi projetado para fins genéricos e será aplicado neste trabalho como ferramenta de extração da expressão-alvo de opinião (OTE).

Etapas seguidas nos experimentos com o classificador.

De acordo com a documentação da ferramenta utilizada, tanto o arquivo de treinamento quanto o arquivo de teste precisam estar em um formato específico para que o CRF ++ funcione corretamente (Fig. 6). De um modo geral, o arquivo de treinamento e teste deve consistir de vários tokens. Além disso, um token consiste em várias colunas. Cada token deve ser representado em uma linha, com as colunas separadas por espaços em branco (espaços ou caracteres tabulares). Uma sequência de tokens se torna uma sentença. Para identificar o limite entre as sentenças, uma linha vazia é colocada.

Algumas “exigências” da aplicação também precisaram ser satisfeitas, já que há um tipo de semântica entre as colunas do arquivo de entrada. A primeira coluna, por exemplo, é a “palavra" em si, a segunda, obrigatoriamente o *PoS-tag*. A última coluna, por sua vez, representa a etiqueta de resposta que será treinada pelo CRF.

Configurando o CRF++

Como o CRF ++ é projetado como uma ferramenta de uso geral, foi necessário especificar os modelos (*templates*) de características antecipadamente. É este template que define quais características serão utilizadas para treinamento e teste[6]. A figura 7 apresenta ums egmento de um dos modelos de template gerados nos experimentos.

[6] Descrição do procedimento baseada na documentação do CRF++

```
# Unigram

#combinações| não executadas:#

U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
#U05:%x[-1,0]/%x[0,0]
#U06:%x[0,0]/%x[1,0]

U07:%x[-2,1]
U08:%x[-1,1]
U09:%x[0,1]
U10:%x[1,1]
U11:%x[2,1]
#U12:%x[-2,1]/%x[-1,1]
#U13:%x[-1,1]/%x[0,1]
#U14:%x[0,1]/%x[1,1]
#U15:%x[1,1]/%x[2,1]

#U16:%x[-2,1]/%x[-1,1]/%x[0,1]
#U17:%x[-1,1]/%x[0,1]/%x[1,1]
#U18:%x[0,1]/%x[1,1]/%x[2,1]

U19:%x[-2,2]
U20:%x[-1,2]
U21:%x[0,2]
U22:%x[1,2]
U23:%x[2,2]

U24:%x[-2,3]
U25:%x[-1,3]
U26:%x[0,3]
U27:%x[1,3]
U28:%x[2,3]
```

Figura 10: Trecho de um dos templates utilizados na fase de treinamento do classificador CRF++.

Modelo básico e macro

Cada linha no arquivo de templates denota um template. Em cada template, um macro especial %x[linha, coluna] era usado para especificar um token nos dados de entrada. A linha especifica a posição relativa do token atual e a coluna especifica a posição absoluta da coluna, ou seja a posição de cada uma das características. As figuras 8 e 9 ilustram essa configuração:

```
Input data

F-1         F-2    F-3
He          PRP    B-NP
reckons     VBZ    B-VP
the         DT     B-NP << CURRENT TOKEN
current     JJ     I-NP
account     NN     I-NP
```

Figura 11: Ilustração representativa do token corrente e dos condicionais acima e abaixo.Fonte: https://taku910.github.io/crfpp/ (adaptado pelo autor)

template	expanded feature
%x[0,0]	the
%x[0,1]	DT
%x[-1,0]	reckons
%x[-2,1]	PRP
%x[0,0]/%x[0,1]	the/DT
ABC%x[0,1]123	ABCDT123

Figura 12: Trecho ilustrativo do arquivo de template. O token atual "the" congrega informações (características) marginais, inclusive de outros tokens. Fonte: https://taku910.github.io/crfpp/

Tipos de Template

Nas configurações do CRF++, há a possibilidade de se aplicarem dois tipos de modelos (*templates*):

- **Unigrama** (U): Este é um modelo para descrever os recursos de um unigrama. Quando é definido um *template* do tipo "U01:% x [0,1]", o CRF ++ gera automaticamente um conjunto de funções de *características* (func1... funcN) como demonstrado abaixo:

func1 = if (output = B-NP and feature="U01: DT") return 1 else return 0
func2 = if (output = I-NP and feature="U01: DT") return 1 else return 0
func3 = if (output = O and feature="U01: DT") return 1 else return 0
....
funcXX = if (output = B-NP and feature="U01:NN") return 1 else return 0
funcXY = if (output = O and feature="U01:NN") return 1 else return 0

O número de funções de recursos geradas por um modelo equivale a (**L * N**), em que L é o número de classes de saída e N é o número de sequência exclusiva expandida a partir do modelo fornecido.

- **Bigrama** (B): Com o modelo configurado como Bigrama, uma combinação do token de saída atual e do token de saída anterior (bigrama) é gerada automaticamente. Este tipo de *template* gera um total de características distintas (L

* L * N), em que L é o número de classes de saída e N é o número de características exclusivas geradas pelos modelos. Quando o número de classes é grande, esse tipo de modelo produziria uma quantidade enorme de características distintas, que causariam ineficiência tanto na fase de treinamento, quanto de teste.

Apesar de o número de classes de saída do sistema desenvolvido ser de duas (detecção se o token é ou não um Termo de Opinião), optou-se pela utilização apenas de um *template* de bigrama, "B", ou seja, apenas as combinações do token de saída anterior e do token atual são usadas como recursos do Bigrama.

De forma simplificada:

unigram: |output tag| x |all possible strings expanded with a macro|

bigrama: |output tag| x |output tag| x |all possible strings expanded with a macro|

Deste modo, foi gerado um *template* que selecionava para treinamento todas as 22 características de cada *token* no arquivo de entrada, levando também em consideração as mesmas características dos 2 tokens anteriores e 2 posteriores ao corrente para a classificação deste.

4.6.1.1. Etapa de treinamento

Tanto os arquivos de treinamento, quanto os de teste passaram pelo mesmo processo de extração de características.

Para execução do processo de aprendizagem do classificador, utilizou-se o CRF++ via *prompt* de comando. Os passos constam da chamada do método de treinamento e carregamento de 2 arquivos para cada ensaio realizado: template e banco de dados de treinamento (matriz de características gerada). A etapa de treinamento culmina com a geração de um modelo de aprendizado. Este modelo é **específico** para o dataset e template tomados como entrada. Ele apresenta todas as funções que o classificador vai aplicar no conjunto de teste para inferência dos resultados.

4.6.1.2. Etapa de classificação

Posteriormente à etapa de treinamento, um novo arquivo contendo dados de teste de Restaurantes do SemEval 2014 foi alimentado para o classificador. Esse dataset contém 23 colunas (22 características + tag de marcação) com 800 sentenças diferentes, num total de aproximadamente 9.000 tokens. Nenhuma das sentenças do conjunto de treinamento está presente no arquivo de testes, assim, o *cross-validation* não se fez necessário. Ainda, uma vez que o processo de aceitação e validação de resultados de uma pesquisa passa por etapas de comparação com trabalhos relacionados no estado da arte, foi preferível seguir-se à risca toda metodologia sugerida pelo SemEval2014, a título de tornar possível essa fase de validação dos resultados.

Para possibilitar que o CRF++ executasse os testes de classificação, o arquivo supracitado foi carregado no classificador. Nesta etapa, o modelo de treinamento recém-criado precisou ser invocado. Um detalhe a se observar é que tanto o conjunto de treinamento, quanto o de testes é composto das 23 colunas (22 características mais a **tag de classificação do arquivo de treinamento.** No CRF++, a 23ª coluna (*tag*) ela não deve ser suprimida, a própria aplicação desconsidera tal coluna na etapa de predição.

Como saída do classificador, é gerado um arquivo de mesma estrutura do de entrada, no entanto com **uma coluna a mais** (predição do classificador) (Tabela 3). A aplicação gera uma última coluna (CRF++) com a classificação obtida. Os valores com a tag de anotação estão listados na coluna F-23 (BIO). Valores iguais entre as duas colunas indicam um acerto do classificador. Neste trecho, especificamente, o único não acerto foi na última linha da região da tabela.

Tabela 3: Tabela parcial tratada de arquivo de saída do classificador CRF++. F...-22 (Características extraídas), F-23(BIO - tag de marcação da classe), CRF++ (predição do classificador).

F-18	F-19	F-20	F-21	F-22	F-23	
ANTONYM	IS_SPEC_CHARACT	DIST_ADJ_ADV	IS_StopWord	W2Vector	BIO	CRF++
NULL	0	1	0	0	O	O
NULL	0	2	0	2	B	B
differ	0	1	1	0	O	O
bottom	0	2	0	0	O	O
NULL	0	1	0	1	O	O
NULL	0	1	1	0	O	O
ill	0	1	0	0	O	O
NULL	0	NULL	0	0	O	O
lack	0	NULL	1	0	O	O
NULL	0	NULL	1	0	O	O
NULL	0	NULL	0	0	O	O
NULL	0	NULL	1	0	O	O
lack	0	NULL	1	0	O	O
NULL	0	NULL	0	0	O	O
NULL	0	NULL	1	0	O	O
NULL	0	NULL	1	0	O	O
slow	0	NULL	0	0	O	O
NULL	0	NULL	0	1	B	B
NULL	0	0	0	0	I	O

Apesar de a aplicação gerar um arquivo que contém todas as colunas, inclusive as de tag de marcação e a de predição, é curioso, no entanto, que a própria ferramenta não traga as medidas de desempenho como uma de suas funcionalidades. Deste modo, o arquivo de saída recém-gerado foi transferido para um formato de planilha eletrônica para a etapa de avaliação dos resultados.

5

Resultados

Este capítulo apresenta os resultados da avaliação. Ele será dividida em 4 subseções:

- **Medidas de avaliação de desempenho;**
- **Avaliação da importância das características;**
- **Predições do classificador;**
- **Análise dos dados.**

5.1. Medidas de Avaliação de desempenho

As medidas de avaliação são fundamentais para verificar se o modelo proposto apresenta resultados satisfatórios. Com o uso das métricas de Precisão, Cobertura e Medida F (*Precision*, Recall e F-*measure*), tem-se uma visão clara dos resultados e uma melhor interpretação de como o modelo de classificação está funcionando dentro do experimento proposto.

Para um melhor esclarecimento de cada uma das 3 medidas, algumas terminologias são fundamentais:

- **Verdadeiro Positivo (VP):** indica uma classificação correta da classe positivo. Por exemplo, a classe real é Positiva e o modelo classificou como Positivo.

- **Verdadeiro Negativo (VN):** indica uma classificação correta da classe negativo. Por exemplo, a classe real é Negativa e o modelo classificou como Negativo.

- **Falso Positivo (FP):** indica uma classificação errada da classe positivo. Por exemplo, a classe real é Negativa e o modelo classificou como Positivo.

- **Falso Negativo (FN) :** indica uma classificação errada da classe negativo. Por exemplo, a classe real é Positiva e o modelo classificou como Negativo.

Os conceitos de cada uma das 3 medidas de avaliação são descritos abaixo. Uma vez que este trabalho foi direcionado à identificação da **presença** de um termo de opinião, e devido ao fato de o número de tokens com valores positivos ser bem menor que o total de negativos, os cálculos das medidas restringir-se-ão aos valores positivos de precisão, cobertura e medida F :

- **Precisão (P):** de todas as instâncias que foram classificadas como pertencentes a uma determinada classe, quantas, na verdade, pertencem a essa classe.

Ou seja, o número de vezes que uma classe foi predita corretamente dividida pelo número de vezes que a classe foi predita.

- **Precisão de positivos (P_+)**

$$P_+ = \frac{VP}{VP + FP}$$

- **Cobertura / Revocação / Recall (C) :** de todas as instâncias que pertencem a uma certa classe, quantas foram classificadas como pertencentes a essa classe.

Ou seja, o número de vezes que uma classe foi predita corretamente (TP) dividido pelo número de vezes que a classe aparece no dado de teste (FN).

- **Cobertura de Positivos (C_+)**

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

De forma simples, por exemplo, seria o número de vezes que a classe Positivo foi predita corretamente dividido pelo número de classes Positivo que contém no dado de teste.

- **F-measure – Medida F (F-1):** é uma medida que relaciona a precisão e a cobertura por uma média harmônica entre essas medidas.

Utilizando-se a F-measure é possível conhecer o desempenho do classificador com apenas um indicador. Uma vez que essa métrica é uma média das duas anteriores, ela fornece uma visão mais exata da eficiência do classificador do que apenas a precisão ou a cobertura.

$$F1 = 2 \cdot \frac{P_+ \cdot C_+}{P_+ + C_+}$$

5.2. Avaliação da importância das características

Esta seção apresenta detalhes sobre a avaliação da importância das características para o problema de extração de aspectos. Para essa etapa foram utilizados os métodos de IG e CHI, explicados na seção 4.5. O Quadro 1 apresenta um resumo das características avaliadas, já descritas na seção 2.3.3.2.

Quadro 1: Descrição dos códigos das características (*Features*) utilizadas

Feature	Descrição	Feature	Descrição
F1	Token (t)	**F12**	Objeto direto
F2	POS-tag t	**F13**	Objeto indireto
F3	POS-tag t+1	**F14**	Verbo de ligação
F4	Lemma	**F15**	Conjunção
F5	Stem	**F16**	Conjunção coordenada
F6	Superlativo	**F17**	Sinonímia
F7	Comparativo	**F18**	Antonímia
F8	Negativo em 4	**F19**	Caracter especial
F9	Pontuação positiva	**F20**	Distância Adj / Adv
F10	Pontuação negativa	**F21**	Stopword
F11	Sujeito	**F22**	Word2Vec Similar.

Os resultados das aplicações dos métodos são mostrados nas tabelas 4 e 5. A tabela 4 é composta de 3 colunas. A coluna **Mérito Médio** exibe os resultados da nota de classificação de relevância de cada uma das características (exceto a tag de marcação da classe), acrescida do desvio padrão. A coluna **Rank Médio** rotula a posição em ordem decrescente de relevância das características. Já a coluna **Atributo** define o código da respectiva feature (1F1 = Feature 1, 2F2 = Feature 2, etc.). A coluna **Nome** especifica a característica usada, e **Grupo**, o grupo do qual ela faz parte. As características 18 e 21 ficaram tecnicamente empatadas em grau de importância, o mesmo ocorre com as menos importantes F11 e F9, e F6 e F8. O quadro 1 auxilia, de forma enumerada, na lembrança de cada feature (descritas com mais detalhes nas seções de Método e Referencial Teórico).

Tabela 4: Resultados da aplicação de Information Gain sobre as características utilizadas.

Information Gain				
Mérito Médio	**Rank Médio**	**Atributo (Feature)**	**Nome**	**Grupo**
0.613 + 0.001	1 + 0	**F1**	Token	Palavra
0.607 + 0.001	2 + 0	**F4**	Lemma	Palavra
0.585 + 0.001	3 + 0	**F5**	Stem	Palavra
0.52 + 0.002	4 + 0	**F17**	Sinonímia	Dicionário
0.269 + 0.001	5 + 0	**F2**	POS-tag	Palavra
0.204 + 0.001	6 + 0	**F22**	W2VecSimilar	Dicionário
0.112 + 0	7.3 + 0.46	**F18**	Antonímia	Dicionário
0.112 + 0	7.7 + 0.46	**F21**	StopWord	Palavra
0.072 + 0.001	9 + 0	F3	POS-tag +1	Palavra
0.022 + 0	10.2 + 0.4	F11	Sujeito	Sintaxe
0.022 + 0	10.8 + 0.4	F9	Pontuação Positiva	Dicionário
0.02 + 0	12 + 0	F12	Objeto Direto	Sintaxe
0.01 + 0	13 + 0	F10	Pontuação Negativa	Dicionário
0.008 + 0	14 + 0	F16	Conj.Coord	Marginal
0.005 + 0	15 + 0	F20	Dist.Adj/Adv	Marginal
0.002 + 0	16 + 0	F15	Connjunção	Marginal
0.001 + 0	17.1 + 0.3	F6	Superlativo	Palavra
0.001 + 0	17.9 + 0.3	F8	Negativo em -4	Palavra
0.001 + 0	19 + 0	F7	Comparativo	Palavra
0 + 0	20.1 + 0.3	F14	Verbo de Ligação	Sintaxe
0 + 0	20.9 + 0.3	F13	Objeto Indireto	Sintaxe
0 + 0	22 + 0	F19	Caracter Especial	Outras

De forma análoga à tabela 4, mas com o algoritmo de análise de ganho de informação CHI quadrado, tem-se: *Average merit*: nota de classificação de relevância, acrescida do desvio padrão. Rank: posição em ordem decrescente de relevância. *Attribute*: Código da feature (1F1 = Feature 1, 2F2 = Feature 2, etc.).

Tabela 5: Chi quadrado. Classificação pelo método Chi quadrado das características mais relevantes

	CHI Quadrado			
Mérito Médio	**Rank Médio**	**Atributo (Feature)**	**Nome**	**Grupo**
34973.719 +101.539	1 + 0	F1	Token	Palavra
34127.725 +104.381	2 + 0	F4	Lemma	Palavra
32262.409 +108.391	3 + 0	F5	Stem	Palavra
28567.904 +105.823	4 + 0	F17	Sinonímia	Dicionário
10928.756 +58.869	5 + 0	F2	POS-tag	Palavra
10259.168 +44.289	6 + 0	F22	W2VecSimilar	Dicionário
4597.5 +44.584	7 + 0	F18	Antonímia	Dicionário
3302.705 +14.8	8 + 0	F21	StopWord	Palavra
2866.279 +45.761	9 + 0	F3	POS-tag +1	Palavra
1088.893 +20.897	10.1 + 0.3	F11	Sujeito	Sintaxe
1027.263 +21.061	10.9 + 0.3	F12	Objeto Direto	Sintaxe
607.902 + 6.312	12 + 0	F9	Pontuação Positiva	Dicionário
303.936 + 8.243	13 + 0	F10	Pontuação Negativa	Dicionário
212.334 + 8.309	14.1 + 0.3	F20	Dist.Adj/Adv	Marginal
201.917 + 5.04	14.9 + 0.3	F16	Conj.Coord	Marginal
88.093 + 6.569	16 + 0	F15	Connjunção	Marginal
37.37 + 3.25	17 + 0	F8	Negativo em -4	Palavra
28.582 + 1.031	18 + 0	F6	Superlativo	Palavra
23.183 + 1.34	19 + 0	F7	Comparativo	Palavra
0 + 0	20 + 0	F13	Objeto Indireto	Sintaxe
0 + 0	21 + 0	F14	Verbo de Ligação	Sintaxe
0 + 0	22 + 0	F19	Caracter Especial	Outras

Diante das análises dos resultados dos métodos *Information Gain* e CHI-quadrado, pôde-se observar que 8 características tiveram uma importância bem mais significativa para o processo de classificação do que as demais. Analisando-se a tabela 4, percebe-se que a partir do rank 10, as características têm um valor de mérito de relevância inferior a 2%. A característica 9, um pouco mais bem ranqueada, não chegou a 10% de mérito. Em contrapartida, as características F1, F4 e F5 apresentaram um alto grau de relevância média, tanto quando da aplicação do método Information Gain, quanto do CHI quadrado. Analogamente, em relação às características menos relevantes, percebemos que a F13, F14 e F19 simplesmente parecem não contribuir para a classificação. É mister esclarecer que não significa que essas características são irrelevantes para um processo de extração de informação, mas que dentro de um contexto de 22 características aplicadas de forma conjunta, a presença das 3 últimas não pareceu relevante (segundo os métodos utilizados).

5.3. Resultados da classificação de aspectos

Diante dos resultados obtidos pelos algoritmos *Information Gain* e CHI quadrado, foram aplicados 3 experimentos no classificador CRF:

- Treinamento e teste com todas as características;
- Treinamento e teste com todas as características com exceção das 3 pior ranqueadas pelos algoritmos IG e CHI quadrado;
- Treinamento e teste apenas com as 8 características mais bem ranqueadas nos testes de IG e CHI quadrado.

Todos os 3 ensaios utilizaram como dados de treinamento a matriz de características gerada pelo sistema relativa a esse grupo (matriz de características do dataset de treinamento) e para dados de teste, a matriz de características do dataset de teste.

O primeiro ensaio utilizou como dado de treinamento a matriz de características desse grupo, aplicando todas as 22 características a fim de se treinar o classificador. A 23ª "feature" é a tag de marcação e não é considerada pelo classificador no momento do treinamento.

A primeira análise (treinamento e teste com todas as características) já mostrou resultados relevantes. A aplicação das 22 características ao dataset de restaurantes, combinada ao modelo CRF resultou numa precisão, cobertura e *F-measure* de 0,7830, 0,9075 e 0,8407, respectivamente.

Após essa análise inicial, como já foi dito, foram realizados outros dois testes com conjuntos diferentes de características, baseados nos resultados da seção 5.2. O primeiro consistiu em se executarem os eventos de treinamento e teste com o banco de entrada **sem a utilização das 3 características mais mal ranqueadas**. Assim, a ferramenta de classificação só utilizaria 19 das 22 características para geração do modelo de aprendizado e posterior inferência dos dados. Para o último teste foram selecionadas apenas as **8 características mais bem ranqueadas** para uso no classificador. Os resultados comparativos dos novos testes encontram-se na tabela 6.

Tabela 6: Comparação das métricas de avaliação em 3 cenários diferentes: todas as 22 características utilizadas no classificador, remoção das 3 piores e utilização apenas das 8 mais relevantes, de acordo com os resultados do Information Gain e CHI quadrado.

Características	Precisão	Cobertura	F-1
Todas as 23 Características	0,783065513	0,907593123	0,840743198
20 Características (F13, F14, F19 removidas)	0,778121137	0,906407487	0,837379448
8 melhores (F1, F4, F5, F17, F2, F22, F18, F21)	0,793572311	0,924406048	0,854007316

A análise da tabela mostra que embora as características F13, F14 e F19 tenham sido as menos relevantes, segundo os dois métodos de avaliação de redundância e importância de informação (Information Gain e CHI quadrado), após a remoção dessas características, os resultados não melhoraram, como esperado, mas decaíram nas 3 medidas de avaliação (precisão, cobertura e F-1). Por outro lado, ao se removerem outras características, mantendo-se só as 8 melhores, houve uma melhora de 1,31% na precisão, 1,82% na cobertura e consequentes 1,55% na medida F.

Diante desses resultados, os recém-obtidos valores foram comparados com os resultados do SemEval2014 (tabela 7).

Tabela 7: Comparação dos resultados de precisão, cobertura e medida-F com os obtidos na SemEval2014

Equipe	Versão	Precisão
COMMIT-P1WP3	V1	0.90909094
NILCUSP	V1	0.8772727
SeemGo	V1	0.86624205
XRCE	V1	0.8624535
IHS_RD_Belarus	V1	0.8606557
DLIREC	V2	0.85350317
DLIREC	V1	0.85077345
NRC-Canada	V1	0.8440546
DLIREC	V1	0.84040403
SAP_RI	V1	0.83687943
UWB	V1	0.8328109
UWB	V1	0.82695985
UNITOR	V1	0.8244631
ECNU	V1	0.8215325
JU_CSE-Patra	V1	0.8184855
UNITOR	V1	0.8131267
lsis_lif	V1	0.8120045
R3-UFRPE	**V2**	**0.79357**
USF	V1	0.78265524
DMIS	V1	0.78194726
IIT_Patan	V1	0.7797903
UBham	V1	0.77952754
UBham	V1	0.77408636
Blinov	V1	0.7056277
EBDG	V1	0.6934629
SINAI	V1	0.5960841
V3	V1	0.5715409
SNAP	V1	0.5714286
UFAL	V1	0.49577805
iTac	V1	0.37076795

Equipe	Versão	Cobertura
R3-UFRPE	**V2**	**0.92440**
DLIREC	V2	0.8271605
DLIREC	V1	0.824515
XRCE	V1	0.81834215
UNITOR	V1	0.7865961
UNITOR	V1	0.77865964
NRC-Canada	V1	0.7636684
UWB	V1	0.76278657
ECNU	V1	0.74691355
IHS_RD_Belarus	V1	0.7407407
DLIREC	V1	0.7336861
SAP_RI	V1	0.72839504
SINAI	V1	0.7248677
UFAL	V1	0.7248677
IIT_Patan	V1	0.72134036
SeemGo	V1	0.7195767
Blinov	V1	0.71869487
UWB	V1	0.70282185
EBDG	V1	0.6922399
DMIS	V1	0.6798942
JU_CSE-Patra	V1	0.6481481
USF	V1	0.64462084
V3	V1	0.6410935
Isis_lif	V1	0.6322751
UBham	V1	0.6164021
UBham	V1	0.6111111
iTac	V1	0.39594355
SNAP	V1	0.3915344
COMMIT-P1WP3	V1	0.38800704
NILCUSP	V1	0.340388

Equipe	Versão	F-1
R3-UFRPE	**V2**	**0.85400**
DLIREC	V2	0.84012544
XRCE	V1	0.83981895
DLIREC	V1	0.8374384
NRC-Canada	V1	0.8018518
UNITOR	V1	0.8009071
UNITOR	V1	0.7996414
IHS_RD_Belarus	V1	0.79620856
UWB	V1	0.79357797
SeemGo	V1	0.78612715
DLIREC	V1	0.78342754
ECNU	V1	0.782448
SAP_RI	V1	0.7788779
UWB	V1	0.7623147
IIT_Patan	V1	0.74942744
DMIS	V1	0.7273585
JU_CSE-Patra	V1	0.72342515
Blinov	V1	0.7121013
lsis_lif	V1	0.71095675
USF	V1	0.70696324
EBDG	V1	0.6928508
UBham	V1	0.6863034
UBham	V1	0.6851211
SINAI	V1	0.65419817
V3	V1	0.60432255
UFAL	V1	0.5888252
COMMIT-P1WP3	V1	0.54388136
NILCUSP	V1	0.49047014
SNAP	V1	0.4646782
iTac	V1	0.38294244

Análise dos resultados

O *framework* recém-desenvolvido neste trabalho utiliza algumas das características mais aplicadas no estado da arte para detecção de aspectos em revisões de restaurantes. A aplicação de diversas características de análise a nível de palavra, combinadas a um classificador condicional conseguiu alcançar uma taxa de F-measure de 85,40%, alcançando a primeira posição quando comparado às equipes mais bem colocadas no SemEval 2014, seguindo os mesmos critérios de avaliação deste. O sistema também conseguiu resultados sutilmente melhores que os recentemente obtidos por Xiang (2018) [69]. A literatura e as diretrizes do SemEval consideram como métrica principal de avaliação a medida F, uma vez que esta leva em consideração os resultados tanto de precisão, quanto de cobertura. Esta última métrica de desempenho teve destaque no projeto.

Quando utilizarmos apenas as características (F1, F4, F5, F17, F2, F22, F18 e F21), o sistema apresentou uma cobertura de mais de 92%, que quando combinada à precisão de 79,35%, culminou com bom resultado da medida F. Vale destacar que a alta taxa de cobertura indica que a aplicação conseguiu extrair uma média de 9 em cada 10 aspectos da classe de testes.

Com uma Precisão de 79% (valor relativamente significativo, dados os demais resultados para essa métrica), o sistema incluía alguns dados como falso-positivos, o que pode ter ocorrido pela identificação de alguns termos substantivos em excesso dada as características melhor ranqueadas. Uma outra explicação é a baixa taxa de Ganho de Informação obtida pelas características responsáveis pelas relações sintáticas de complementos verbais, como objetos direto e indireto, ocasionada possivelmente por uma baixa eficiência das ferramentas responsáveis pela extração dessas relações. No entanto, os resultados obtidos alcançaram o estado da arte para a tarefa de identificação de termos de opinião do SemEval 2014, o que contribui para a comunidade científica.

6

Conclusão

Neste trabalho foram pesquisadas e estudadas as principais características para extração de aspectos de forma automática. Um classificador CRF foi aplicado, obtendo, em combinação com um processo de seleção de ganho de informação, ótimos resultados no estado da arte para extração de aspectos na área de Análise de Sentimento. Deste modo, podemos elencar algumas das principais contribuições deste trabalho:

Elaboração de um *framework* para aplicação de características em um conjunto de dados de revisões de restaurantes em ambiente de programação Python;

Geração de duas matrizes de características para treinamento e teste em modelo de classificação CRF;

Escolha das características mais relevantes pela aplicação de algoritmos de avaliação de ganho de informação como o *Information Gain*;

Desenvolvimento de testes no classificador e obtenção de resultados que alcançaram o estado da arte na tarefa de extração de aspectos em Análise de Sentimento;

Limitações e trabalhos futuros

Apesar dos bons resultados conquistados no desenvolvimento deste trabalho, podemos destacar algumas limitações. A primeira delas foi a combinação limitada de características, onde talvez um número maior de combinações poderia gerar resultados ainda melhores. Em segundo lugar, o fato de apenas um dataset ter sido utilizado (revisões de restaurantes) não permitiu aos autores verificar a possibilidade de o sistema ser eficiente em outros domínios. Por último, testes em outro classificador, diferente do CRF, possibilitaria uma comparação mais aprofundada com outros sistemas na literatura.

Em relação aos trabalhos futuros, os autores têm as seguintes propostas:

- Realizar novas combinações de características e/ou bibliotecas para melhorar os resultados de precisão;
- Executar as tarefas de definição de categoria do aspecto e atribuição de polaridades;
- Realizar testes do sistema com uso de dois outros bancos de dados, um semelhante (hotéis) e outro diferente (*laptops*);
- Realizar testes com outros classificadores;
- Aperfeiçoar o *framework* desenvolvido para que seja possível a integração do sistema de extração de características e classificador em um único;

Bibliografia

1. RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: Tasks, approaches. **Knowledge-Based Systems**, 2015. 33.
2. LI, F. et al. Structure-Aware Review Mining and Summarization. **Proceedings of the 23rd International Conference on Computational Linguistics** , Beijing, 2010.
3. CAMILO, A. O. . S. J. C. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.. **Instituto de Informática. Universidade de Goiás**, 2009.
4. RANA, T. A.; CHEAH, Y.-N. Aspect extraction in sentiment analysis: comparative. **Springer Science+Business Media Dordrecht 2016**, 2016.
5. WEISS, S. M. E. A. Text mining: predictive methods for analyzing unstructured information. **Springer Science & Business Media**, 2010.
6. WITTEN, I. H.; FRANK, E. **Data Mining:** Practical machine learning tools and techniques. [S.l.]: Morgan Kaufmann, 211.
7. HAN, J.; KAMBER, M. Data Mining: Concepts and Techniques.. **Elsevier**, 2006.
8. WANG, J.; HU, X.; ZHU, D. Data Mining in Public Administration, n. p. 556–567, 2008.
9. MINERAÇÃO DE TEXTO: ANÁLISE COMPARATIVA DE ALGORITMOS -. Revista SQL Magazine 138. **Dev Media**, 2018. Acesso em: 2018.
10. BIRD, S. . K. E. . A. L. E. Natural Language Processing with Python.. **O´Reilly PublishinG**, 2009.
11. BRITO, E. M. N. **Mineração de Textos:** Detecção automática de. Belo Horizonte: [s.n.], 2016.
12. AQUARELA. Datasets, o que são e como utilizá-los, 2018. Disponivel em: <https://aquare.la/datasets-o-que-sao-e-como-utiliza-los/#Porque_dataset_e_nao_conjunto_de_dados>. Acesso em: jul. 2018.
13. WANNIARATCHY, P. Sentiment Analysis. Capeesh! Disponivel em: <https://www.lexalytics.com/technology/sentiment>. Acesso em: 20 mar. 2017.
14. PANG, B. . L. L. Opinion mining and sentiment analysis. **Found Trends Inform. Retrieval 2**, 1-135, 2008.
15. COLLOMB, A. . C. C. . J. D. . H. O. . B. L. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation., 2015.
16. BECKER, K.; TUMITAN, D. **Introdução à Mineração de Opiniões:** Conceitos, Aplicações e Desafios. [S.l.]: [s.n.], 2015.
17. LIU, B.; ZHANG, L. A Survey on Opinion Mining and Sentiment Analysis. Chicago: [s.n.], 2012.
18. STONE, P. .; DUNPHY, D. . C.; SMITH, M. . S. The general inquirer - A computer approach to content analysis, 1966.
19. WIEBE, J. A. R. E. Creating subjective and objective sentence classifiers from unannotated texts. In Computational Linguistics and Intelligent Text Processing, 2005.
20. STRAPPARAVA, C. A. V. A. Wordnet affect: an affective extension of wordnet. **In LREC**, 4, 2004.

21. BACCIANELLA, S. . E. A. . A. S. F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. **In LREC**, 10, 2010. 2200-2204.
22. HU, M. A. L. B. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. **ACM.**, 2004. 168–177..
23. FELLBAUM, C. WordNet. **Springer**, 2010.
24. RUSSELL, S. &. N. P. **Artificial Intelligence:** A Modern Approach. [S.l.]: New Jersey: Pearson Education, Inc, 2003.
25. BOIY, E. . H. P. . D. K. &. M. M.-F. **"Automatic Sentiment Analysis in OnLine Text**. Proceedings of the Conference on Electronic Publishing (ELPUB-2007), p. 349-360. [S.l.]: [s.n.]. 2007.
26. SCHRAUWEN, S. **MACHINE LEARNING APPROACHES TO SENTIMENT ANALYSIS USING THE DUTCH NETLOG CORPUS**. [S.l.]: [s.n.]. 2010.
27. HE, X.; ZEMEL, R. S.; CARREIRA-PERP, M. A. **Multiscale conditional random fields for image labelling**. Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: [s.n.]. 2004.
28. A. BERNAL, K. C. A. H. A. F. P. Global discriminative learning for higher-accuracy computational gene prediction,". **PLoS Computational Biology,** , v. 3, 2007.
29. CAMBRIA, E.; PORIA, S.; GELBUKH, A. Sentiment Analysis Is a Big Suitcase. **IEEE Intelligent Systems**, v. 32, 2017.
30. SUTTON, C.; MCCALLUM, A. An Introduction to Conditional Random Fields. **Now - the essence of knowledge**, 2012.
31. MCCALLUM, F. P. A. A. Accurate information extraction from research papers using conditional random fields. **in Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)**, 2004.
32. SETTLES, B. Abner: An open source tool for automatically tagging genes, proteins, and other entity names in text. **Bioinformatics**, 21, 2005.
33. SHA, F.; PEREIRA, F. Shallow parsing with conditional random fields. **Conference on Human Language Technology and North American Association for Computational Linguistics (HLT-NAACL)**, 2003.
34. ARNAB, A. et al. **Higher Order Conditional Random Fields in Deep Neural Networks**. [S.l.]: [s.n.]. 2016.
35. CHANG, K. Y. et al. Analysis and Prediction of the Critical Regions of Antimicrobial Peptides Based on Conditional Random Fields. **PLOS One**, 2015.
36. TASKAR, B.; ABBEEL, P.; KOLLER, D. **Discriminative probabilistic models for relational data**. Conference on Uncertainty in Artificial Intelligence (UAI). [S.l.]: [s.n.]. 2002.
37. STEINBERGER, J. . B. T. . K. M. Aspect-Level Sentiment Analysis in Czech, 2015.
38. FALK, S. . R. A. . K. R. Know-Cebter at SemEval-2016 Task 5: Using Words Vectors with Typed Dependencies for Opinion Target Expression Extraction. **SemEval 2016**, 2016.

39. YANASE, T. . Y. K. . S. M. . M. T. . N. Y. bunji at Semeval-2016 Task 5: Neural and Syntatic Models of Entity-Attribute Relationship for Aspect Based Sentiment Analysis. **SemEval 2016**, 2016.
40. KAUER, A. U. **Análise de Sentimentos baseada em Aspectos e Atribuição de Polaridade**. UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. Porto Alegre. 2016.
41. HU, M.; LIU, B. **Mining and summarizing customer reviews.** CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 10.. New York NY, USA: ACM. 2004. p. 168–177.
42. QIU, G. E. A. **Opinion word expansion and target extraction through double propagation.** Computational Linguistics. [S.l.]: MIT Press, v. 37, n. 1. 2011. p. 9–27,.
43. PONTIKI, M. E. A. **Semeval-2015 task 12:** Aspect based sentiment analysis. WORKSHOP ON SEMANTIC EVALUATION. DenverColorado: Association for Computational Linguistics. 2015. p. 486–495.
44. SANTANA, M. D. S. **Dois olhares sobre as palavras cognatas**. GMHP - Grupo de Morfologia Histórica do Português. [S.l.]. 2002.
45. LUCCA, J. L. D.; NUNES, M. G. V. **Lematização versus Stemming**. São Carlos, SP. 2002.
46. TEXT MINING ONLINE. Dive Into NLTK, Part IV: Stemming and Lemmatization. **Text Mining Online**. Disponivel em: <https://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization>. Acesso em: 2018.
47. THE COMPARATIVE AND THE SUPERLATIVE. **Education First**, jul. 2018. Disponivel em: <https://www.ef.com/english-resources/english-grammar/comparative-and-superlative/>. Acesso em: 10 jul. 2018.
48. KREUTZER, J.; WITTE, N. Opinion Mining Using SentiWordNet, 2013. Disponivel em: <http://stp.lingfil.uu.se/~santinim/sais/Ass1_Essays/Neele_Julia_SentiWordNet_V01>. Acesso em: 16 jul. 2018.
49. RAJARAMAN, A. . &. U. J. Data Mining. In: _____ **Mining of Massive Datasets**. [S.l.]: Cambridge: Cambridge University Press, 2011.
50. PERRY, P. O. Positive Words. **ptrckprry.com**, 2018. Disponivel em: <http://ptrckprry.com/course/ssd/data/positive-words.txt>. Acesso em: Julho 2018.
51. BECHARA, E. **Moderna Gramática Portuguesa**. [S.l.]: Nova Fronteira, 2015.
52. COSERIU, E. **Fundamentos e Tarefas da Sócio e Etnolinguística**. João Pessoa. 1990. Atas do 1º Congresso de Sócio e.
53. CUNHA, C.; CINTRA, L. **Nova Gramática do Português Contemporâneo**. 5. ed. Rio de Janeiro: Lexikon, 2008.
54. GRAMMARLY. **https:** //www.grammarly.com. Acesso em: 08 jul. 2018.
55. ENGLISH EXPERTS. https://www.englishexperts.com.br/pronomes-no-ingles/. **www.englishexperts.com.br**, 2018. Acesso em: 20 jul. 2018.
56. INGLESWINNER. https://ingleswinner.com/blog/as-20-preposicoes-em-ingles-mais-importantes/. **https:** //ingleswinner.com, 2018. Acesso em: jul. 2018.

57. FERREIRA, A. B. H. **Novo dicionário da língua portuguesa**. 2a. ed. Rio de Janeiro.: Nova Fronteira., 1986.
58. GRAMMARLY. https://www.grammarly.com/blog/articles/. **https:** //www.grammarly.com. Acesso em: 08 jul. 2018.
59. UOTTAWA. The Writting Centre. **uOttawa**, 2017. Disponivel em: <https://arts.uottawa.ca/writingcentre/en/hypergrammar/the-parts-of-the-sentence>. Acesso em: 2018.
60. STANOJEVIć, M. **COGNITIVE SYNONYMY: A GENERAL OVERVIEW**. College of Applied Vocational Studies, Vranje, Serbia. [S.l.]. 2009.
61. LI, F. et al. Structure-aware review mining and summarization. In: Proceedings of the 23rd international conference on computational linguistics. **Association for computational linguistics,** p. 653–661., 2010.
62. JAKOB, N.; GUREVYCH, I. **Extracting opinion targets in a single-and cross-domain setting with conditional random fields.** Proceedings of the 2010 conference on empirical methods in natural language processing. Association for computational linguistic. [S.l.]: [s.n.]. 2010.
63. ZHUANG, L.; JING, F.; ZHU, X.-Y. **Movie review mining and summarization**. Proceedings of the ACM 15th Conference on Information and Knowledge Management. Arlington, Virginia: [s.n.]. 2006. p. pages 43–50.
64. CHOI, Y.; CARDIE, C. Hierarchical Sequential Learning for Extracting Opinions and their Attributes. **ACL 2010**, 2010.
65. HUANG, S. et al. Fine-grained Product Features Extraction and Categorization in Reviews Opinion Mining. **2012 IEEE 12th International Conference on Data Mining Workshops**, 2012.
66. YANG, B.; CARDIE, C. Joint Inference for Fine-grained Opinion Extraction. **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**, 2013.
67. AL, C. E. Comparison of feature-level learning methods for mining online consumer reviews. **Expert Systems with Applications**, 2012.
68. BELLOT, P.; HAMDAN, H.; BECHET, F. Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis. **9th International Workshop on Semantic Evaluation (SemEval 2015)**, 2015. 753-758.
69. XIANG, Y.; HE, H.; ZHENG, J. Aspect Term Extraction Based on MFE-CRF. **MDPI**, 2018.
70. CHEN, L.; QI, L.; WANG, F. Comparison of Feature-Level Learning Methods for Mining Online Consumer Reviews. **Expert Systems with Applications**, 2012.
71. STENETORP, P. et al. BRAT: a web-based tool for NLP-assisted text annotation. **In Proceedings of EACL**, 2012.
72. PONTIKI, M. et al. SemEval-2014 Task 4: Aspect Based Sentiment Analysis - Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 2014.
73. PIPERIDIS, S. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions.. **Proceedings of LREC-2012**, pages 36–42, 2012.

74. NLTK PROJECT. nltk.org, 2017. Acesso em: jul. 2018.
75. MIKOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space, 2013.
76. EVALITA. **Evalita - Evaluation of NLP and Speech Tools for Italian**, 2009. Disponivel em: <http://www.evalita.it/2009/tasks/entity>. Acesso em: 2018.
77. JOACHIMS, T. **Text categorization with support vector machines:** learning with many relevant features. Proceedings of ECML98, 10th European conference on machine learning. [S.l.]: [s.n.]. 1998.
78. LEWIS, D. D. . &. R. M. **A comparison of two learning algorithms for text categorization.** Proceedings of SDAIR--94, 3rd annual symposium on document analysis and information retrieval. [S.l.]: [s.n.]. 1994.
79. SCHUTZE, H. . H. D. A. . &. P. J. O. **Toward optimal feature selection.** Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval. [S.l.]: [s.n.]. 1995.
80. YANG, Y. . &. P. J. O. **A comparative study on feature selection in text categorization. In Proceedings of ICML-97**. 14th international conference on machine learning. [S.l.]: [s.n.]. 1997.
81. KONKOL, M. . B. T. . S. J. UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis, 2014.
82. CAMBRIA, E. . S. R.; HAVASI, C.; HUSSAIN, A. SenticNet: A Publicly Available Semantic Resource for Opinion Mining, 2010.
83. STRAPPARAVA, C. . V. A. WordNet-Affect: an affective extension of WordNet. **Proceedings of LREC**, vol. 4, 2004, pp. 1083–1086.
84. CAMBRIA, E. . F. J. . B. F. . S. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. **AAAI**, pp. 508–514, 2015.
85. THE COMPARATIVE AND THE SUPERLATIVE. **Education First**, jul. 2018. Disponivel em: <https://www.ef.com/english-resources/english-grammar/comparative-and-superlative/>. Acesso em: 10 jul. 2018.
86. E. COSERIU [ECS.12, 1.; ECS.1, 2. [S.l.]: [s.n.].
87. MIKOLOV, T.; AL., E. "Efficient Estimation of Word Representations in Vector Space", 2013.

ANEXO

1. Tabela de Características por grupo.

Features for Supervised Methods CRF	Feature ID
Word Feature	
Token	1
Lemma	2
Word type / shape (capital letter, small, digit, etc.)	3
NER	4
Chunk (sintagmas)	5
Prefixes (length 1 to 4)	6
Lower case and upper case combination	7
Sufixe (length 1 to 4)	8
Stop word	9
Stem	10
Token polarity	11
POS	12
Previous Token, Lemma, POS	13
Next Token, Lemma, POS	14
Negative in previous 4 words	15
Is superlative	16
Is comparative	17
Intensifier	18
Diminisher	19
Modal Verb	20
Semantic frame	21
Dictionary	
WordNet Similarity	22
WordNet Synonym	23
WordNet Antonym	24
WordNet Hyperonym	25
Subjectivity	26
SentiWordNet Prior Polarity	27
Sentence Feature	

N. Positive Words in SentiWordNet	28
N. Negative Words in SentiWordNet	29
N. of Negation Words	30
Syntatic Features	
Parent Word	31
Parent SentiWordNet Prior Polarity	32
In subject	33
In copular	34
In object	35
Edge Feature	
Conjunction Word	36
Syntatic relationship	37
Word distance	38
Short dependency path (dependency tree)	39
Distance from adjective / adverb from target words	40
Other	
Opinion sentence	41
Opinion Lexicon	42
Shallow-Parser CASS	43
Prior Polarity / Intensity	44
Expression Polarity / Intensity	45
Count of Strong / Weak	46
Expression Span (boolean)	47
Distance to Exp Span (0-3+)	48
Rule-based	49
Symbols removal	50
Self tagging	51
Word clustering	52

Word Feature	Dictionary	Sentence Feature	Syntatic Features	Edge Feature	Other
Token	WordNet Similarity	N. Positive Words in SentiWordNet	Parent Word	Conjuction Word	Opinion sentence
Lemma	WordNet Synonym	N. Negative Words in SentiWordNet	Parent SentiWordNet Prior Polarity	Syntatic relationship	Opinion Lexicon
Word type / shape (capital letter, small, digit, etc)	WordNet Antonym	N. of Negation Words	In subject	Word distance	Shallow-Parser CASS
NER	WordNet Hyperonym		In copular	Short dependency path (dependency tree)	Prior Polarity / Intensity
Chunk (sintagmas)	Subjectivity		In object	Distance from adject / adverb from target words	Expression Polarity / Intensity
Prefixes (length 1 to 4)	SentiWordNet Prior Polarity				Count of Strong / Weak
Lower case and upper case combination					Expression Span (boolean)
Sufixe (length 1 to 4)					Distance to Exp Span (0-3+)
Stop word					Rule-based
Stem					Symbols removal
Token polarity					Self tagging
POS					
Previous Token, Lemma, POS					
Next Token, Lemma, POS					
Negative in previous 4 words					
Is superlative					
Is comparative					
Intensifier					
Diminisher					
Modal Verb					
Semantic frame					