Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Avaliação de técnicas para extração de características em textos opinativos

Raphael Nunes de Almeida

JUIZ DE FORA
NOVEMBRO, 2017

Avaliação de técnicas para extração de características em textos opinativos

Raphael Nunes de Almeida

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Jairo Francisco de Souza

JUIZ DE FORA

NOVEMBRO, 2017

Avaliação de técnicas para extração de características em textos opinativos

Raphael Nunes de Almeida

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Jairo Francisco de Souza
Doutor em Informática / PUC-Rio

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação / UFRJ

Fabrício Martins Mendonça
Doutor em Ciência da Informação / UFMG

JUIZ DE FORA

16 DE NOVEMBRO, 2017

Resumo

A Análise de Sentimentos tem por objetivo analisar, extrair e classificar opiniões sobre uma entidade ou produto. Inúmeras informações sobre produtos estão disponíveis na *web*, porém, a extração de opiniões por uma pessoa é extremamente custosa. Este trabalho tem por objetivo avaliar as técnicas da Análise de Sentimentos relacionadas a extração de opiniões explícitas. Os textos utilizados nesse trabalho foram extraídos do perfil da *Apple Support* do *Twitter*. Este trabalho avalia os resultados através de técnicas que fazem uso da frequência das palavras e técnicas que utilizam relações de dependências das palavras. Duas abordagens foram propostas através de adaptações de técnicas existentes na literatura.

Palavras-chave: Frequência, Relação de Dependência, Opinião Explícita, Análise de Sentimento.

Abstract

Sentiment Analysis aims to analyze, extract and classify opinions about an entity or product. Countless product information is available on *web*, but the extraction of opinions by a person is extremely costly. This work aims to evaluate the techniques of the Sentiment Analysis related to the extraction of explicit opinions. The texts used in this work were taken from the Apple's support profile on Twitter. This work evaluates the results through techniques that make use of word frequency and techniques that use relations of words dependences. Two approaches were proposed through adaptations of existing techniques in the literature.

Keywords: Frequency, Dependency Relation, Explicit Opinion, Sentiment Analysis.

Agradecimentos

Agradeço ao apoio dos meus pais. Principalmente ao da minha mãe, por sempre ter me incentivado, me dado todo apoio quando precisei, e não ter me deixado desistir ao longo do curso. Agradeço a minha esposa, por seu auxílio e sua paciência que nunca faltaram.

A todos meus parentes, por sempre torcerem por mim.

Agradeço ao professor Jairo, que durante o curso contribuiu para o meu crescimento pessoal e profissional. Obrigado Jairo por sua amizade, orientação e acima de tudo, sua paciência, sem a qual esse trabalho não teria se concluído.

Agradeço especialmente a Deus, que durante o curso me deu saúde para correr atrás de meus sonhos.

Projeto Final

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

API	Application Programming Interface
AS	Análise de Sentimentos
HAC	High Adjective Count (HAC)

1 Introdução

O surgimento da *Web* mudou a maneira das pessoas se comunicarem, (Siqueira et al, 2010). Conversas de textos que antes eram feitas apenas por cartas ou SMS tornaram-se obsoletas. Hoje em dia, os meios de comunicações que as pessoas mais utilizam são aplicativos de mensagens como *Whatsapp* e principalmente redes sociais como *Facebook*, *Twitter*, *Google+*, dentre outros. Com os inúmeros avanços nos meios comunicações existentes, tornou-se fácil para uma pessoa com acesso a internet ter acesso às redes sociais, e com isso, trocar mensagens sobre um fato que aconteceu na sua vida, em seu serviço, ou sobre algum produto que comprou e não gostou ou algum serviço por ela contratado.

A *Web*, além de facilitar os meios de comunicação entre pessoas, ela também proporciona o avanço e o crescimento de várias lojas virtuais. Hoje, através de sites de compras, existe a facilidade de comprar produtos sem sair de casa. Esses sites são conhecidos como *e-commerce*. Em alguns *e-commerce*, como o da DELL, é possível, por meio de mensagens de texto, conversar com vendedores por meio de *chats* para auxiliar o cliente durante a compra. Em *e-commerce* também existe a possibilidade de um cliente deixar mensagens de texto de sugestões, reclamações e opiniões para a loja através de áreas do *e-commerce* que possuem por finalidade permitir que clientes da loja se comuniquem expressando sua opinião sobre o produto comprado.

Pelo fato do crescente número de usuários que utilizam redes sociais e da facilidade de acesso devido aos avanços nos meio de comunicações, algumas grandes marcas, donas de produtos famosos passaram a oferecer formas de serem contactadas sobre seus produtos em redes sociais. O contato é realizado através das redes sociais por um de seus clientes e tem por finalidade obter ajuda sobre como utilizar o produto, fornecer *feedback* e relatar *bugs* encontrados no produto fornecido ou serviço prestado por essas grandes empresas. Com isso as informações dos produtos não ficam restritas apenas aos sites de *e-commerce*. Pode-se notar que a maioria dos clientes que entram em contato com as empresas através do uso de rede social tem por finalidade relatar problemas.

Uma das formas de contactar algumas marcas em redes sociais é através de um

perfil que é criado pela empresa, cujo único objetivo é oferecer suporte do produto aos clientes. Através desse perfil, torna-se possível ao cliente fazer reclamações e pedidos de ajuda.

As informações a respeito de um produto, de acordo com Tuarob et al (2015), são valiosas para clientes e donos de produtos. Pelo fato de um perfil empresarial ser público, outros clientes podem utilizar respostas á dúvidas de problemas similares que estão tendo, possíveis clientes podem pesquisar as principais reclamações e decidir se adquirem um produto da empresa. Já para a empresa, a informação possui um grande valor, pois permite saber a recepção dos seus produtos do ponto de vista dos clientes. Com isso, a empresa pode tomar suas decisões sobre a próxima versão do produto, o que deve ser melhorado ou adicionado, permitindo direcionar de maneira mais objetiva os esforços de sua equipe.

Realizar a análise de mensagens, de uma rede social, que são direcionadas a uma empresa, é uma tarefa inviável para o cliente e para o dono do produto. Do ponto de vista do cliente esta é uma tarefa custosa visto que nem sempre existe um agrupamento por característica do produto ou serviço, sendo necessário ler todas as mensagens a fim de achar alguma de seu interesse. Em relação a empresa esse problema é ainda mais custoso, visto que essa tende a ter vários perfis em várias redes sociais.

Para tentar resolver o problema de analisar milhares de opiniões, surgiu uma área de pesquisa dentro da Ciência da Computação chamada Análise de Sentimentos (AS). A área de AS tem por objetivo analisar a opinião de um consumidor sobre um produto ou serviço e dizer qual o seu sentimento pelo produto ou serviço adquirido. De acordo com Siqueira et al (2010), o processo de AS pode ser dividido em quatro etapas: (1) Detectar se o texto expressa uma opinião; (2) Extrair as características do produto e a opinião a respeito delas; (3) Detectar o sentimento expresso em cada opinião; (4) Realizar a sumarização das opiniões.

No escopo do presente trabalho serão abordados os passos (1) e (2) da AS. As opiniões utilizadas são opiniões extraídas da rede social *twitter*.

1.1 Justificativa

Em um perfil de uma rede social de uma marca de grande porte, apenas um produto específico pode conter centenas a milhares de comentários. Para o cliente interessado no produto, é desgastante ler centenas de comentários para encontrar informações sobre algum aspecto do produto ou serviço. Ainda existe a possibilidade de não existir comentários sobre o aspecto de interesse, fazendo assim o consumidor gastar um tempo que poderia ser alocado em outra atividade. Existe ainda a possibilidade das opiniões encontradas sobre a característica não serem satisfatórias para o cliente tomar uma decisão se deve ou não comprar o produto. Por exemplo, grandes marcas como *Samsung*, *Apple*, *Motorola* e *Nokia* possuem vários modelos de celulares com grande número de usuários, o que gera um grande número de comentários na rede e se torna inviável a análise manual.

Textos opinativos são textos em linguagem natural. Eles possuem uma opinião a respeito de uma entidade ou a uma de suas características. O nível de formalidade do texto varia de acordo com o domínio. Textos de redes sociais são mais informais, pois são escritos com maior número de abreviações e gírias. Com isso, a AS durante a etapa de extração de características tem que lidar com problemas de processamento de linguagem natural como, ironia, gírias e características implícitas (Yan et al, 2015). Os problemas descritos torna a AS um problema não trivial.

1.2 Objetivos

O objetivo principal do trabalho é avaliar as técnicas de extração de características de produtos em textos opinativos do perfil de *Twitter* da *Apple*. Duas técnicas avaliadas foram retiradas da literatura, e duas técnicas foram propostas a partir da junção de trabalhos existentes.

Os objetivos específicos do trabalho são: implementar um algoritmo para cada técnica, realizar testes de precisão e cobertura e avaliar cada uma das abordagens.

1.3 Metodologia

A natureza do trabalho consiste na realização de uma pesquisa aplicada. O trabalho a ser desenvolvido tem por objetivo aplicar técnicas da Análise de Sentimentos para desenvolver uma aplicação capaz de extrair características e sentimentos de textos opinativos, classificar o sentimento em relação a característica, e apresentar as informações para usuário de forma que ele possa tomar decisões.

O trabalho consiste em pesquisas bibliográficas e de pesquisa experimental. A pesquisa bibliográfica para realizar o trabalho teve por objetivo a busca pelo conhecimento das técnicas básicas de extração de características, como a extração de características utilizando a frequência, e teve também por objetivo a extração de características e sentimentos através da análise sintática utilizando *parser tree*. A pesquisa bibliográfica foi útil para a descoberta de problemas existentes na AS, como o uso de gírias, palavras escritas na forma coloquial.

A pequisa experimental foi realizada com o intuito de aplicar os métodos encontrados na literatura e analisar os resultados.

Os dados para a realização da pesquisa são *tweets* de usuários do *twitter*. Os *tweets* utilizados foram extraídos do perfil da *AppleSupport*. Para realizar a captura dos *tweets* é utilizada a API(*Application Programming Interface*) *LINQ to Twitter*. A API utilizada permite capturar todos *tweets* que são direcionados diretamente a um usuário qualquer. A API também permite capturar todas as respostas ao *tweet* do usuário.

1.4 Estrutura do Trabalho

Este trabalho está organizado da seguinte forma. O capítulo 1 representa a parte introdutória, onde é apresentado a área de pesquisa, a motivação e a proposta pela qual o trabalho foi realizado.

O Capítulo 2 representa o estado da arte. Onde é definido os conceitos básicos da AS e citado os principais trabalhos relacionados.

O Capítulo 3 representa as abordagens usadas nesse trabalho. Nele é detalhado cada uma das 4 abordagens utilizadas.

O Capítulo 4 mostra a criação do *benchmark*, as métricas de avaliação, como os experimentos foram realizadas, os resultados alcançados e por último a análise dos resultados.

O Capítulo 5 mostra as conclusões, e as sugestões de melhorias.

2 Análise de Sentimentos

A Análise de Sentimentos(AS) consiste no estudo de métodos computacionais que sejam capazes de entender sentimentos e emoções expressas através de meios de comunicações (Lima, 2011).

Um sentimento pode expressar valores positivos, negativos ou neutros sobre uma determinada entidade. A Análise de Sentimentos basicamente tenta extrair os valores e as entidades contidas no textos (Lima, 2011). De acordo com AS, um sentimento é positivo quando alguém expressa uma opinião boa a respeito de algo (*e.g "A pizza estava ótima."*). Um sentimento é negativo quando alguém expressa uma opinião ruim a respeito de algo (*e.g "A pizza estava ruim."*). E neutro quando não expressa uma opinião (*e.g "Sempre peço a mesma pizza nesta pizzaria."*).

A seguir são apresentados os principais conceitos que serão utilizados neste trabalho.

2.1 Conceitos Básicos

Segundo o dicionário Aurélio, uma opinião é definida como "Modo de ver pessoal" ou "Juízo que se forma de alguém ou de alguma coisa". conceito de opinião descrito acima, não pode ser aplicado em AS, pois é um conceito vago. Uma opinião consiste em um alvo, do que se fala, e dos sentimentos associados a ele (Cambria et al, 2013). O conceito de "opinião", no sentido de AS, pode ser visto nas sentenças (2) e (3) do exemplo 2.1. A sentença (1), não é uma opinião, e sim um fato.

"(1)Comprei esse celular há alguns meses. (2)Qualidade da foto é ótima. (3)Mas o aparelho é muito caro."

Exemplo 2.1

Opiniões são expressas sobre um determinado objeto de forma implícita ou explícita (Barros et al, 2012). Observando a frase do Exemplo 2.1, a opinião é composta por 3 sen-

tenças. A sentença (1) não expressa nenhuma opinião sobre o celular adquirido, ela apenas relata um fato. Já a sentença (2) contém uma opinião positiva em relação ao aparelho. A sentença (3) expressa uma opinião negativa.

Pela sentença (2) do exemplo 2.1 pode-se ver que "qualidade da foto" é o objeto de opinião, e aparece de forma explícita. Na sentença (3), "preço" é o objeto da opinião aparece de forma implícita no texto. Então, pelas sentenças (2) e (3), nota-se que sempre existe um alvo e ao menos um sentimento associado ao alvo.

2.1.1 Objeto e Características

Para Lima (2011), um objeto pode ser um produto, pessoa, evento, serviço ou tópico. Portanto, objetos possuem características que podem ser vistas como componente, parte ou atributo do objeto. Logo, cada objeto possui características que o tornam único.

Como citado acima, o objeto pode possuir uma ou mais características, assim é possível estruturar o objeto através de uma árvore de características, conforme ilustrado na figura 2.1.

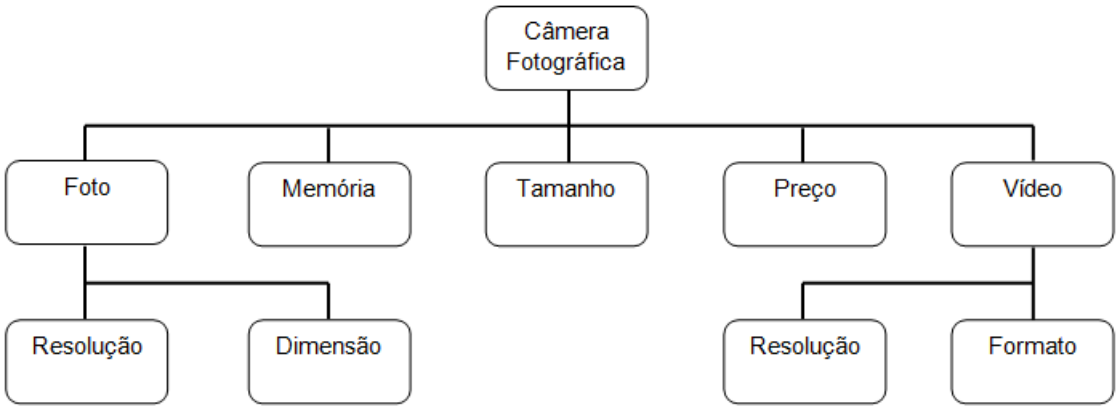

Figura 2.1: Árvore de características
Fonte: Lima, 2011, p. 5.

Uma característica pode ser classificada como implícita ou explícita. Uma característica é explícita quando ela ou algum de seus sinônimos aparecem em uma sentença (*e.g "O monitor possui ótima resolução."*). Uma característica é dita implícita, quando ela ou algum de seus sinônimos não aparecem na sentença (*e.g "Celular extremamente caro (preço)."*).

Em textos opinativos podem existir sentimentos tanto em relação ao objeto quanto à algumas de suas características, conforme mostra o exemplo 2.2.

"(1)Ótimo celular. (2)Câmera com alta definição, o que permite tirar ótimas fotos. (3) Bateria possui ótima durabilidade, o que permite utilizar o aparelho durante todo o dia. (4) Porém, ele é grande. (5) Mas, infelizmente o *wifi* não funciona."

Exemplo 2.2

Analisando o texto do exemplo 2.2, percebe-se que a sentença (1) expressa um sentimento em relação ao objeto celular. As sentenças (2) e (3), expressam sentimentos sobre as características "câmera" e "bateria" respectivamente, sendo escritas de forma explícita. Na sentença (4) existe um sentimento que faz menção à característica "tamanho" do objeto celular, porém, essa característica só pode ser identificada devido ao adjetivo "grande".

2.1.2 Palavras Opinativas

Palavras opinativas são utilizadas em textos opinativos que expressam o sentimento do dono da opinião em relação a um objeto ou a uma de suas características. Geralmente, palavras opinativas (sentimentos ou opiniões) são descritas por adjetivos e/ou advérbio, mas também podem ser descritas utilizando substantivos ou verbos.

Ainda pelo exemplo 2.2 observa-se que as palavras opinativas das sentenças (1) e (4) são respectivamente "ótimo" e "grande", todos exemplos de uso de adjetivos. Na sentença (1) tem-se que a palavra opinativa se refere à opinião sobre o objeto celular. Nas demais sentenças as palavras opinativas demonstram opiniões em relação à característica do produto. A sentença (2), tem como como opinião "alta resolução", que, ao contrário das outras sentenças, é formada pelo advérbio "alta" e pelo substantivo "definição". Na sentença (3), a opinião é formada por "ótima durabilidade", que é um adjetivo seguido de um substantivo. Já na sentença (5), a opinião é formada por "não funciona". A opinião é formada pelo advérbio "não", seguido por "funcionando", que é o gerúndio do verbo funcionar.

2.2 Extração de Características e Opiniões

Os vários trabalhos existentes na literatura(Yan et al, 2015; Hu et al, 2004; Barros et al, 2012; Somprasertsri et al, 2010; Saranya, 2010; Cambria et al, 2013) que tratam o tema abordado possuem em comum dois objetivos. O primeiro é a extração de características explícitas em textos opinativos. O segundo consiste na extração de características implícita. Nas seções seguintes serão apresentadas as abordagens que tratam dessas extrações de características.

2.2.1 Pré-Processamento

Para obter maior ganho na extração de características, é necessário realizar o pré-processamento do texto, cuja a finalidade é corrigir palavras que foram escritas em linguagem coloquial, tais como gírias ou abreviações. Então, converte-se as palavras na forma coloquial para sua forma correta na linguagem formal. Além disso, tenta-se corrigir palavras que foram escritas de maneira errada. Textos vindo de redes sociais possuem um maior número gírias.

A verificação de gírias é realizada através da apuração de cada palavra do texto identificando se está contida como gíria em uma base de dados que contém a gíria e a forma correta em linguagem formal da palavra (Barros et al, 2012). A grande desvantagem da abordagem é que a base de dados é criada de forma manual.

2.2.2 Candidatos de Características Explícitas

Para a identificação de características explícitas, alguns trabalhos geram uma lista de candidatos. Por exemplo, nos trabalhos de Siqueira et al (2010) e Barros et al (2012), a lista de candidatos é gerada utilizando o cálculo da frequência. Na abordagem de Siqueira et al (2010), são escolhidos os substantivos que estão entre os 3% mais frequentes.

Outros trabalhos, como o de Yan et al (2015), utilizam a estrutura gramatical para gerar a lista de candidatos. A estrutura gramatical é analisada através de uma *Parse tree*. A *Parse tree* é uma árvore gerada a partir das relações entre as palavras e sentenças da frase. A figura 2.2 exemplifica uma *Parse tree* baseada em dependência, da frase "*Bell,*

based in Los Angeles, makes and distributes electronic, computer and building products.".

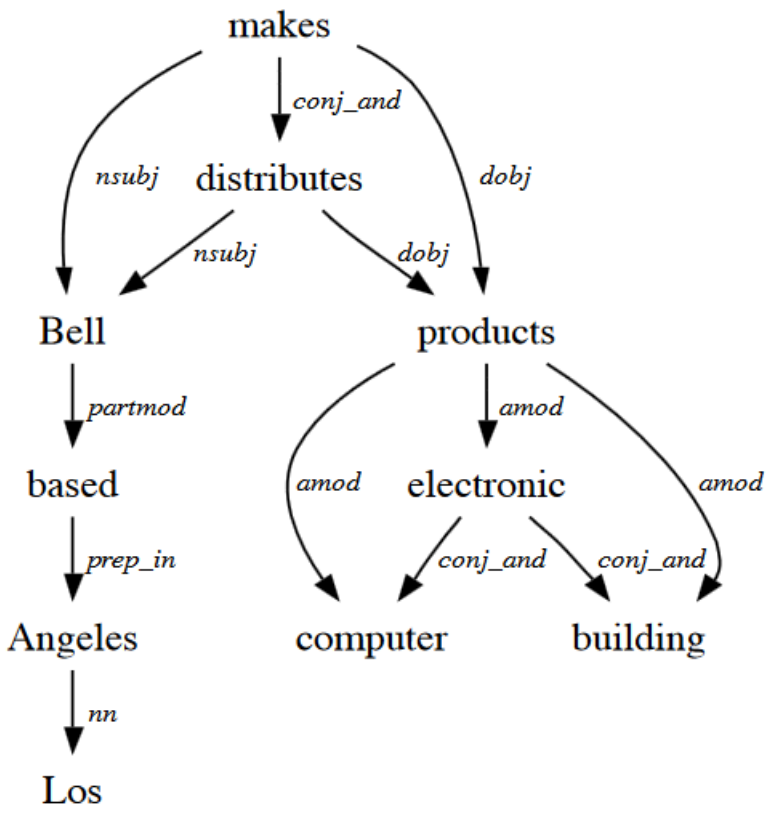

Figura 2.2: *Parser tree* (Marneffe et al, 2008)

Na árvore da Figura 2.2, os nós são formados por palavras e arestas são dependências gramaticais de duas palavras. O significado de uma frase não depende apenas do significado de cada palavra. Caso as palavras sejam mudadas de posição, a frase pode tornar agramatical. Nas frases, algumas palavras se ligam mais facilmente com outras. Palavras ligadas formam uma unidade da frase. Essa unidade recebe o nome de constituinte. Então, uma frase é o resultado da propensão que constituintes e palavras possuem para se ligar. A árvore da Figura 2.2 é útil, pois permite saber como as palavras se ligam, ou seja, como uma palavra depende da outra.

Quando é feita a análise de cada palavra da frase, captura-se como candidatas a características as palavras que possuem relação de *nsubj, amod* ou *dobj* com um nó filho (sentimento) (Yan et al, 2015; Somprasertsri et al, 2010). De acordo com Yan et al (2015), utilizando apenas os 3 tipos de relações é possível obter bons resultados na extração de características e sentimentos. As relações acima são explicadas na seção 3.2.1.

2.2.3 Candidatos a Palavras Opinativas

A maneira como o conjunto de palavras opinativas é criado varia de acordo com a abordagem escolhida.

Técnicas de frequência

Para técnicas de AS que utilizam a frequência, palavras que expressam sentimento são adjetivos (Hu et al, 2004). Então, o conjunto de sentimentos é formado por todos os adjetivos existentes. Para capturar os adjetivos, é necessário descobrir a classe gramatical de cada palavra, processo conhecido como *Pos-Tagging*. Classes gramaticais são conhecida como substantivo, adjetivo, numeral, pronome, verbo e etc. Então a criação do conjunto de sentimentos é feito aplicando-se o *Pos-Tagging* e selecionando apenas adjetivos.

Contudo, essas técnicas se divergem na escolha dos adjetivos que pertencem ao conjunto de sentimento. O conjunto pode ser formado por adjetivos que são anteriores ou posteriores as características ou separados das características apenas por uma preposição ou *StopWord* (Siqueira et al, 2010). O conjunto de sentimentos também pode ser formado pelos n adjetivos com maior *score* (Eirinaki et al, 2012). O *score* de cada adjetivo começa em zero e, para cada vez que o adjetivo é o mais próximo da característica do produto, acrescenta-se 1 em seu *score*.

Técnicas de análise gramatical

Em técnicas de análise gramatical, verbos e adjetivos representam opiniões (Somprasertsri et al, 2010). Para capturar os verbos e adjetivos é necessário realizar o *Pos-Tagging*, como em técnicas de frequência. Então, é criado um conjunto inicial O de opinião, onde O contém todos verbos e adjetivos. São removidas de O todas as palavras que não possuem alguma das relações de dependências consideradas úteis por Yan et al (2015).

2.2.4 Formação de Opiniões

A opinião demonstra o sentimento em relação a um objeto ou a alguma de suas características. Assim, uma opinião o é formada pela tupla $\{c, s\}$, onde c representa o

objeto ou alguma de suas características, e s representa uma ou mais palavras utilizadas para representar um sentimento. O sentimento, neste trabalho, pode ser considerado uma qualidade intrínseca do objeto ou uma qualidade aferida pelo usuário. Na frase "Notebook com processador rápido, porém a tela é LCD.", tem-se ("tela LCD") como uma qualidade intrínseca e ("processador rápido) como uma qualidade aferida pelo usuário.

A criação do conjunto de opinião $O = \{o_1, o_2, o_3, ..., o_n\}$, varia de acordo com a abordagem utilizada.

Técnicas de frequência

Quando a abordagem leva em consideração o cálculo da frequência, as palavras com maior frequência não serão necessariamente uma característica do objeto. Com a finalidade de resolver esse problema, é realizado um filtro onde é verificado se cada palavra contida na lista de candidatos a característica é imediatamente anterior a um adjetivo ou separado apenas por uma preposição. As palavras que não satisfazem essa condição são retiradas imediatamente da lista de candidatos. Assim, cada palavra como candidata a característica do produto é unida ao adjetivo mais próximo, formando uma lista de opiniões.

Algumas características de um produto podem não ser tão comentadas quanto outras. Com isso, elas podem não estar entre as mais frequentes (Barros et al, 2012). Para resolver esse problema, é realizado um ranqueamento dos adjetivos, capturando os adjetivos mais frequentes. De posse dos adjetivos mais frequentes, é realizado o processamento no conjunto de opiniões onde o intuito é capturar todos os substantivos anteriores aos adjetivos mais relevantes ou separados apenas por uma preposição.

Técnicas de análise gramatical

Em análise gramatical as opiniões são formadas pelo *head* e *dependent* das relações de dependência. Através de alguns tipos de relação ou sequências de tipo, é possível identificar de forma direta características e sentimentos (Kumar et al, 2013). Então, utiliza os *heads* e *dependents*, que foram extraídos nos passos anteriores, como o conjunto inicial de opinião. Em trabalhos como o de Yan et al (2015), apenas as relações dos tipos

nsubj, amod e *dobj* são consideradas como opiniões. Então, todas as opiniões que não são formadas por alguma das relações acima são removidas. No trabalho de Somprasertsri et al (2010), não é verificado o tipo de relação de dependência, mas sim, mantido como opiniões as relações formadas por um substantivo e verbo ou por um substantivo e adjetivo.

2.2.5 Filtragem

Independente da abordagem escolhida, a filtragem dos pares candidatos à opinião é necessária, pois pares que formam uma opinião podem ter sido escolhidos erroneamente. O objetivo desse passo consiste em remover pares que possuem características e sentimentos fora do contexto.

Quando é utilizada a frequência para detectar características do produto, é necessário validar se a opinião é formada por um substantivo que corresponde a uma característica do objeto. Para isso, pode ser utilizada a medida PMI-IR (Siqueira et al, 2010), conforme abaixo.

$$PMI_{\mathrm{IR}}(t_1, t_2) = \log_2 \frac{Hits(t_1 \wedge t_2)}{Hits(t_1) * Hits(t_2)} \tag{2.1}$$

Onde $Hits(t_1)$ e $Hits(t_2)$ representam, respectivamente, o total de páginas que contém uma característica t_1 e um sentimento t_2, sendo $Hits(t_1 \wedge t_2)$ o número de páginas que contém os dois termos. É calculado o PMI-IR de cada candidato a opinião. Os candidatos a opiniões com valor de PMI-IR abaixo do limiar pré-definido, são removidos.

Já em (Barros et al, 2012), a verificação do par característica e opinião, representada pela tupla(x,y), é realizada através da medida NGD.

$$NGD(x, y) = \frac{max\{\log f(x), \log f(x)\} - \log f(x, y)}{\log N - min\{\log f(x), \log f(y)\}} \tag{2.2}$$

Onde $f(x)$ é a quantidade de resultados retornados para a pesquisa de uma palavra x, $f(y)$ é a quantidade de resultados retornados para a pesquisa de uma palavra y, sendo que $f(x, y)$ é a quantidade de resultados retornados para a pesquisa de ambas as palavras e N é um fator normalizador. O fator normalizador é o número total de páginas indexadas por um motor de busca. Esse valor pode ser obtido através de endereços como

o WorldWideWebSize.com.

Quando a busca de características é realizada através de uma *parse tree*, a filtragem pode ser feita removendo todos substantivos candidatos a característica que se encontram dentro de algum dos tipos de substantivos pré-determinados. Em (Yan et al, 2015) é apresentado uma tabela de exemplo conforme a tabela 2.1.

Tabela 2.1: Exemplos de substantivos que não são características

Tipos	Exemplos
(1) Nomes próprios	Setembro, Beijing, Tom
(2) Nomes de marcas	Canon, Samsung, Apple
(3) Substantivo verbais	Ouvinte, Andador
(4) Nomes pessoais	Amigo, Pai

Após a realização do filtro, geralmente é feito um ranqueamento das palavras candidatas a características do produto. Para realizar o ranqueamento, é usado o método de *NodeRank*. Esse método se baseia no algoritmo de *PageRank* para definir o grau de relevância de um vértice. Para realizar o ranqueamento é gerado um grafo direcionado com os pares (característica, opinião), sendo calculado o *NodeRank* de cada par. O valor achado é utilizado para calcular a importância da opinião (Yan et al, 2015).

2.2.6 Características implícitas

Características implícitas podem ser mapeadas através de indicadores no texto (Barros et al, 2012). Por exemplo, na opinião *"Celular caro"*, a característica *"preço"* não é mencionada. Mas, através do adjetivo *"caro"*, é possível saber que o autor da opinião está se referindo à característica *"preço"*. As listas de indicadores são criadas de forma manual (Barros et al, 2012; Silva et al, 2013).

A extração pode ser feita através de uma matriz de co-ocorrências entre palavras (Zhang et al, 2013; Schouten et al, 2014). A matriz de co-ocorrência é criada com pares características e sentimento extraídos de uma base de treinamento. As linhas da matriz são sentimentos e as colunas são palavras que se relacionam com sentimentos extraídos. Dessa maneira, características implícitas em novos textos são inferidas da matriz.

A Tabela 2.2 representa uma matriz de co-ocorrência. É possível ver que as palavras "Caro" e "Preço" aparecem juntas em 150 textos. Já as palavras "Caro" e "Tela"

Tabela 2.2: Matriz de Co-Ocorrência

	Preço	Caro	Barato	Tela	Boa
Preço	-	150	160	1	0
Caro	150	-	0	0	0
Barato	160	0	-	2	0
Tela	1	0	2	-	5
Boa	0	0	0	5	-

aparecem juntas em apenas 2 textos. Então, quando o sentimento "Caro" aparecer em um texto, torna-se possível extrair a característica "Preço".

Para abordagens que utilizam relação de dependência, a extração de características explícita pode ser feita no final do processo. Seleciona-se frases opinativas com adjetivos e verbos, que foram classificados como sentimento na etapa anterior. Então, os adjetivos e verbos são associados ao vértice vizinho do grafo com maior valor de *NodeRank*, formando uma opinião (Yan et al, 2015).

2.3 Aplicações de análise de sentimentos

Existem diversos trabalhos na literatura para análise de sentimentos sobre produtos. Nesta seção, abordaremos os trabalhos que usam algumas das técnicas mencionadas anteriormente para realizar extração de características em diferentes aplicações.

A aplicação mais comum é a encontrada nos trabalhos de (Hu et al, 2004; Siqueira et al, 2010; Eirinaki et al, 2012), cujo o objetivo é a geração de um sumário que permite a visualização dos sentimentos expressos para cada característica do objeto. Com isso, o usuário tem a opção de realizar um filtro buscando apenas os sentimentos bons ou ruins a respeito das características de seu interesse.

O trabalho de Tuarob et al (2015) tem como objetivo melhorar o produto através da identificação de comentários de *leadusers* aplicando a extração de características em redes sociais. *Leaduser* são usuários que podem acrescentar ganhos para o produto ou serviço vendido. Esses usuários geralmente deixam comentários em redes sociais dizendo quais características os produtos devem ter e quais características não são necessárias para o produto. Para descobrir quais características o produto deve possuir, os autores utilizam técnicas de AS sobre os comentários de *leaduser* em redes sociais, o resultado é

comparado a características já existentes do produto que são encontradas em manuais. Com as opiniões dos *leadusers*, os autores esperam que as empresas consigam colocar características novas em seus produtos com uma maior aceitação.

3 Extração de características e sentimentos

Neste trabalho, foram utilizadas quatro abordagens para extração de características e comparados os resultados. A abordagens estão divididas em extração por frequência das palavras e extração por grafo de dependências.

Na extração por frequência, a abordagem 1 foi adaptada utilizando como base os trabalhos de (Eirinaki et al, 2012; Barros et al, 2012; Siqueira et al, 2010; Hu et al, 2004). Por sua vez, a abordagem 2 foi implementada a partir do trabalho de (Eirinaki et al, 2012). Na extração por grafo de dependências, a abordagem 3 foi adaptada utilizando como base (Yan et al, 2015; Somprasertsri et al, 2010). Por fim, a abordagem 4 foi implementada a partir (Yan et al, 2015). As seções seguintes detalham as abordagens e as adaptações criadas neste trabalho.

3.1 Extração de características utilizando frequência

Algoritmos de AS que se baseiam em frequência consideram, inicialmente, que o conjunto das palavras que representam características do produto é formado por palavras pertencentes à classe gramatical dos substantivos (Barros et al, 2012). Neste trabalho, para auxiliar na identificação das classes gramaticais, foi utilizada a ferramenta *Stanford POS Tagger*[1]. Com a utilização dessa ferramenta, é possível descobrir a qual classe gramatical pertence cada palavra do texto. Neste trabalho foram utilizados *tweets* como texto, pois eles são de fácil acesso. Assim, é possível criar um *dataset* com vários textos heterogêneos.

3.1.1 Extração por frequência dos substantivos e proximidades dos adjetivos

Essa abordagem é formada por 5 passos, que serão exemplificados abaixo. Esses passos foram criados a partir da junção dos conceitos citados por (Eirinaki et al, 2012;

[1] https://nlp.stanford.edu/software/tagger.shtml

Barros et al, 2012; Siqueira et al, 2010; Hu et al, 2004).

Antes de iniciar o primeiro passo da abordagem é necessário que cada *tweet* da base de dados esteja normalizado.

A sentença (1), do exemplo 3.1.1, mostra uma frase antes de ser normalizada. A sentença (2), mostra a frase (1) após ser normalizada. Percebe-se que na sentença (2) a abreviação "*pls*" foi substituída pela forma normal "*please*". Pelo fato da sentença (1) ser um *tweet*, a palavra "*@AppleSupport*" é removida durante o processo de normalização. A remoção ocorreu pelo fato da palavra indicar um perfil do *Twitter*.

(1)"*My phone keeps turning off randomly, @AppleSupport pls I need the next update to fix this*"

(2)"*My phone keeps turning off randomly, please I need the next update to fix this*"

Exemplo 3.1.1

Após a normalização é realizado o *POS-Tagging*. O primeiro passo do algoritmo consiste no cálculo da frequência de cada substantivo (Eirinaki et al, 2012). A frequência é dada pela soma das ocorrências do substantivo em cada *tweet*.

O segundo passo consiste na filtragem dos substantivos. O intuito do filtro é realizar a remoção de candidatos que não sejam características do produto sobre análise (Hu et al, 2004). Primeiro, remove-se dentre os substantivos mais frequentes aqueles que são *stop words*. São *stop words* palavras mais comuns da língua como artigos, preposições, dentre outras. Em seguida, é calculada a frequência média dos substantivos restantes. Então, seleciona-se os substantivos que possuem uma frequência maior do que a frequência média. Com isso, tem-se os substantivos que possuem maior probabilidade de serem candidatos a características, pois uma característica do produto tende a ser comentada mais vezes (Siqueira et al, 2010).

Portadores de opiniões utilizam adjetivos para demonstrar como se sentem a respeito de determinada característica do produto (Saranya, 2010). O terceiro passo do algoritmo toma a afirmação anterior como verdadeira. Essa etapa consiste em selecionar todos os adjetivos da base de *tweets*.

Após a captura de todos os substantivos candidatos e de encontrar todas palavras

que representam sentimentos do usuário em relação a alguma das características, pode-se então realizar o quinto passo do algoritmo. O passo consiste em analisar cada palavra de todos os *tweets* da base de dados. Para cada palavra do *tweet* é verificado se essa palavra pertence ao conjunto de substantivos candidatos a características do produto. Caso a palavra sobe análise seja uma candidata a característica do produto, é feita a busca pelo adjetivo mais próximo no mesmo *tweet*. Com o passo anterior, torna-se possível determinar o conjunto $O = \{o_1, o_2, ..., o_n\}$ formado por todas as opiniões contidas no *tweet*, onde o_n é formada pelo par característica c e sentimento s, ou seja, $o=(c,s)$. Após a criação do conjunto O, é aplicado um filtro cujo o objetivo é remover as opiniões associadas a mais de uma característica. Para filtrar as opiniões, verifica-se se um sentimento s_i pertence a mais de uma opinião o_i. Caso existam opiniões o_i e o_j onde $o_i = (c_i, s_i)$ e $o_j = (c_j, s_i)$, então é removido do conjunto O a opinião onde a característica possui a maior distância do adjetivo. A distância é calculada como o total de palavras que separa a característica do sentimento.

A tabela 3.1, mostra os passos executados por (Eirinaki et al, 2012; Barros et al, 2012; Siqueira et al, 2010; Hu et al, 2004) que correspondem, respectivamente, ao autor 1, autor 2, autor 3 e autor 4. O autor 5 corresponde a abordagem 3.1.1, desenvolvida nesse trabalho a partir dos autores 1, 2, 3 e 4.

Após a realização dos passos acima, é possível obter um conjunto formado por características e sentimentos sobre o domínio que está sendo analisado. A figura 3.1 mostra de forma exemplificada a extração utilizando essa abordagem.

3.1.2 HAC

A ideia principal do algoritmo de Eirinaki et al (2012) é de que substantivos que aparecem com uma maior frequência nos textos são mais prováveis de serem características importantes do produto.

O algoritmo inicialmente identifica todos os substantivos, que não são *StopWords*, e adjetivos no *dataset*. Em seguida é atribuído um peso (inicialmente igual a zero) para cada substantivo. Então, cada adjetivo é associado ao substantivo mais próximo. De acordo com Eirinaki et al (2012), esse adjetivo provavelmente expressa uma qualidade do

Tabela 3.1: Comparação entre os passos das abordagens de frequência existentes

Passo	autor 1	autor 2	autor 3	autor 4	autor 5
Normalização	x	x	x	x	x
POS-Tagging	x	x	x	x	x
Uso de substantivo como candidato a característica	x	x	x	x	
Uso de adjetivos como opiniões	x	x	x	x	x
Uso de *score* para substantivos	x				
Incrementar *score* em 1, caso o substantivo seja o mais próximo do adjetivo	x				
Cálculo da frequência dos substantivos			x	x	x
Seleção dos substantivos mais frequentes			x	x	x
Seleção dos substantivos com maiores *scores*	x				
Associação da característica ao adjetivo mais próximo e com maior frequência					x
Associação do adjetivo mais próximo e com maior *score*	x				
Extração das características	x	x	x	x	x
Extração dos sentimentos		x	x	x	x
Extração dos pares de opiniões		x	x	x	x
Extração de características implícitas		x	x		
Filtragem das opiniões utilizando a medida *PMI-IR*			x		
Filtragem das opiniões utilizando a medida *NGD*		x			
Uso de uma lista com características pré-definidas					x

substantivo. Caso um substantivo i seja o substantivo mais próximo de um adjetivo, o peso de i é incrementado em uma unidade. Pela sentença (1), do Exemplo 3.1.1, é possível ver que as palavras "com" e "preço" possuem a mesma distancia do adjetivo "ótimo". Porém a palavra "com" é uma *StopWords*, então o peso de "preço" é incrementado em uma unidade. Na sentença (2), do Exemplo 3.1.1, é possível ver que "qualidade" é o substantivo mais próximo de "excelente" e "preço" é o substantivo mais próximo de "ótimo".

(1) "*Celular com ótimo preço.*"

(2) "*Celular de excelente qualidade, e ótimo preço.*"

Exemplo 3.1.1

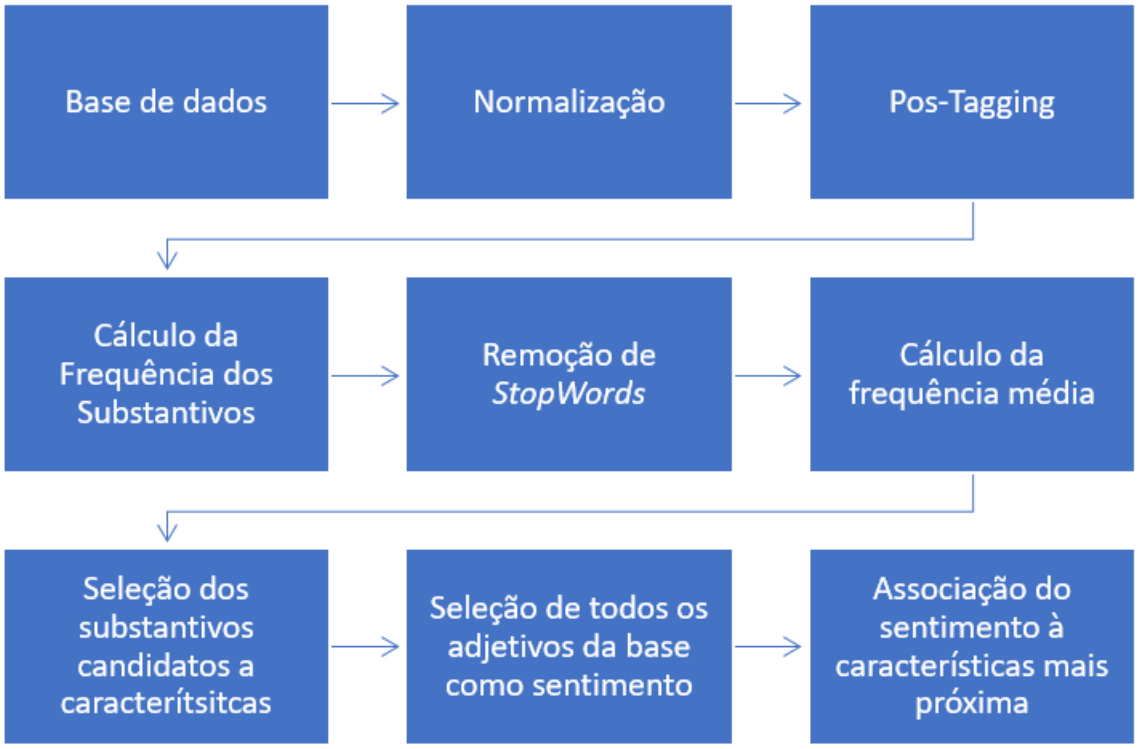

Figura 3.1: *Visão geral do processo de extração de característica*

Ao final é possível criar um *rank*, onde os substantivos que possuem maior peso estarão no topo. O *rank* do Exemplo acima pode ser visto na Tabela 3.2.

Tabela 3.2: *Rank*

Palavra	Peso
Preço	2
Qualidade	1
Celular	0

Então, um limiar pode ser utilizado para a escolha dos substantivos que serão utilizados como características. Apenas os substantivos com peso maior que o limiar são considerados como características. O valor do limiar pode ser obtido baseado em experimentos, ou pela experiência obtida na utilização da abordagem em outros *datasets*.

Para obter as opiniões, realiza-se novamente o processamento de cada texto da base. Associando cada adjetivo à característica mais próxima, formando assim uma opinião. Essa abordagem é conhecida como *High Adjective Count (HAC)*.

3.2 Extração de características utilizando relação de dependência

O principal problema da área de AS consiste na extração dos pares características e sentimentos. Para tentar resolver este problema, de acordo com Somprasertsri et al (2010), pode-se fazer uso de abordagens que se baseiam na identificação, análise e aplicação de regras entre as relações de dependências de substantivos (*nouns*) e sintagmas nominais (*noun phrases*) com adjetivos e verbos.

Uma relação de dependência é considerada como uma relação binária assimétrica entre uma palavra, chamada *head* ou *governor*, e outra, chamada *dependent* (Marneffe et al, 2008). Uma relação binária é assimétrica quando, para todo a e b pertencente a um conjunto X, se a está relacionado com b, então b não está relacionado com a. No contexto de AS a relação binária é útil, visto que um sentimento expresso no texto pode estar associado a várias características do produto. Na extração de características por frequência, não é verificado se o sentimento está relacionado com mais de uma característica, devido ao fato de escolher a característica com menor distância do adjetivo.

3.2.1 Extração utilizando *PageRank* com grafo não orientado

Os passos dessa abordagem serão exemplificados abaixo. Esses passos foram criados a partir da junção dos conceitos citados por (Yan et al, 2015; Somprasertsri et al, 2010).

Passo 1: Pré-Processamento

Assim como nas abordagens anteriores, o pré-processamento possui etapas básicas, tais como a realização de correções de palavras que estão escritas na forma de gírias e remoção de ruídos.

Para esta abordagem é executado um passo a mais. Esse passo consiste em realizar o *parsing* do texto, cujo objetivo é identificar as relações de dependências existentes para cada palavra e a sua classe gramatical.

Para encontrar as relações de dependências no *parsing*, foi necessário fazer uso

do *Stanford Parser*. Nesta etapa, obtém-se como retorno um arquivo *xml* que contém todas relações de dependência. Para identificar a classe gramatical de cada palavra após o *parsing*, é realizado o *POS tagging* utilizando também o *Stanford Parser*.

A tabela 3.3 apresenta as relações de dependência obtidas ao aplicar o *parsing* para a frase ("*My new iphone 7 camera is suddenly blurry in the corners when I take a picture why?*").

Na tabela 3.3, a relação de *nmod:poss(camera, my)* mostra uma relação de posse, pois a *camera* pertence ao dono da sentença. Na relação *nmod(blurry, camera)*, a palavra *blurry* indica um defeito da característica *camera*. As relações consideradas importantes serão detalhadas nas seções seguintes.

Tabela 3.3: Relações de dependências obtidas ao aplicar o *parsing*

Dependência	Governor	Dependent
root	ROOT	blurry
nmod:poss	camera	My
amod	camera	new
compound	camera	iphone
nummod	camera	7
nsubj	blurry	camera
cop	blurry	is
advmod	blurry	suddenly
case	corners	in
det	corners	the
nmod	blurry	corners
advmod	take	when
nsubj	take	I
advcl	blurry	take
det	picture	a
dobj	take	picture
dep	picture	why
punct	blurry	?

Passo 2: Candidatos a características e sentimentos

Portadores de opiniões utilizam substantivos ou sintagmas nominais para descrever palavras que representam características do produto. Para expressar seu sentimento em relação ao produto são utilizados adjetivos e/ou advérbios, substantivos ou verbos (Somprasertsri et al, 2010).

Ao contrário da abordagem anterior, a escolha dos candidatos é realizada através

das relações de dependências entre as palavras (Yan et al, 2015). Apesar de existirem várias relações de dependência em uma frase, apenas algumas são úteis para identificar sentimentos (Wu et al, 2009). De acordo com Yan et al (2015), as principais dependências são *'nsubj'*, *'amod'* e *'dobj'* que correspondem, respectivamente, às relações *subjectpredicate relations*, *adjectival modifying relations* e *verb-object relations*. Essas dependências são extraídas após o pré-processamento.

De acordo com Kumar et al (2013), temos as seguintes definições para as relações de dependência acima. Existe uma relação de *nsubj* quando um termo substantivo tem o seu sentido complementado por um verbo ou adjetivo. Na frase *"The Wifi works and the appearance looks very beautiful"*, o adjetivo *"beautiful"* complementa o sentido de *"appearance"* e o verbo *"works"* complementa o sentido de *"Wifi"*. A dependência *'amod'* é a dependência entre um substantivo e adjetivo. Um texto possui essa relação caso exista qualquer adjetivo que se refira a um substantivo. Na frase *"There is a great camera."*, existe uma relação *'amod'* formada pela tupla *(camera, great)*. A dependência *'dobj'*, é formado pelo verbo (ação) da sentença e por alguém ou algo que recebe a ação. Na frase *"I like this camera."*, a relação *'dobj'* é formada pela tupla *(like, camera)*.

Com as relações de dependências, torna-se possível obter as palavras correspondentes aos termos *head* e *dependent*. Para o par de palavras ser candidato a característica e sentimento, é necessário que uma palavra pertença à classe gramatical de substantivo e outra à alguma das classes gramaticais que representam um sentimento.

Passo 3: Geração dos pares

Neste trabalho, a geração dos pares candidatos e sentimentos é realizada em duas etapas. Na primeira etapa, é gerado um grafo com as relações de dependência e identificados os vértices mais relevantes. Em seguida, cada texto é processado e as relações relevantes contidas no grafo são verificadas e extraídas.

Passo 4: Geração do grafo de dependências

O algoritmo *PageRank* foi criado com o objetivo de medir a importância das

páginas *Web* (Page et al, 1998). O algoritmo computa um *ranking* para cada página da *Web* baseado no grafo de ligações das páginas. O grafo da *Web* é formado considerando que cada página é um vértice e as aresta são *links* de um site para outro, gerando um grafo orientado. Uma aresta a possui direção $a = \{s_i, s_j\}$ quando um site s_i possui um *link* que direciona para o site s_j. São consideradas como páginas importantes aquelas que são muito *linkadas* por outras página ou, então, páginas que possuem *links* para sites muito *linkados*. O valor do *PageRank* de uma página é calculado por:

$$PR(s) = c \sum_{v \in B_u} \frac{PR(v)}{N_v} \tag{3.1}$$

Onde, s e v são páginas *Web*, B_u é o conjunto de páginas que apontam para s, N_v é número de *links* que uma página v faz referência, $PR(s)$ é o valor de *PageRank* de uma página s, e c é um fator normalizador.

Utilizando o conceito do algoritmo *PageRank* descrito acima, é possível utilizar a mesma abordagem para a seleção de características e sentimentos em textos opinativos (Yan et al, 2015). Para isso, é utilizado *head* e *dependent* como vértices e como aresta é utilizada a relação de dependência.

Neste trabalho, de posse das relações de dependências selecionadas no passo anterior, escolhe-se as relações que não possuem *stop words* como *head* ou *dependent*. Então, gera-se o grafo utilizando os *governors* e *dependents* das dependências restantes. O grafo da figura 3.2 foi construído utilizando as dependências da tabela 3.3. Os pares candidatos a opiniões da tabela são *(camera - new), (blurry - camera), (take - I) e (take - picture)*. Contudo, as palavras *new, I* e *take* são *stop words*. Assim, as relações que possuem essas palavras são desconsideradas na criação do grafo.

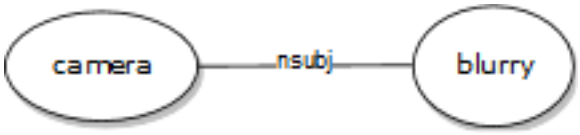

Figura 3.2: Grafo de dependência gerado a partir da tabela 3.3

Em (Yan et al, 2015) foi gerado um grafo orientado através do produto cartesiano do conjunto de opiniões pelo conjunto de características. O grafo sugerido por Yan et al

(2015) tende a aumentar com o tamanho da base de treinamento, devido ao produto cartesiano. Então, para tentar se diferenciar do proposto por Yan et al (2015), este trabalho propôs um grafo não direcionado. Cujo intuito é obter resultados semelhantes, utilizando uma estrutura menos custosa. O grafo gerado (figura 3.2) não é orientado, o que faz essa abordagem se diferenciar da proposta por Yan et al (2015). O grafo desse trabalho não é gerado pelo produto vetorial dos *heads* e *dependents*. Neste trabalho cada *heads* e *dependents* extraídos formam um vértice. Os vértices que possuem ao menos uma relação de dependência, possuem uma aresta ponderada com valor inicial igual a 1. Esta abordagem se diferencia do de Yan et al (2015) por causa da estrutura do grafo. Além disso, o valor de *PageRank* inicial de cada vértice e aresta é 1. Quando existem mais de uma relação do mesmo tipo com o mesmo *governor* e *dependent*, acrescenta-se 1 no peso da aresta. A figura 3.3 exemplifica o grafo deste trabalho.

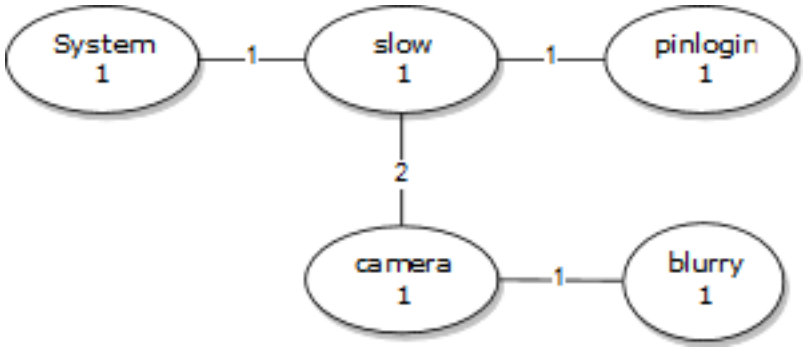

Figura 3.3: Grafo de dependência

Na figura 3.3, o valor de *PageRank* está abaixo do nome do vértice e o peso da aresta entre os vértices. O valor de *PageRank* dos vértices é recalculado, devido ele ser o mesmo para todos vértices. O cálculo do *PageRank* utiliza a equação abaixo:

$$PR(s) = c\sum_{v \in B_u} \frac{P_{(v,s)} * PR(v)}{P_v} \tag{3.2}$$

Essa equação é uma adaptação da equação 3.1, onde o termo $P_{(v,s)}$ é o peso da aresta que liga o vértice v ao vértice s, P_v representa o peso de todas as arestas que estão ligadas a v. A figura 3.4 mostra o valor de cada vértice após o cálculo do *PageRank* utilizando a equação acima.

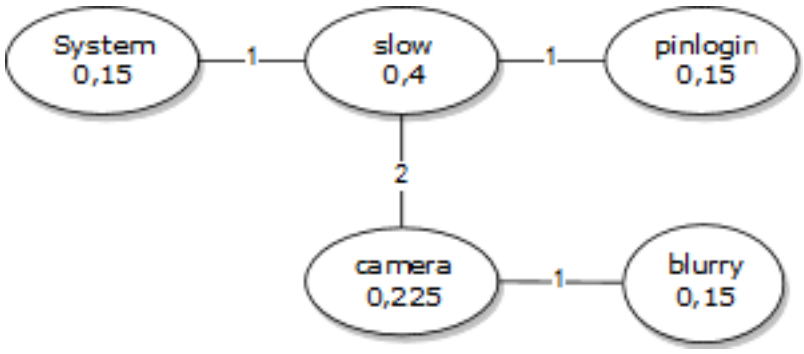

Figura 3.4: Grafo de dependência após calcular o *PageRank* de cada vértice

Passo 5: Extração dos pares

A partir do grafo gerado na etapa anterior, é criado o conjunto N formado por todos os vértices que possuem seu valor de *PageRank* acima de um limiar α. Neste trabalho, α é a média dos valores de *PageRank* dos vértices.

Para extrair os pares, é realizado o *parsing* de dependência do texto. Então seleciona-se cada nó n_i da árvore de *parsing*, tal que $n_i \in N$. Para cada n_i, cria-se o conjunto de vizinhos V_{n_i} de n_i, onde $V_{n_i} = \{v_1, v_2, ..., v_n\}$ e v_i é um vizinho de n do grafo. Então, seleciona-se os vizinhos de n_i na *Parser tree* que também são vizinhos no grafo, formando assim os pores de opinião (n_i, v_i).

No exemplo 3.2.2, a sentença (1) possui a palavra *slow*. Através da figura 3.4, tem-se que *slow* é um vértice do grafo e $PR(slow) = 0,4$. Como $\alpha = 0,215$ e $PR(slow) >= \alpha$, verifica-se se alguma outra palavra do texto é vizinha de *slow* no grafo. Como a palavra *system* é um nó vizinho de *slow* no grafo, o par *(system, slow)* é selecionado como opinião. A sentença (2), de acordo com a abordagem 3, não possui opinião a ser extraída. Pelo fato dos vértices do grafo da Figura 3.4 não serem formados pelas palavras da sentença (2).

(1)"*Also, I can't work because of a very slow system.*"
(2)"*My phone keeps turning off randomly, please I need the next update to fix this*"

Exemplo 3.2.2

Uma visão esquemática do processo de extração utilizando *PageRank* pode ser visto na figura 3.5.

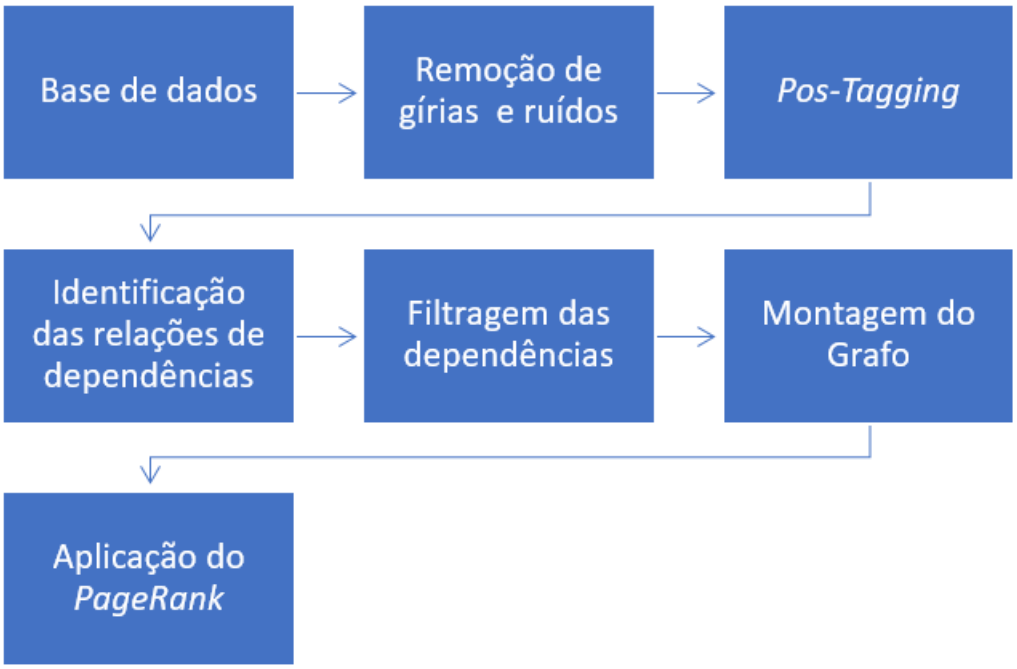

Figura 3.5: *Visão geral do processo de extração de característica utilizando PageRank*

A tabela 3.4, exemplifica os passos executados por Yan et al (2015); Somprasertsri et al (2010) que correspondem respectivamente ao autor 1 e 2. O autor 3 corresponde a abordagem 3, desenvolvida a partir dos autores 1 e 2.

3.2.2 EXPRS

Essa seção tem por objetivo apresentar, de forma simplificada, a ideia principal do algoritmo *EXPRS*, proposto por Yan et al (2015).

Inicialmente, é feito o *POS tagging* de cada sentença da base. Com base no resultado do *POS tagging*, é derivado um conjunto com todos os substantivos. Em seguida é feito um *parsing* de cada sentença para obter as relações de dependências existentes. Em (Yan et al, 2015), apenas as relações *nsubj, amod, rcmod* e *dobj* são consideradas como importantes. Então, seleciona-se apenas os substantivos que possuem alguma das relações consideradas como importantes.

Em seguida, é realizada uma filtragem nas relações escolhidas. Então, remove-se as relações de dependências onde os substantivos representam nomes de marca, nome próprio e etc.

A partir das relações de dependências selecionadas, cria-se um grafo direcionado.

Tabela 3.4: Passos das abordagem que utilizam relação de dependência

Passo	autor 1	autor 2	autor 3
Normalização	x	x	x
Utilizar substantivos como características	x	x	x
Utilizar adjetivos como opiniões	x	x	x
Utilizar verbos como opiniões		x	x
POS-Tagging	x	x	x
Extração das relações de dependência	x	x	x
Filtragem das relações de dependência	x		x
Geração do grafo direcionado através das dependências	x		
Geração do grafo não direcionado através das dependências			x
Uso de modelos probabilísticos para validação das relações extraídas como sentimentos		x	
Aplicação do *PageRank*	x		x
Seleção dos pares com maior valor *PageRank*	x		x
Extração de características implícitas	x		
Criação de uma lista com características pré-definidas			x

Assumindo que W representa o conjunto formado pelas relações de dependências. F é o conjunto das características (*heads*) em W. S é o conjunto o conjunto de todos os sentimentos (*dependents*) existentes em W. F é o produto cartesiano de F e V. Os nós do grafo são formados pelos elementos de V. Um sentimento s de um nó v_1, e uma característica f de um nó v_2 compõe um par (f, s). Se $(f, s) \in W$, então, v_1 e v_2 são conectados por um aresta $a_{1-2}(v1, v2)$.

De acordo com Yan et al (2015), quanto mais frequente é o substantivo, mais provável ele é de ser uma característica. Então, a importância $p(i)$ de um nó do grafo é definida como:

$$p(i) = (1 - \alpha)H(i) + \alpha \sum_{(j,i)\in E} \frac{p(j)}{O_j} \tag{3.3}$$

Onde α representa o fator de amortecimento, $p(j)$ é valor de *PageRank* do nó j. O_j é arestas que j *links*. E é o conjunto de arestas existentes. $H(i)$ é a função de frequência de ocorrencia de i, que é defina por:

$$H(i) = \frac{N \ln f(i)}{\ln(\prod_{j=1}^{N} f(j))} \tag{3.4}$$

Onde N é o total de nós existentes no grafo. $f(i)$ é a frequência de um nó i na base. A função de log tem por objetivo diminuir o impacto de vértices com pouca relevância.

Após as dependências serem extraídas e o grafo ser gerado, são escolhidos como candidatos a características e sentimentos as relações que possuem sua importância maior que um limiar. Então, para cada texto existente do *dataset*, é verificado se ele possui alguma das relações de maior importância. Caso ele possua, essa relação é extraída como característica (*head*) e sentimento (*dependent*).

4 Desenvolvimento

Para avaliação das quatro abordagens, foi criada uma base com *tweets* do perfil da empresa Apple com a finalidade de simular o uso das técnicas de extração de características de produtos em um ambiente real. Na seção 4.1 é discutida a criação da base de avaliação. Na seção 4.3 são apresentados os resultados e estes são analisados na seção 4.4.

4.1 Criação do *benchmark*

Os dados escolhidos para realizar AS são *tweets* direcionados ao *Twitter* da *Apple Support*. Os *tweets* direcionados ao suporte da *Apple* tem como finalidade buscar ajuda de como utilizar os produtos, reclamações, sugestões de melhorias, elogios e críticas para o SAC da empresa.

A escolha dos *tweets* para formar a base de dados levou em consideração a enorme quantidade e variedade de *tweets*, aliado à facilidade de obter esses dados reais. Também foi levado em consideração que os trabalhos de (Eirinaki et al, 2012; Barros et al, 2012; Siqueira et al, 2010; Hu et al, 2004; Yan et al, 2015; Somprasertsri et al, 2010) obtiveram ótimos resultados com textos bem estruturados e com poucos ruídos. Então, foi verificado como que estas heurísticas se comportam em uma base com um número maior de ruídos.

Para coletar os dados foi utilizado o *framework LINQ to Twitter*. O *framework* realiza a integração com a *API* fornecida pelo próprio *Twitter*. O *framework* permite a realização de consultas no site para recuperar as informações de interesse do desenvolvedor.

Por se tratar de uma rede social, vários usuários não se preocupam em escrever da maneira correta. Então, torna-se necessário realizar um pré-processamento na base dados. O *Twitter* é considerado um *DataSet* cheio de ruídos, ou seja, textos com erros gramaticais e um dialeto simbólico único, criado pela própria comunidade. Essas características dificultam a normalização das palavras (Han et al, 2011). Outro problema que favorece a escrita coloquial é regra imposta pelo próprio site, onde cada *tweet* não pode ultrapassar o tamanho de 140 caracteres. Devido à essa restrição, vários usuários passam

a utilizar gírias com o intuito de conseguir se expressar mais utilizando menos palavras.

A normalização é realizada através da remoção de menção a outros usuários (i.e. @AppleSuport), *hashtags* e *urls* (Han et al, 2011). A correção de gírias, abreviações e símbolos foi realizada através da utilização de um dicionário de gírias. Foi utilizado o dicionário Noslang[2], que contém sinônimos para 5512 gírias em língua inglesa.

Para realizar a avaliação dos algoritmos, foi utilizado um total de 22.580 *tweets*. A coleta dos dados foi realizada dos dias 14/10/2016 a 29/05/2017. Os dados coletados foram divididos entre a base de treinamento e a de testes. A base de testes possui 51 *tweets*. Os *tweets* dessa base foram selecionados de forma quase aleatória. Não foi possível fazer um processo totalmente aleatório, pois vários *tweets* destinados ao perfil *@AppleSuport* eram *spams*. O processo de seleção dos *tweets* foi realizado da seguinte maneira: primeiro foram escolhidos 100 *tweets* da base da dados de forma aleatória. Em seguida, foram selecionados 3 *tweets* dentre os 100, também de forma aleatória. Através de uma análise manual, verificava-se os 3 *tweets*. Caso nenhum deles fossem considerados como *spam* e representassem um texto opinativo, então selecionava-os para a base de validação. Caso contrário, selecionava-se novamente mais 3 *tweets*, dentro os 100. O processo acima foi repetido até gerar uma base de teste com 51 *tweets*.

Após selecionar os *tweets* da base de teste, foi necessário reanalisar cada *tweet* manualmente. Neste passo, o objetivo foi de extrair as características do produto e os sentimentos expressos para que fosse possível calcular a acurácia do algoritmo. A base de teste foi utilizada para calcular a precisão, cobertura e a medida F das abordagens. A base de treinamento é constituída pelos *tweets* restantes.

4.2 Métricas de avaliação

A avaliação do *data set* foi realizada utilizando as medidas de precisão, cobertura e medida F. A precisão P é calculada dividindo os resultados retornados que estão corretos n, pelo total de resultados retornados T_r.

[2]http://www.noslang.com/

$$P = \frac{n}{T_{\mathrm{r}}} = \frac{\text{opiniões relevantes} \cap \text{opiniões recuperadas}}{\text{opiniões recuperados}} \tag{4.1}$$

A cobertura R é a fração de resultados corretos retornados n sobre o total de resultados corretos da base T_c.

$$R = \frac{n}{T_{\mathrm{c}}} = \frac{\text{opiniões relevantes} \cap \text{opiniões recuperadas}}{\text{opiniões relevantes}} \tag{4.2}$$

Opiniões relevantes formam o conjunto de opiniões corretas. Já as opiniões recuperadas são o conjunto que corresponde a todas as opiniões retornadas pelo algoritmo. Medida F é uma média harmônica que combina precisão e cobertura.

$$F = 2\frac{P * R}{P + R} \tag{4.3}$$

4.3 Experimentos realizados

Um total de 600 *tweets*, foram escolhidos de forma aleatória para formar a base de testes. Cada algoritmo foi executado uma vez em 3 cenários diferentes. No primeiro cenário, o *data set* está normalizado, e todas *StopWord* são removidas. No segundo cenário, o *data set* não está normalizado, porém todas *StopWord* são removidas. No terceiro cenário, o *data set* está sem normalização e com *StopWords*.

Para as abordagens 1 e 3, foi criada uma lista com características pré-definidas dos produtos. Então, essas abordagens foram executadas novamente no cenário descrito acima.

Nos gráficos tem-se que a sigla *CC* diz que abordagem foi executada utilizando uma lista pré-definida de características do produto. As siglas Ab1, Ab2, Ab3 e Ab4 representam respectivamente as abordagens 1, 2, 3 e 4.

Os resultados alcançados podem ser visualizados, em forma de tabelas, no apêndice A.1.

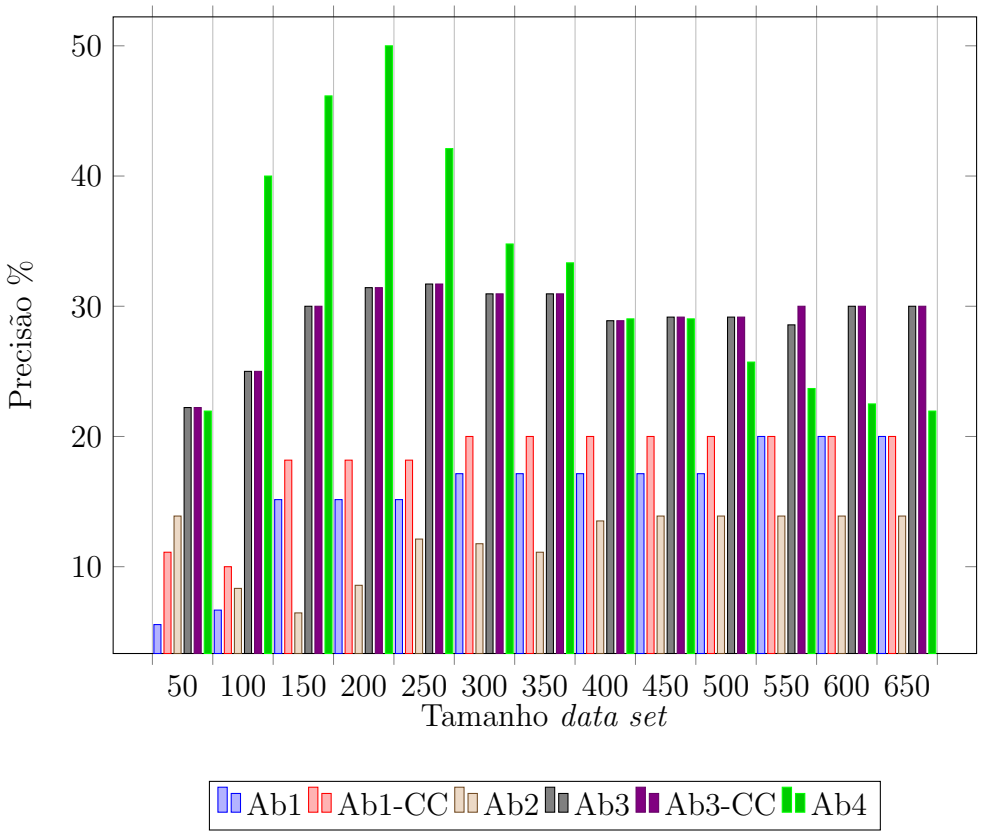

Figura 4.1: Precisão das abordagens com o *data set* normalizado e sem *Stop Words*

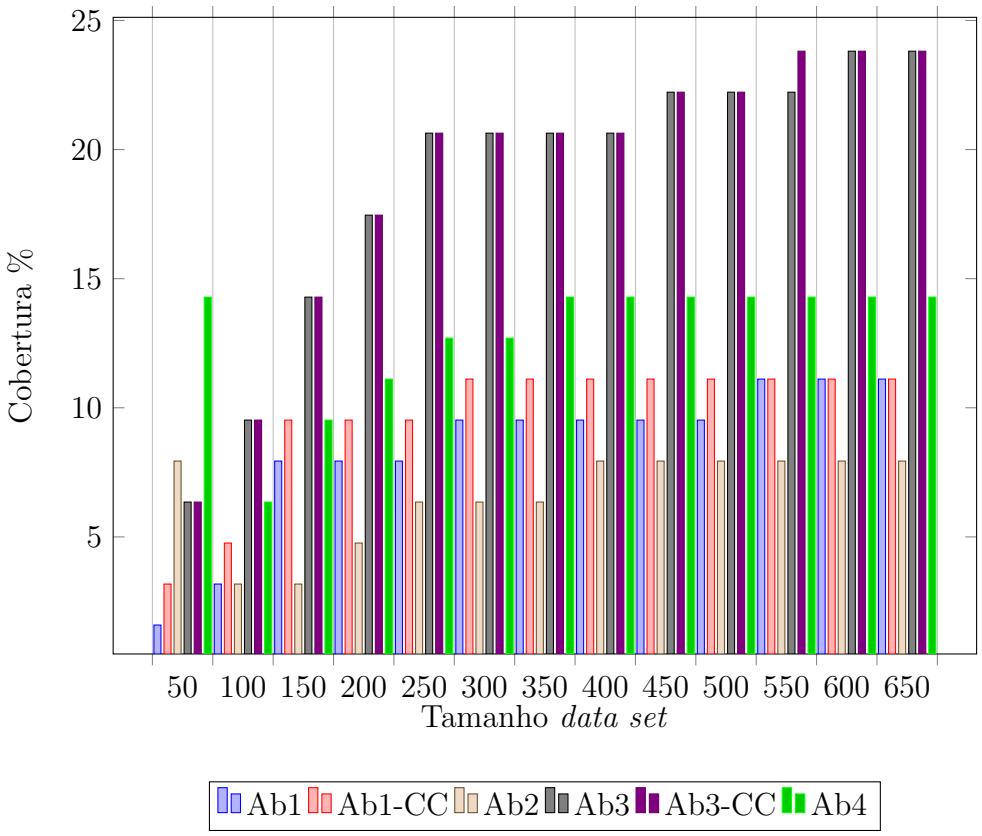

Figura 4.2: Cobertura das abordagens com o *data set* normalizado e sem *Stop Words*

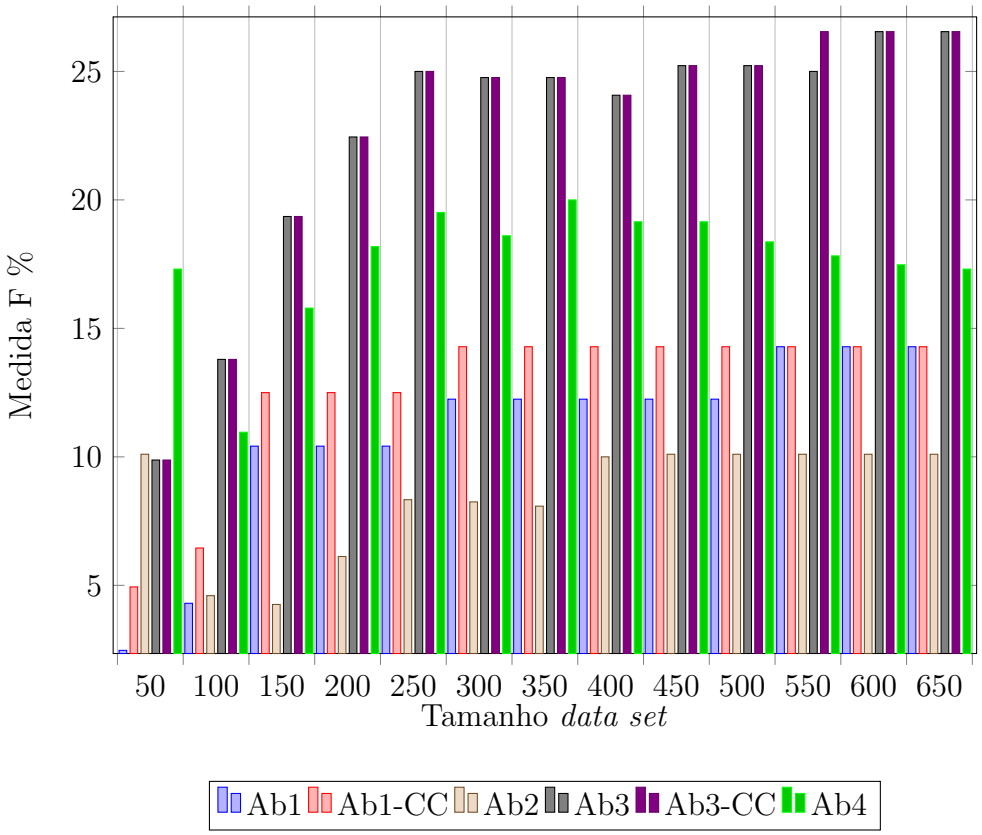

Figura 4.3: Medida F das abordagens com o *data set* normalizado e sem *Stop Words*

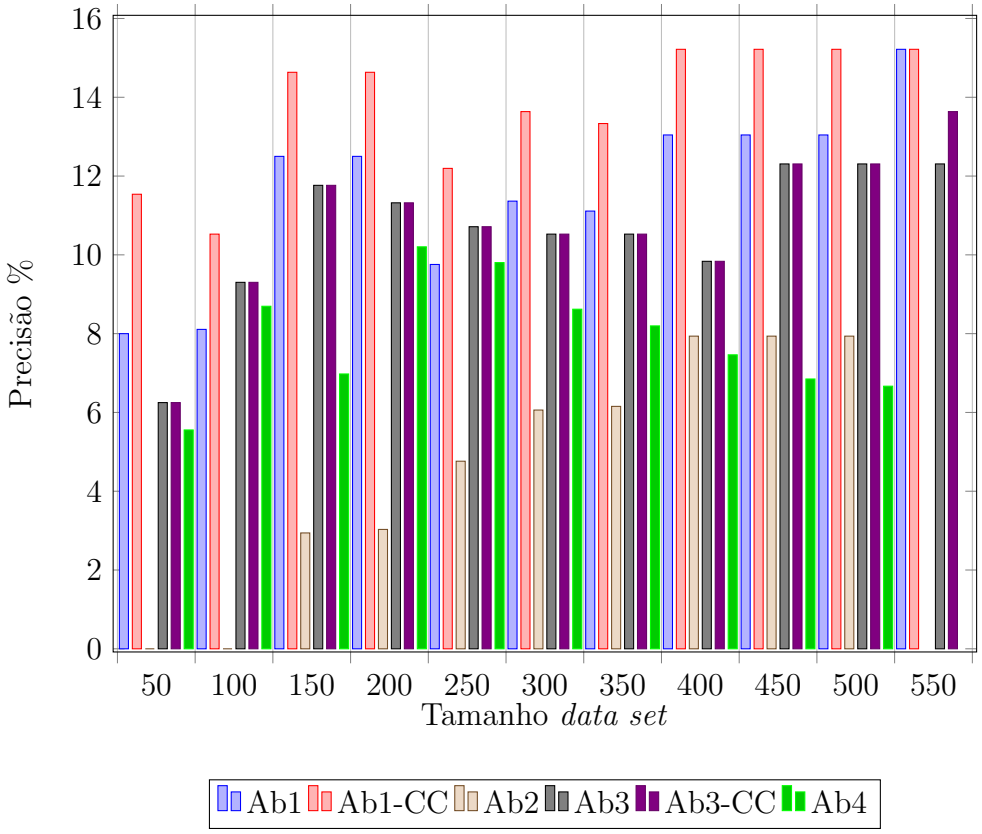

Figura 4.4: Precisão das abordagens com o *data set* não normalizado e sem *Stop Words*

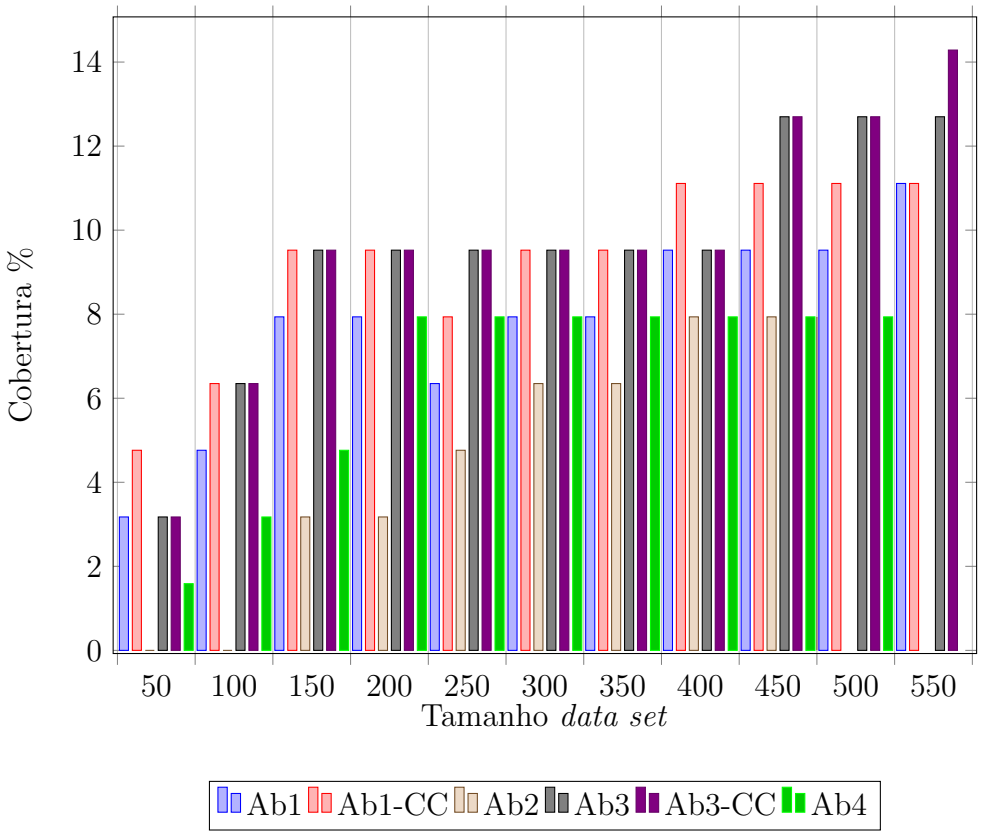

Figura 4.5: Cobertura das abordagens com o *data set* não normalizado e sem *StopWords*

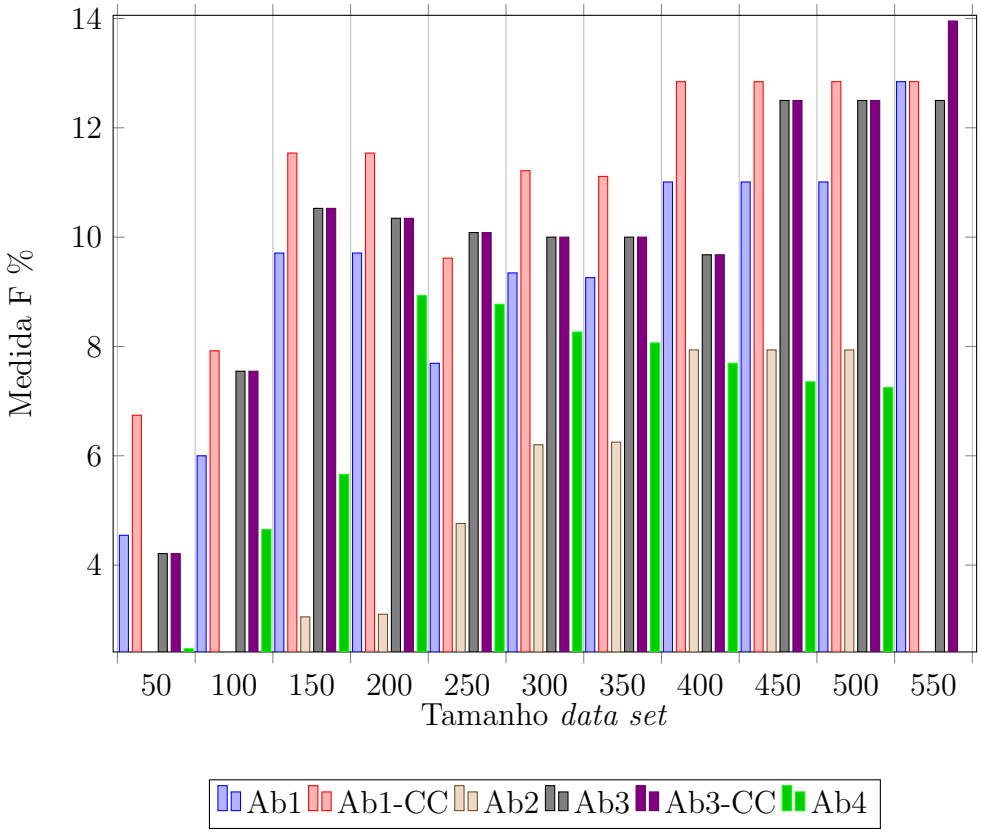

Figura 4.6: Medida F das abordagens com o *data set* não normalizado e sem *StopWords*

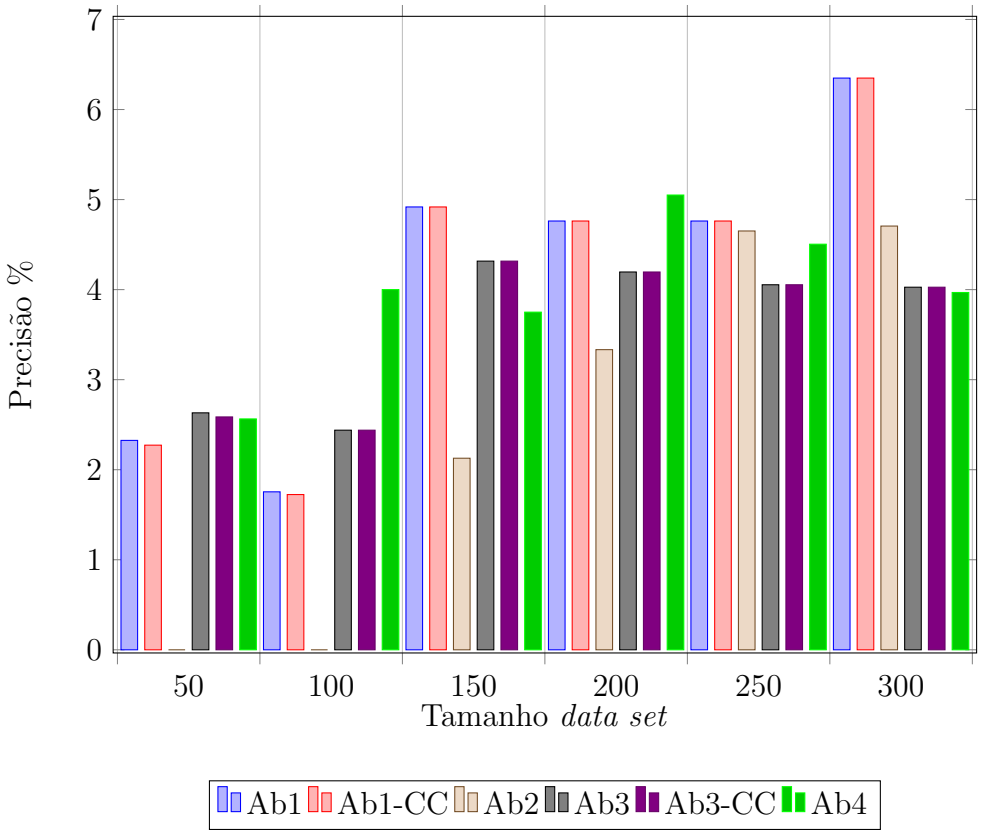

Figura 4.7: Precisão das abordagens com o *data set* não normalizado e com *StopWords*

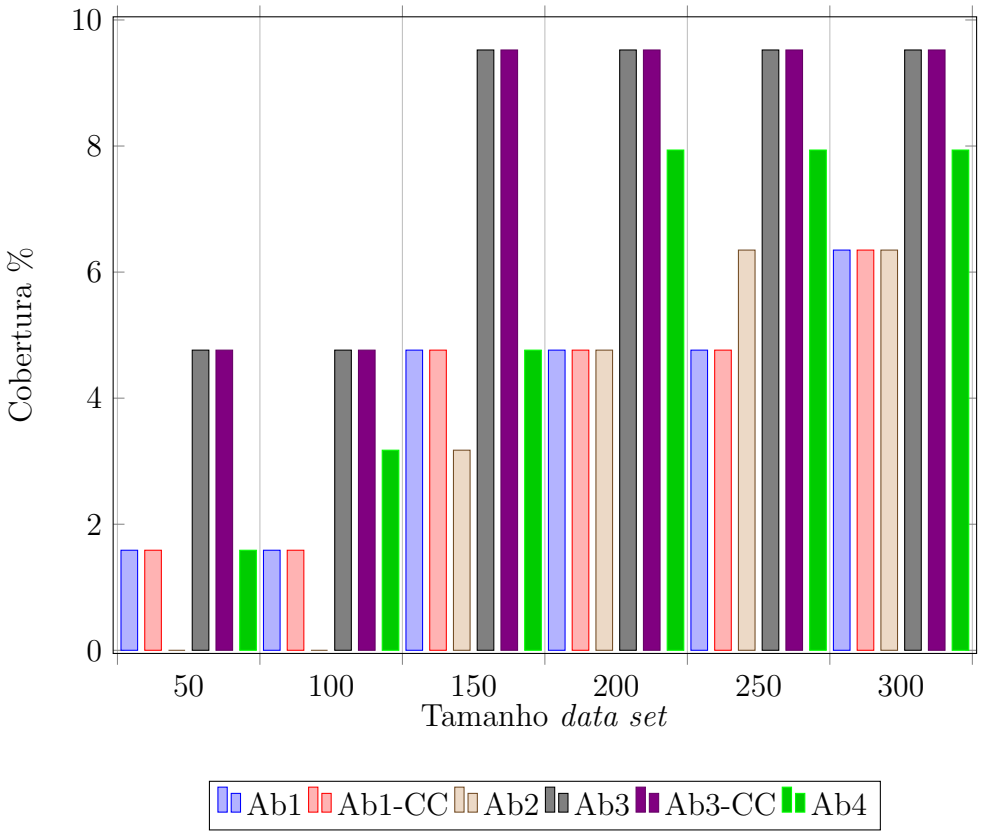

Figura 4.8: Cobertura das abordagens com o *data set* não normalizado e com *StopWords*

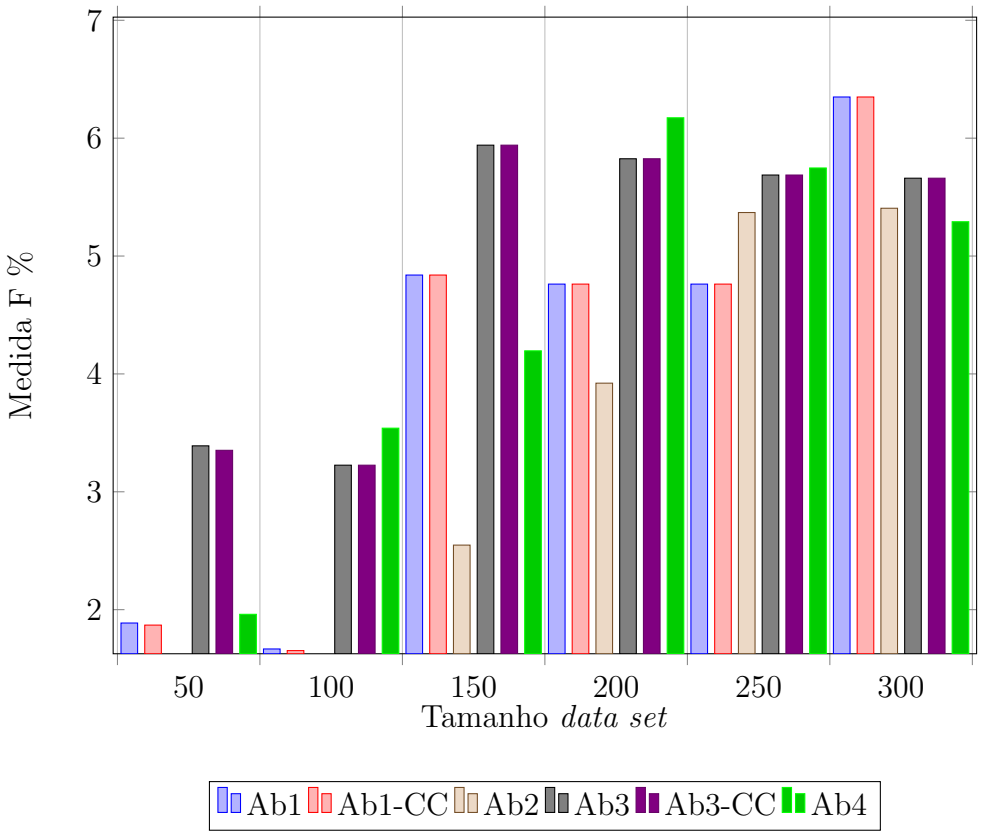

Figura 4.9: Medida F das abordagens com o *data set* não normalizado e com *StopWords*

4.4 Análise dos resultados

Os gráficos das Figuras 4.1, 4.2 e 4.3, mostram os resultados obtidos utilizando um *data set* normalizado e sem *StopWord*.

As abordagens que fazem uso das relações de dependências não verificam apenas se a característica é um substantivo. O diferencial dessas abordagens está em verificar como os substantivos se comportam na sentença. Pode-se ver esse diferença analisando a figura 4.1, onde as abordagens que utilizaram relações de dependências alcançaram maior precisão.

Para a abordagem 1, a implementação de uma lista com características pré-definidas, se mostrou útil quando o *data set* possui menos de 550 *tweets*. Para a abordagem 3, a lista de características mostrou-se útil apenas para um *data set* com 550 *tweets*. Independente da abordagem, a lista de características perde valor a medida que o tamanho do *data set* aumenta. Através dos gráficos das figuras 4.1, 4.2 e 4.3, é possível ver que a precisão, cobertura e medida F se mantém as mesmas, independente se a abordagem utiliza a lista com as características pré-definidas.

Para um *data set* normalizado e sem *StopWord*, é possível ver que a abordagem 3 apresentou os melhores resultados de precisão, cobertura e medida F.

O processo de normalização nesse trabalho contou com a remoção de citações a perfis do *twitter* e *links*. Gírias e abreviações foram transformadas para a sua forma normal durante a normalização. Através dos gráficos das figuras 4.4, 4.5 e 4.6 é possível ver como a acurácia das abordagens diminui.

Após a normalização, é calculada a frequência das palavras para as abordagens 1 e 2. Nas abordagens 3 e 4 é gerado o grafo a partir do tipo de ligação das palavras. Qualquer que seja a abordagem, caso uma palavra seja uma *StopWord* ela é removida. A remoção de *StopWords* é útil, pois palavras fora do contexto ou irrelevantes não são processadas. Quando essas *StopWord* são consideradas como características e sentimentos, tem-se uma diminuição na acurácia das abordagens, o que poder ser visto através dos gráficos das figuras 4.7, 4.8 e 4.9.

5 Conclusões

Este trabalho teve como objetivo realizar extração de características e sentimentos em textos do Twitter. Foi utilizada uma base real de textos oriundos do perfil de suporte da *Apple*. Foram realizados experimentos com quatro abordagens distintas adaptadas neste trabalho. Porém, os resultados apresentados nesse projeto não foram tão promissores. Pois uma série de fatores contribuíram para que a precisão, cobertura e medida F apresentasse baixa acurácia. Um dos fatores foi a exatidão exigida na comparação dos resultados obtidos com os esperados, com isso, características formadas por palavras compostas não são consideradas como corretas caso apenas uma palavra seja classificada. Outro fator foi a qualidade e a complexidade do texto utilizado para treinar os algoritmos. Através de uma breve análise foi verificado que na base de treinamento existiam vários *spam*.

Nos trabalhos estudados, os *data sets* não possuíam algumas peculiaridades como a desse trabalho. Os textos desse trabalho possuem tamanho máximo de 140 caracteres, ocasionando um maior número de abreviações e gírias. Outra característica desse trabalho, é que opiniões são formadas por adjetivos e verbos, na maioria dos trabalhos as opiniões são apenas adjetivos. Devido ao tipo de texto utilizado, as abordagens 2 e 4 não apresentaram resultados semelhantes ao mostrados pelos autores.

5.1 Sugestões de melhorias

Para fazer melhor proveito das abordagens existentes para realização de AS em *tweets* é necessário fazer algumas adaptações nas abordagens existentes.

Para as abordagens que utilizam frequência, uma opção de melhoria é adaptar o algoritmo para considerar palavras opinativas como sendo adjetivos e verbos. Com isso, frases como *"wifi isn't connecting"* teriam *Wifi* como característica, e *isn't connecting* como sentimento.

Nas abordagens que fazem uso do *PageRank*, pode-se adicionar mais tipos de

relações de dependência, tais como ligação do tipo (substantivos, verbos) ou (substantivos, adjetivos). Outra melhoria é procurar por relações que formam palavras compostas e então extrair palavras compostas como apenas uma característica.

A ferramenta *Stanford Parser* em alguns casos não classificou as palavras e as relações de dependência de forma correta. A classificação errada ocorre principalmente quando o *tweet* não possui pontuação correta. Através de uma breve análise, foi possível perceber que vários *tweets* não possuíam ponto de interrogação, prejudicando assim o resultado do *parser*. Então, outra forma de obter melhores resultados é através do treinamento da árvore de *parser* e do *POS-tagging*. O ideal é que a ferramenta seja treinada de acordo com o domínio trabalhado.

Durante o processo de normalização, palavras como {*IPhone4, IPhone5, IPhone5s, IPhone5c, smartphone*} foram consideradas como palavras distintas. Apesar da escrita distinta das palavras acima, elas possuem o mesmo significado. O fato de não substituir essas palavras por um mesmo sinônimo é considerado um problema. Nas abordagens de frequência, ao não utilizar um sinônimo, essas palavras terão menor frequência. Caso essas palavras fossem substituídas por um sinônimo em comum, elas teriam a mesma frequência. Então, seria possível alcançar uma melhor acurácia. Já nas abordagens que utilizam grafo, ao considerar essas palavras como distintas, tem-se um menor valor de *PageRank*.

Porém, algumas palavras que são consideradas como *Stop Words* são úteis para a AS. A palavra *useful* é considerada como uma *Stop Word*. Porém, em textos como *"Iphone is useful"*, *"Airdrop is not useful"* ou *"Make apple maps more useful"*, ligações importantes são desconsideras por causa da palavra *useful*.

Devido aos problemas relatados acima, os resultados obtidos não foram satisfatórios. Casos os problemas citados sejam resolvidos e seja utilizado um *data set* rotulado, espera-se obter melhores resultados.

A Apêndice

A.1 Resultados

As tabelas abaixo mostram de forma mais exata os resultados alcançados na seção 4.3.

Tabela A.1: Abordagem 1 com normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	5,55555	1,58730	2,46913
100	6,66666	3,17460	4,30107
150	15,15151	7,93650	10,41666
200	15,15151	7,93650	10,41666
250	15,15151	7,93650	10,41666
300	17,14285	9,52380	12,24489
350	17,14285	9,52380	12,24489
400	17,14285	9,52380	12,24489
450	17,14285	9,52380	12,24489
500	17,14285	9,52380	12,24489
550	20	11,11111	14,28571
600	20	11,11111	14,28571
650	20	11,11111	14,28571

Tabela A.2: Abordagem 1 sem normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	8	3,17460	4,54545
100	8,10810	4,76190	6
150	12,5	7,93650	9,70873
200	12,5	7,93650	9,70873
250	9,75609	6,34920	7,69230
300	11,36363	7,93650	9,34579
350	11,11111	7,93650	9,25925
400	13,04347	9,52380	11,00917
450	13,04347	9,52380	11,00917
500	13,04347	9,52380	11,00917
550	15,21739	11,11111	12,84403

Tabela A.3: Abordagem 1 sem normalização e com *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	2,32558	1,58730	1,88679
100	1,75438	1,58730	1,66666
150	4,91803	4,76190	4,83870
200	4,76190	4,76190	4,76190
250	4,76190	4,76190	4,76190
300	6,34920	6,34920	6,34920

Tabela A.4: Abordagem 1 com lista de características, com normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	11,11111	3,17460	4,93827
100	10	4,76190	6,45161
150	18,18181	9,52380	12,5
200	18,18181	9,52380	12,5
250	18,18181	9,52380	12,5
300	20	11,11111	14,28571
350	20	11,11111	14,28571
400	20	11,11111	14,28571
450	20	11,11111	14,28571
500	20	11,11111	14,28571
550	20	11,11111	14,28571
600	20	11,11111	14,28571
650	20	11,11111	14,28571

Tabela A.5: Abordagem 1 com lista de características, sem normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	11,53846	4,76190	6,74157
100	10,52631	6,34920	7,92079
150	14,63414	9,52380	11,53846
200	14,63414	9,52380	11,53846
250	12,19512	7,93650	9,61538
300	13,63636	9,52380	11,21495
350	13,33333	9,52380	11,11111
400	15,21739	11,11111	12,84403
450	15,21739	11,11111	12,84403
500	15,21739	11,11111	12,84403
550	15,21739	11,11111	12,84403

Tabela A.6: Abordagem 1 com lista de características, sem normalização e com *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	2,27272	1,58730	1,86915
100	1,72413	1,58730	1,65289
150	4,91803	4,76190	4,83870
200	4,76190	4,76190	4,76190
250	4,76190	4,76190	4,76190
300	6,34920	6,34920	6,34920

Tabela A.7: Abordagem 2 com normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	13,88889	7,93651	10,10101
100	8,33333	3,1746	4,5977
150	6,45161	3,1746	4,25532
200	8,57143	4,7619	6,12245
250	12,12121	6,34921	8,33334
300	11,76471	6,34921	8,24743
350	11,11111	6,34921	8,08081
400	13,51351	7,93651	10
450	13,88889	7,93651	10,10101
500	13,88889	7,93651	10,10101
550	13,88889	7,93651	10,10101
600	13,88889	7,93651	10,10101
650	13,88889	7,93651	10,10101

Tabela A.8: Abordagem 2 sem normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	0	0	NaN
100	0	0	NaN
150	2,94118	3,1746	3,05344
200	3,0303	3,1746	3,10077
250	4,7619	4,7619	4,7619
300	6,06061	6,34921	6,20155
350	6,15385	6,34921	6,25
400	7,93651	7,93651	7,93651
450	7,93651	7,93651	7,93651
500	7,93651	7,93651	7,93651

Tabela A.9: Abordagem 2 sem normalização e com *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	0	0	NaN
100	0	0	NaN
150	2,12766	3,1746	2,54777
200	3,33333	4,7619	3,92156
250	4,65116	6,34921	5,36913
300	4,70588	6,34921	5,40541

Tabela A.10: Abordagem 3 com normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	22,22222	6,34920	9,87654
100	25	9,52380	13,79310
150	30	14,28571	19,35483
200	31,42857	17,46031	22,44897
250	31,70731	20,63492	25
300	30,95238	20,63492	24,76190
350	30,95238	20,63492	24,76190
400	28,88888	20,63492	24,07407
450	29,16666	22,22222	25,22522
500	29,16666	22,22222	25,22522
550	28,57142	22,22222	25
600	30	23,80952	26,54867
650	30	23,80952	26,54867

Tabela A.11: Abordagem 3 sem normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	6,25	3,17460	4,21052
100	9,30232	6,34920	7,54716
150	11,76470	9,52380	10,52631
200	11,32075	9,52380	10,34482
250	10,71428	9,52380	10,08403
300	10,52631	9,52380	10
350	10,52631	9,52380	10
400	9,83606	9,52380	9,67741
450	12,30769	12,69841	12,5
500	12,30769	12,69841	12,5
550	12,30769	12,69841	12,5

Tabela A.12: Abordagem 3 sem normalização e com *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	2,63157	4,76190	3,38983
100	2,43902	4,76190	3,22580
150	4,31654	9,52380	5,94059
200	4,19580	9,52380	5,82524
250	4,05405	9,52380	5,68720
300	4,02684	9,52380	5,66037

Tabela A.13: Abordagem 3 com lista de características, normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	22,22222	6,34920	9,87654
100	25	9,52380	13,79310
150	30	14,28571	19,35483
200	31,42857	17,46031	22,44897
250	31,70731	20,63492	25
300	30,95238	20,63492	24,76190
350	30,95238	20,63492	24,76190
400	28,88888	20,63492	24,07407
450	29,16666	22,22222	25,22522
500	29,16666	22,22222	25,22522
550	30	23,80952	26,54867
600	30	23,80952	26,54867
650	30	23,80952	26,54867

Tabela A.14: Abordagem 3 com lista de características, sem normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	6,25	3,17460	4,21052
100	9,30232	6,34920	7,54716
150	11,76470	9,52380	10,52631
200	11,32075	9,52380	10,34482
250	10,71428	9,52380	10,08403
300	10,52631	9,52380	10
350	10,52631	9,52380	10
400	9,83606	9,52380	9,67741
450	12,30769	12,69841	12,5
500	12,30769	12,69841	12,5
550	13,63636	14,28571	13,95348

Tabela A.15: Abordagem 3 com lista de características, sem normalização e com *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	2,58620	4,76190	3,35195
100	2,43902	4,76190	3,22580
150	4,31654	9,52380	5,94059
200	4,19580	9,52380	5,82524
250	4,05405	9,52380	5,68720
300	4,02684	9,52380	5,66037

Tabela A.16: Abordagem 4 com normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	21,95121	14,28571	17,30769
100	40	6,34920	10,95890
150	46,15384	9,52380	15,78947
200	50	11,11111	18,18181
250	42,10526	12,69841	19,51219
300	34,78260	12,69841	18,60465
350	33,33333	14,28571	20
400	29,03225	14,28571	19,14893
450	29,03225	14,28571	19,14893
500	25,71428	14,28571	18,36734
550	23,68421	14,28571	17,82178
600	22,5	14,28571	17,47572
650	21,95121	14,28571	17,30769

Tabela A.17: Abordagem 4 sem normalização e sem *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	5,55555	1,58730	2,46913
100	8,69565	3,17460	4,65116
150	6,97674	4,76190	5,66037
200	10,20408	7,93650	8,92857
250	9,80392	7,93650	8,77192
300	8,62068	7,93650	8,26446
350	8,19672	7,93650	8,06451
400	7,46268	7,93650	7,69230
450	6,84931	7,93650	7,35294
500	6,66666	7,93650	7,24637

Tabela A.18: Abordagem 4 sem normalização e com *stop words*

Tamanho *data set*	Precisão	Cobertura	Medida F
50	2,56410	1,58730	1,96078
100	4	3,17460	3,53982
150	3,75	4,76190	4,19580
200	5,05050	7,93650	6,17283
250	4,50450	7,93650	5,74712
300	3,96825	7,93650	5,29100

Bibliografia

Barros, F.; Silva, N. R. ; Lima, D. **Sapair: Um processo de análise de sentimento no nível de característica**. In: Brazilian Conference on Intelligent System, 2012.

Cambria, E.; Schulle, B.; Xia, Y. ; Havasi, C. **New avenues in opinion mining and sentiment analysis**. In: IEEE Intelligent Systems, volume 31, p. 15–21. IEEE Computer Society, 2013.

Eirinaki, M.; Pisal, S. ; Singh, J. **Feature-based opinion mining and ranking**. In: Journal of Computer and System Sciences, volume 78, p. 1175–1184. Academic Press, Inc. Orlando, FL, USA, 2012.

Han, B.; Baldwin, T. **Lexical normalisation of short text messages: Makn sens a #twitter**. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, volume 1, p. 368–378. Association for Computational Linguistics Stroudsburg, PA, USA ©2011, 2011.

Hu, M.; Liu, B. **Mining and summarizing customer reviews**. In: KDD '04 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Knowledge Discovery and Data Mining, p. 168–177. ACM New York, 2004.

Kumar, R.; Raghuveer, K. **Dependency driven semantic approach to product features extraction and summarization using customer reviews**. In: Advances in Computing and Information Technology, volume 178, p. 225–238. Springer, Berlin, Heidelberg, 2013.

Lima, D. C. L. A. **Pairextractor: Extração de pares livre de domínio para análise de sentimentos**, 2011.

Marneffe, M.-C.; Manningi, C. D. **Stanford typed dependencies manual**. https://nlp.stanford.edu/software/dependencies_manual.pdf, 2008.

Page, L.; Brin, S.; Motwani, R. ; Winograd, T. **The pager-ank citation ranking: Bringing order to the web**. In: Tech. report, Stanford University, 1998.

Saranya, T. Mining features and ranking products from online customer reviews. **International Journal of Engineering Research & Technology (IJERT)**, v.2, 2013.

Schouten, K.; Frasincar, F. **Implicit feature detection for sentiment analysis**. In: Proceedings of the 23rd International Conference on World Wide Web, p. 367 – 368. ACM New York, NY, USA ©2014, 2014.

Silva, N. G. R. **Pairclassif - um método para classificação de sentimentos baseado em pares**, 2011.

Siqueira, H.; Barros, F. A feature extraction process for sentiment analysis of opinions on services. **III International Workshop on Web and Text Intelligence**, 2010.

Somprasertsri, G.; Lalitrojwong, P. Mining feature-opinion in online customer reviews for opinion summarization. **Journal of Universal Computer Science**, v.16, p. 938–955, 2010.

Tuarob, S.; Tucker, C. S. **Automated discovery of lead users and latent product features by mining large scale social media networks**. In: JOURNAL OF MECHANICAL DESIGN, volume 137, 2015.

Wu, Y.; Zhang, Q.; Huang, X. ; Wu, L. **Phrase dependency parsing for opinion mining**. In: EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, volume 3, p. 1533–1541. Association for Computational Linguistics Stroudsburg, PA, USA ©2009, 2009.

Yan, Z.; Xing, M.; Zhang, D. ; Maa, B. **Exprs: An extended pagerank method for product feature extraction from online consumer reviews**. In: Information & Management, volume 53, p. 850–858. Elsevier, 2015.

Zhang, Y.; Zhu, W. **Extracting implicit features in online customer reviews for opinion mining**. In: Proceedings of the 22nd International Conference on World Wide Web, p. 103 – 104. ACM New York, NY, USA ©2013, 2013.