Pós-Graduação em Ciência da Computação

DIANA CABRAL CAVALCANTI

MINERAÇÃO DE OPINIÕES BASEADA EM ASPECTOS PARA REVISÕES DE MEDICAMENTOS

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2017

Diana Cabral Cavalcanti

Mineração de Opiniões baseada em aspectos para revisões de medicamentos

Este trabalho foi apresentado à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR(A): **Prof.. Ricardo Bastos Cavalcante Prudêncio**

RECIFE
2017

Catalogação na fonte
Bibliotecário Jefferson Luiz Alves Nazareno CRB 4-1758

C376m Cavalcanti, Diana Cabral.
Mineração de opiniões baseada em aspectos para revisões de medicamentos / Diana Cabral Cavalcanti – 2017.
168.: fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, Recife, 2017.
Inclui referências e apêndices.

1. Inteligência artificial. 2. Minerações de opiniões. 3. Recuperação da informação. 4. Revisões de medicamentos. I. Prudêncio, Ricardo Bastos Cavalcante. (Orientador). II. Titulo.

006.3 CDD (22. ed.) UFPE-MEI 2017-185

Diana Cabral Cavalcanti

"**Mineração de Opiniões Baseada em Aspectos para Revisões de Medicamentos"**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutora em Ciência da Computação

Aprovado em: 14/08/2017

_____

Orientador: Prof. Dr. Ricardo Bastos Cavalcanti Prudêncio

BANCA EXAMINADORA

_____

Profa. Dra. Patricia Cabral de Azevedo Restelli Tedesco
Centro de Informática / UFPE

_____

Prof. Dr. Sergio Ricardo de Melo Queiroz
Centro de Informática / UFPE

_____

Profa. Dra. Solange Oliveira Rezende
Departamento de Ciência de Computação e Estatística / USP

_____

Prof. Dr. Renato Fernandes Corrêa
Departamento de Ciência da Informação / UFPE

_____

Prof. Dr. Rafael Ferreira Mello
Departamento de Estatística e Informática / UFRPE

Agradecimentos

Agradeço ao meu orientador Ricardo Prudêncio, pela dedicação e paciência durante todos os anos de desenvolvimento deste trabalho. Agradeço pelo aprendizado e todo o suporte e apoio na elaboração deste trabalho.

A professora Flávia Barros, pelas suas importantes dicas e disponibilidade sempre que precisei de auxílio.

Ao CESAR, como instituição, aos gerentes de projeto Felipe Furtado e Carlos Ferreira, por permitirem flexibilidade nos horários de trabalho para que eu pudesse conciliar o trabalho com minha formação pessoal. E pelo incentivo e apoio por parte da instituição e de amigos.

A professora Diana Inkpen, pela oportunidade e todo aprendizado vivenciado na Ottawa University.

Aos professores Solange Oliveira Rezende, Rafael Ferreira Mello, Renato Fernandes Corrêa, Sérgio Ricardo de Melo Queiroz e Patricia Cabral de A. Restelli Tedesco que fizeram parte da minha banca e deram, cada um, excelentes contribuições ao meu trabalho.

Enfim, sou grata a todos aqueles que direta ou indiretamente contribuíram nesta jornada de conhecimento.

Resumo

Mineração de Opinião baseada em Aspectos pode ser aplicada para extrair informações relevantes expressas por pacientes em comentários textuais sobre medicamentos (por exemplo, reações adversas, eficácia quanto ao uso de um determinado remédio, sintomas e condições do paciente antes usar o medicamento). Este novo domínio de aplicação apresenta desafios, bem como oportunidades de pesquisa em Mineração de Opinião. No entanto, a literatura ainda é escassa sobre métodos para extrair múltiplos aspectos relevantes presentes em análises de fármacos. Nesta tese foi desenvolvido um novo método para extrair e classificar aspectos em comentários opinativos sobre medicamentos. A solução proposta tem duas etapas principais. Na extração de aspectos, um novo método baseado em caminhos de dependência sintática é proposto para extrair pares de opiniões em revisões de medicamento. Um par de opinião é composto por um termo de aspecto associado a um termo opinativo. Na classificação de aspectos, propõe-se um classificador supervisionado baseado em recursos de domínio e de linguística para classificar pares de opinião por tipo de aspecto (por exemplo, Condição clínica, Reação Adversa, Dosagem e Eficácia). Para avaliar o método proposto, foram realizados experimentos em conjuntos de dados relacionados a três diferentes condições clínicas: ADHD, AIDS e Ansiedade. Para o problema de extração foi realizado avaliação comparativa com outros dois métodos, onde o método proposto atingiu resultados competitivos, alcançando precisão de 78% para ADHD, 75,2% para AIDS e 78,7% para Ansiedade. Enquanto para o problema de classificação, resultados promissores foram obtidos nos experimentos e várias questões foram identificadas e discutidas.

Palavras-chave: Mineração de Opinião. Extração de Aspectos. Classificação de Aspectos. Processamento de Linguagem Natural. Aprendizagem de máquina. Comentários sobre medicamento.

Abstract

Aspect-based opinion mining can be applied to extract relevant information expressed by patients in drug reviews (e.g., adverse reactions, efficacy of a drug, symptoms and conditions of patients). This new domain of application presents challenges as well as opportunities for research in opinion mining. Nevertheless, the literature is still scarce of methods to extract multiple relevant aspects present in drug reviews. In this thesis we propose a new method to extract and classify aspects in drug reviews. The proposed solution has two main steps. In the aspect extraction, a new method based on syntactic dependency paths is proposed to extract opinion pairs in drug reviews, composed by an aspect term associated to opinion term. In the aspect classification, a supervised classifier is proposed based on domain and linguistics resources to classify the opinion pairs by aspect type (e.g., condition, adverse reaction, dosage and effectiveness). In order to evaluate the proposed method we conducted experiments with datasets related to three different diseases: ADHD, AIDS and Anxiety. For the extraction problem, a comparative evaluation was performed with two other methods, the proposed method obtained competitive results, obtained an accuracy of 78% for ADHD, 75.2% for AIDS and 78.7% for Anxiety. For the classification problem, promising results were obtained in the experiments and various issues were identified and discussed.

Keywords: Opinion mining. Aspect extraction. Aspect classification. Natural language processing. Machine learning. Drugs reviews.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ADE: Adverse Drug Event

ADHD: Attention Deficit Hyperactivity Disorder

ADR: Adverse Drug Reaction

AIDS: Acquired Immune Deficiency Syndrome

API: Application Programming Interface

ASUM: Aspect and Sentiment Unification Model

COSTART: Coding Symbols for Thesaurus of Adverse Reaction Terms

CRF: Conditional Random Field

DDI: Drug-drug interactions

DRNN: Deep Recurrent Neural Networks

FDA: Food and Drug Administration

FRRF: Feature Representational Richness Framework

HDL: High-density Lipoprotein

HMM: Hidden Markov Models

KNN: K-Nearest Neighbors

LDL: Low-density Lipoprotein

LR: Logistic regression

ME: Máxima Entropia

MNB: Multinomial Naïve Bayes

MRF: Markov Random Field

NB: Naïve Bayes

NER: Name Entity recognition

PGM: Probabilistic graphical model

PLN: Processamento de Linguagem Natural

PMI: Pointwise Mutual Information

PMI-IR: Pointwise Mutual Information and Information Retrieval

POS: Tagging: Part-of-speech Tagging

RPPCA: Regressional probabilistic principal component analysis

SIDER: Side Effect Resource

SL: Subjectivity Lexicon

SVM: Support Vector Machines

SWN: SentiWordNet

SVR: Support Vector Regression

TF-IDF: Term Frequency Document Inverse

TF: Term Frequency

UMLS: Unified Medical Language System

VADER: Valence Aware Dictionary and sEntiment Reasoner

WHO: World Health Organization

WSD: Word Sense Disambiguation

Sumário

1 INTRODUÇÃO

Mineração de texto é definida como o processo extração de conhecimento a partir de grandes volumes de textos estruturados ou não estruturados utilizando métodos computacionais (HARPAZ et al., 2014). Muitos cientistas têm concentrado suas pesquisas para o desenvolvimento de técnicas de mineração de textos na área médica e farmacêutica a partir de dados publicamente disponíveis na web com objetivo de identificar informações que podem trazer melhorias e contribuir no campo da medicina e cuidados para saúde (TOMAR; AGARWAL, 2013; PAUL et al., 2016; DENECKE; DENG, 2015; ALI et al., 2013; DENECKE; NEJDL, 2009; SARKER et al., 2015b).

Informações textuais podem ser classificadas em textos que relatam fatos e textos que relatam opiniões. Fatos são expressões objetivas sobre entidades ou eventos, enquanto opiniões são normalmente expressões subjetivas onde atitudes, emoções, opiniões, sentimentos ou comparações são descritos sobre entidades (DENECKE, 2015a).

Minerar sentimentos e opiniões em mídias sociais médicas podem ter múltiplas aplicações, podendo ser estudadas opiniões e experiências sobre tratamentos para evidências clínicas, investigar relações entre sintomas, investigar estilo de vida, eficácia ou ineficiência de um medicamento no tratamento de uma doença, como também podem ser estudados os efeitos de complicações sobre o resultado de uma cirurgia ou tratamento (DENECKE; DENG, 2015; DENECKE, 2015a).

Assim como também o sentimento pode ser relatado através de uma mudança no estado da saúde do paciente, por exemplo, “o paciente se sente melhor ou pior”, também pode ser relatado através da descrição do resultado ou eficácia de um tratamento, por exemplo, “o procedimento cirúrgico foi um sucesso”, ou através do relato de experiências ou opiniões sobre um tratamento ou um medicamento específico, por exemplo, “... tomei este medicamento e estou dormindo muito bem.” (DENECKE; DENG, 2015; DENECKE, 2015a).

Mineração de Opinião ou Análise de Sentimento lidava originalmente com a classificação de opiniões quanto à sua polaridade, ou seja, classificar se um texto é uma citação positiva, negativa ou neutra sobre a entidade comentada. Atualmente tem sido

interpretada de forma mais ampla como o tratamento computacional de opiniões, sentimento e subjetividade em texto que pode incluir tarefas de seleção, extração, classificação em nível de documento, sentença ou aspecto. Outras tarefas também são estudadas em Mineração de Opinião, são elas: detecção de emoção (emotion detection), transferência de conhecimento entre domínios (cross-domain), transferência de conhecimento entre diferentes linguagens (cross-language), extração de entidades, extração de frases opinativas, aspecto ou características sobre entidades, classificação de tipo de aspecto, construção de léxicos de sentimento, mineração de sentenças comparativas sobre entidades ou aspectos e detecção de ironia e sarcasmo em comentários (LIU, 2012; MEDHAT et al., 2014; RAVI; RAVI, 2015; JUSOH; ALFAWAREH, 2014).

Pesquisas de Mineração de Opinião relacionadas à saúde é bastante recente (DENECKE; DENG, 2015). Os trabalhos mais recentes têm sido aplicados sobre dados de saúde on-line, como blogs médicos ou fóruns. E também, a partir de conteúdo médico para medir a qualidade sobre entidades citadas ou ajudar a avaliar as preocupações do paciente, sentimento e emoções sobre determinadas marcas, medicamentos ou procedimentos (SHARIF et al., 2014).

Outros trabalhos têm focado no domínio de revisões de medicamentos a fim de minerar reações adversas ao uso de remédios ou citações de sintomas relacionadas à condição no qual se encontrava o usuário ao usar o medicamento, ou também citações sobre dosagem e benefícios do uso do medicamento em questão.

Trabalhos anteriores sobre dados de citações de remédio sugerem que a previsão inicial de reações adversas a medicamentos pode ser essencial para prevenir a exposição de pacientes e prejuízos financeiros para as empresas farmacêuticas (SHARIF et al., 2015). Mineração de Opinião no contexto médico apresenta várias aplicações significativas, porém não é uma tarefa trivial e envolve vários desafios, no capítulo 3 discutimos sobre alguns deles (DENECKE; DENG, 2015; DENECKE, 2015a).

1.1 Objetivo

O objetivo principal desta tese é apresentar um método original para extração de aspectos relacionados a termos opinativos em comentários de medicamentos e classificá-los em uma das categorias de aspectos: Condição, Efeito Adverso, Dosagem, ou Eficácia.

Assim, o objetivo nesta pesquisa é identificar em revisões de medicamentos citações em texto que possam estar associados a aspectos de interesse como, reações adversas e eficácia quanto ao uso do medicamento, condições do paciente, entre outros. As condições do paciente, por exemplo, uma vez extraídas, podem ser úteis para revelar os motivos de uso do medicamento pelo paciente, como também revelar usos errados de um medicamento entre os pacientes. Outros tipos de aspectos podem proporcionar uma melhor avaliação das experiências relatadas dos usuários, bem como um melhor resumo das avaliações associadas a um determinado medicamento.

A implementação das tarefas de extração e classificação de tipo de aspecto e termos opinativos é de grande relevancia uma vez que corresponde à implementação parcial de um sistema de Mineração de Opinião. Este pode ser composto por tarefas de extração de aspectos, classificação de tipos de aspectos, classificação de polaridade e por fim, tarefas de sumarização a fim de organizar e apresentar de forma intuiva e objetiva ao usuário as informações detectas pelo sistema. Nesta pesquisa não são tratadas as tarefas de classificação de polaridade e sumarização.

Um sistema de Mineração de Opinião para avaliar *feedback* de usuários quanto ao uso de medicamentos pode ser adotada como potencial ferramenta de suporte a órgãos regulamentadores para monitoramento de novas ADRs e eficácia de medicamentos disponibilizados no mercado.

De modo a alcançar esse objetivo geral, o trabalho possui os seguintes objetivos específicos:

- Investigar métodos de extração de aspecto e termo opinativo na tarefa de Mineração de Opinião e sua eficácia no domínio de revisões de medicamentos;

- Investigar recursos linguísticos e de domínio para construção de um classificador de tipos de aspectos.

1.2 Problemas das abordagens atuais

Observamos que trabalhos que focam na tarefa de extração de aspectos em comentários sobre fármacos são limitados. Os trabalhos de EBRAHIMI et al. (2015), KORKONTZELOS et al. (2016) e MUKHERJEE et al. (2014), CHENG et al. (2014) se concentram na extração do aspecto "ADR". Enquanto MOON; BORKAR (2016) focou na extração dos tipos de aspecto "idade" e "gênero".

Foi verificado que estes trabalhos não focam em extrair aspectos ligados com um termo opinativo. Observe o comentário "heart palpitations decreased" citado em um comentário de medicamento, ao extrair o aspecto "heart palpitations" isoladamente podemos considerá-lo como um aspecto do tipo condição clínica ou então uma ADR causada pelo medicamento. Porém, ao extrair o aspecto ligado ao termo opinativo "decreased" podemos considerar um aspecto de eficiência positiva do medicamento.

Outra lacuna observada neste domínio é referente à classificação de tipos de aspectos em comentários de medicamentos. Os estudos identificados concentram apenas na classificação do tipo de aspecto ADR. Os trabalhos de EGGER et al. (2016), PATKI et al. (2014), SAHANA; GIRISH (2015), SARKER; GONZALEZ (2015) se dedicaram a classificar se comentários mencionam ADRs, categorizando comentários em sua maior parte em ADR ou Non-ADR. O trabalho de EBRAHIMI et al. (2015) se destaca dos demais, pois tratou a categorização de aspectos em ADR ou sintoma.

Segundo, NA et al. (2012), NA; KYAING (2015), (NIKFARJAM et al., 2015; NIKFARJAM, A. et al., 2015) em um comentário de medicamento podem haver citações referente a outros tipos de aspectos, além de ADR, como dosagem, eficácia, custo, entre outros. Na seção 3.3 apresentamos uma discussão mais detalhada.

1.3 Trabalho realizado

Para resolver os problemas descritos na seção anterior, esta tese apresenta uma nova abordagem de extração e classificação de tipo de aspectos em revisões de medicamentos.

Para a tarefa de extração de aspecto, um método linguístico, simples e eficiente, baseado na combinação de caminhos de dependência representados em árvore sintática é adotado para extrair pares de opinião. Neste estudo, um par de opinião é composto por um termo que representa um tipo de aspecto e seu respectivo termo opinativo.

Além da extração de pares de opinião, o método também extrai termos negativos ligados a termos opinativos que pode inverter o sentido da opinião dada. Por exemplo, adicionando o termo "not" no mesmo exemplo citado na seção anterior temos "heart palpitations not decreased", onde pode apontar para uma ineficiência do medicamento. Assim como também, o método identifica termos de intensidade ligados a termos opinativos, por exemplo, as expressões "pain less intense" e "pain very intense" apresentam grau de intensidade diferente, pois possuem os termos "less" e "very" ligados ao termo opinativo.

Para a tarefa de classificação, o algoritmo de aprendizagem supervisionada *Random Forest* é adotado para classificar pares de opinião entre quatro tipos de aspectos que são bastante frequentes em comentários textuais sobre medicamentos: Condição clínica ou sintoma, ADR, Dosagem ou Eficácia. Um conjunto de sete recursos foi combinado para construção de um modelo preditivo de tipo de aspectos, incluindo recursos linguísticos e de domínio.

Experimentos foram realizados em três bases de condições clínicas: ADHD, AIDS e Anxiety. Para cada base um conjunto de pares de opinião válidos foi extraído manualmente, e também para cada par de opinião foi atribuído um rótulo de tipo aspecto. Estes conjuntos são utilizados para comparar a eficácia do método de extração automatizado e avaliar a performance quanto a predição da classe alvo (o tipo de aspecto de cada par de opinião) do classificador.

O método de extração proposto foi comparado a outros três estudos quanto à precisão, cobertura e F-Measure, onde superou estes estudos em medida de F-Measure e

atingiu precisão acima 0.75% para as três bases de experimentos. Também foi avaliada a taxa de acerto e cobertura para cada combinação de caminho de dependência sugerido.

O modelo de classificação apresentou nos experimentos resultados promissores, várias questões foram identificadas e são discutidas neste estudo. Detalhes da solução proposta são apresentados no capítulo 4 e os resultados obtidos estão descritos no capítulo 5.

1.4 Contribuições

O trabalho atual preencheu uma lacuna na literatura ao investigar e propor métodos efetivos para mineração de múltiplos aspectos em textos opinativos sobre medicamentos. Assim, podemos citar as seguintes principais contribuições:

- Proposta de um novo método para extrair aspectos em revisões de medicamentos baseado em caminhos de dependência sintática;

- Investigação de técnicas anteriores que usam recursos linguísticos para extração de aspecto e sua adequação para o domínio de revisões de medicamentos;

- Produção de um conjunto de dados composto de revisões de usuários sobre o uso de medicamentos etiquetados por tipo de aspectos, dos quais podem ser adotadas para novos experimentos;

- Proposta de um novo método supervisionado baseado em recursos específicos de domínio e recursos linguísticos para a classificação de tipo de aspecto;

- Resultados apresentados à comunidade científica em:

  - “Unsupervised aspect term extraction in Online Drugs Reviews” em *The 30th International Florida Artificial Intelligence Research Society Conference* (CAVALCANTI; PRUDENCIO, 2017b) (Anexo A);

- "Aspect-based Opinion Mining in Drug Reviews" em *The 18th EPIA Conference on Artificial Intelligence* (CAVALCANTI; PRUDENCIO, 2017a) (Anexo B).

1.5 Organização da Tese

Além dessa introdução, a tese está organizada em mais seis capítulos conforme apresentado abaixo:

- **Capítulo 2 – Mineração de Opinião baseada em Aspecto:** Nesse capítulo, são apresentados os conceitos e terminologias aplicados à Mineração de Opinião baseada em Aspecto, descreveremos sobre a aplicabilidade do tema e tarefas relacionadas.

- **Capítulo 3 – Mineração de Opinião em contexto médico:** Neste capítulo, abrimos a discussão sobre Mineração de Opinião em domínio médico e farmacêutico. É apresentado como Mineração de Opinião tem sido aplicada para este contexto, desafios e trabalhos relacionados.

- **Capítulo 4 – Solução proposta:** Nesse capítulo, é descrito detalhadamente as tarefas de extração e classificação de aspecto em comentários de medicamentos proposto nesta pesquisa.

- **Capítulo 5 – Resultados:** Este capítulo apresenta o conjunto de dados adotado para realização de experimentos e os resultados obtidos a partir dos experimentos realizados para as tarefas descritas no Capítulo 4;

- **Capítulo 6 – Conclusões:** Nesse capítulo, concluímos a tese. Descrevemos um breve resumo dos resultados obtidos, apontando as contribuições do trabalho realizado, além das possibilidades de trabalhos futuros.

2 MINERAÇÃO DE OPINIÃO BASEADA EM ASPECTO

Mineração de Opinião corresponde ao estudo computacional de opiniões, avaliações, atitudes e emoções dadas por pessoas sobre determinados aspectos em relação a uma ou mais entidades em textos estruturados ou não estruturados (LIU, 2010; LIU, 2012; MEDHAT et al., 2014; RAVI; RAVI, 2015; ZHANG; LIU, 2014). No artigo de CHEN; ZIMBRA (2010), Mineração de Opinião é uma subdisciplina de mineração de dados e linguística computacional, que aponta para várias técnicas computacionais para extração, classificação, compreensão e avaliação de opiniões expressas em várias fontes de texto como notícias, comentários em meios de comunicação social e outros tipos de conteúdo gerado pelo usuário na web.

ESULI (2008) menciona em seu trabalho que a denominação Mineração de Opinião não é a única referenciada na literatura. Outros nomes têm sido empregados como Análise de Sentimento, Classificação de Sentimento, Análise de Opinião e Análise Afetiva. O autor também cita que estes nomes podem ser considerados aproximadamente como sinônimos para denominar a disciplina geral e que possivelmente indicam algumas subtarefas específicas.

Assim como ESULI (2008) comenta, LIU (2007) também inclui em seu trabalho que a denominação Análise de Sentimento tem sido aplicada como foco na classificação de opiniões quanto à sua polaridade. Mas que atualmente também tem sido interpretado de forma mais ampla para definição do tratamento computacional de opiniões, sentimento e subjetividade, que pode incluir tarefas de seleção, extração, classificação e comparação de opiniões em frases que expressam comparações entre pessoas, produtos, entre outros. Análise de Sentimento também pode ser considerada como uma etapa da Mineração de Opiniões. Deste modo, os termos Mineração de Opinião e Análise de Sentimento indicam o mesmo campo de pesquisa. Para este trabalho será adotada a denominação de Mineração de Opinião.

Pesquisas relacionadas à Mineração de Opinião em textos iniciaram sobre textos de avaliações e comentários relacionados ao consumo de bens e serviços, principalmente sobre produtos ou serviços, como por exemplo, hotéis, restaurantes, filmes, produtos como celular ou câmera fotográfica, entre outros. Posteriormente pesquisas expandiram para outros contextos relacionados, como a questões políticas a

respeito de opiniões sobre candidatos concorrentes a eleições (JI et al., 2013; YANG; YANG, 2015).

Mineração de Opinião também é relacionada aos sistemas de recomendação filtrando produtos citados positivamente, pois pode não ser conveniente que tal sistema recomende produtos ou serviços que recebam uma grande quantidade de avaliações negativas. Também passou a ser aplicada para criar resumos de opiniões ou a sumarização de múltiplas revisões de pontos de vistas semelhantes sobre um time de futebol, uma companhia aérea, um artista, um programa de tv, entre outros. Ser capaz de reconhecer esse tipo de informação em uma forma sistemática nos dá uma melhor imagem da opinião pública, além de evitar que leitores leiam dezenas de comentários similares (LIU, 2012; MEDHAT et al, 2014; RAVI; RAVI, 2015; JUSOH; ALFAWAREH, 2014).

Mineração de Opinião atua sobre fragmentos de textos de qualquer tamanho ou formato. De acordo com LIU (2012), um texto opinativo é formado por pelo menos dois componentes: um objeto e um sentimento sobre este objeto. A identificação de sentimento em textos pode ocorrer em diferentes níveis (LIU, 2012; BECKER; TUMITAN, 2013; FELDMAN, 2013; NA et al., 2012; NA; KYAING, 2015):

- **Documento**: Pretende localizar o sentimento geral do autor em um documento texto opinativo. Os primeiros estudos em Mineração de Opinião focavam principalmente em detectar a polaridade de um documento, se positiva ou negativa. Mas estudos posteriores demonstraram quando se necessita uma análise mais profunda de texto, a investigação de sentimento no nível do documento é menos eficaz.

- **Sentença**: Visa identificar o sentimento em uma sentença existente no texto. Inicialmente o texto é subdivido em sentenças, sendo possível classificá-las em subjetivas ou objetivas. Em seguida, cada sentença subjetiva é analisada individualmente. Esta abordagem permite uma análise mais profunda no texto, pois um texto pode apresentar múltiplas opiniões.

- **Aspecto**: Mineração de Opinião em nível de sentença permite analisar cada sentença quanto a sua polaridade. No entanto, em uma mesma sentença pode conter múltiplas opiniões sobre uma mesma ou entidades diferentes. Com a

análise textual sobre aspecto é possível avaliar cada atributo ou entidade descrito em uma sentença.

Minerar opinião em nível de documento ou sentença é muitas vezes insuficiente para aplicações. Ambas as análises não apresentam o que exatamente o autor gosta e não gosta sobre uma entidade alvo, pois um documento ou uma sentença pode citar diferentes entidades e/ou expressar diferentes opiniões sobre aspectos de uma ou mais entidades (LIU, 2012). A extração e classificação de aspectos e entidades em textos é chamada de *Análise de Sentimento baseada em aspecto* (Aspect-based) ou também *Mineração de Opinião baseado em características* ou *aspectos* (Feature-based) (ZHANG et al., 2015).

Segundo LIU (2012), Mineração de Opinião em nível de aspecto é baseada na premissa de que uma opinião consiste de um termo opinativo (positivo ou negativo) e um alvo (opinion target). Um alvo é descrito por entidades e seus respectivos aspectos. Logo, visa classificar o sentimento em relação aos aspectos específicos de cada entidade. Uma opinião identificada sem o seu respectivo alvo tem uso limitado.

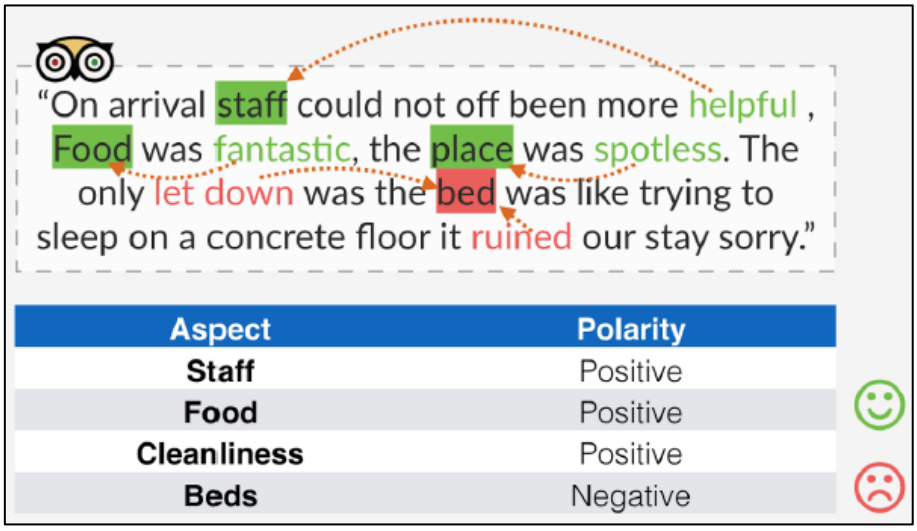


Figura 2.1: Exemplo de sentenças e atributos relacionados à Mineração de Opinião baseada em aspecto (RUDER et al., 2016)

A Figura 2.1 consiste de um trecho de opinião sobre um hotel (entidade), extraído do website *tripadvisor.com*. Quatro aspectos sobre a entidade hotel são avaliados: "staff", "food", "place (cleanliness)" e "bed" e dada sua respectiva opinião:

“helpful”, “fantastic”, “spotless” e “ruined”. E, conforme apresentado na figura, a opinião sobre cada aspecto pode ser classificada como positivo ou negativo.

Mineração de Opinião baseada em aspectos foi discutida pela primeira vez no trabalho de HU; LIU (2004). Neste trabalho, foi desenvolvido um método para minerar e sumarizar somente características de produtos (product features) no qual clientes expressam opiniões positivas ou negativas. Regras de associação foram utilizadas para minerar frases nominais frequentes como potenciais características de produtos. Os adjetivos mais frequentes são minerados em sentenças que foram localizadas características de produtos como potenciais termos opinativos. Em seguida, para cada característica de produto, o adjetivo mais próximo foi tratado como seu efetivo termo opinativo. Sinônimos e antônimos existentes no léxico WordNet (MILLER, 1995) é agregado a um conjunto *seed* de comuns adjetivos a fim de expandir o conjunto de termos opinativos. Por fim, termos opinativos são utilizados para minerar infrequentes características.

Para um melhor entendimento descrevemos abaixo definições relacionadas à Mineração de Opinião baseada em aspecto (JUSOH; ALFAWAREH, 2014; LI et al., 2015; JO; OH, 2011; WANG; ESTER, 2014):

- **Titular da Opinião:** Pessoa ou organização (opinion holder) que expressa uma opinião específica sobre um objeto particular.

- **Comentário:** pode consistir de um documento de texto, uma palavra, uma frase, ou um parágrafo.

- **Entidade:** Indica um produto alvo, serviço, tema, assunto, pessoa, organização ou evento discutido em um comentário. Uma entidade pode conter um conjunto de componentes e atributos e, cada componente da entidade pode ter seus próprios subcomponentes e atributos. Exemplo: restaurante, telefone móvel, hotel, Barack Obama.

- **Aspecto:** Uma palavra ou uma colocação que determina um atributo de uma entidade. Por exemplo: “price”, “value”, and “worth” pode caracterizar o aspecto “price” para telefone móvel. A sentença “The phone is great but the battery life is too short” possui dois aspectos “phone” e “battery life”.

- **Termo opinativo:** Corresponde a termos que expressam o sentimento ao avaliar um aspecto. Exemplo: a sentença “The phone is great but the battery life is too short”; o termo opinativo “great” para o aspecto “phone” e termo opinativo “too short” para o aspecto “battery life” para a entidade “mobile phone”. Um termo opinativo também pode estar ligado a um termo modificador. Exemplo “less expensive”, “extremely clear”, “very affordable”, os termos “less”, “extremely” e “very” são modificadores de palavras opinativas.

- **Polaridade:** Indica se uma opinião é positiva, negativa ou neutra. Exemplo: a sentença “The phone is great but the battery life is too short”; possui comentário positivo “great” para o aspecto “phone” e comentário negativo “too short” para o aspecto “battery life”.

- **Par de opinião:** Composto de um aspecto e um termo opinativo. Exemplos: { battery life; too short }, { phone; great }.

- **Tempo:** Um termo ou uma cláusula que descreve situações, estado ou condições da avaliação. Exemplo: “for outdoors”, “on the freeway”, “at night” (ZHANG et al., 2011).

- **Opinião:** Segundo LIU (2012), uma opinião é definida por uma quíntupla (e_i, a_{ij}, s_{ijkl}, h_k, t_l) onde:

    - $\mathbf{e_i}$: corresponde a uma entidade
    - $\mathbf{a_{ij}}$: corresponde a aspectos da entidade $\mathbf{e_i}$
    - $\mathbf{s_{ijkl}}$: corresponde ao termo opinativo sobre o aspecto $\mathbf{a_{ij}}$ da entidade e_i
    - $\mathbf{h_k}$: corresponde ao titular da a opinião
    - $\mathbf{t_l}$: corresponde o tempo no qual a opinião foi expressa por $\mathbf{h_k}$

2.1 Aspectos explícitos e implícitos

Um aspecto correspondente a cada componente ou atributo de uma entidade pode ser mencionado explicitamente ou implicitamente. Observe as seguintes sentenças abaixo:

(1) “The staff was very nice but Dr. Greenberg seemed as if he was too busy.”

(2) “... prescribed most expensive medication after asking if I had drug coverage”.

Na sentença (1) podemos identificar explicitamente os aspectos “The staff” e “Dr. Greenberg”. Já na sentença (2) podemos inferir o aspecto implícito “price” a partir da citação “most expensive medication” (MARRESE-TAYLOR et al., 2014; MOGHADDAM; ESTER, 2012).

Aspectos explícitos denotam explicitamente o alvo em sentenças opinativas e são claramente mencionados na sentença. Em geral, substantivos e sentenças nominais são minerados como potenciais aspectos. *POS Tagging*, *n-gramas*, *bag of words e similaridade* de termos são as técnicas mais frequentes adotadas para identificar *tokens* em sentenças, e também o uso de *árvore de dependência sintática* são usadas para análise de dependência entre termos (LIU, 2012; MEDHAT, W. et al, 2014; MAHADIK; BHARAMBE, 2015). Identificar aspecto explícito pode contribuir indiretamente na identificação de aspectos implícitos citados em sentenças (LAGUITAN et al., 2015).

Minerar aspectos implícitos é uma tarefa mais desafiadora, porque estes não podem ser identificados por somente métodos simples e relação sintática de termos. Aspectos podem ser derivados do uso de dicionário, por exemplo, o termo “expensive” pode estar relacionado à “high price”. O número de aspectos relacionados pode ser incrementado através da utilização da relação entre múltiplos dicionários (LAGUITAN et al., 2015). Detecção de categoria de aspecto também pode contribuir com o problema de extração de aspecto implícito (ZHOU et al., 2015).

Aspectos explícitos podem ser indicados por diferentes classes gramaticais, como adjetivos, advérbios, verbos, verbos frasais, expressões, entre outros (NOFERESTI; SHAMSFARD, 2015b). HU; LIU (2004) focaram em seu trabalho apenas na extração de aspectos explícitos. A pesquisa de MANKAR; INGLE (2015) apresentou um método supervisionado chamado *Implicit Aspect Indicators (IAI)* que trata o problema de extração de aspectos implícitos e explícitos usando *Conditional Random Fields*. Outros trabalhos relacionados à extração de aspectos (ZHANG; ZHU, 2013; LAL; ASNANI, 2014; ZENG; LI, 2013; WANG et al., 2013; SU et al., 2006; VIVEKANANDAN; ARAVINDAN, 2014).

NOFERESTI; SHAMSFARD (2015b) descreveram em seu trabalho sobre opiniões implícitas e explícitas em outra perspectiva, denominando de opinião direta e indireta, onde opinião direta é expressa diretamente sobre uma entidade ou um dos seus aspectos e, opinião indireta é expressa em uma entidade com base no seu efeito sobre outras entidades.

2.2 Etapas da Mineração de Opiniões baseada em Aspectos

Mineração de Opinião baseada em aspecto envolve três principais tarefas (Figura 2.2): (1) extrair aspectos e identificar termos opinativos sobre aspectos que usuários têm expressado suas opiniões, (2) classificação, e (3) sumarização (BRYCHCÍN et al., 2014; PONTIKI et al., 2014; ALGHUNAIM, 2015).

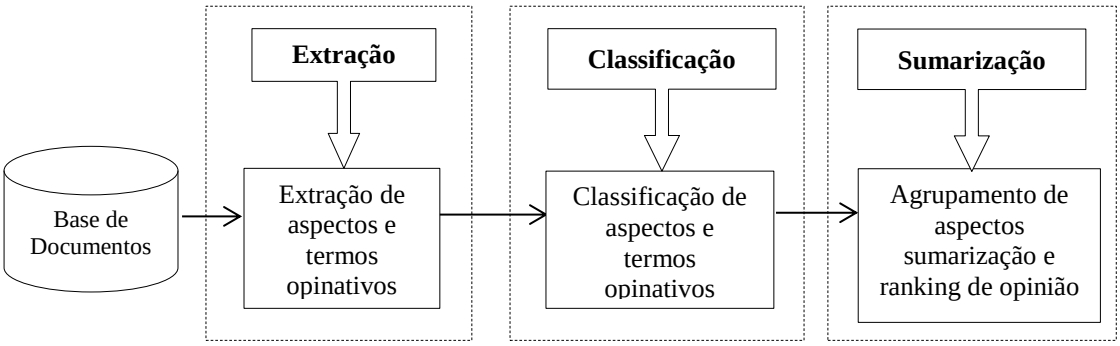


Figura 2.2: Etapas da Mineração de Opinião baseada em aspecto

A extração de aspecto pode ser dividida em subtarefas que incluem extração de aspectos implícitos e explícitos e agrupamento de aspectos por similaridade. E também, a identificação de termos opinativos sobre aspectos. A classificação pode envolver tarefas de identificar a categoria de um aspecto, assim como também abranger a tarefa de predição da polaridade dos termos opinativos sobre aspectos, tais como positivo, negativo ou neutro. Enquanto a sumarização corresponde à apresentação do resumo geral baseado nos aspectos e opiniões extraídas. Outras tarefas, como a identificação de sentenças comparativas, também pode ser consideradas (BRYCHCÍN et al., 2014; PONTIKI et al., 2014; ALGHUNAIM, 2015).

Nas próximas seções, apresentamos uma discussão mais detalhada sobre as tarefas citadas na Figura 2.2 e exemplos de trabalhos.

2.2.1 Extração de aspectos e termos opinativos

O processo de identificar características citadas sobre uma entidade é chamado de extração de aspectos (KHAN et al., 2014). A principal premissa é que uma opinião sempre tem um alvo. Um alvo é muitas vezes um aspecto ou tópico comentado no texto. Isto é importante para reconhecer cada opinião expressa e seu alvo em uma sentença (*multi aspect*) (LIU, 2012). Como já citamos anteriormente (seção 2.1), uma expressão de opinião pode indicar sentimento positivo ou negativo e citar o alvo explicitamente ou implicitamente no comentário.

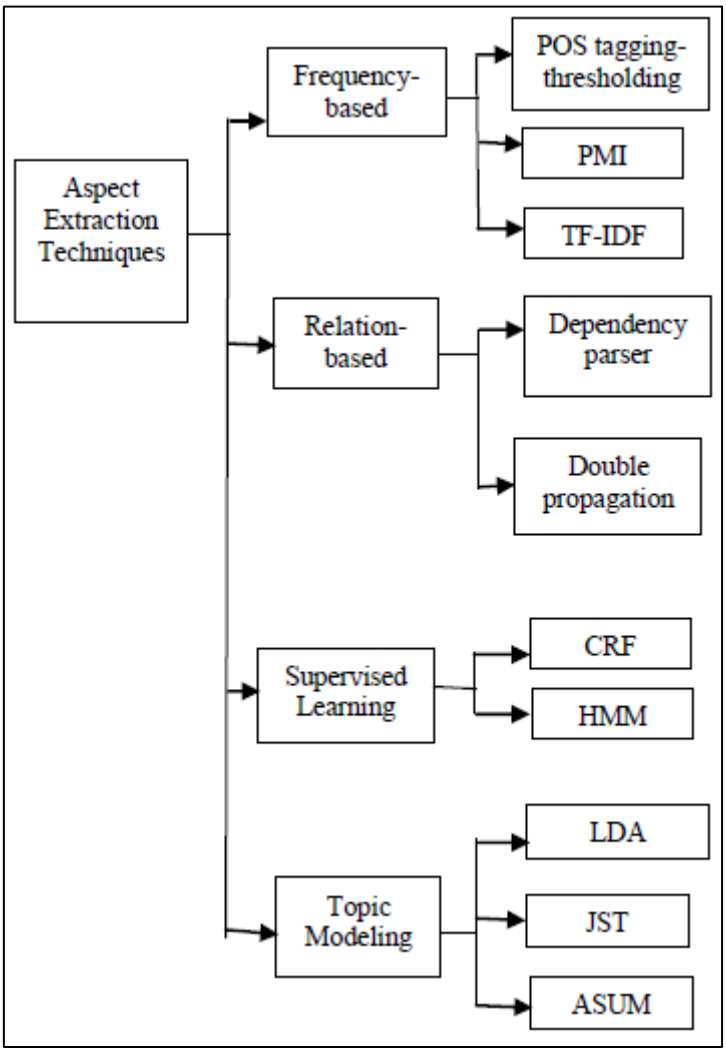


Figura 2.3: Diferentes abordagens para extração de aspecto (MORE; BORKAR, 2016)

Diversos estudos revelam que o processo de extração de aspecto poder envolver várias técnicas de processamento de linguagem natural, algoritmos supervisionados e não supervisionados, agrupamento, grafo, uso de recurso léxico, e também uso de ferramentas como geração de *tokens*, *parte-of-speech tagging*, remoção de ruído, *bag-of-word*, *n-gramas*, similaridade e sinônimos de termos, entre outros (LIU, 2012; BANCKEN et al., 2014).

De acordo com (MORE; BORKAR, 2016), as técnicas para extração de aspectos são agrupadas em quatro tipos de abordagens (Figura 2.3): baseada em *Frequência* de termos (Frequency-based), baseada em *Relações* (Relation-based), *Aprendizagem Supervisionada* (Supervised Learning) e *Modelagem de Tópicos* (Topic Modeling). Cada abordagem possui um grupo de técnicas que utilizam algoritmos específicos (MORE; BORKAR, 2016). Logo a seguir apresentamos uma breve introdução a estas abordagens.

2.2.1.1 Abordagem baseada em Frequência de termos

A abordagem baseada em Frequência (Frequency-based) envolve localizar substantivos, ou frases nominais, que são frequentes em textos como aspectos (MORE; BORKAR, 2016). Segundo MAHADIK; BHARAMBE (2015), um aspecto pode ser expresso por um substantivo, adjetivo, verbo ou advérbio. Porém, recentes pesquisas demonstraram que 60 a 70% dos aspectos correspondem a substantivos explícitos.

Este método é habilitado, em sua maioria, para extrair aspectos explícitos em um determinado domínio. A abordagem se baseia em que o usuário ao comentar sobre aspectos de uma determinada entidade, o vocabulário utilizado é semelhante e convergem (MORE; BORKAR, 2016).

POPESCU; ETZIONI (2005) desenvolveram um método para extrair aspectos analisando a relação entre aspectos e produtos. Frases nominais com alta frequência são extraídas como candidatas a aspectos. A medida de similaridade PMI é utilizada para computar uma pontuação a cada candidato a partir da sua frequência e coocorrência. Se o valor de PMI do candidato for muito baixo, este não é considerado aspecto do produto, pois tem baixa frequência.

Em KU et al. (2006), termos frequentes são extraídos utilizando a medida TF-IDF considerando termos no nível do documento e nível de sentença. No trabalho LI et al. (2015), foi proposto um método baseado em frequência para extração de aspectos. Regras são implementadas para compactar e remover redundâncias de termos. Relevantes aspectos são selecionados baseados em ordenação e na medida de similaridade PMI-IR.

Esta abordagem é considerada simples de implementar e bastante eficaz. Entretanto, tende a gerar muitos aspectos não relacionados ao domínio, e também, aspectos que apresentam baixa frequência no conjunto de dados são desconsiderados nesta técnica. (MORE; BORKAR, 2016; MAHADIK; BHARAMBE, 2015).

2.2.1.2 Abordagem baseada em Relações

A abordagem baseada em Relações Sintáticas (Relation-based) é considerada um método não supervisionado e tem como objetivo analisar a relação entre aspectos e termos opinativos para identificar aspectos. Relações gramaticais e padrões sintáticos entre palavras de aspecto e de opinião são investigados para localizar regras de extração (MORE; BORKAR, 2016). Por exemplo, a Figura 2.4 demonstra a relação de dependência sintática entre palavras para a sentença “The phone has a good screen”.

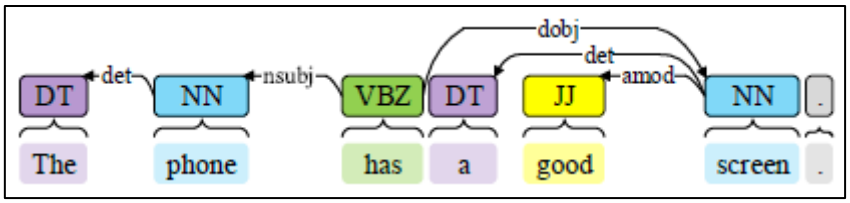


Figura 2.4: Relações de dependência em “The phone has a good screen.” (LIU; GAO et al., 2016)

O adjetivo “good” é considerado um termo de opinião e o substantivo “screen” é considerado um aspecto sobre um determinado objeto e estes termos estão relacionados pela relação *amod*. Então, dado um conjunto de termos opinativos é possível extrair um conjunto de aspectos usando relações sintáticas. Ou seja, através de aspectos é possível extrair um conjunto de termos opinativos, por exemplo, termos ligados por conjunção.

Nesta abordagem, para produzir bons resultados, as regras de extração devem ser selecionadas cuidadosamente, pois padrões sintáticos podem extrair termos irrelevantes (LIU; GAO et al., 2016).

Em ZHUANG et al. (2006), uma lista de palavras chaves referente ao domínio de filmes é construída a partir da frequência de termos no conjunto de dados e dados do *WordNet*. Em seguida, esta lista é adotada pra extrair aspectos e opiniões a partir da análise de algumas relações de dependência sintática, como <NN-noun, JJ-adjective>, <NP-noun frase, JJ-adjective>.

WU et al. (2009) propuseram um método combinando resultados de analisadores léxicos (shallow parser e lexical dependency parser) para extrair aspectos e opiniões sobre produtos. Este método foca somente em substantivos como aspectos e verbos como opiniões. Frases adjetivas são excluídas do método.

HTAY; LYNN (2013) focaram na extração de aspectos e opiniões através de uma de lista padrões de termos de ocorrência sequencial em sentenças. Por exemplo, <JJ + NN/NNS>, onde o primeiro termo ocorrido na sentença deve ser um adjetivo e o segundo termo em sequencia deve ser um substantivo. Este método possui um conjunto de regras muito limitado e, a ocorrência de aspectos e termos que não se enquadram nas regras definidas não será extraída por este método.

No trabalho de BANCKEN et al. (2014) foi desenvolvido um algoritmo chamado *Aspectator* para detectar automaticamente e avaliar aspectos de produtos em comentários de usuários. O algoritmo consiste em analisar caminhos de dependência em árvore sintática para encontrar opiniões expressas sobre aspectos candidatos. A etapa de seleção e extração de dados consiste na extração de pares de termos ou expressões chamados de *opinion pair*, onde o primeiro termo é chamado de *Sentiment modifier*, palavra em torno do aspecto que expressa uma opinião e o segundo termo chamado de *Aspect mention*, é a menção de um aspecto. Termos de negação que alteram o sentido da opinião, como *not, no, never*, entre outros, também são extraídos com este método. E, termos que alteram a intensidade da opinião, como *less, very, much*, também são extraídos. Este método não requer etiquetagem de dados e conhecimento específico de domínio.

ZHENG et al. (2014), desenvolveram uma abordagem não supervisionada baseada em análise de dependência para extrair expressões opinativas chamada de *Appraisal Expression Patterns* (AEPs) em comentários considerados como uma representação de relação sintática entre aspecto e termos de sentimento. No trabalho, o AEP é aplicado para representar a relação sintática entre aspecto e termo de sentimento usando *Shortest Dependency Path* (SDP) no qual conecta duas palavras em um grafo de dependência. Comentários sobre hotel, restaurante, leitor de MP3 e câmeras foram coletados para experiências.

No trabalho de SAMHA (2016) é realizada a extração de aspectos em nível de sentença usando técnicas de PLN. A abordagem é baseada em técnicas que envolveram pré-processamento de texto, lematização e POS *Tagging* para obter a estrutura sintática de sentenças por meio de regras de relações em árvore de dependência. É explorado um conjunto de regras sintáticas e relações que foram observadas a partir de um conjunto de dados sobre produtos.

BANCKEN et al. (2014), ZHENG et al. (2014) e SAMHA (2016) aplicaram análise de dependência sintática para extração de aspectos. No entanto em BANCKEN et al. (2014) é aplicado um número maior de regras sintáticas, o método extrai negação, termos aspectos compostos (mais de um termo) e termos que alteram a intensidade de termos opinativos.

Outro algoritmo nesta abordagem é chamado de *Double propagation*, onde relações de dependência são identificadas usando analisador de dependência e depois são expandidos usando um pequeno léxico (seed) de opinião para extrair aspecto. *Double propagation* corresponde à propagação de informação entre termos opinativos e aspectos (QIU, 2011). Este trabalho tem como limitação o processo de propagação. O processo de propagação termina quando palavras de opinião ou recursos não podem ser mais encontradas. Ou seja, a qualidade da expansão depende da qualidade do conjunto *seed* e também depende da relação de dependência entre aspectos e opiniões para propagar informações. Assim, como também havendo ruído no conjunto de documento e relações com estes termos, o método irá propagar ruído.

Comparado a métodos baseados em frequência, abordagens baseadas em relações podem localizar aspectos de baixa frequência e funcionam bem para conjunto

de dados de tamanho médio. Além, de ser um modelo não supervisionado, evitando trabalho de rotulagem de dados.

Uma limitação desta abordagem é que, na maioria dos trabalhos, aspectos são considerados apenas do tipo substantivo e termos opinativos apenas do tipo adjetivo (ATOLE, 2015; MAHADIK; BHARAMBE, 2015; MORE; BORKAR, 2016). Outra limitação, é que para grandes conjuntos este método pode extrair muitos termos que não são relevantes. Estratégicas de ranking de características para computar relevância de termos podem amenizar parte do problema de cobertura (ATOLE, 2015).

2.2.1.3 Abordagem baseada em Aprendizagem Supervisionada

Abordagens baseadas em Aprendizado Supervisionado (Supervised Learning) envolve construir um modelo, a partir de dados de treinamento, e aplicá-lo em conjuntos de dados sem rótulo. A identificação de aspectos, opinião e sua polaridade podem ser vista como um problema de rotulagem onde padrões são aprendidos a partir de dados rotulados e aplicados a dados não marcados.

Os modelos de aprendizagem mais comuns para extração de aspectos são o *Hidden Markov Models* (HMM) e *Campos Aleatórios Condicionais* (CRF) (MORE; BORKAR, 2016). Outras abordagens também aplicam algoritmos com SVM, NB, KNN, entre outros (DE CLERCQ et al., 2015; GUHA, et al., 2015; JEYAPRIYA; SELVI, 2015; PEKAR et al., 2014).

JIN et al. (2009a) e JIN et al. (2009b) desenvolveram um método para identificar aspectos e opiniões sobre produtos. Também classifica se uma sentença contém ou não contém citações de aspecto e opinião e categoriza os termos de opinião como positivo ou negativo. Conjuntos de marcadores são usados para etiquetar manualmente cada sentença representando padrões entre aspectos e opiniões. Em seguida, os dados etiquetados são treinados no algoritmo HMM. Sentenças que contêm pares de opinião são identificadas.

Segundo (LIU; GAO et al., 2016), uma limitação para HMM é que é assumida independência de cada palavra para o contexto, o que no contexto de mineração de aspecto e opinião termos são amplamente dependentes ao domínio. O autor ainda cita

que, por este problema, o HMM pode não representar adequadamente os recursos necessários para extração de aspecto e assim prejudicar o desempenho. Para resolver essa limitação, alguns pesquisadores adotam o algoritmo CRF, que é um modelo de sequência não direcionado e pode introduzir mais recursos que o HMM (LIU; GAO et al., 2016).

JAKOB; GUREVYCH (2010) aplicou o algoritmo CRF para treinar um conjunto de dados etiquetados a partir de dados de diferentes domínios para extração de aspectos independente de domínio. Termos são etiquetados quanto ao seu POS Tagging. Também são etiquetados termos que tem relação direta de dependência com uma expressão de opinião em uma sentença a partir de relações de arvore sintática. No trabalho de RUBTSOVA; KOSHELNIKOV (2015) foi utilizado CRF para treinar um conjunto etiquetado baseado em POS Tagging e lematização de termos em sentenças.

LI et al. (2010) desenvolveram um método chamado *Skip-Tree CRF*. O método tem como base o algoritmo CRF para extrair aspectos e sentimentos relacionados e, também identificar a polaridade de termos em comentários sobre produtos. Dois métodos são propostos: *Skip-chain CRF* e *Tree CRF*. O primeiro, *Skip-chain CRF*, assume que termos ou frases são conectados pela conjunção “and”, e que em suma estes termos possuem a mesma polaridade. A suposição inversa é dada para termos conectados pela conjunção “but”, onde a polaridade do termo é invertida. O segundo método, *Tree-chain CRF*, considera a estrutura de árvore sintática em sentenças para prover a relação sintática entre aspectos e sentimentos. Os dois modelos são unificados, compondo o modelo *Skip-Tree CRF*. O método assume para cada termo informação de POS Tagging e lematização. O dicionário *WordNet* é adotado para identificar relações de sinônimos e antônimos para cada palavra e o *SentiWordNet* para adquirir a polaridade dos termos. Um conjunto de comentários sobre produtos e filmes é etiquetado manualmente e treinado no *Skip-Tree CRF* proposto.

Uma das principais desvantagens do CRF, citado pelos autores MAATEN et al. (2011), é a natureza linear de termo dependente de dados, se os fatores de transição de estado fornecem uma entrada uniforme, o modelo se reduz a uma coleção de regressores de logísticas lineares simples. O algoritmo também é conhecido por ter lenta convergência durante o treinamento, logo adicionar novos dados ao conjunto de dados de treinamento é necessário retreinar todo o modelo CRF e isto

pode ser muito demorado devido à alta complexidade da fase de treinamento do algoritmo devido a muitas outras operações como escalonamento numérico, suavização, entre outros (SUTTON; MCCALLUM, 2012; WALLACH, 2002).

2.2.1.4 Abordagem baseada em Modelagem de Tópicos

A modelagem chamada de Modelagem de Tópicos (Topic Modeling) é considerada um método de aprendizagem não supervisionada para descobrir tópicos em documentos textuais considerando que documentos consistem de uma mistura de tópicos e cada tópico é a distribuição de probabilidade sobre as palavras. Ou seja, é considerado que documentos possuem uma lista de termos relacionados a um tópico e sua respectiva distribuição de frequências. A probabilidade de um documento tratar de um ou mais tópicos é vista a partir da coocorrência das frequências destes termos.

LIU; GAO et al. (2016) comentam que a modelagem baseada em tópicos muitas vezes apenas retornam tópicos irregulares em um conjunto de documentos em vez de precisos aspectos, um termo como tópico não necessariamente significa representar um aspecto. E cita como exemplo, em tópicos sobre bateria, o modelo pode localizar os termos “battery”, “life”, “day” e “time”, entre outros, que estão relacionados com a vida útil da bateria, mas nem todas as palavras individuais podem ser um aspecto. Os dois modelos frequentes empregados são o LDA e pLSA. Já para os autores MORE; BORKAR (2016), identificar tópicos através de modelos de tópicos corresponde a identificar aspectos. E a modelagem de tópicos ajuda no agrupamento de aspectos e abrange palavras de opinião e de aspecto (MORE; BORKAR, 2016, WANG, 2015; ZHENG et al., 2014; YU et al., 2013).

Latent Dirichlet Allocation (LDA) é um modelo probabilístico para coletas de dados discretos (dados separados individualmente, distintos e finitos), tais como conjuntos de texto. LDA é considerado um modelo hierárquico *Bayesiano*, no qual cada item de uma coleção é modelado como uma mistura finita sobre um conjunto subjacente de tópicos (MORE; BORKAR, 2016; BLEI et al., 2003). O modelo LDA é semelhante ao modelo pLSA (DING et al., 2015) porém, neste modelo o número de parâmetros cresce linearmente de acordo com o tamanho do conjunto, podendo causar *over-fitting.* No LDA, o número de parâmetros necessários é reduzido e também é utilizado *dirichlet*

priors (distribuição de probabilidades) evitando problema de *over-fitting* e assim proporcionando melhores resultados (ONEATA, 1999, DING et al., 2015).

Segundo CASCAVEL (2016), conforme ilustrado na Figura 2.5, para os documentos é assumido que existe um determinado número de tópicos (Figura 2.5 à esquerda). Para cada documento é gerado uma distribuição de tópicos, conforme observado no histograma (Figura 2.5 à direita). Em seguida, cada palavra é etiquetada a um tópico (ver círculos na Figura 2.5), um termo correspondente a um tópico é destacado no texto.

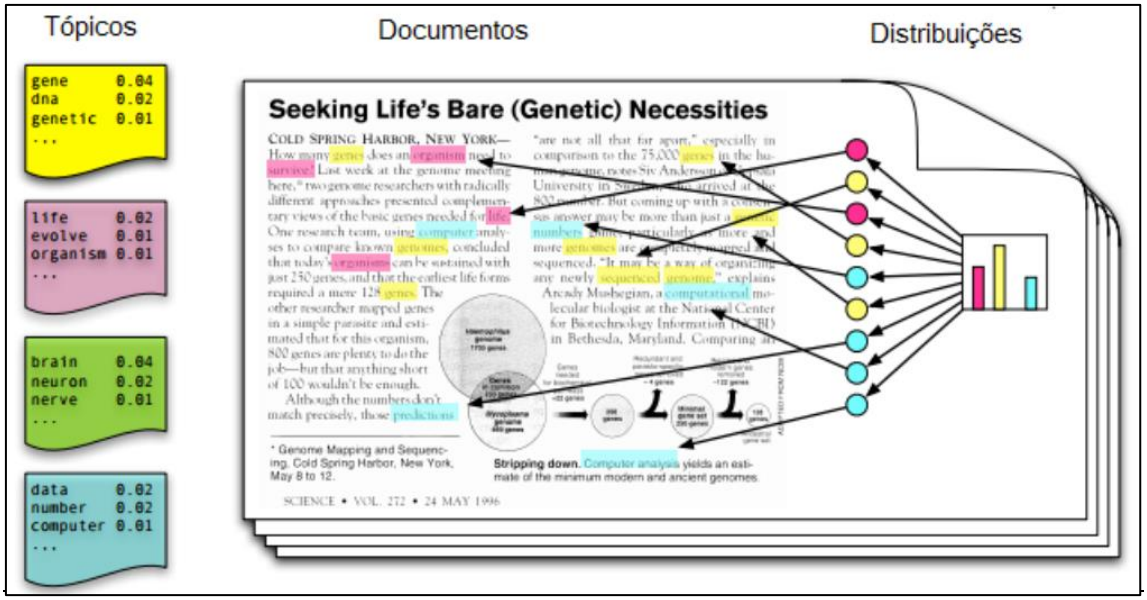


Figura 2.5: Ilustração do modelo LDA (CASCAVEL, 2016)

MEI et al. (2007) propuseram, baseado no algoritmo pLSA, o *Topic-Sentiment Mixture* (TSM). O modelo corresponde a método probabilístico para capturar a mistura entre termos tópicos e termos de sentimento simultaneamente. ZHAO et al. (2010) desenvolveram um modelo híbrido chamado de *MaxEnt-LDA* para descobrir simultaneamente aspectos e termos opinativos. A informação de POS Tagging é utilizada para discriminar termos entre aspectos e opiniões.

TITOV; MCDONALD (2008) desenvolveram um modelo estatístico para extrair aspectos e termos opinativos. O modelo é dividido em global e local. O modelo global descobre entidades enquanto o modelo local extrai aspectos usando um conjunto de

sentenças. O modelo assume que termos assinalados a um conjunto de tópicos podem representar um aspecto. Termos distintos que são relacionados são automaticamente agrupados sob o mesmo aspecto. Segundo LIU (2012), este modelo não separa aspectos de termos de opinativos.

BRODY; ELHADAD (2010) apresentam um modelo não supervisionado para extrair aspectos e determinar sentimento em textos. O trabalho foca em modelo de tópicos, baseado no algoritmo LDA, onde trabalha em nível de sentença e explora um pequeno conjunto de tópicos para automaticamente identificar aspectos. Para detecção de sentimento, o modelo deriva de um conjunto *seed* de adjetivos etiquetados como positivos ou negativos.

Segundo TANG et al. (2014), LDA tem dificuldade para localizar tópicos em conjuntos que possuem volume reduzido de documentos e também em documentos de textos curtos. Termos infrequentes são dados baixa prioridade pelo algoritmo, ocorrendo esparsidade de relações entre os termos (um texto pode não ter sentido quando analisado separadamente), levando o LDA a não construir relações entre determinados termos e conduzindo o resultado ao erro. Logo, LDA ainda enfrenta problema de *over-fitting* quando se trata de textos curtos, uma vez que vocabulário é esparso. Ainda, sobre modelos de extração de tópicos, estes são modelos probabilísticos que tem custo computacional alto (DING et al., 2015; NEVES, 2016).

De acordo com NEVES (2016), as abordagens *Topic Mapping* (Mapeamento de Tópicos) (LANCICHINETTI et al., 2014) e ADVF *Topic Modeling* (focado em redes sociais) (KIDO et al., 2016) para descoberta de tópicos baseadas em grafos tendem a minimizar a aleatoriedade dos resultados e também reduzir ruídos devido a incidência de termos coloquiais nos textos. O modelo *Topic Mapping* é descrito como um modelo baseado na modelagem em grafos de termos, onde no grafo cada vértice representa um termo e cada aresta representa a coocorrência entre termos. E em seguida, um método de *cluster* é aplicado para agrupar termos segundo suas correlações, resultando ser mais eficiente que o LDA em textos longos, mas não foi avaliado em textos curtos (NEVES, 2016).

De acordo com (DING et al., 2015), a modelagem baseada em tópicos apenas leva em consideração a relação entre termos e sua frequência e probabilidade, mas

ignora a influência que a polaridade do sentimento impõe à modelagem de tópicos. Mediante esta limitação foi desenvolvimento o modelo *Joint Sentiment Topic* (JST), este modelo integra fator de sentimento no processo de modelagem de tópicos, e assume que todas as palavras em um documento incorporam tendência de sentimento, então cada tópico tem seus atributos próprios de sentimento.

O modelo supervisionado JST é aplicado na pesquisa de LIN; HE (2009). O trabalho corresponde a um modelo probabilístico baseado no LDA, onde detecta, simultaneamente, sentimento em nível do documento e extrai tópicos de textos. Neste modelo, cada documento corresponde a uma distribuição de sentimentos. O JST tem o diferencial de incluir uma camada de sentimento adicional entre o documento e a camada de tópicos. Além, tópicos são relacionados com rótulos de sentimento e termos são associados com rótulos de sentimento e tópicos. O modelo tem a limitação de que sentimentos e tópicos não são explicitamente distinguidos e termos opinativos são predefinidos por léxicos de sentimento.

O autor DING et al. (2015) cita que um sentimento forte influenciaria a preocupação das pessoas em relação ao tema e, consequentemente, alteraria a distribuição do tópico e a distribuição das palavras. No entanto, no modelo JST a evolução da polaridade do sentimento de um tópico não é levado em consideração. E, esta limitação induz o algoritmo a não se adaptar com a variação de polaridade quando se lida com dados de mídia social uma vez que são expressos com variações afetivas diversas (DING et al., 2015).

Outro modelo desenvolvido pra tratar o problema de extração de aspecto versus termo de sentimento é o *Aspect and Sentiment Unification Model* (ASUM), que consiste, também, de um modelo probabilístico para automaticamente descobrir aspectos e diferentes sentimentos sobre o aspecto. O modelo incorpora sentimento e aspectos juntos para localizar em aspecto textos. ASUM tem como desvantagem inicializar baseado em conjuntos *seeds* onde, nestes conjuntos, termos gerais são marcados como positivos ou negativos (JO; OH, 2011). Assim, o método é dependente da qualidade do conjunto *seed* e também um mesmo conjunto *seed* por não ser adaptável para outro domínio. Ainda sobre JST e ASUM, os autores (YUAN; WU, 2017) comentam como desvantagem é que estes modelos de extração não separam as palavras de aspecto e de opinião.

Como podemos observar existem diversas abordagens para o problema de extração de aspectos e termos opinativos. Podemos notar também, que cada uma das abordagens possuem pontos positivos e limitações. Na tabela 2.1, são apresentadas as abordagens para extração de aspectos citadas acima agrupadas a partir da observação das vantagens e limitações técnicas (MORE; BORKAR, 2016).

Abordagem	**Vantagens**	**Limitações**
Baseada em Frequência	Método simples e bastante eficaz.	1) Gera muitos não-aspectos e não abrange aspectos de baixa frequência. 2) Necessita de sintonização manual do número de parâmetros, dificultando o porte para outro conjunto de dados.
Baseada em Relações	1) Não há necessidade de dados rotulados manualmente. 2) Ajuda a encontrar aspectos de baixa frequência.	Pode produzir muitos não-aspectos que correspondem ao padrão sintático pré-definido.
Baseada em Aprendizagem Supervisionada	Supera a limitação dos métodos baseado em Frequência e em Relação por meio do aprendizado de parâmetros do modelo a partir do conjunto de dados de treinamento.	1) Há necessidade de dados rotulados manualmente. 2) A precisão do modelo aprendido depende da precisão com que os dados de treinamento são rotulados para aspectos e não-aspectos.
Baseada em Modelagem de Tópicos	1) Não há necessidade de dados rotulados manualmente. 2) Executa extração de aspecto e tarefa de agrupamento simultaneamente de forma não supervisionada.	1) Requer um grande volume de dados. 2) Não distingue aspecto de termo opinativo.

Tabela 2.1: Vantagens e limitações de técnicas de extração de aspecto (adaptado) (MORE; BORKAR, 2016).

2.2.2 Classificação de aspectos

Uma coleção de documentos pode ser composta por uma coleção de sentenças e cada sentença pode conter uma coleção de aspectos. Aspectos podem ser identificados por meio de diversas abordagens conforme discutido na seção anterior. Após a extração de aspectos tarefas de classificação podem ser aplicadas. O processo de classificação de

aspectos pode ser dividido em duas categorias: classificação de sentimento (sentiment prediction) e identificação da categoria do aspecto (category detection) (Figura 2.6).

A predição de sentimento tem como objetivo classificar sentimento sobre aspectos em uma das escalas de polaridade, como positivo, negativo ou neutro, dado um conjunto de aspectos em um texto. Ou também, classificar em categorias emotivas, como, alegria, tristeza, desespero, medo, entre outros (LIU, 2012, SHARIF et al. (2014)).

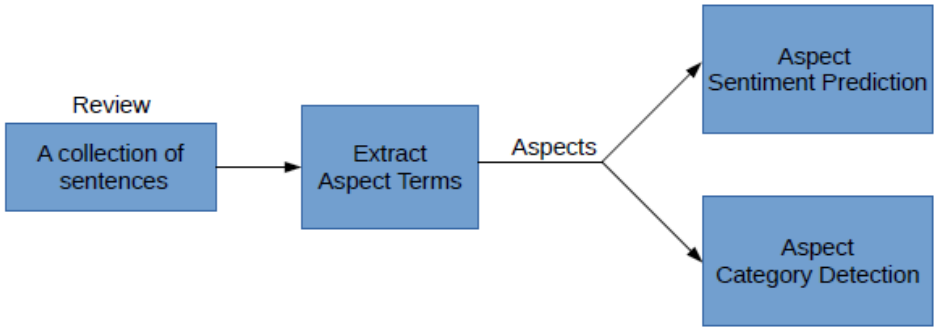


Figura 2.6: Classificação de aspectos (ALGHUNAIM, 2015)

Além da classificação de termos outras particularidades devem ser tratadas como palavras que podem mudar o sentido da orientação de sentimento, conhecidas como V*alence Shifters*. Existem vários tipos, os mais comuns são: *not, never, none, nobody, nowhere, neither,* e *cannot* (LIU, 2012; CASTAÑEDA et al., 2014; BORNEBUSCH et al. , 2014; NA et al. (2012); ZHANG et al. (2015); WANG; ESTER, 2014; ZHANG et al., 2011).

A categorização de aspectos envolve detectar qual a classe majoritária de um aspecto de acordo com o domínio. Uma categoria não necessariamente ocorre no texto (ALGHUNAIM, 2015). Por exemplo, a entidade “restaurante” pode ter diversas categorias de aspectos, como “comida”, “ambiente”, “serviço”, a sentença “O garçom é atencioso e atende rápido” contem citações opinativas sobre a categoria “serviço” do restaurante.

Para automatizar o processo de classificação de opinião e categorização de aspectos em documentos textuais, diversas abordagens têm sido adotadas. Dentre essas

abordagens, algumas englobam recursos linguísticos através de técnicas de processamento de linguagem natural ou também adotam recursos léxicos. Outras abordagens, a fim de classificar opiniões, adotam algoritmos de aprendizagem de máquina, denominadas abordagem supervisionada ou não supervisionada (LIU, 2010; LIU, 2012; MEDHAT et al., 2014; RAVI; RAVI, 2015).

HU; LIU (2004) apresentaram um método baseado em léxico para classificação em nível de aspecto, o método também pode determinar a orientação de sentenças. O método utiliza um conjunto *seed* de termos positivos e negativos e relações de sinônimos e antônimos do *WordNet*. A orientação de sentimento de uma sentença é determinada pelo soma de sentimento de todos os termos da sentença. Termos de negação também são considerados.

ZHANG et al. (2015) descrevem um método para detectar categoria de aspectos e polaridade de sentimento. Para detecção de categoria é combinado um classificador SVM com indicadores de aspecto implícitos. Para classificação de polaridade de sentimento é combinado um classificador SVM com um léxico de polaridade. Experimentos são realizados em comentários sobre laptops e restaurantes.

No trabalho de TOH; SU (2015) é modelado um problema de multi-classe de aspecto onde classificadores binários são treinados para prever as categorias de aspecto. CRF é utilizado para treinar classificadores para extração de sentimento sobre aspectos. Léxicos e recursos sintáticos são utilizados para construção de modelos.

BRUN et al. (2014) apresentaram experimentos sobre um conjunto de comentários de restaurantes para as tarefas de detecção de aspecto e categoria e detecção de polaridade de termo. Informações léxicas são utilizadas para detectar termos de domínio, e regras foram desenvolvidas para produzir dependências semânticas que associam polaridade a termos e categorias.

Nos trabalhos de PATKI et al. (2014), SARKER; GONZALEZ (2015), SAHANA; GIRISH (2015) e EGGER et al. (2016) técnicas de Mineração de Opinião são aplicadas para categorizar comentários quanto a menção de aspectos de efeito colateral. Já o trabalho de EBRAHIMI et al. (2015) utiliza Mineração de Opinião para classificar termos em uma das categorias: doença ou efeito colateral. Discutimos mais sobre estes trabalhos na seção 3.3.5.

O processo de classificação de aspectos incluem diversas abordagens e são bastante distintas entre elas. Tanto para a classificação de polaridade quanto a categorização de aspectos podem envolver técnicas supervisionadas ou não supervisionadas.

A classificação supervisionada depende da qualidade dos dados de treinamento e demanda tempo para treinar o conjunto. E, também está sujeito a excesso de treinamento. E que, além disso, o desempenho do classificador depende muito da seleção de características, e isto está relacionado com o domínio em que se está trabalhando. Outro fator importante, relacionado ao domínio de dados, é que uma vez selecionado e treinado um conjunto de treinamento pertencente a um domínio e aplicá-los em outro domínio pode não obter mesma eficácia, sendo necessário reajustá-los (LIU, 2007; LIU, 2010; LIU, 2012).

Algoritmos não supervisionados têm apresentado resultados bem sucedidos e são essenciais em situações em que não se tem um conjunto de dados com dimensão significativa para etiquetar, uma vez que algoritmos supervisionados exigem uma coleção de dados eficaz para se obter uma boa classificação. Esta abordagem também é promissora devido a haver menor dependência de domínio, problema este normalmente associado a métodos supervisionados por manter um conjunto fixo de dados etiquetados (LIU, 2007; LIU, 2010; LIU, 2012).

2.2.3 Sumarização de aspectos

Usuários geralmente usam diferentes palavras ou frases para descrever um mesmo aspecto. Após a extração de aspecto e classificação há necessidade de agrupar os resultados a fim de gerar uma visão sintetizada das informações extraídas. A tarefa de sumarização baseada em aspecto difere da sumarização tradicional, pois não sumariza agrupando comentários ou sentenças. O objetivo é obter um resumo estruturado formado por todos os aspectos citados sobre a entidade e também suas respectivas opiniões. Um exemplo é ilustrado na Figura 2.7.

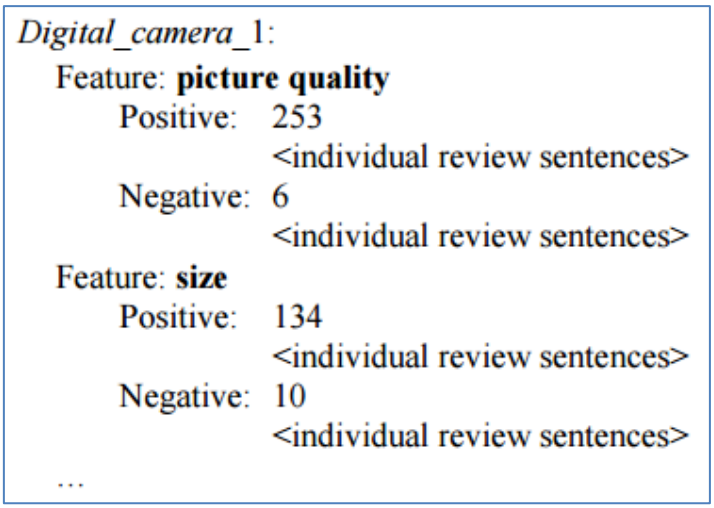


Figura 2.7: Exemplo de sumarização de aspectos (HU; LIU, 2004)

A sumarização pode ser dividida em duas categorias: agregar aspectos por categoria (aspect categorization) e agrupar aspectos quanto ao sentimento opinado (aspect rating). O agrupamento por categoria busca unir termos que podem ser representados por um aspecto, por exemplo, os termos “chicken”, “steak” e “fish” podem ser substituídos pelo aspecto “food”. É bastante comum o uso de dicionário e relações de sinônimos para tratar o problema. O agrupamento quanto ao sentimento é computado termos opinativos ou pares de opinião sobre cada aspecto quanto a sua polaridade (HU; LIU, 2004; ZHANG et al., 2015; WANG; ESTER, 2014; ZHOU et al., 2015; BANCKEN et al., 2014; THOTA; MEENAKUMARI, 2015; RANA; CHEAH, 2015).

2.3 Desafios

O maior desafio na Mineração de Opinião baseada em aspecto é detectar aspectos relevantes relacionados à entidade opinada (CHE et al., 2015). Mineração de Opinião vem sendo bastante investigado, porém existem muitos desafios a serem superados. Destacamos, por exemplo, a existência de termos coloquiais em textos que podem ocasionar resultados inconsistentes e erros de analisador sintático. Destacamos também que pessoas possuem certas percepções sobre palavras ou frases em textos

opinativos, uma vez que atuais analisadores sintáticos não podem lidar com determinadas percepções humanas (RANA; CHEAH, 2015).

Geralmente em comentários, as pessoas tendem a usar diferentes palavras para citar o mesmo aspecto, por exemplo, no domínio câmera, “picture” e “photo” refere-se ao mesmo objeto. Porém no domínio de filme estas mesmas palavras podem ter significados diferentes. Assim, como palavras que não são sinônimos pode representar mesmo aspecto, por exemplo, “appearance” e “design” no domínio de telefone (GUPTA et al., 2015).

Termos opinativos também podem variar sua polaridade de acordo com o domínio. Por exemplo, no domínio restaurante, a palavra opinativa “cheap” pode inferir polaridade positiva para o aspecto “food”, mas para o domínio de ambiente e decoração pode inferir conotação negativa (KHAN et al., 2014; TOMAR; AGARWAL, 2013).

Abordagens tradicionais na tarefa de extração apresentam resultados relevantes e são geralmente baseadas na frequência de termos, mas em corpus com baixa frequência de termos podem falhar. E também, a maioria dos recursos utilizados para extração explora recursos lexicais, sintáticos e semânticos baseada em aprendizagem supervisionada ou não supervisionada. Tais características utilizadas para um domínio muitas vezes não alcançam bom desempenho aplicado em outros domínios (KHAN et al., 2014; TOMAR; AGARWAL, 2013).

2.4 Considerações finais

Neste capítulo foram apresentados pontos importantes relacionados à Mineração de Opinião baseada em aspecto, abordando-se alguns conceitos básicos. Destacamos a importância do uso da área exemplificando com possíveis aplicações e em distintos contextos. Também vimos que diversas tarefas podem estar associadas à Mineração de Opinião baseada em aspecto, entre elas a extração, a classificação e a sumarização de aspectos e termos opinativos. Para cada tarefa foi apresentado alguns trabalhos relacionados. Outras tarefas podem ser aplicadas como a detecção de entidades e a detecção de ironia e sarcasmo (LIU, 2012). Em seguida, foram apresentados desafios a serem considerados no processo de Mineração de Opiniões.

Nesta tese, é proposta a extração de aspectos em comentários de medicamentos, para isso adotamos um método baseado em *Relações Sintáticas*. Como já discutimos acima, a abordagem baseada em *Frequência* tem como desvantagem ignorar aspectos de baixa frequência. De acordo com a base de experimentos utilizada neste trabalho (descrito na seção 5.1.), um dos conjuntos possui um número relativamente baixo de comentários, o que seria inviável trabalhar com frequência de termos em um conjunto de tamanho limitado. Assim como também, no domínio em questão termos como baixa frequência tem importância a sua extração, visto que pode revelar ocorrência de novas ADRs a um medicamento.

Em relação à abordagem de *Aprendizagem Supervisionada*, verificamos que devido às diferenças observadas entre as bases de experimentos selecionadas seria um processo muito custoso etiquetar dados para um processo de treinamento, pois um conjunto etiquetado não poderia ser reutilizado para outra base de conhecimento. E ainda, quanto à abordagem baseada em *Modelagem de Tópicos*, esta requer um grande volume de dados. Como já citamos, o conjunto de experimentos deste trabalho possui bases de comentários como tamanho limitado.

Logo, foi definida para este trabalho a aplicação de um método baseado em relações sintáticas. Foi proposto a extensão do algoritmo *Aspectator* (BANCKEN et al., 2014) para abranger particularidades do domínio médico. Com este modelo não é necessário a etiquetagem de dados, o método atendente a extração identificando distintamente aspectos e termos de sentimento. Além de também, atender a extração de termos de negação de termo opinativo e termos que apresentam variação de intensidade de termos de opinião. Na seção 4.2, discutimos com detalhes sobre o método de extração de aspecto e termo de opinião proposto nesta tese. Na seção 5.1, apresentamos detalhes do conjunto de experimentos utilizado neste trabalho. E na seção 5.2, discutido os resultados obtidos no experimento de extração.

Neste trabalho também tratamos o problema de classificação em comentários de experiências de usuários quanto ao uso de medicamento. O classificador é focado na identificação em uma das categorias de aspecto: Condição (C), Efeito Adverso (ADR), Dosagem (D) ou Eficácia (E). Um classificador construído a partir do uso de aprendizagem supervisionada foi adotado. Na seção 4.3 discutimos detalhes sobre a solução proposta.

No próximo capítulo apresentamos uma discussão sobre o estudo de Mineração de Opinião no domínio da medicina. Descrevemos aspectos relacionados à mineração de conhecimento médico e farmacêutico, em seguida apresentamos trabalhos relacionados ao tema.

3 MINERAÇÃO DE OPINIÃO EM CONTEXTO MÉDICO

Mineração de dados é uma das áreas de investigação mais motivadoras com o objetivo de encontrar informações significativas em enormes conjuntos de dados. Pesquisas para minerar dados na área de saúde têm progredido significamente nos últimos anos porque existe a necessidade de eficientes metodologias analíticas para a detecção de desconhecidas e valiosas informações em dados de saúde. Para a indústria de saúde a mineração de dados oferece vários benefícios como a detecção da fraude em seguro de saúde, soluções médicas para os pacientes à menor custo, a detecção de causas de doenças, identificação de métodos para tratamento médico, e construção de sistemas de recomendação de remédios, entre outros (PAUL et al., 2016).

Plataformas de mídia social têm visto um crescimento sem precedentes em todo o mundo. As redes sociais formam uma plataforma para as pessoas compartilharem e discutirem os seus pontos de vista e opiniões. Cada vez mais, consumidores estão se voltando para sites de saúde para procurar assistência médica e também para relatar suas experiências. Muitos têm compartilhado informações de saúde em meios de comunicação social, tais como Twitter e Facebook, ou redes sociais relacionadas à saúde, tais como *DailyStrength.org*, *Ehealthforum.com*, *MedHelp.org*, *Askapatient.com*, *Webmd.com*, *Medicinenet.com*, *Drugs.com*, entre outros.

Estas comunidades e fóruns são fontes ricas de informações sobre o feedback de usuários em relação a um determinado tratamento, a respeito de descrições de condições e sintomas sobre uma determinada doença, descrição de procedimentos médicos, estado de saúde, atendimento recebido, descrição sobre reações adversas ou eficácia sobre o uso de um determinado medicamento, entre outros (PAUL et al., 2016; DENECKE; NEJDL, 2009; SARKER et al., 2015b). Estas informações on-line podem atingir um público mais amplo e também servir como complemento aos profissionais de saúde em suas tomadas de decisões e diagnósticos e também a fabricantes de medicamentos uma vez que uma determinada reação adversa sobre um remédio, que ainda não foi relacionada oficialmente ao medicamento, esteja sendo bastante citada nestas fontes e assim, pode ser rastreada mais facilmente a partir destas informações publicamente disponíveis.

Trabalhos têm sido publicados recentemente, incluindo estudos sobre monitoramento de reações adversas causadas por medicamentos, identificação de círculos sociais com experiências comuns, monitoramento de negligência e rastreamento de propagação de doenças infecciosas (DENECKE; DENG, 2015).

Métodos de Mineração de Opinião para domínio médico não tem sido investigado intensivamente comparado a outros domínios gerais como comentários de produtos, restaurantes, hotéis, entre outros (DENECKE; DENG, 2015; DENECKE, 2015a). Estender Mineração de Opinião para documentos clínicos traz casos de uso adicionais para investigação, citamos alguns abaixo (MELZI et al., 2014):

- Monitorar ou avaliar uma mudança no estado de saúde;
- Monitorar eventos críticos, situações inesperadas ou condições médicas específicas positivas ou negativas;
- Avaliar o resultado ou a eficácia de um tratamento ou diagnóstico positivo ou negativo;
- Opiniões e experiências em direção a algum tratamento ou medicamento podem ser estudadas para evidencias clínicas;
- Avaliar relações entre sintomas;
- Julgar o impacto de uma condição médica;
- Minerar efeitos de um medicamento no tratamento de uma doença.

Entity	**Possible sentiment values**
Health status	Improve, worsen
Medical condition	Present, improve, worsen
Diagnosis	Certain, uncertain, prelimiray
Effect of a medical event	Critical, non-critical
Medical procedure	Positive or negative outcome, sucessful or unsucessful
Medication	Helpful, useless, serious adverse events

Tabela 3.1: Entidades e eventos no domínio médico e possíveis características de sentimento (DENECKE, 2015a)

Avaliações “good”, “bad”, “positive” ou “negative” no contexto da medicina sugere inferência sobre o estado de saúde, uma condição médica ou um tratamento. A Tabela 3.1 apresenta exemplos de eventos e entidades, e possíveis valores de sentimento sobre atributos para o contexto médico.

Observações e opiniões sobre tratamentos ou medicamentos expressados em narrativas clínicas ou meios de comunicação online fornecem uma nova particularidade de sentimento (para pesquisas em Mineração de Opinião) no contexto da medicina (DENECKE; DENG, 2015). Sentimento em contexto médico pode ser visto como um reflexo do estado de saúde de um paciente que pode ser bom, ruim ou normal em algum intervalo de tempo (DENECKE; DENG, 2015; DENECKE, 2015a).

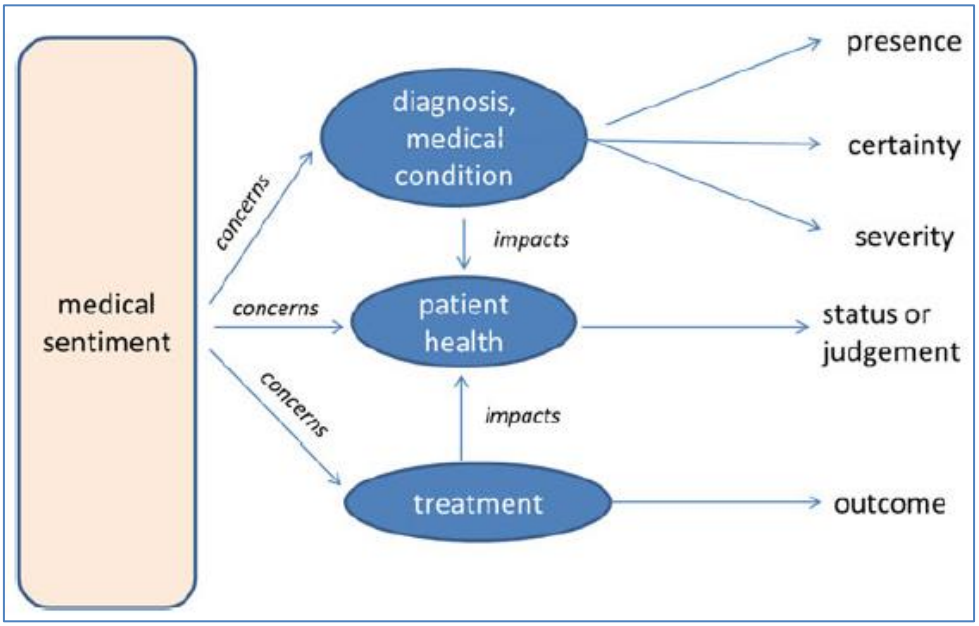


Figura 3.1: Visão geral de sentimento em contexto médico (DENECKE, 2015a)

Aspectos e seus termos opinativos impactam no estado de saúde e qualidade de vida de um paciente, a expressão “severe pain” influencia mais na vida do paciente do que “slight pain”. Descrições de status de saúde dizem respeito à citação de sintomas, exemplo, “severe pain”, “extreme weight loss”, “high blood pressure”. Resultados de tratamentos podem ser neutro, positivo ou negativo, exemplo, “surgery was successful or failed” ou através do relato de experiências ou opiniões sobre um tratamento ou um medicamento específico, por exemplo, “... tomei este medicamento e estou dormindo

muito bem". Observe a Figura 3.1 onde é apresentada uma visão geral sobre "sentimento" no contexto médico (DENECKE; DENG, 2015; DENECKE, 2015a).

Torna-se claro que o sentimento ou opinião em mídia social médica pode ser expressa de forma diferente do que sentimentos em notícias ou comentários de produtos. Um sentimento também pode ser descrito por um sintoma refletindo a situação de saúde de uma pessoa, o que não é apenas um sentimento, é caracterizada por sintomas citados por termos patológicos. Assim, extrair opiniões e intenções em narrativas médicas pode ser importante para avaliar dados clínicos, monitorar o estado de saúde de um paciente ou fornecer apoio automatizado à decisão para médicos.

Semelhante a abordagens de Mineração de Opinião em domínio geral existente, pesquisas de Mineração de Opinião no contexto médico podem ser agrupados: de acordo com a fonte textual (textos de origem online, literatura biomédica, notas clínicas, entre outros), em tarefas (análise de polaridade, classificação de resultado, sumarização, etc.), em método (baseado em regras linguísticas, baseado em aprendizagem de máquina, entre outros), em nível (nível de documento, nível de sentença, de aspecto, etc.) (DENECKE; DENG, 2015; DENECKE, 2015a).

De acordo com SARKER et al., (2015b), abordagens tradicionais de Mineração de Opinião para a classificação de polaridade geralmente não são suficientes para prever a polaridade de dados médicos sobre experiências de pacientes e cita três razões:

1) Geralmente termos médicos como "pain" e "depression" são considerados polaridade negativa em muitos recursos léxicos como *SentiWordNet* (ESULI; SEBASTIANI, 2007) e *SenticNet* (CAMBRIA et al., 2010), mas ocorrem com frequência em comentários positivos. Também comenta que verbos desempenham um papel importante na análise de polaridade, a expressão "reduced my pain" é positiva embora "pain" é um termo negativo.

2) Algumas experiências de pacientes não contêm termos opinativos.

3) E, por fim, o domínio médico é altamente técnico.

YAZDAVAR et al. (2017) discutem em seu trabalho que um grande número de pacientes para expressar uma opinião sobre a eficácia ou efeito colateral de

medicamentos usa valores quantitativos. Por exemplo, a sentença, “Using this drug, my cholesterol level went from 518 to 175”, implica sentimento positivo sobre o medicamento, pois o paciente esta informando que houve o índice de colesterol baixo, no contexto médico colesterol alto é um aspecto negativo. Ou a sentença “Taking this drug, my blood pressure rise to 18” demonstra sentimento negativo, pois a pressão arterial (blood pressure) desviou do estado normal.

Os autores YAZDAVAR et al. (2017) identificaram potenciais sentenças objetivas no qual podem expressar sentimento dividido em três categorias:

1) Sentenças que contêm dois valores de um específico termo médico, por exemplo, “NIASPAN, brought my LDL from *175 to 105*”, 175 to 105 indica a redução para LDL.

2) Sentenças que demonstram a mudança em um termo médico específico. Exemplo, “This lowered my "bad" cholesterol *25 pts* in two months”, o colesterol alterou em 25pts. Ou ainda, “First three months, HDL rose 11 points”, indica que o HDL melhorou a partir do uso do medicamento. Também pode ser expresso em porcentagem, por exemplo, “Worked great to reduce my cholesterol by about *20%* within just a few weeks”.

(3) Sentenças que contêm um termo médico específico e um tipo especial de verbo conectado, como “lowered”, “brought”, “fell”, “changed”, “decline”, “dropped”, “rise”, “shot”, “increased”, “decreased”, “down”, “up”, “improve”, “reduced”, etc.

Os autores também discutem que existem numerosas sentenças em revisões de medicamento que não expressam sentimento ou estado sobre fatos. Por exemplo, “My doctor want my cholesterol went down to 150”. Também ressalta que números podem denotar a dosagem do medicamento ou a duração de tempo que é utilizado. (YAZDAVAR et al., 2017). A Figura 3.2 apresenta diferentes tipos de sentenças em que usuários citam termos quantitativos em comentários sobre medicamentos.

	Sentence Type	Numeric Fields	Sample sentences
Potentially opinionated	From *First value* To *Second value*	Cholesterol	My cholesterol went from 279 to 230.
		LDL	LDL Went from 201 to 65.
		HDL	Increased HDL from 51 to 86.
		Triglycerides	Niaspan dropped my Triglycerides level from 311 to 175 over a one month period.
	Change by *Percent*	Cholesterol	The drug did knock my Cholesterol down by 55%, so I have motivation to try and make it work.
		LDL	Improve my LDL 30 %.
		HDL	Lpa from 33 to 5 HDL up 20%.
		Cholesterol, HDL, LDL, Triglycerides	Reduced Total 27%, Trig 40%, LDL 32% and increased HDL 17%.
	Change by *Amount*	Triglycerides	Good stuff, just 500 mg lowered my trig. 30 points in a month.
		Cholesterol	This dropped my Cholesterol by 30 pts w/in a couple of months.
		Cholesterol, HDL, LDL	My cholesterol dropped 5 points, my HDL dropped 3 points, and my LDL raised 1 point.
	Changed into *Final value*	Cholesterol	After 3 months, I lowered my Cholesterol total to 167.
		HDL	After 90 days of 500mg, HDL raised to 34, not good enough.
		LDL	By the way 10 mg dropped my LDL to 77.
Non-opinionated	Facts	Consuming time	I took at 9:30 pm and it is now 8 am
		Dosage	So I decided to cut my pills in half and go back to 25mg.
		Age	I am only 36 and my husband is a young 40
		Duration time	Drug is helping hair loss, was told it would take from 6 to 12 months to start to work.

Figura 3.2: Diferentes tipos de sentenças quantitativas em comentários de medicamentos (YAZDAVAR et al., 2017)

Também destacamos outras limitações da Mineração de Dados em contexto clínico, por exemplo, como obter dados médicos de qualidade e relevantes (PAUL et al., 2016). Dados em mídia social são gerados por usuários, e tendem a apresentar erros de ortografia e expressões não médicas para descrever questões de saúde. Por exemplo, o termo *acidente vascular cerebral* (AVC) ou *encefálico* (AVE) é geralmente citado pelo termo coloquial *derrame*.

SARKER et al. (2015a), SARKER et al. (2015b) e GOSAL (2015) também comentam que na esfera médica é comum usuários expressarem suas opiniões de forma indireta. No domínio de medicamentos, pacientes geralmente expressam suas experiências sobre eficácias de remédios e efeitos colaterais ao invés de expressar uma opinião direta usando sentimento explícito. Experiências de pacientes são frequentemente expressos sem qualquer expressão explícita de opinião.

EGGER et al. (2016) citam as principais áreas investigadas em contexto médico:

1) Minerar e recuperar informações e opiniões de saúde pessoal (JONNAGADDALA et al., 2016; NA et al., 2012; SOKOLOVA et al., 2013; XIA et al., 2009; DENECKE; NEJDL, 2009);

2) Medir a qualidade do conteúdo do documento ou de interações de saúde (CAMBRIA et al., 2012; PESTIAN et al., 2012);

3) Analisar as emoções e estudar efeitos emocionais (SOKOLOVA; BOBICEV, 2013; BIYANI et al., 2013; NIU et al., 2005);

4) Determinar resultados clínicos (SHARIF et al., 2014);

5) Detectar reações adversas sobre medicamentos (NOFERESTI; SHAMSFARD, 2015a).

As principais tarefas e desafios são (EGGER et al., 2016):

1) Modelar contexto clínico implícito e determinar o sentimento implícito;

2) Construir léxico de domínio específico;

3) Determinar a polaridade de sentimento dependente ao contexto;

4) Modelar diferentes aspectos sobre comentários de paciente.

Nos trabalhos de JI et al. (2013) e BEHERA; ELURI (2015) a classificação de sentimento em mensagens do *Twitter* é realizada para medir o grau de preocupação (Degree of Concern - DOC) de pessoas sobre surto de uma doença. *Tweets* são extraídos baseados no tempo e localização geográfica. A abordagem baseada em Recurso Léxico e o algoritmo NB são aplicados para classificar em positivo, negativo ou neutro os *tweets* minerados.

MCCOY et al. (2015) aplicaram um algoritmo de pontuação de sentimento para quantificar sentimento em um corpus de narrativas sobre internações em hospital. DENG et al. (2014) abordam a questão da viabilidade de avaliar julgamento em textos clínicos através de métodos de Mineração de Opinião. SONDHI et al. (2012) apresentaram o *SympGraph* que modela e analisa relações de sintomas em notas clínicas. O grafo é constituído de nós de sintomas e vértices são relações entre sintomas. XU et al. (2015) em seu estudo apresentam classificar a polaridade do sentimento de citações em documentos clínicos.

Nas próximas seções apresentamos discussão sobre Mineração de Opinião em revisões de medicamentos e trabalhos relacionados ao domínio médico em geral e no domínio específico sobre medicamentos.

3.1 Mineração de aspecto em revisões de medicamentos

As atividades relativas à detecção, avaliação, compreensão e prevenção de efeitos adversos relacionados a medicamentos prescritos é conhecido como farmacovigilância (pharmacovigilance ou drug safety-monitoring). Farmacovigilância inicia durante os ensaios clínicos sobre uma droga e continua depois que a mesma é liberada para consumo. Devido às várias limitações em ensaios clínicos, não é possível avaliar plenamente as consequências da utilização de um determinado medicamento antes de ser lançado (GOSAL, 2015; CHENG et al., 2014).

Um Evento Adverso (Adverse Drug Event – ADE) refere-se a qualquer dano causado por uma medicação na dosagem normal ou overdose e qualquer dano associado com o uso do medicamento como a interrupção da terapia. Reações Adversas a

Medicamentos (Adverse Drug Reactions – ADRs) são reações nocivas ou prejudiciais causados pela ingestão de medicação resultando uma intervenção relacionada à utilização do medicamento, que prevê perigo em uso futuro ou tratamento específico, alteração no regime da dosagem ou até a retirada do produto do mercado (YANG; YANG, 2015).

Um efeito colateral corresponde à resposta ou reação indesejada de usuário ao consumir um medicamento, esta reação pode ser positiva ou negativa. ADR é um tipo especial de ADE (FELDMAN et al., 2015). Interações medicamentosas (Drug-drug interactions – DDIs) correspondem a alterações dos efeitos de um fármaco devido ao uso simultâneos de um ou mais medicamentos (YANG; YANG, 2015).

Ressaltamos que os efeitos colaterais sobre uso de medicamentos podem ser tanto positivos ou negativos. No entanto, ADRs são negativas, pois podem afetar gravemente a saúde de pacientes, e até fatalmente também (CHENG et al., 2014).

Segundo FELDMAN et al. (2015), reações adversas a medicamentos compreendem a quarta principal causa de morte nos EUA. GOSAL (2015) também cita que ADRs causadas por medicamentos após sua publicação no mercado é um dos principais problemas de saúde pública com mortes e hospitalização.

Food and Drug Administration (FDA) e *World Health Organization* (WHO) são dois dos principais órgãos internacionais a realizar monitoramento de medicamentos em diversos canais para médicos, farmacêuticos, empresas farmacêuticas e pacientes. Também é responsável por reportar relatórios ADRs sobre medicamentos pós-comercialização que não foram encontrados em testes clínicos. ADRs que não foram encontrados em testes clínicos são frequentemente relatados mais tarde, às vezes, até mesmo anos depois que uma droga vem ao mercado com uma mudança de rótulo do FDA.

Estes, e outros órgãos como *MedWatch* e *Institute of Safe Medication Practices Medication Error Reporting System* (MERP) disponibilizam sistemas para reportar eventos adversos. Porém, depende da ação voluntária de usuários e também estão passíveis a excesso de informação sobre ADRs já conhecidas, dados incompletos, postagens duplicadas e informações desconexas, entre outros (GOSAL, 2015; FELDMAN et al., 2015).

Uma ADR pode não ser detectada antes de o produto ir ao mercado e pode levar tempo após sua venda para rastrear novas ADRs e relacioná-las ao rótulo do (SAMPATHKUMAR et al., 2014) remédio. FELDMAN et al. (2015) comentam que, suponha que um novo medicamento é introduzido no mercado, desenvolver métodos eficientes e com precisão para minerar dados em fóruns online relacionados à saúde pode contribuir a identificar potenciais ADRs que não foram encontrados em ensaios clínicos e prever ADRs antes da notificação pela FDA.

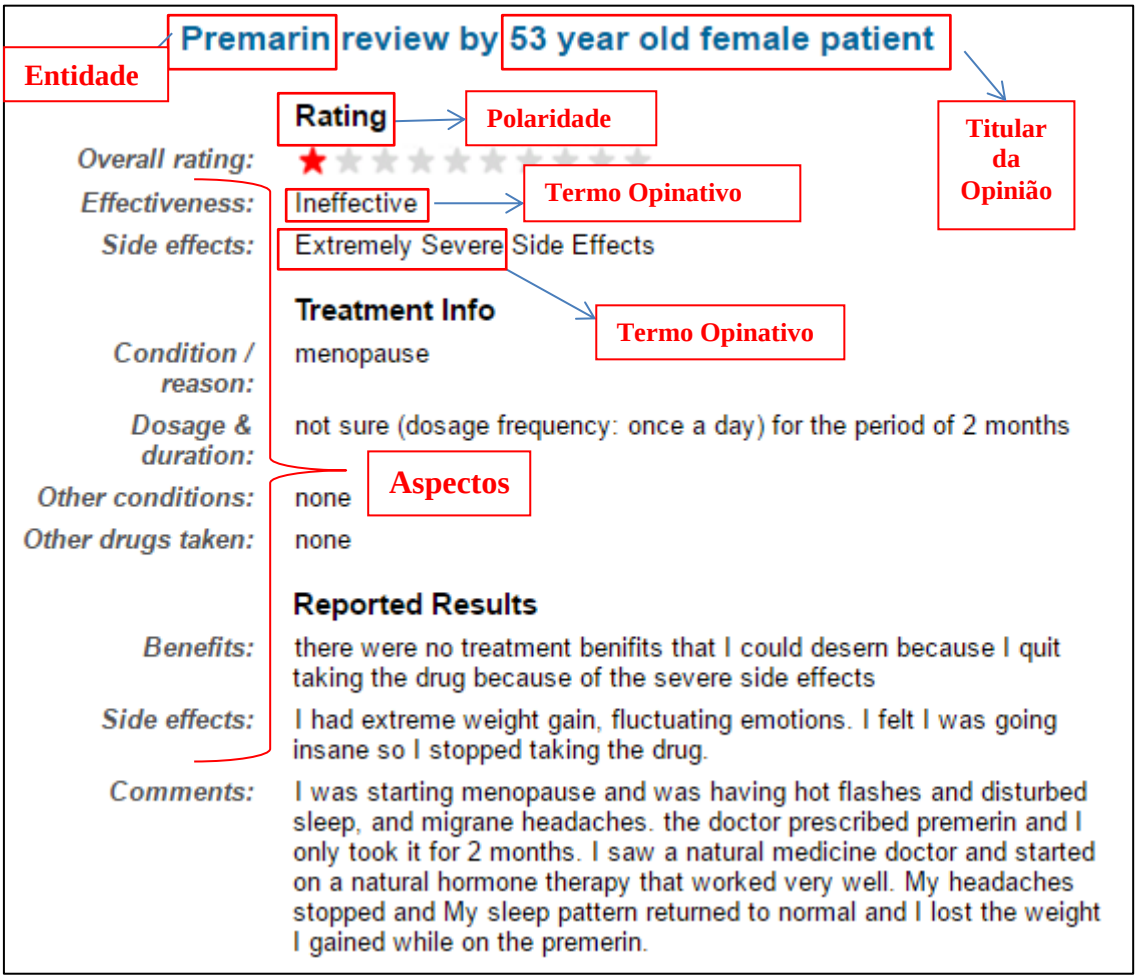


Figura 3.3: Exemplo de comentário do medicamento *Premarin* (*Conjugated Estrogens*) extraído do site *druglib.com*

A dificuldade em lidar com comentários de medicamentos implica que a descrição dada por usuários sobre experiências de eficácia e efeitos colaterais são muito distintas. A ocorrência de efeitos colaterais para cada medicamento é dependente dos sintomas ocorridos, muitos sintomas são aplicáveis a outros medicamentos, mas efeitos

colaterais são dependentes ao medicamento (NOFERESTI; SHAMSFARD, 2015a). Também ocorre em um mesmo comentário sobre um determinado medicamento o autor citar vários aspectos que não está somente relacionado a uma ADR, mas geralmente é citado informações sobre os sintomas ou condições no qual o usuário do medicamento se encontrava antes do uso da medicina (NA; KYAING, 2015).

Apresentamos as Figuras 3.3 e 3.4, que representam respectivamente comentários extraídos do site *druglib.com* e do *site drugs.com.* Consideramos o comentário apresentado na Figura 3.3 estruturado, o usuário para postar um comentário sobre um medicamento no site, preenche em campos específicos de forma que informe em cada campo a condição solicitada. Em seguida o site publica o comentário agrupando em *rating*, informações sobre o tratamento e qual o relato sobre os resultados quanto ao uso do medicamento dado pelo usuário.

Trabalhar com Mineração de dados e Mineração de Opinião sobre textos estruturados como apresentado nesta Figura 3.3 se torna mais simples uma vez que, por exemplo, um trabalho destinado a investigar novas reações adversas será filtrado dado somente no campo “Side effects”, ou investigar benefícios e resultados positivos sobre o medicamento utilizar o campo “Benefits”. Assim, dados estruturados para comentários de medicamento, como apresentamos nesse exemplo, os aspectos já estão filtrados. Apresentamos também na imagem a relação de definições relacionadas à Mineração de Opinião baseado em Aspecto apresentada no capítulo 2 com dados existentes no comentário.

A Figura 3.4, como já citamos, representa opiniões extraídas do site *drugs.com* sobre os medicamentos “Lexapro” e “Xanax”. Classificamos este comentário como não estruturado, destacamos em cada texto alguns possíveis aspectos. Observe que no texto de cada comentário o usuário cita aspectos de sintomas antes de comentar sobre o uso do medicamento, também é citado aspectos de benefícios, eficiência e dosagem quanto ao uso do remédio no mesmo texto.

Consideramos que mineração de aspectos em textos não estruturados abrange desafios mais complexos, por exemplo, como identificar e separar o que é referente a sintoma, ou ADR ou benefício, se termos semelhantes podem ser utilizados para expressar uma das situações. Observe a Figura 3.4a cita “panic attacks” como

sintoma/condição, na Figura 3.4b “panic symptoms” é utilizado para citar a eficácia do medicamento, o mesmo para o termo “anxiety” na Figura 3.4b.

O trabalho de NIKFARJAM et al. (2015) introduziu o chamado *ADRMine*, um sistema baseado em aprendizado de máquina para extração de ADRs utilizando o algoritmo *Conditional Random Fields* (CRFs). Foram coletados posts de usuários sobre medicamentos de duas redes sociais diferentes, a *DailyStrength.com* e *Twitter.com*. Uma equipe de dois anotadores especialistas etiquetaram cada post sob a supervisão de um profissional farmacologista.

KORKONTZELOS et al. (2016) adicionaram ao método o *ADRMine* (NIKFARJAM, A. et al., 2015) ferramentas de Mineração de Opinião a fim de aumentar a performance do classificador.

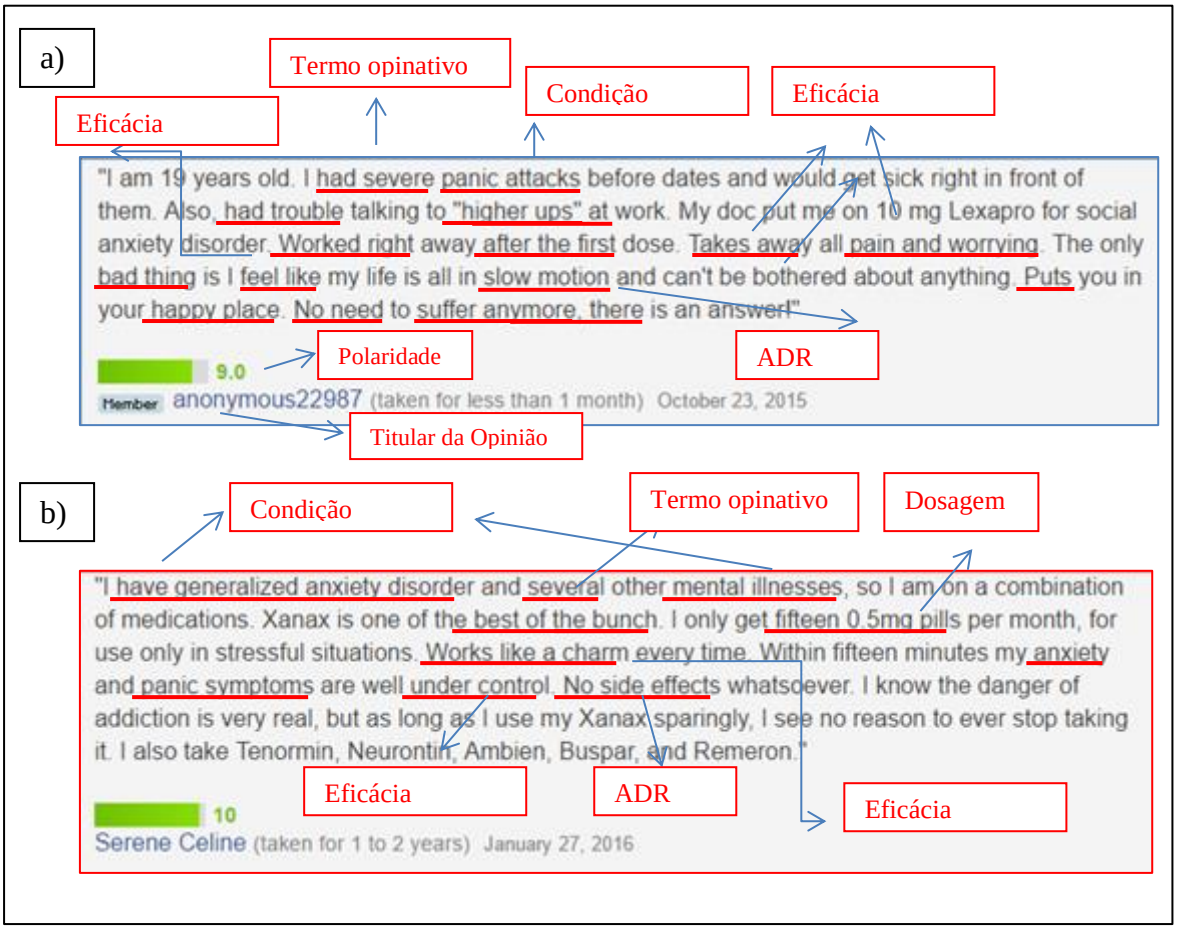


Figura 3.4: Exemplos de comentários dos medicamentos *Lexapro* e *Xanax* extraído do site *drug.com*

A rotulagem dos dados identificou que comentários sobre medicamentos podem apresentar diferentes tipos semânticos ou aspectos listados abaixo (NIKFARJAM et al., 2015):

- **Reação adversa ao medicamento (Adverse drug reaction):** uma reação nociva e não intencional que é considerada **negativa**;
- **(Efeito benéfico) Beneficial effect:** uma inesperada reação **positiva** da medicação;
- **Indicação (Indication):** a razão por qual o paciente está tomando o medicamento;
- **(Other):** qualquer outra menção de ADR ou sintomas.

NA et al. (2012) e NA; KYAING (2015) comentam em seus textos que cada sentença em um texto pode conter múltiplas cláusulas discutindo sobre múltiplos aspectos, e destaca que no contexto de opiniões sobre medicamentos existem seis tipos de aspectos mais comuns:

- **Opinião Geral (Overall):** Corresponde a opinião geral de um medicamento ou quando na cláusula não é citado nenhuma das outras cinco categorias de aspectos;
- **Eficácia (Effectiveness):** Corresponde às mudanças notadas após o uso do medicamento e está diretamente ligada a condição ou doença do paciente;
- **Efeitos colaterais (ADR):** São todas as reações que não estão relacionadas ao medicamento;
- **Dosagem (Dosage):** Relata a quantidade, a frequência ou o período de tratamento no qual o medicamento foi usado;
- **Condição (Condition):** Corresponde a uma descrição sobre a condição do paciente, pode ser uma doença ou problemas de saúde em geral;
- **Custo (Cost):** Relata citações sobre o preço do medicamento;

Na tabela 3.2 são ilustrados exemplos de cada aspecto e de frases ou cláusulas relacionadas a cada aspecto.

Tipo de Aspecto	Exemplos de sentenças
Opinião Geral	-It is great. -I've been pleased with Actoplus Met overall.
Eficácia	-But it didn't work well enough for me.
Efeitos colaterais	-I developed a serious rare stomach disorder from it.
Dosagem	-I take 50 mg as needed.
Condição	-I had some sinus surgery. -I have a history of respiratory problems.
Custo	- It is a fairly expensive medication.

Tabela 3.2: Tipos de aspectos em comentários de medicamentos (adaptado) (NA; KYAING, 2015)

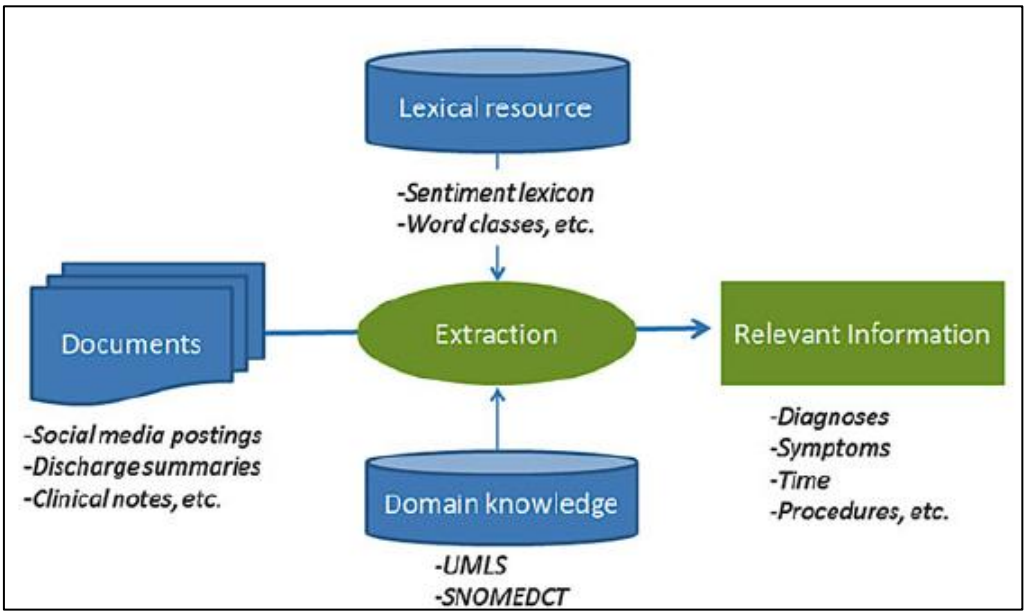


Figura 3.5: Exemplo de processo de extração de informação e recursos de conhecimento de domínio (DENECKE, 2015b)

Podemos observar na Tabela 3.2 que geralmente cada cláusula contém poucos termos subjetivos, como é de costume em sentenças de opinião em outros domínios, e há bastante incidência de termos dependente ao domínio.

3.2 Ferramentas e recursos de conhecimento do domínio

Identificação de informação especializada em um domínio requer recursos léxicos que proporcionam o conhecimento sobre o domínio e termos relacionados. No

domínio médico e farmacêutico vários glossários padronizados e ontologias estão disponíveis. Este conhecimento é explorado pelas ferramentas de extração para identificar frases ou trechos de texto relevantes para tarefa de extração (ver Figura 3.5) e classificação.

Citamos abaixo glossários e ferramentas disponíveis na comunidade científica para explorar informação médica e farmacêutica em sistemas de extração e classificação de conhecimento relacionado ao domínio:

- **Consumer Health Vocabulary (CHV):** Conecta palavras informais e frases comuns sobre saúde aos termos técnicos utilizados pelos profissionais de saúde. Inclui jargões, gírias, termos ambíguos, e palavras com erros ortográficos utilizados pelos consumidores e profissionais de saúde (http://consumerhealthvocab.org/).

- **Unified Medical Language System (UMLS):** É um conjunto de arquivos e software que reúne vocabulários de saúde, e biomédicos. O UMLS consiste nestes componentes principais: *Metathesaurus, Semantic Network* e *Specialist Lexicon* (https://www.nlm.nih.gov/research/umls/).

- **MetaMap**: MetaMap é uma API Java configurável desenvolvido para mapear texto biomédico do UMLS para minerar conceitos em textos (https://metamap.nlm.nih.gov/).

- **MedTagger**: É uma suíte de programas clínico mantido pelo OHNLP Consortium (http://ohnlp.org/index.php/MedTagger).

- **Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART):** Vocabulário criado pela FDA que é utilizado para codificação, arquivamento e recuperação de ADR de medicamentos pós-comercialização e relatos de experiências biológicas.

- **Medical Dictionary for Regulatory Activities (MedDRA):** O COSTART foi substituído pelo MedDRA (http://www.meddra.org/).

- **SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms):** É um vocabulário padronizado e multilingue de terminologias clínicas (http://www.snomed.org/snomed-ct).

- **Side Effect Resource (SIDER):** Contém informações sobre medicamentos comercializados e as suas reações adversas registradas (http://sideeffects.embl.de/).

- **DrugBank:** Um recurso de bioinformática/quimioinformática que combina informação detalhada entre remédios referente a composto químico versus proteína (http://www.drugbank.ca/).

- **MedEx-UIMA:** É uma implementação em Java do MedEx, um sistema de extração de informações relacionado a medicamentos (https://sbmi.uth.edu/ccb/resources/medex.htm).

- **Medical Literature Analysis and Retrieval System Online b(MEDLINE):** Sistema online de busca e análise de literatura médica, base de dados bibliográficos da Biblioteca Nacional de Medicina dos Estados Unidos da América (https://structuredabstracts.nlm.nih.gov/).

- **MedPost/SKR Part of Speech Tagger:** É uma implementação em Java para identificação de POS Tagging em texto biomédico (https://metamap.nlm.nih.gov/MedPostSKRTagger.shtml).

- **GENIA Tagger:** Fornece ferramenta de identificação de POS Tagging, shallow parsing e NER para texto biomédico (http://www.nactem.ac.uk/tsujii/GENIA/tagger/).

- **SPECIALIST dTagger:** Fornece ferramenta em linguagem Java de identificação de POS Tagging em texto biomédico (https://lsg3.nlm.nih.gov/LexSysGroup/Projects/dTagger/current/index.html).

- **Apache cTAKES:** Sistema de processamento de linguagem natural para extração de informações em texto livre de prontuários clínicos eletrônicos (http://ctakes.apache.org/).

- **A Biomedical Named Entity Recognizer (ABNER):** Ferramenta para análise de texto de biologia molecular, como DNA, proteína, tipo de célula, entre outros (http://pages.cs.wisc.edu/~bsettles/abner/).

- **ADRMine:** Provê ferramenta de *Named Entity Recognition* para extração de ADR (https://github.com/azinik/ADRMine) e outras fontes de dados construídas no trabalho (NIKFARJAM et al., 2015) como *ADR Lexicon*, *Drug names*, *Word clusters*, entre outros (http://diego.asu.edu/).

- **Dicionário médico:** Provê a definição de mais 16.000 termos relacionados a situações de saúde. (http://alexabe.pbworks.com/f/Dictionary+of+Medical+Terms+4th+Ed.-+(Malestrom).pdf).

- **ADR Review Dataset:** Fornece um conjunto de dados etiquetado de ADRs (http://ir.cs.georgetown.edu/data/adr/).

- **Recursos online relacionados ao domínio para extração de comentários de usuários:**

    - askapatient.com,
    - cancercompass.com,
    - csn.cancer.org,
    - dailystrength.org,
    - druglib.com,
    - drugs.com,
    - ehealthforum.com,
    - forum.internationaldrugmart.com/drug-information-f5/,
    - forums.studentdoctor.net,
    - healthboards.com,
    - komen.org,
    - medHelp.org,
    - medications.com,
    - medicinenet.com,
    - mediGuard.org,
    - medschat.com,
    - patient.info,
    - patientslikeme.com,
    - pharmacyreviewer.com/forum/,
    - revolutionhealth.com,
    - steadyHealth.com,
    - treato.com,
    - ebmd.com,

- patientopinion.org.uk,
- druginformer.com.

3.3 Trabalhos relacionados

Aplicações de Mineração de Opinião no domínio da medicina se concentram em extrair dados sobre problemas específicos e condições médicas específicas. Outros trabalhos têm desenvolvido ferramentas de uso geral para mineração de opiniões relacionadas com a saúde. A maioria das pesquisas considera dados da web, como blogs médicos ou fóruns para o propósito de mineração ou estudar opiniões de pacientes ou medir a qualidade.

Na seção 2.2 discutimos sobre as principais tarefas desenvolvidas em Mineração de Opinião agrupadas por diversas abordagens e algoritmos. Destacamos as tarefas de extração, classificação e sumarização de aspectos e termos opinativos.

No domínio médico detectamos diversos trabalhos que aplicam técnicas de Mineração de Opinião. Muitos dos trabalhos focaram na detecção de termos opinativos e classificação da polaridade destes termos. Outros trabalhos focaram na classificação de categoria, e se dedicaram principalmente da detecção de menção de ADRs em textos com citações sobre medicamentos. Muitos autores levantaram a hipótese que comentários associados à ADRs geralmente apresentam sentimento negativo, logo minerar termos negativos poderia melhorar a precisão de detecção de citações de ADRs.

Os trabalhos relacionados à extração de aspectos identificados se concentram na extração do tipo aspecto ADR. Localizamos um trabalho que se concentrou na extração de aspectos implícitos para detecção de ADRs e discriminação entre tipos de aspectos ADRs e sintomas. Outro trabalho localizado tratou a extração de aspectos do tipo “idade” e “gênero”.

Também foram identificados alguns estudos que se dedicaram a construção de recursos léxicos de sentimento para o domínio da medicina. A maioria dos trabalhos estenderam léxicos de domínio geral atribuindo adaptações para o domínio em questão.

Nas próximas seções apresentamos os trabalhos localizados durante o desenvolvimento desta tese que aplicaram técnicas de Mineração de Opinião para o domínio médico. Seguindo as tarefas discutidas na seção 2.2 agrupamos os trabalhos da

seguinte forma: (1) Extração de aspectos. (2) Classificação de sentimento (predição de polaridade). Agrupamos em duas abordagens: estudos que aplicaram técnicas de aprendizagem supervisionada e estudos que aplicaram técnicas de aprendizagem não supervisionada. (3) Classificação de tipos de aspectos (categorização). E por fim, (4) Construção de léxico de sentimento.

Os estudos apresentados não se restringem ao contexto de revisões de medicamentos. Apresentamos também pesquisas que realizaram experimentos em dados de relatórios clínicos, mensagens *tweets*, entre outros.

3.3.1 Extração de aspecto

A tarefa de extração de aspectos pode envolver técnicas de linguagem natural que consideram informações contextuais, métodos estatísticos que utilizam medidas como a frequência de dados, entre outros, ou adotar o uso de recursos léxicos ou aprendizagem de máquina.

A seguir citamos alguns trabalhos. Nesta pesquisa consideramos trabalhos que focam na extração de citações de ADRs, condição clínica, entre outros, como trabalhos relacionados à mineração de aspectos.

Os trabalhos relacionados à extração de aspectos diferem bastante entre si quanto às técnicas e algoritmos adotados. Em relação às tarefas, os estudos de EBRAHIMI et al. (2015), KORKONTZELOS et al. (2016) e MUKHERJEE et al. (2014) se concentram na extração do aspecto “ADR”. O trabalho de MOON; BORKAR (2016) focou na extração dos tipos de aspecto “idade” e “gênero”.

Outro estudo identificado é o de CHENG et al. (2014) que desenvolveu método para extração de aspectos em comentários de medicamentos. O autor cita que neste tipo de comentário é possível encontrar os tipos de aspectos: “price”, “ease of use”, “dosages”, “effectiveness”, “side effects”, “gender” e “people’s experiences”. Os autores focaram nos aspectos de satisfação e gênero feminino.

EBRAHIMI et al. (2015) tratou o problema de extração de aspectos implícitos. Enquanto todos os demais estudos citados acima trabalham com extração de aspectos explícitos.

Com relação às técnicas empregadas EBRAHIMI et al. (2015) adotaram uma abordagem baseada em léxico que usa a combinação de algumas expressões regulares simples e regras semânticas. Em KORKONTZELOS et al. (2016), foi adotado o algoritmo CRF. CHENG et al. (2014) adotaram um modelo probabilístico que segundo os autores pode ser considerado como uma abordagem de modelo de tópicos. No estudo MOON; BORKAR (2016) foi aplicado um model baseado em frequência de termos. E, MUKHERJEE et al. (2014) desenvolveram um modelo de grafo probabilístico baseado no algoritmo MRF.

A seguir descrevemos com mais detalhes as abordagens destacadas acima para extração de aspectos em comentários relacionados à medicina. Na Tabela 3.3 apresentamos um resumo dos trabalhos identificados por tarefa, recursos externos, e ferramentas adotadas, método e dados de experimento.

EBRAHIMI et al. (2015) proporam um modelo de extração de opinião implícita em comentários de medicamentos. ADRs são considerados citações de opinão implícita. A técnica de extração de efeitos colaterais implícitos é incorporada a um sistema de Mineração de Opinião para classificar termos em uma das categorias: doença (disease-Manifestation Related Symptom - MRS) ou efeito colateral (drug-Adverse Drug Event – ADE). Para extração técnicas de PLN são adotadas e um classificador SVM é utilizado para classificar efeito colateral e sintomas da doença.

KORKONTZELOS et al. (2016) formularam a hipótese que, em postagens on-line, citações de ADR são associadas com sentimento negativo. No estudo é avaliada a contribuição de recursos de Mineração de Opinião para a extração de menções de ADR e distingui-los de menções de citações indicação de sintomas. O método estende o *ADRMine* (NIKFARJAM, A. et al., 2015), um sistema de extração baseado em aprendizado de máquina que utiliza o algoritmo CRF. Ferramentas de Mineração de Opinião são adicionadas ao *ADRMine* a fim de aumentar a performance do classificador. O recurso UMLS é adotado para marcação de termos médicos. Um conjunto de ferramentas linguísticas é utilizado para extrair características das bases. A

análise de polaridade é baseada em cinco léxicos de sentimento. O estudo demonstra que adicionar funcionalidades de Mineração de Opinião pode melhorar o desempenho de métodos de extração de ADRs.

Em CHENG et al. (2014) foi proposto o algoritmo chamado de *Probabilistic Aspect Mining Model* (PAMM), um método para minerar significantes tópicos/aspectos em revisões de medicamentos. PAMM foca em localizar aspectos referindo-se a apenas uma classe, em vez de localizar aspectos para todas as classes simultaneamente em cada execução, isso ajuda a reduzir a possibilidade de ter aspectos formados a partir da mistura de conceitos de diferentes classes. Unigramas e bigramas foram formados, termos que apresentaram frequência inferior a cinco foram desconsideradas. Comentários relacionados à *score* de 1 a 2 foram receberam etiqueta de "dissatisfaction" e score de 4 a 5 receberam a etiqueta "satisfaction". Matrizes de dados de comentários foram formadas com o uso de *bag-of-words e* TF-IDF. O método é comparado a diversas abordagens de modelagem de tópicos como sLDA, LDA, entre outros. A performance é avaliada usando o método PMI e classificação SVM.

O trabalho de MOON; BORKAR (2016) focam em três tarefas. Extrair aspectos do tipo idade e gênero, e extrair palavras de opinião (adjetivo) em comentários sobre medicamentos do website *webMd.com*. Em seguida, a orientação semântica é identificada para definir se uma sentença tem orientação positiva ou negativa. Por último, os dados são sumarizados. Os sinônimos e antônimos da lista de adjetivos do dicionário *WordNet* (MILLER, 1995) é usado para análise sentimental. A classificação final é computada manualmente pela soma da polaridade dos termos. O sistema retorna a sumarização dos dados agrupados por aspectos/medicamento.

MUKHERJEE et al. (2014) focaram a tarefa de extração de raros e desconhecidos efeitos adversos de medicamentos. Um modelo baseado em grafo probabilístico é aplicado a partir do algoritmo *Markov Random Field* (MRF), um modelo linguístico e características de usuários extraídas da comunidade online *helthboards.com* a fim de capturar importantes interações entre postagem e usuários. O léxico *WordNet-Affect* (STRAPPARAVA et al., 2004) é adotado para mapear características afetivas em cada postagem. O modelo supera métodos de *baseline* baseados em frequência e no classificador supervisionado SVM. O autor cita que o método se baseia em um processo automatizado de extração de informações

relativamente simples para identificar efeitos colaterais candidatos, e é bastante propenso a erros.

Extração					
Artigo	**Tarefa**	**Recursos externos**	**Ferramentas**	**Método**	**Dados de Experimento**
EBRAHIMI et al. (2015)	Extração de opiniões implícitas e discriminação entre ADRs e sintomas.	MetaMap	Token, POS Tagging	Baseado em regras e SVM	Comentários de medicamentos (drugRatingz.com)
KORKONTZELOS et al. (2016)	Extração de aspecto: ADRs.	UMLS, ADR Lexicon, H&L, SL, NRC, NRC# e S140	POS Tagging, lematização, negação, clusterização, n-gramas	CRF	Fórum (dailystrength.org) e Tweet
CHENG et al. (2014)	Extração tópicos para identificar aspectos: satisfação e gênero.	WordNet	Lematização, stop words, n-gramas, TFI-DF	Probabilístico	Comentários de medicamentos (webmd.com)
MOON; BORKAR (2016)	Extração de aspecto: idade e gênero.	WordNet	Relações de sinônimos e antônimos	Frequência de termos	Comentários de medicamentos (webmd.com)
MUKHERJEE et al. (2014)	Extração de aspecto: ADRs raras ou desconhecidas.	WordNet-Affect	Frequência de termos	Grafo probabilístico baseado em MRF	Fórum (healthboards.com)

Tabela 3.3: Estudos que empregam tarefas de extração de aspectos em domínio médico

Podemos observar poucos trabalhos focaram na tarefa de extração de aspecto com emprego de técnicas de Mineração de Opinião. Sobre os trabalhos citados, a maioria focou apenas na extração de tipo de aspecto ADR. E, quanto às técnicas empregadas são bastante distintos.

Conforme já discutimos na seção 2.2, trabalhos que usam abordagem de frequência de termos, como MOON; BORKAR (2016), ignora termos de baixa frequência, e no contexto médico estes termos podem ser relevantes extrair. CHENG et al. (2014) adotaram um método probabilístico, porém é bastante complexo de implementar. KORKONTZELOS et al. (2016) adotou CRF que possui lenta convergência durante o processo de treinamento. Além disso, todos os trabalhos não apresentam um método eficiente de extração de termos aspecto e seu termo opinativo relacionado.

Mediante estas observações, nesta tese é proposto um método eficiente, de fácil implementação e de baixo custo de processamento para extração de aspectos e termos

opinativos em comentários de medicamentos. O método é baseado na extensão do algoritmo *Aspectator* (BANCKEN et al., 2014). Com este modelo não é necessário a etiquetagem de dados, o método atendente a extração identificando distintamente aspectos e termos de sentimento. Além de também, o método atender a extração de termos de negação de termo opinativo e termos que apresentam variação de intensidade de termos de opinião. Na seção 4.2, apresentamos com detalhes a solução proposta.

3.3.2 Classificação

Identificar a polaridade de sentimento em termos pode parecer trivial para a mente humana. No entanto, automatizar este processo pode não ser tão simples. Assim como também categorizar termos automaticamente pode ser bastante complexo, uma vez que um mesmo termo pode pertencer a diversas categorias.

Trabalhos envolvendo a tarefa de classificação em geral abordam recursos linguísticos e estatísticos aplicando técnicas de processamento de linguagem natural (PLN) ou técnicas de aprendizado de máquina. Alguns métodos de classificação de polaridade se baseiam em *score*, *g*eralmente esta técnica classifica textos opinativos baseado na soma do total de características extraídas determinadas como positivas ou negativas. Ou adotam recursos léxicos de sentimento existentes.

Quanto à categorização de termos muitos trabalhos se baseiam na identificação de diversas características semânticas e domínio a fim de localizar uma categoria mais próxima ao atributo em questão.

No domínio médico identificamos diversos estudos que se concentraram na classificação de polaridade, negativa, positiva, ou neutra, de termos. Ou também, na classificação baseada em emoção, como “tristeza”, “alegria”, entre outros. Para esta tarefa a maioria dos trabalhos desenvolveram pesquisas adotando aprendizagem supervisionada. Os algoritmos empregados mais comuns são o SVM e o NB.

3.3.3 Classificação de polaridade baseada em abordagem supervisionada

Quanto a esta tarefa, identificamos estudos que focaram na classificação de termos ou comentários de acordo com o sentimento expresso pelo usuário, geralmente positivo, negativo ou neutro. A classificação não se restringe a comentários de medicamentos, onde também incluem comentários sobre hospitais, documentos clínicos, mensagens *tweets*, entre outros.

Identificamos os seguintes trabalhos que se concentram na classificação de polaridade, categorizando em positivo, negativo ou neutro: ALI et al. (2013), BEHERA; ELURI (2015), DANIULAITYTE et al. (2016), DENECKE; NEJDL (2009), GREAVES et al. (2013), ISAH et al. (2014), JI et al. (2013), MANEK et al. (2015), MELZI et al. (2014), MERIS et al. (2016), MONDAL et al. (2016b), NIU et al. (2005), NIU et al. (2006),SARKER et al. (2011), SHARIF et al. (2014), SWAMINATHAN et al. (2010), XIA et al. (2009), XU et al. (2015).

Estes trabalhos se diferenciam no tipo de algoritmo supervisionado selecionado para treinamento. O algoritmo SVM é um dos mais utilizados nos estudos. Outra característica que difere estes trabalhos são os tipos de ferramentas adotadas para seleção e etiquetagem de características. As ferramentas mais utilizadas são POS Tagging, lematização, e *stemming* e n-grams.

Uma característica comum entre estes trabalhos é que para atribuição de polaridade a maioria adota algum léxico de sentimento. Ou seja, nenhum dos trabalhos desenvolveu alguma técnica de Mineração de Opinião para atribuição de polaridade. O léxico *SentiWordNet* é o mais utilizado entre as abordagens. Na Tabela 3.4 apresentamos um resumo sobre os trabalhos citados acima.

ALI et al. (2013), DENECKE; NEJDL (2009), DENG et al. (2014), NIU et al. (2005), e SOKOLOVA; BOBICEV (2013) aplicam Mineração de Opinião em mensagens postadas em fóruns médicos para identificar que tipo de sentimento pode ser expresso em documentos de domínio médico.

Além da classificação baseada em polaridade positiva, negativa ou neutra, localizamos alguns trabalhos que também executam tarefas para classificar emoção, (exemplo: medo, alegria, tristeza) em comentários médicos.

DENECKE; NEJDL (2009) classifica weblogs médicos em informativos ou afetivos. MELZI et al. (2014) extrai citações de emoção (alegria, raiva, surpresa). PESTIAN et al. (2012) classifica emoções encontradas em notas de suicídio. SHARIF et al. (2014) extrai citações afetivas para localizar indicadores de ADRs a medicamentos. SOKOLOVA; BOBICEV (2013) categoriza posts em: encorajamento, gratidão, confusão, fatos e fatos + sentimentos.

Ainda sobre classificação supervisionada localizamos os trabalhos de BEHERA; ELURI (2015) e JI et al. (2013), que além da classificação de polaridade, medem o grau de preocupação público sobre doenças a partir de citações de doenças, tempo e localização geográfica em mensagens *tweets*.

A seguir descrevemos com mais detalhes alguns dos trabalhos citados acima. A Tabela 3.4 apresenta um resumo dos trabalhos, identificando os recursos externos, nível de classificação (A - aspecto, S – sentença ou D- documento), ferramentas adotadas, método e dados de experimento utilizados em cada estudo.

SHARIF et al. (2014) consideraram que a identificação de sentimento negativo relacionado a um medicamento pode potencialmente colaborar para a detecção de reações adversas a medicamentos. Para isso foi proposto um framework de classificação de sentimento chamado *Feature Representational Richness Framework* (FRRF) para detecção de ADRs. O framework é baseado nas seguintes características: *n-grams*, semântica (POS *Tagging*, sinônimos e hiperônimos), termos relacionados à emoção, sentimento (atribuição de polaridade a *n-grams*), emoção (atribuição de categorias afetivas: “despair”, “sadness”, “fear”, “joy”, “liking” e “affection”) e características específicas de domínio (medicamento, ADR e anatomia). Para rotular termos a suas correspondentes tags de sentimento foram utilizados os léxicos *WordNet* (MILLER, 1995), *SentiWordNet* (ESULI; SEBASTIANI, 2007) e *WordNet-Affect* (STRAPPARAVA et al., 2004). O conjunto de características é treinado em classificador supervisionado SVM.

MISHRA et al. (2015) concentram em analisar comentários sobre medicamentos oncológicos onde é desenvolvido um sistema para mapear intervenções de contra indicações e sintomas. Recursos como *MedDRA* e *SIDER* são adotados para identificar termos médicos. Um léxico específico também é construído para mapear termos de sintomas, entre outros, relacionados a medicamentos oncológicos. Os léxicos *WordNet* e *SentiWorNet* são

utilizados para extrair características dos comentários. A polaridade global da frase é calculada tomando a soma das polaridades individuais das palavras. Um classificador SVM é utilizado para classificar o sentimento sobre os fármacos. A classificação em nível de sentença obteve 59.72% de precisão.

MERIS et al. (2016) proporam um sistema para categorizar fármacos baseado na análise de polaridade positiva ou negativa de mensagens do T*witter*. Os *tweets* classificados são analisados com base em polaridade das palavras como "bom", "mau", "não", entre outros. Os dados são processados usando classificador SVM.

Em MANEK et al. (2015), um algoritmo é proposto, onde revisões de medicamentos são coletadas a partir da web e classificadas como positivas e negativas. O método proposto é testado com os algoritmos KNN, Naive Bayes e SVM. Os resultados e análise de desempenho mostra melhor desempenho com SVM com precisão de 97.46%.

ISAH et al. (2014) desenvolveram um framework para classificar a polaridade (positivo, negativo ou neutro) de experiências de usuários sobre marcas populares de cosméticos e medicamentos. Termos são etiquetados baseados em léxico de sentimento e ferramentas de características sintáticas. Os dados são treinados em classificador NB. O classificador obteve uma precisão de 83%.

SHARIF et al. (2015) proporam um classificador baseado em sentimento para a previsão de recorrentes tópicos de comentários (*viral trends*) em fóruns on-line sobre saúde. Segundo os autores, tendências virais nas mídias sociais online possuem duas características essenciais: volume e velocidade. E ainda, para eles minerar opinião adiciona um valioso indicador para identificar longas conversações, pois observaram que postagens com alto grau de sentimento possuíam tópicos mais longos de conversações. O método adotada modelagem de tópico, o léxico *WordNet* para identificar características semânticas e o léxico *SentiStrength* para identificar polaridade de termos. Em seguida são treinados em vários classificadores supervisionados.

JI et al. (2013) e BEHERA; ELURI (2015) focaram na classificação de sentimento para mensurar a preocupação (*Degree of Concern* - DOC) de usuários sobre doenças em mensagens *tweets*. A abordagem envolve duas etapas. Inicialmente é distinguido *tweets* pessoais de *tweets* de notícias (não pessoal). Em seguida, Mineração

de Opinião é aplicada em *tweets* pessoais para identificar *tweets* negativos. O número *tweets* negativos é utilizado para calcular o grau de preocupação. A abordagem é avaliada em diferentes métodos de aprendizagem de máquina supervisionada, onde *Multinomial Naïve Bayes* alcançou melhores resultados e levou menos tempo para construir o classificador comparado a outros métodos de aprendizagem.

XU et al. (2015) proporam classificar citações de sentimento quanto a polaridade em documentos clínicos científicos utilizando o algoritmo supervisionado SVM incorporando características extraídas do contexto da citação. O autor conclui que a maioria das citações em literatura biomédica é neutra em relação ao sentimento. Também cita que a polaridade do sentimento era frequentemente expressa como uma sentença comparativa e que os cientistas são relutantes em serem críticos quando não conseguem reproduzir resultados anteriores. Em alguns casos, o sentimento expresso da citação é muito implícito sem quaisquer pistas linguísticas explícitas, tais como palavras negativas. E por fim, o escopo de uma citação varia amplamente, podendo ser citado em uma única sentença ou cláusula ou estar distribuído em várias cláusulas ou sentenças.

Classificação de polaridade: Abordagem supervisionada						
Artigo	**Tarefa**	**N**	**Recursos externos**	**Ferramentas**	**Método**	**Dados de Experimento**
ALI et al. (2013)	Classificação de polaridade: positivo, negativo ou neutro.	S	SL, SWN	POS Tagging, lematização	NB, SVM, LR	Comentários de medicamentos (medhelp, alldeaf, hearingaidforums).
BEHERA; ELURI (2015)	Classificação de polaridade. Mede o grau de preocupação público sobre doenças	D	Parametros tweet (Keyword, time e geo-code).	Remoção de caracteres especiais, URLs e Tweets duplicados, stop words	Estatístico, Léxico, NB.	Tweet
DANIULAITYTE et al. (2016)	Classificação de polaridade: positivo, negativo ou neutro.	D	VADER	N-gramas, TF-IDF	LR, NB, SVM.	Tweets (cannabis e synthetic cannabinoid tweets)
DENECKE; NEJDL (2009)	Classifica weblogs médicos em informativos ou afetivos: Positividade, negatividade e objetividade.	D	SWN, UMLS	Frequência de termos, SeReMed	LR, NB	Blogs
GREAVES et al. (2013)	Classificação de polaridade: positivo, negativo ou neutro.	D		Bag-of-words, n-gramas	NBM, árvore de decisão, SVM	Comentários sobre hospitais
ISAH et al. (2014)	Classificação de polaridade: positivo, negativo ou neutro.	D	Léxico de domínio e Léxico de domínio geral.	Remoção de caracteres especiais, stop words, radicalização, bag-of-words, TF-IDF	NB	Comentários de medicamentos e cosméticos (Tweet, Facebook).

JI et al. (2013)	Classificação de polaridade: positivo, negativo ou neutro. Mede o grau de preocupação público sobre doenças.	D	Re-tweets	Remoção de caracteres especiais, bag-of-words, radicalização, n-gramas, TF-IDF	NB, SVM	Tweet
MANEK et al. (2015)	Classificação de polaridade: positivo ou negativo.	D		Remoção de caracteres especiais, stop words, token, radicalização	W-Bayesian LR, SVM	Comentários de medicamentos (drugs.com)
MELZI et al. (2014)	Classificação de polaridade: positivo ou negativo. Extrai citações de emoção (joy, anger, surprise).	D	Emolex	Remoção de caracteres especiais e termos de gírias	SVM	Fórum
MERIS et al. (2016)	Classificação de polaridade: positivo ou negativo.	D		Stop words	SVM	Tweet
MISHRA et al. (2015)	Classificação de polaridade: positivo, negativo ou neutro. Detecção e monitoramento de ocorrência de novas ADRs	D	MedDRA, SIDER, WordNet, SWN	Token, POS tagging, radicalização, stop words, correção ortográfica, remoção de caracteres especiais, word sense tagging, n-gramas	SVM	Várias fontes
MONDAL et al. (2016b)	Classificação de polaridade: positivo ou negativo.	A	SenticNet, SWN, WME2.0	Stop-words, radicalização	PLN, SVM, NB.	Documentos clínicos
NIU et al. (2005)	Classificação de polaridade: positivo, negativo ou neutro.	D	UMLS	N-gramas, radicalização, negação	SVM	Literatura Biomedica: resultados clínicos em textos médicos
NIU et al. (2006)	Classificação de polaridade: positivo, negativo ou neutro ou sem resultados (No Outcomes).	D	UMLS	N-gramas, radicalização, negação, características semânticas	SVM	Dados clínicos
PESTIAN et al. (2012)	Classifica emoções encontradas em notas de suicídio.	S A		POS Tagging, frequência de termos, anotação emocional.	PLN e aprendizagem de máquina	Notas de suicídio
SARKER et al. (2011)	Classificação de polaridade: positivo, negativo ou sem resultados (No Outcomes).	D	MetaMap, UMLS, NegEx, General Inquirer	N-gramas, radicalização, negação,	NB, Bayes Net, SVMs, C4.5 árvore de decisão	Artigo medicos (Abstracts)
SHARIF et al. (2014)	Classificação de polaridade: positivo, negativo ou neutro. Classificação de citações de emoção.	D	SWN, léxicos médicos, WordNet, AffectWordNet.	Token, POS Tagging, negação, n-gramas, sinônimo, hiperônimo, NER.	SVM	Fórum (askaPatient.com, Pharma)
SHARIF et al. (2015)	Classificação de polaridade: positivo ou negativo. Previsão de threads virais e identificação de ADRs.	D	SentiStrength, WordNet, Lista personalizada de medicamentos e ADRs	PMI	Modelagem de Tópico, J48, RandomTree, REPTree, BayesNet, NB, LR, RBFNetwork	Comentários de medicamentos (Drugs.com e Telco dataset do Digitalhome)
SOKOLOVA; BOBICEV (2013)	Categoriza posts em: Encorajamento, gratidão, confusão, fatos e fatos + sentimentos (positivo ou negativo).	D	WordNet-Affect	N-gramas, PMI.	NB, KNN	Fórum

SWAMINATHAN et al. (2010)	Classificação de polaridade: Positivo, negativo, neutro e sem relacionamento (no-relationship).	S	UML, WordNet	POS Tagging, n-gramas, coreference resolution, características semânticas	SVM , SVR	1000 abstracts do PUBMED
XIA et al. (2009)	Classificação de polaridade: positivo ou negativo.	A		Frequência de termos, radicalização, n-gramas, token, stop words	MNB	Blog Patient Opinion
XU et al. (2015)	Classificação de polaridade: positivo ou negativo.	D	Léxico de sentimento Biomédico	N-gramas, negação, relações comparativas	SVM	Documentos de ensaios clínicos em domínio biomedical

Tabela 3.4: Estudos que empregam tarefas de classificação de polaridade em domínio médico usando aprendizagem supervisionada

3.3.4 Classificação de polaridade baseada em abordagem não supervisionada

Identificamos que os trabalhos que desenvolveram métodos de classificação não supervisionada em sua maioria aplicam experimentos em comentários sobre medicamentos.

Assim como citamos na seção anterior sobre a classificação de emoção, localizamos o trabalho de CAMBRIA et al. (2012) que focou na classificação afetiva. O trabalho foca em determinar a intensidade de expressões de emoção classificando em: agradável, atenção, sensibilidade e aptidão. Além de também efetuar classificação de polaridade (positiva ou negativa).

Os demais trabalhos focaram na classificação de polaridade. E diferentemente dos trabalhos que aplicaram técnicas supervisionadas, observamos que estes estudos não incluem a classificação “neutra”, focaram apenas na classificação de polaridade “positiva” ou “negativa”.

DENG et al. (2014) e MCCOY et al. (2015) aplicam Mineração de Opinião em mensagens postadas em fóruns médicos para identificar que tipo de sentimento pode ser expresso em documentos de domínio médico.

BIYANI et al. (2013) analisa a polaridade de postagens usuários de uma comunidade on-line de apoio ao câncer. DENG et al. (2014) analisam a aplicabilidade de métodos de análise de sentimento em narrativas clínicas. São avaliados quais são as

características léxicas existentes em comentários clínicos e quais são as características de sentimento existentes.

Apresentamos na Tabela 3.5 o resumo das abordagens não supervisionadas identificadas durante o desenvolvimento desta tese. A seguir descrevemos detalhes sobre alguns trabalhos destacados na Tabela 3.5.

CHENG et al. (2012) proporam uma abordagem não supervisionada chamada *Regressional Probabilistic Principal Component Analysis* (RPPCA) para classificar revisões de medicamento como positivas ou negativas. Unigrama e bigrama são analisados em revisões de medicamentos a partir do dicionário *MedDRA*. Segundo o autor, o uso do dicionário *MedDRA* é insuficiente para marcação e processamento de documentos médicos.

NA et al. (2012) e NA; KYAING (2015) desenvolveram uma abordagem não supervisionada de classificação de aspectos quanto a polaridade (positivo, negativo ou neutro) em revisões de medicamentos extraídas de fóruns. O algoritmo é aplicado em nível de cláusula e é baseado puramente em regras linguísticas. O método adota os recursos *Subjectivity Lexicon*, *SentiWordNet* (ESULI; SEBASTIANI, 2007) e o *MetaMap*. Os resultados experimentais mostraram que o método proposto apresentou melhores resultados que abordagens baseadas em aprendizado de máquina.

O trabalho de NOFERESTI; SHAMSFARD (2015a) apresenta um modelo não supervisionado baseado em regras para classificação de polaridade em comentários de medicamentos. O método concentra na identificação de dados baseados em duas categorias: (1) Dados que expressam relações polarizadas entre entidades (exemplo, *Piroxicam is used to reduce the pain*). (2) Termos ou frases que contêm polaridade (exemplo, *severe abdominal pain, dry lips*). O método extraiu 9.703 de cláusulas com polaridade com previsão 92.26% e os experimentos em comentários de medicamentos demonstram que a abordagem pode atingir 79.78% de precisão para a tarefa de classificação de polaridade.

NOFERESTI; SHAMSFARD (2015b) apresentam o *OpinionKB*, uma abordagem para construir uma base de conhecimento de opiniões indireta do qual pretende ser um recurso para classificar automaticamente as opiniões de usuários sobre medicamentos. O módulo de extração de opiniões indireta consiste do uso de *Named*

entity recognition através do *MetaMap*. A definição da polaridade para as opiniões indiretas localizadas é baseada em regras podendo ser classificada como positiva ou negativa. A classificação de opiniões indireta obteve uma precisão de 88.09%.

O trabalho de MAJETHIA et al. (2016) descreve o *PeopleSave*, um sistema de recomendação e feedback baseado em comentários contextuais de pacientes diabéticos extraídos da internet sobre medicamentos. Os comentários são processados com técnicas de linguagem natural para extração de citação de efeitos colaterais. A análise automática do sentimento foi alcançado através da experimentação com as APIs *SentiStrength* e *AlchemyAPI*.

YAZDAVAR et al. (2017) realiza estudo sobre implícitas expressões quantitativas que expressam sentimento em comentários sobre medicamentos. A determinação da polaridade para cada sentença é construída a partir de conhecimento de lógica fuzzy. Experimentos demonstram que o modelo proposto obteve de 72% de F1.

MCCOY et al. (2015) aplicam Mineração de Opinião em 2.484 notas de alta hospitalar sobre indivíduos de uma unidade de internação psiquiátrica e 20.859 notas de sobre indivíduos de unidades médicas gerais. O método investiga a associação entre características sociodemográficas e sentimento em documentos de internação psiquiátrica. Sentimento negativo e positivo foi analisado em um conjunto de várias características como idade, gênero, seguro de saúde, transtorno psicótico, etnia, entre outros.

Classificação de polaridade: Abordagem não supervisionada						
Artigo	**Tarefa**	**N**	**Recursos externos**	**Ferramentas**	**Método**	**Dados de Experimento**
BIYANI et al. (2013)	Classificação de polaridade: positivo ou negativo.	D	SL Adaptado, Wikipédia, SentiStrength	Negação, remoção de caracteres especiais	Co-training (semi-supervisionado).	Fórum
CAMBRIA et al. (2012)	Classificação de polaridade: positivo ou negativo. Classificação de emoção: Pleasantness, Attention, Sensitivity, Aptitude	D	ConceptNet, WordNet-Affect	Lematização, n-gramas, analisador semântico	Baseado em regras e agrupamento	Questionários
CHENG et al. (2012)	Classificação de polaridade: positivo ou negativo.	A	MedDRA	Correção ortográfica, lematização, stop words, n-gramas, bag-of-words, TFI-DF	RPPCA, SVM	Comentários de medicamentos (webmd.com)
DENG et al. (2014)	Classificação de polaridade: positivo ou negativo. Analisa a	D	SL	POS Tagging, stop words	PLN	Relatório de radiologia, Carta de enfermeiras,

	aplicabilidade de métodos para análise de sentimento em narrativas clínicas.					entrevista de Slashdot.
GOEURIOT et al. (2011)	Classificação de polaridade: positivo ou negativo. Minera informações e opiniões sobre doenças e tratamentos.	A	UMLS, Metamap	Radicalização, frequência de termos, POS Tagging, stop words	Marcação semântica,	Comentários de medicamentos (druglib.com, drugs-expert.com, onlinemedsreview .com)
MAJETHIA et al. (2016)	Classificação de polaridade: positivo ou negativo. Sistema de recomendação de medicamentos.	A	SentiStrength, AlchemyAPI	Token, n-gramas	Similaridade e frequência de termos	Comentários de medicamentos (webmd.com)
MCCOY et al. (2015)	Classificação de polaridade: Positividade e Negatividade e subjetividade versus objetividade	A		POS Tagging	PLN	Notas sobre admissão e alta hospitalar.
MOON; BORKAR (2016)	Classificação de polaridade: positivo ou negativo. Sumarização de comentários.	D	WordNet		Probabilístico	Comentários de medicamentos (webmd.com)
NA et al. (2012)	Classificação de polaridade: positivo ou negativo.	A	SWN, SL, UMLS, MetaMap.	Correção ortográfica, POS Tagging, relações gramaticais	PLN	Comentários de medicamentos (druglib.com)
NA; KYAING (2015)	Classificação de polaridade: positivo ou negativo.	A	SWN, SL, UMLS, MetaMap	Correção ortográfica, POS Tagging, relações gramaticais	PLN	Comentários de medicamentos (webmd.com)
NOFERESTI; SHAMSFARD (2015a)	Classificação de polaridade: positivo ou negativo.	S	UMLS	Token, lematização, POS Tagging, relações gramaticais, coreference resolution, n-gramas	Linked Data Sources	Comentários de medicamentos (druglib.com, askapatient.com)
NOFERESTI; SHAMSFARD (2015b)	Detecção de polaridade de opinião indireta.	S	Léxico de domínio, MetaMap, UMLS	Token, lematização, POS tagging, relações gramaticais, coreference resolution, n-gramas	NER, extrações de relações, probabilístico	Comentários de medicamentos (druglib.com, askapatient.com)
WILEY et al. (2014)	Categorização e classificação de polaridade.	D	MetMap, UMLS, SWN	POS Tagging, ccorreção ortográfica, radicalização	Estatístico, regras de associação e similaridade	Comentários de medicamentos (Várias fontes)
YAZDAVAR et al. (2017)	Identificação de implícitos sentimentos. Classificação de polaridade.	S	Gazetteer, JAPER	Token, separação de sentenças	Lógica Fuzzy	Comentários de medicamentos (askapatient.com)

Tabela 3.5: Estudos que empregam tarefas de classificação de polaridade em domínio médico usando aprendizagem não supervisionada

3.3.5 Classificação de categoria

Os trabalhos identificados relacionados à classificação de categoria no domínio da medicina que aplicam técnicas de Mineração de Opinião se concentraram na tarefa de categorizar comentários quanto à menção de ADRs e sintomas.

Os estudos de PATKI et al. (2014), SAHANA; GIRISH (2015) e SARKER; GONZALEZ (2015) aplicaram técnicas para localizar citações explícitas de ADRs. Enquanto no trabalho de EBRAHIMI et al. (2015) foi tratado o problema de classificação de citações implícitas de opiniões para detecção de ADRs. E também, técnicas foram adotadas para diferenciar ADRs de sintomas, pois um mesmo termo pode ocorrer em ambas as categorias. Já no trabalho de EGGER et al. (2016) mensagens *tweets* foram classificadas quanto a menção de ADRS, sintomas ou nomes de medicamentos.

Os trabalhos citados adotaram técnicas de classificação supervisionada a partir da aplicação de algoritmos como SVM, NB entre outros. No trabalho de EBRAHIMI et al. (2015) foi adotado uma abordagem híbrida baseada em um conjunto de regras semânticas e classificação supervisionada. No entanto, EBRAHIMI et al. (2015) utilizaram o conjunto de regras para extração de aspectos e a categorização foi aplicada através de classificação supervisionada com aplicação do algoritmo SVM para treinamento. A Tabela 3.6 apresenta um resumo sobre este trabalhos citados. A seguir, apresentamos alguns detalhes sobre cada abordagem.

PATKI et al. (2014) apresentaram um problema de classificação binária para categorização de comentários e medicamentos a partir da polaridade de termos. Os autores sugerem que comentários associados à ADRs geralmente apresentam sentimento negativo. Inicialmente, um comentário é classificado se inclui uma menção de uma reação adversa ao medicamento (categoria: “ADR” ou “noADR”). Em seguida, é verificado se as combinações dos comentários indicam uma incidência excessiva de reações adversas (“red flag”) para o medicamento (categoria: “normal” ou “blackbox”). Os recursos *COSTART, SIDER, UMLS* e *MedEffect* são adotados para rotular características de domínio. E os léxicos *Wordnet* (MILLER, 1995) e *SentiWords* são utilizados para atribuição de informação semântica e de polaridade. Para a classificação binária foram adotados os algoritmos supervisionados MNB e SVM. O melhor classificador atinge uma precisão de 82% e *F-Score* de 0,652.

O trabalho de SARKER; GONZALEZ (2015) também tratou o problema de detecção automatizada de menções de ADR em textos. É endereçada a tarefa de classificação binária quanto à presença (“ADR”) ou ausência (“non-ADR”) de citação de ADR em comentários. São adotadas as ferramentas de POS Tagging, TF-IDF, modelagem de tópicos e, os recursos *Wordnet* (MILLER, 1995), *MetaMap* e um léxico de sentimento para rotular dados. Em seguida, são treinados em classificadores supervisionados SVM, NB e ME. O método atingiu F-score de 0.812.

Em SAHANA; GIRISH (2015) um conjunto de ferramentas de PLN e um léxico de polaridade de termos são utilizadas para etiquetar dados de comentários com objetivo de classificar comentários quanto à menção de ADRs (“ADR” ou “non-ADR”). Foram utilizados para experimentos dados clínicos do *Electronic Health Records* (EHR) e mensagens do *Twitter*. O conjunto de características é treinado em um classificador SVM.

EGGER et al. (2016) descreveram um classificador para determinar se uma mensagem *tweet* menciona uma reação adversa sobre medicamentos. O trabalho levantada a suposição que há existência de uma correlação entre sentimento e ADRs. *Tweets* foram separados em *tokens*, e etiquetados quanto a características semânticas. Diversos léxicos de sentimentos foram adotados para atribuição de polaridade. O pacote SIDER foi utilizado como recurso de domínio. Para treinamento é adotado classificador supervisionado SVM. *Tweets* são classificados em uma das categorias: “contains drug mention”, “contains symptom mention”, “contains both” e “contains none”.

No estudo de EBRAHIMI et al. (2015) é desenvolvido um sistema de Mineração de Opinião para classificar termos em uma das categorias: doença (disease-Manifestation Related Symptom - MRS) ou efeito colateral (drug-Adverse Drug Event – ADE). É proposto um modelo de extração de opinião implícita em comentários de medicamentos técnicas baseada em regras e um classificado SVM é utilizado para classificar efeito colateral e sintomas da doença.

Como podemos observar, poucos trabalhos se concentraram na tarefa de classificar tipos de aspectos em documentos relacionados à medicina. Nesta tese, além do problema de extração de aspecto, abordamos o problema de categorização de aspecto. No entanto, abordamos a classificação de múltiplas categorias em comentários

sobre medicamentos, diferentemente dos trabalhos acima que focaram apenas nas categorias ADR e sintomas.

Consideramos no método proposto nesta tese as categorias de aspecto: condição clínica, ADR, dosagem e eficácia. Assim, preenchemos uma lacuna identificada na literatura para o problema de classificação de várias categorias de aspectos em comentários de medicamentos não estruturado. Consideramos a classificação de aspectos proposto um problema de Mineração de Opinião, uma vez que a classificação abrange um par de opinião, ou seja, o conjunto de um termo de aspecto e um termo opinativo. Justificamos com o exemplo "stomach pain reduced", citado em um comentário de medicamento, o termo "stomach pain" isolado pode representar um aspecto do tipo condição clínica ou ADR, porém ao ser ligado ao termo opinativo "reduced" podemos categorizá-lo como aspecto de eficiência do medicamento. Na seção 4.3 apresentamos com detalhes o modelo proposto.

Classificação de categoria					
Artigo	**Tarefa**	**Recursos externos**	**Ferramentas**	**Método**	**Dados de Experimento**
EBRAHIMI et al. (2015)	Detecção de opiniões implícitas e classificação de ADRs e sintomas.	MetaMap	Token, POS Tagging	Baseado em regras e SVM	Comentários de medicamentos (drugRatingz.com)
EGGER et al. (2016)	Identifica se tweets mencionam ADRs, nome de medicamento e sintomas.	NRC Canada, Bing Liu's Sentiment Words, Sen140 Lexicon, MPQA Lexicon, SIDERS	Token, negação, n-gramas, hiperonímia, NER, word embedding	SVM	Tweet
PATKI et al. (2014)	Identifica comentários que mencionam ADRs. Sumarização.	COSTART, SIDER, CHV, SentiWord, WordNet	Lowercasing, radicalização, n-gramas, bag-of-words	MNB, SVM	Comentários de medicamentos (IMS Health's), Fórum (dailystrength.org)
SAHANA; GIRISH (2015)	Identifica comentários que mencionam ADRs. Identificação de polaridade.		Token, POS Tagging, remoção de caracteres especiais	SVM	Tweet, Relatórios clínicos
SARKER; GONZALEZ (2015)	Identifica comentários que mencionam ADRs.	UMLS, MetaMap, NegEx, WordNet	Token, POS Tagging, lowercasing, radicalização, n-gramas, remoção de caracteres especiais, TF-IDF	NB, SVM, ME	Comentários de medicamentos Tweet, Fórum (dailystrength.org), Relatórios clínicos

Tabela 3.6: Estudos que empregam tarefas de classificação de categoria em domínio médico

3.3.6 Construção de léxico

De acordo com ESULI (2008), um dos fatores principais para qualquer trabalho relacionado à Mineração de Opinião é a necessidade de identificar quais elementos da linguagem contribuem para expressar subjetividade em texto. Essa identificação pode ser realizada através de recursos léxicos que listam relevantes propriedades relacionadas à opinião de itens lexicais.

Os itens lexicais em recurso léxico podem ser palavras únicas ou sequências de palavras. Um léxico é compreendido como uma estrutura sistemática que define o significado das palavras. Cada termo de um léxico consiste de uma forma ortográfica e fonológica com um formato de representação de significado (ESULI, 2008).

No contexto de Mineração de Opinião, léxicos têm sido construídos ou adaptados com intuito de aplicá-los em pesquisas para detecção de subjetividade em textos e classificação de sentimento. Alguns recursos são compostos por um conjunto de termos, relacionados a categorias, por exemplo, positiva ou negativa. (ESULI, 2008).

Os recursos lexicais para Mineração de Opinião de domínio geral mais conhecidos são: o Subjectivity Lexicon (SL)[1], SentiWordNet (SWN)[2], WordNet Affect[3], General Inquirer[4] (GI), Emolex[5], Sentiment140[6], VADER[7], SenticNet[8], SentiStrength[9], ANEW (Affective Norms for English Words) [10] e SentiWords[11].

Descrevemos a seguir alguns trabalhos desenvolvidos para construção de recursos léxicos para o domínio da medicina.

GOEURIOT et al. (2012) apresentam a construção de um recurso léxico de domínio geral baseado em merge de termos existentes nos léxicos *Subjectivity Lexicon* (RILOFF; WIEBE, 2003) e o *SentiWordNet* (ESULI; SEBASTIANI, 2007). E, também apresentam a

[1] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[2] http://sentiwordnet.isti.cnr.it/
[3] http://wndomains.fbk.eu/wnaffect.html.
[4] http://www.wjh.harvard.edu/~inquirer/
[5] http://www.purl.org/net/emolex
[6] https://github.com/felipebravom/StaticTwitterSent/tree/master/extra/Sentiment140-Lexicon-v0.1
[7] https://github.com/cjhutto/vaderSentiment
[8] http://sentic.net/
[9] http://sentistrength.wlv.ac.uk/
[10] http://csea.phhp.ufl.edu/media/anewmessage.html
[11] http://hlt-nlp.fbk.eu/technologies/sentiwords

construção de léxico para o domínio médico baseado em corpus de comentários sobre medicamentos usando informações estatísticas. O trabalho demonstra que alguns termos têm polaridade diferente no domínio geral e no domínio específico e que muitos termos que possuem polaridade geralmente neutra tornam termos opinativos no domínio médico.

O método em ASGHAR et al. (2014a) propõe um léxico subjetivo a partir da mineração de dados em comentários sobre medicamentos disponíveis em diferentes fóruns relacionados. Um conjunto *seed* correspondente a termos médicos (substantivo, verbo, advérbio e adjetivo) é criado usando um corpus de 10.000 revisões de medicamentos anotados manualmente quanto à polaridade negativa ou positiva. Em seguida, o conjunto *seed* é expandido a partir de informações de sinônimos e antônimos do *SentiWordNet*, anexando *score* de polaridade para cada termo do léxico. Os resultados de avaliação revelam que a abordagem proposta proporciona melhores em comparação com os outros métodos existentes.

A técnica proposta em ASGHAR et al. (2014b) baseia-se em um modelo incremental e em corpus de revisões de saúde para criação de um léxico de polaridade para termos médicos. Em cada incremento, o vocabulário do léxico é aumentado sistematicamente e anexado uma polaridade ao termo. No primeiro incremento cada termo existente no corpus é comparado a um dicionário especial de termos médico para checar se o termo é um termo médico válido. No próximo incremento termos são expandidos usando um *Web Thesaurus*. Termos próximos são identificados usando uma medida de similaridade (Lin's measure) e adicionado ao léxico, o processo é repetido até convergir todos os termos. No terceiro incremento é calculado o score de polaridade para todos os termos existentes com seu respectivo *POS Tag* usando um léxico de opinião. Técnica de *Word sense disambiguation* é aplicados para isolar os termos médicos mais relevantes. Por fim, termos semanticamente relacionados são filtrados usando a medida *PMI* e anexado ao novo léxico. Os resultados da avaliação revelam que a abordagem proposta obtém melhores resultados em comparação com outros métodos existentes e alcança uma precisão de 82% e 74% no corpus de treinamento e teste respectivamente.

O trabalho de ASGHAR et al. (2016) apresenta uma abordagem híbrida para criar um léxico específico para o domínio de saúde para classificação e pontuação de sentimento em textos opinativos. O método é baseado em conceito de *boot-strapping*

(criação de lista *seed*, expansão de léxico, filtragem de termos redundantes, *co-reference resolution* e medida PMI). O conjunto *seed* é construído a partir de vários recursos disponíveis online. A extensão do léxico é conduzida tomando cada entrada do conjunto *seed* perfomando operações *boot-strap* e relações de sinônimos extraídas de um léxico online. O *score* é definido usando o *SentiWordNet* (ESULI; SEBASTIANI, 2007) e estratégicas específicas de domínio.

Neste trabalho, MONDAL et al. (2015), é desenvolvido um léxico chamado *WordNet for Medical Events* (WME) que usa informação contextual para apresentar definição de eventos, senso baseado na descrição contextual, polaridade, entre outros, de termos médicos. O trabalho inicia com uma lista *seed* de 1.654 termos médicos relacionados com seu contexto e polaridade. O *WordNet* (MILLER, 1995) é utilizado para determinar *POSTag* dos termos, sinônimos, etc. Um dicionário médico é incorporado para melhor entendimento de terminologias médicas. Algoritmo de desambiguação de sentido foi utilizado para extrair o sentindo mais relevantes para eventos médicos, três léxicos de polaridade são adotados no método: *SentiWordNet* (ESULI; SEBASTIANI, 2007), *Affect Word List* (STRAPPARAVA et al., 2004) *e Taboda's adjective list*.

Com podemos notar, diversos trabalhos tem focado no estudo para construção de léxicos que podem ser empregados no contexto de Mineração de Opinião no domínio específico. Cada um possui particularidades que diferem tanto no número de termos analisados quanto na quantidade de termos lexicais pertencentes a cada base. Diferem também quanto aos tipos de classes gramaticais que compõem cada recurso. E diferem também na forma como os dados foram etiquetados, onde empregam técnicas manualmente trabalhadas ou adotam uma combinação de regras automatizadas. A Tabela 3.7 apresenta um resumo dos léxicos localizado na literatura para o domínio médico.

Construção de Léxico					
Artigo	**Tarefa**	**Recursos externos**	**Ferramentas**	**Método**	**Dados de Experimento**
ASGHAR et al. (2014a)	Construção de léxico de polaridade sobre dados de medicamentos.	SWN	POS Tagging	Baseado em corpus, PMI-IR	Fórum (askapatient.com)
ASGHAR et al. (2014b)	Construção de léxico de polaridade sobre dados de medicamentos.	Snomed ACT, Mesh, UMLS, SL, Web Thesaurus, SWN, WordNet	Correção ortográfica, remoção de caracteres especiais, stop word, POS Tagging, shallow parsing	Medida de similaridade Lin's, medida de relação semântica baseada em vetor de contexto, PMI	Comentários de medicamentos (askapatient.com, drugratingz.com)
ASGHAR et al. (2016)	Construção de léxico de polaridade sobre dados de medicamentos.	UMLS, SWN, AlchemyAPI, MULTUM, NIH, WebMD, MediLexicon, General Inquirer	Token, stop word, lematização, correção ortográfica, co-reference resolution	PMI	Comentários de medicamentos (druglib.com), fórum (edrugsearch.com)
GOEURIOT et al. (2012)	Construção de léxico de polaridade de domínio geral e de domínio médico.	SWN, SL	Lematização, POS Tagging, frequência de termos	Baseado em regras	Comentários de medicamentos (Drugs Expert)
MONDAL et al. (2015)	Desenvolveu o WordNet para eventos Médicos (WME)	WordNet , SWN, Affect Word List, Taboda's adjective list, English medical dictionary	Word Sense Disambiguation, POS, sinônimos, n-gramas	Estatístico, similaridade, DRNN	
SOKOLOVA; BOBICEV (2013)	Consrtrução do léxico HealthAffect.	WordNet-Affect	N-gramas	PMI, NB, KNN	Fórum

Tabela 3.7: Estudos que empregam construção de léxico de polaridade em domínio médico

3.4 Considerações finais

Neste capítulo foi apresentada uma discussão sobre Mineração de Opinião focada em conteúdo textual relacionado a dados de saúde incluindo citações de medicina em geral e citações sobre medicamento. Podemos observar que citações opinativas em textos relacionados à medicina diferem em comparação a citações opinativas sobre produtos em geral.

Destacamos a importância da aplicação de Mineração de Opinião na área exemplificando como pode ser aplicada e suas possíveis contribuições. Trabalhos aplicados neste contexto foram apresentados e foi observado que abordagens de Mineração de Opinião tradicionais necessitam adaptações para alcançar melhores resultados no domínio em questão.

Ao investigarmos sobre trabalhos relacionados à Mineração de Opinião aplicada a citações textuais de medicina verificamos a ausência de trabalhos dedicados à extração automatizada de citações de aspectos em comentários sobre o uso de medicamentos. Assim como também, métodos automatizados para extrair a relação de termos aspectos e seus respectivos termos opinativos. Outra lacuna corresponde a métodos para classificar múltiplos tipos de aspectos, onde a maioria dos recentes estudos focou apenas na extração e classificação de aspecto do tipo ADR. Assim, este trabalho é focado na extração e classificação de múltiplos tipos de aspectos em comentários online de usuários sobre uso de medicamentos. Nos próximos capítulos apresentamos a solução proposta, os experimentos realizados e os resultados obtidos.

4 SOLUÇÃO PROPOSTA

Foi verificada que a extração de aspectos e opiniões expressas em revisões de remédios é uma carência na literatura atual. Assim, este trabalho aborda o problema de extração e classificação de múltiplos tipos de aspectos e seus respectivos termos opinativos no domínio de revisões de medicamentos. Ressaltamos que a identificação de outros tipos de aspectos, além de ADR, podem proporcionar uma melhor avaliação das experiências relatadas de usuários, e também um melhor resumo das avaliações associadas a um determinado medicamento.

O método proposto neste trabalho é composto de três principais etapas: Pré-processamento, Extração de Aspecto e Classificação de Tipo de Aspecto. Na Figura 4.1, é apresentada uma visão geral das etapas desenvolvidas.

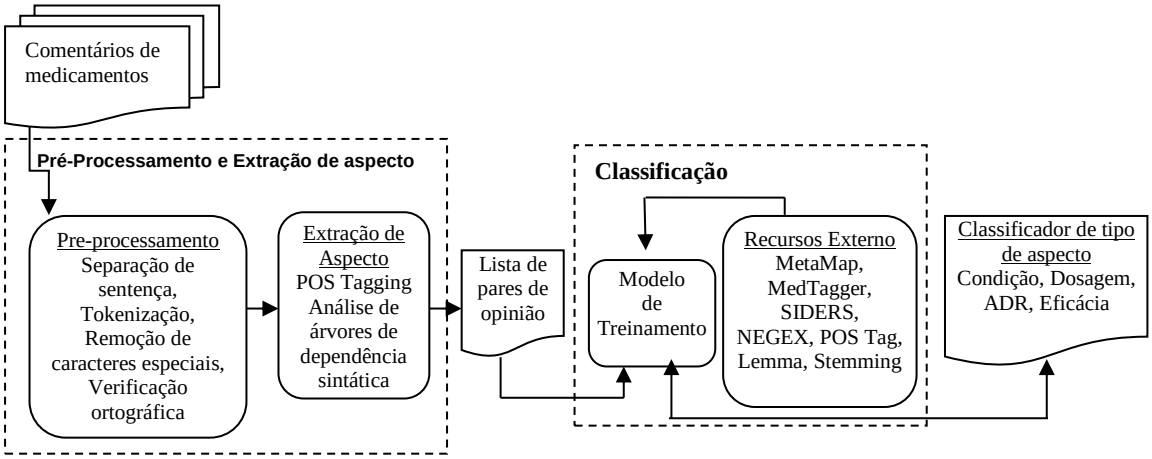


Figura 4.1: Uma visão geral da abordagem proposta

Neste trabalho, definimos uma opinião em texto dado por um usuário como uma tupla *T* = (entidade, aspecto, palavra opinativa), onde um medicamento é definido como uma entidade, um aspecto é definido como um alvo citado no texto e o termo opinativo corresponde a uma palavra que referencia uma opinião negativa ou positiva sobre o termo alvo. A tupla (Adderall, stomach pain, severe) corresponde a uma citação de ADR para o medicamento *Adderall*, o aspecto citado é *stomach pain*, e o termo *severe* corresponde a uma opinião negativa para o aspecto.

Um novo método para extrair e classificar aspectos em comentários opinativos sobre medicamentos é apresentado nesta pesquisa. Inicialmente, dado um texto como entrada, na etapa de Pré-processamento o texto é normalizado. Cada comentário coletado é dividido em sentenças e *tokens*, ferramentas linguísticas (*Apache Commons Lang e Google spell checking API*) são adotadas para realizar a correção ortográfica e remoção de caracteres especiais.

Na etapa de Extração de aspecto, um novo algoritmo baseado em relações é proposto, onde caminhos de dependência representados em árvores sintáticas de cada texto opinativo são adotados para extrair pares de opinião, ou seja, um aspecto e termo opinativo. Esta abordagem não requer treinar um modelo de extração, por exemplo, usando um algoritmo de aprendizagem. Pois, construir modelos para aprendizagem supervisionada é considerado um processo custoso.

Adicionalmente, esta abordagem é bastante útil para minerar aspectos que tem baixa frequência de ocorrência. Uma vez que, este trabalho é focado em domínio específico, citações de uso de medicamentos de baixa frequência sobre uma determinada reação adversa pode ser tão relevante quanto outra citação de maior frequência.

Por fim, na etapa de Classificação de tipo de aspecto, cada par de opinião é classificado em um dos quatros tipos de aspectos: Condição, ADR, Dosagem e Eficácia. Os aspectos “Opinião Geral” e “Custo”, citados por NA et al. (2012) e NA; KYAING (2015), não foram considerados neste trabalho.

O aspecto do tipo “Custo” foi desconsiderado devido a sua baixa frequência na base de experimentos. Quanto ao tipo “Aspecto Geral”, durante o processo de etiquetagem manual de tipos de aspectos para construção do modelo de referencia para a tarefa de classificação, foi observado que alguns termos aspectos poderiam tanto ser identificados como “Aspecto Geral” e também como “Eficácia”. Logo, foi definido etiquetar termos do tipo “Aspecto Geral” como do tipo “Eficácia”.

Para a classificação é adotado o algoritmo de aprendizagem supervisionada *Random Forest* (BREIMAN, 2001). Durante a investigação de trabalhos relacionados a esta tese não foi localizado trabalhos anteriores que identifique múltiplos aspectos em revisões de medicamentos e seus termos opinativos. Pesquisas anteriores apenas se

concentraram em classificação de ADRs, o que é limitado em comparação a este trabalho. Nas seções a seguir, descrevemos detalhes de implementação para cada etapa abordada nesta pesquisa.

4.1 Pré-processamento

Nesta etapa, a entrada corresponde a um conjunto de textos opinativos sobre experiências de uso de medicamentos. Como já citamos em capítulos anteriores, textos coletados em mídias sociais podem conter erros de digitação, inserção de caracteres especiais, entre outros, que pode levar a erros na etapa extração de características.

Para melhorar a qualidade na etapa de extração, é tratada nesta etapa a remoção ou substituição de caracteres. Cada texto é dividido em sentenças e *tokens* utilizando a ferramenta *Stanford CoreNLP*[12] em linguagem *Java.* A *Tokenization* é o processo de dividir um texto de entrada em pequenos pedaços ou *tokens,* cada *token* é representado por um termo contido nas sentenças. Em cada *token* da respectiva sentença, foram aplicadas ferramentas de Processamento de Linguagem Natural, o resultado nesta etapa são textos revisados quanto à remoção ou substituição de:

- Remoção de símbolos e caracteres especiais, por exemplo, !,@,#,$,%,^,+,-*,/;
- Conversão de termos em letra maiúscula para minúscula. Tarefa processada com a biblioteca *Apache Commons Lang*[13]*;*
- Remoção de consecutivos espaços em branco entre termos;
- Verificação e correção ortográfica de termos. A correção ortográfica para minimizar a ocorrência de palavras que foram registradas com erros pelos usuários é realizada através da ferramenta *Google spell checking* API[14].

[12] https://stanfordnlp.github.io/CoreNLP/
[13] https://commons.apache.org/proper/commons-lang/
[14] code.google.com/archive/p/google-api-spelling-java/

Por fim, os termos “meds” e “med” são frequentemente citados no conjunto de dados de experimentos por usuários para referenciar um medicamento adotado. Desta forma, foram substituídos pelo termo “medicine”.

4.2 Extração de aspectos

A fase de extração tem como entrada o conjunto de comentários revisados na etapa de pré-processamento. O objetivo nesta etapa é, dado um conjunto de revisões de medicamentos, extrair, em cada documento, pares de opinião, aspectos e seu respectivo termo opinativo (“opinion pair” (BANCKEN et al., 2014)), que são correspondentes a um sintoma ou condição descrito pelo usuário no qual levou ao uso do medicamento, ou a uma ADR, dosagem ou uma eficácia sobre o uso do fármaco. A saída resultante nesta etapa é uma lista de pares de opinião.

Na seção 2.2.1, abordamos uma discussão sobre diversas abordagens para extração de aspectos que incluem algoritmos baseados em aprendizagem supervisionada e não supervisionada. Diversos autores já citaram que Mineração de Opinião em textos é bastante sensível ao domínio, ou seja, um termo aplicado a um domínio pode ter sentido diferente se aplicado a outro domínio. Um termo também pode ter sentido alterado se ligado por termos de negação, assim como também a intensidade da polaridade é influenciada de acordo com o contexto do domínio.

Durante o desenvolvimento deste trabalho, para os experimentos foi selecionado documentos de opiniões de medicamentos referentes às condições clínicas “ADHD”, “AIDS” e “Anxiety”. Na seção 5.1 descrevemos detalhes sobre a base de experimentos.

Além da premissa de que Mineração de Opinião em textos é sensível ao contexto do domínio, observamos que no domínio da medicina e medicamentos, citações opinativas de medicamentos são sensíveis ao contexto da doença e ao tipo do medicamento em que estão sendo citados. Por exemplo, um determinado medicamento que é fornecido apenas em formato de cápsula possivelmente não haverá citações sobre a dosagem por militros (ml) e sim somente por miligramas (mg).

Assim como também termos específicos referentes às condições clínicas diferem entre si. A condição clínica “AIDS” há frequentes citações sobre a carga viral (viral

load - VL) e sobre a célula sanguínea T-C4, estes termos geralmente não ocorrem em citações sobre as condições clínicas “ADHD” e “Anxiety”.

Na tabela 2.1, seção 2.2.1, são apresentadas vantagens e limitações sobre as diversas abordagens para extração de aspectos. Além das particularidades do domínio citadas acima, observamos também que a frequência de citações de reações adversas e condições clínicas, e outros tipos de aspectos, são bastante distintas para cada medicamento de cada condição clínica e também consideramos que citações mesmo com baixa frequência, por exemplo, citações de reações adversas, tem importante relevância para o domínio.

Na seção 3.3.1, foram descritos diversos trabalhos que concentram na extração de aspectos. Conforme comentamos, citações de novas ADRs por usuários podem ser esparsas, ou seja, ter baixa frequência de ocorrência. Assim, para este trabalho não foi adotada a abordagem de extração de aspecto baseada em *Frequência*. Esta abordagem tem como desvantagem ignorar termos que possuem baixa frequência no conjunto de dados em análise. Entretanto, consideramos que ADRs com baixa frequência tem relevância para o domínio tanto quanto termos com alta frequência. A extração de ADRs pode, por exemplo, contribuir com o trabalho da FDA para acompanhar a ocorrência destas ADRs em novos usuários do medicamento em questão.

Sobre a abordagem baseada em *Aprendizagem Supervisionada* uma das desvantagens, conforme apresentado na Tabela 2.1, é o esforço necessário de construção de conjuntos de dados etiquetados para gerar um modelo supervisionado. Desta forma, também não optamos por este modelo para o problema de extração de aspecto apresentado neste trabalho.

Conforme citamos nos parágrafos anteriores, no domínio médico cada condição clínica possui particularidades de termos, assim como cada medicamento para uma respectiva doença pode ter diferentes citações. Por exemplo, para uma busca simples sobre os medicamentos “Adderall”, “Concerta” e “Daytrana” no site *druglib*, é possível observar que os registros de reações adversas e dosagens para cada remédio diferem.

Assim para este trabalho, consideramos que para trabalhar com um algoritmo supervisionado seria necessário obter dados etiquetados para cada remédio da condição clínica existente no conjunto de experimento. E, isto seria um trabalho muito custoso,

além de o modelo exigir uma grande quantidade de dados para alcançar um bom desempenho e ser dependente da qualidade do conjunto etiquetado.

Quanto à abordagem baseada em *Modelagem de Tópicos*, tem como desvantagem requerer um grande volume de dados. A base de experimento referente à condição clínica "AIDS", por exemplo, possui um conjunto de dados pequeno podendo, não ser suficiente para adotar esse modelo. Além disso, a abordagem é baseada em modelos probabilísticos, o que exige alto custo computacional.

Assim de acordo com as limitações sobre cada abordagem citadas na tabela 2.1, foi selecionado para este trabalho um modelo baseado em relações sintáticas. Uma vez que não é necessário o uso de dados etiquetados para treinamento de modelos e, além disso, possui baixo custo computacional.

A extração de pares de opinião deste trabalho se concentra através da extensão do algoritmo *Aspectator*[15] descrito no trabalho de BANCKEN et al. (2014). O algoritmo usa regras linguísticas a partir da análise de caminhos de dependência em árvore sintática e *POS Tagging* para encontrar opiniões expressas sobre aspectos candidatos. O modelo atende à extração de termos aspectos e de opinião. Também atende quanto à extração de termos de negação que modificam o sentido da opinião sobre um aspecto e captura termos ligados a termos de sentimento que modificam a intensidade de termos opinativos.

A etiquetagem morfossintática realizada por uma ferramenta POS Tagging consiste em identificar a classe gramatical de cada palavra de um texto, colocando uma etiqueta (tag) junto ao termo (LIU, 2007). A técnica POS Tagging é bastante representativa no contexto de Mineração de Opinião, uma vez que uma única palavra pode pertencer a diferentes classes gramaticais. Exemplificamos com a seguinte frase: "I was also very depressed". Através de uma ferramenta POS Tagging, a frase citada pode ser representada da seguinte forma: "I/PRP was/VBD also/RB very/RB depressed/JJ", onde PRP indica pronome, VBD indica verbo, RB indica advérbio e JJ indica adjetivo. Diversas ferramentas POS Tagging adotam um padrão de conjunto de *tags*. A tabela 4.1 apresenta alguns exemplos de *tags* e palavras para a língua inglesa.

[15] https://github.com/spdiana/OpinionPairExtraction

POS TAG	Descrição para a POS TAG	Exemplo/palavra
CC	Conjunção	And
CD	Número Cardinal	third
JJ	Adjetivo	green
JJR	Adjetivo comparativo	greener
JJS	Adjetivo superlativo	greenest
NN	Substantivo singular	table
NNS	Substantivo plural	tables
NNP	Substantivo próprio singular	John
NNPS	Substantivo próprio plural	Vikings
VB	Verbo, infinitivo	take
VBD	Verbo, pretérito	took
VBG	Verbo, gerúndio	taking
VBN	Verbo, particípio pretérito	*t*aken
VBZ	Verbo, presente singular, em 3ªpessoa	takes

Tabela 4.1: Exemplos de *tags* adotadas por ferramentas *POS Tagging* (LIU, 2007)

A etapa de seleção e extração de dados proposta consiste na extração de pares de termos chamados de *opinion pair*, proposto no algoritmo *Aspector* original, conforme ilustrado na Figura 4.2. O primeiro termo é chamado de *Aspect mention*, corresponde a menção de um aspecto e o segundo termo é chamado de *Sentiment modifier*, palavra em torno do aspecto que expressa uma opinião.

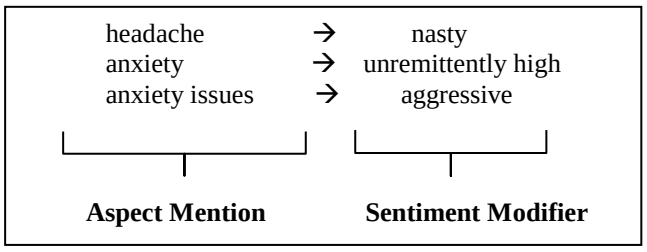


Figura 4.2: Exemplo de "opinion pair" (adaptado) (BANCKEN et al., 2014)

O *Aspectator* é descrito como um algoritmo passível de extrair termos candidatos a aspectos simplesmente combinando determinados caminhos de dependência sintática. Ou seja, este método comparado a outras abordagens não requer etiquetagem de dados. O algoritmo também não requer conhecimento específico de domínio. Em abordagens que exigem rotulagem de dados, um conjunto de dados etiquetados para um domínio específico pode ser não útil para outro domínio.

Na primeira etapa, o *Aspectator* original extrai aspectos candidatos em sentenças através do uso de caminhos de dependência sintática, a partir da ferramenta *Stanford dependency parser*[16], utilizando dois conjuntos: o primeiro conjunto (Tabela 4.2) corresponde ao caminho principal, onde em cada sentença é detectado um *aspect* (A) e um *sentiment modifier* (M) para formar o par {M; A}.

O segundo conjunto (Tabela 4.3) corresponde a extensões de caminhos de dependência da Tabela 4.2 para lidar com *multi-word aspects* (A) e *multi-word sentiment modifiers* (M), e também para capturar negação em sentenças. Este conjunto é dependente dos caminhos principais. Na tabela 4.4, apresentamos exemplo de pares de opinião gerados a partir dos exemplos de sentenças citados nas tabelas 4.2 e 4.3. Nesta tese adotaremos a citação "termo opinativo" para a citação do termo *sentiment modifier*.

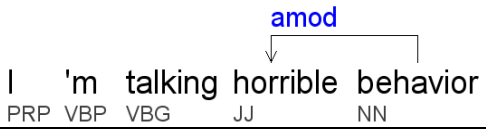
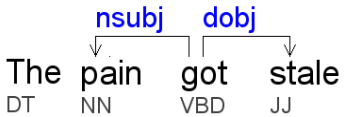
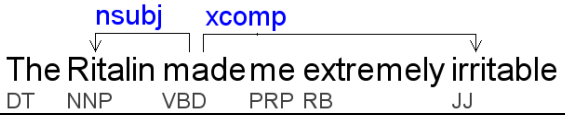
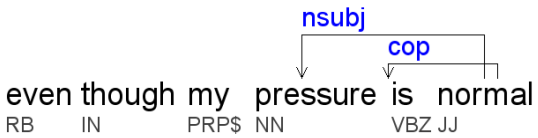
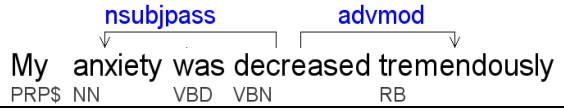
Principais caminhos de dependência	
Caminhos	**Exemplo**
amod	
nsubj_dobj	
nsubj_xcomp	
nubj_cop	
nsubjpass_ advmod	

Tabela 4.2: Principais caminhos de dependência (BANCKEN et al., 2014)

[16] https://nlp.stanford.edu/software/lex-parser.shtml

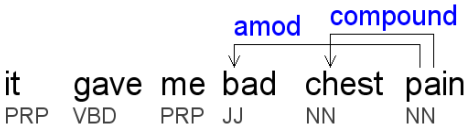
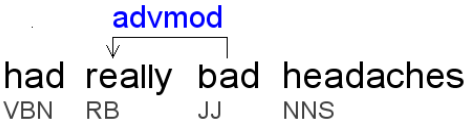
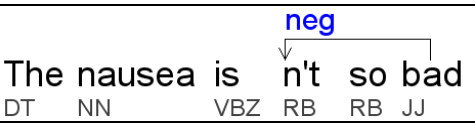
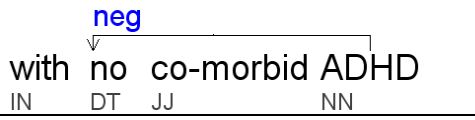
Extensões de caminhos de dependência	
compound noun	
adverbial modifier	
simple negation	
Negation through “no” determiner	

Tabela 4.3: Extensões de caminhos de dependência (BANCKEN et al., 2014)

Abordagens baseadas em regras linguísticas são uma alternativa interessante para a mineração de aspectos. Técnicas anteriores utilizadas com sucesso em outros domínios (como comentários em produtos, filmes, entre outros) podem ser investigadas no domínio das revisões de fármacos, preenchendo assim uma lacuna identificada na literatura.

Par de Opinião		Caminhos de dependência
Aspecto	**Termo Opinativo**	
behavior	horrible	amod
pain	stale	nsubj_dobj
Ritalin	**extremely** irritable	nsubj_xcomp + adverbial modifier
pressure	normal	nubj_cop
anxiety	**tremendously** decreased	nsubjpass_ advmod + adverbial modifier
chest pain	bad	amod + compound
headaches	**really** bad	amod + adverbial modifier
nausea	**n’t so** bad	nubj_cop + simple negation + adverbial modifier
ADHD	**no** co-morbid	amod + negation through “no” determiner

Tabela 4.4: Pares de opinião resultante dos exemplos das Tabelas 4.2 e 4.3

Conforme citamos anteriormente, Sarker et al. (2011) em seu trabalho comenta que a classe gramatical verbo desempenha um papel importante na análise de polaridade

em domínio médico. Geralmente quando vamos fazer um comentário sobre determinado remédio usamos similares tempos verbais.

O algoritmo *Aspectator* (BANCKEN et al.,2014) original está habilitado para extrair pares de opiniões correspondentes a aspecto do tipo substantivo e um termo opinativo do tipo adjetivo. A partir desta premissa, foi realizada uma investigação sobre verbos e seus respectivos tempos verbais citados em comentários em revisões de medicamentos a fim de identificar novos caminhos de dependência para extrair pares de opinião. Também foi observado que o algoritmo não está habilitado para extrair citações opinativas que estão ligadas pela conjunção "and". Vamos exemplificar, na Figura 4.3 é apresentada a árvore sintática de duas frases: (4.3a) "Anxiety was greatly reduced" e (4.3b) "Medicine is very effective and completely worth".

Observe a sentença "Medicine is very affective and completely worth", na Figura 4.3b. Na implementação original do algoritmo será extraído apenas o par de opinião {*very effective, Medicine*}, pois o algoritmo ignora termos que estão ligados pela conjunção "and", que neste caso corresponde a um complemento da citação opinativa sobre o aspecto, o que significa perda de informação. Ao adicionar o tratamento da conjunção "and" ao algoritmo resultaria a extração dos pares {*very effective, Medicine*} e {*completely worth, Medicine*}.

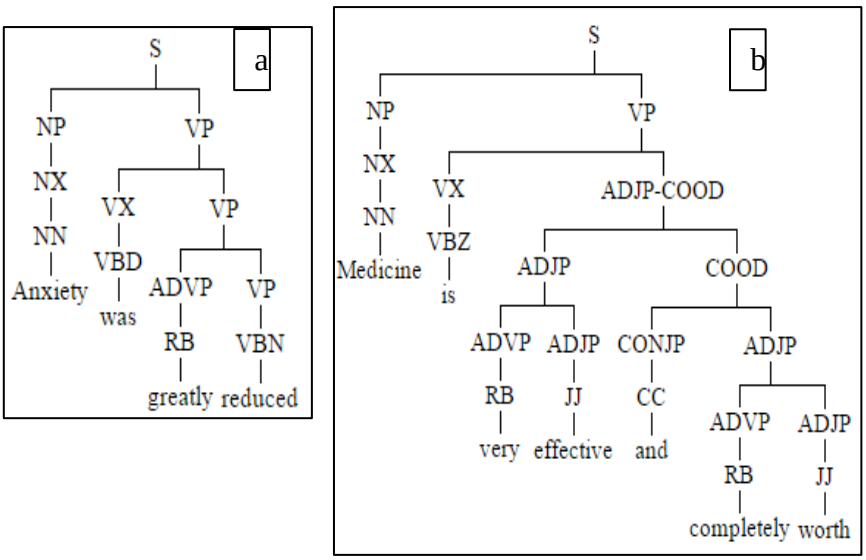


Figura 4.3: Exemplos de árvore sintática

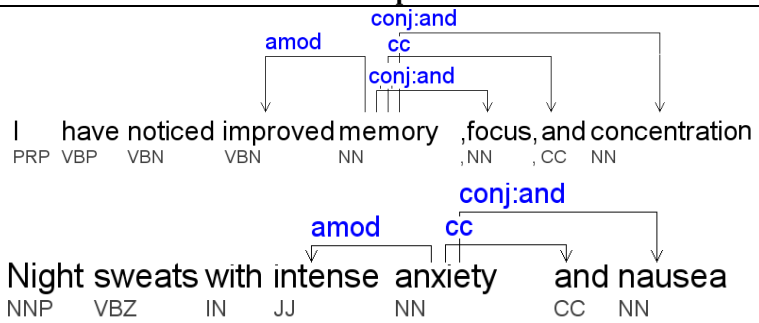
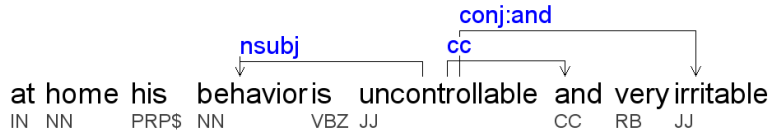
Caminho de dependência	Exemplos
amod_conj	
nsubj_conj	

Tabela 4.5: Adaptação do algoritmo *Aspectator*: tratamento da conjunção “and”.

Apresentamos na Tabela 4.5 os novos caminhos de dependência propostos neste trabalho para atender a questão referente a termos ligados pela conjunção “and”. Na primeira coluna, são descritos os tipos de combinações de relações sintáticas e na segunda coluna são dados exemplos de sentenças para cada nova regra sugerida. Em seguida, na tabela 4.6, são descritos os exemplos de pares de opinião extraídos a partir destes novos caminhos.

Par de Opinião		Caminho de dependência
Aspecto	**Termo Opinativo**	
concentration	improved	amod_conj
focus	improved	
memory	improved	
nausea	intense	amod_conj
anxiety	intense	
behavior	uncontrollable	nsubj_conj
behavior	**very** irritable	nsubj_conj + adverbial modifier

Tabela 4.6: Pares de opinião resultante dos exemplos da Tabela 4.5

O *Aspectator* está habilitado para extrair pares de opiniões correspondentes a um aspecto representado por um substantivo e um termo de sentimento representado por um adjetivo. Mas como já citamos em revisões de medicamentos, pertencente ao domínio

médico, é bastante dependente da classe gramatical verbo, havendo frequente ocorrência de pares de opinião onde o termo opinativo é um verbo.

Observe na Figura 4.3a, a sentença “Anxiety was greatly reduced”, ilustrada por meio de árvore sintática, expressa um sentimento sobre o termo *anxiety* através do verbo *reduced* (verbo não relacionado por um advérbio como sugerido no caminho “nsubjpass_ advmod” na tabela 4.2), o que poderia resultar o par de opinião {*Anxiety; greatly reduced*}, onde o termo *anxiety* representaria um aspecto e *greatly reduced* corresponderia ao termo opinativo de opinião sobre o aspecto *anxiety*. Sendo assim, uma possível citação de eficácia do fármaco não seria identificada pelo algoritmo original *Aspectator*.

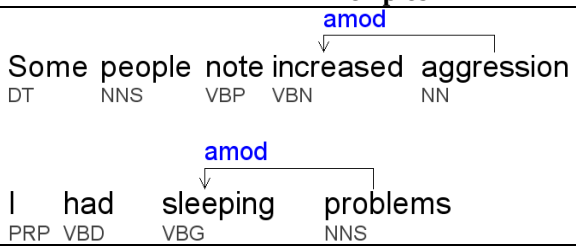
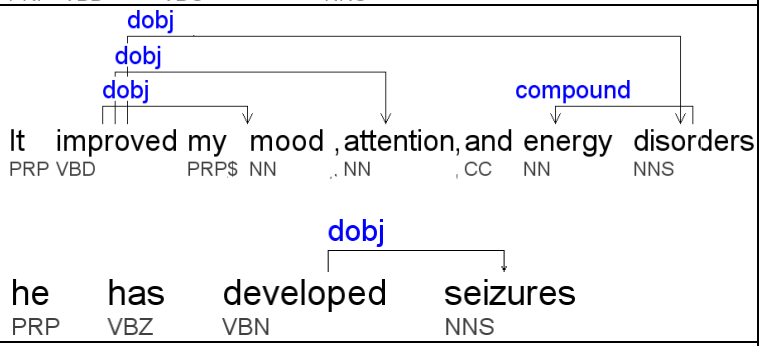
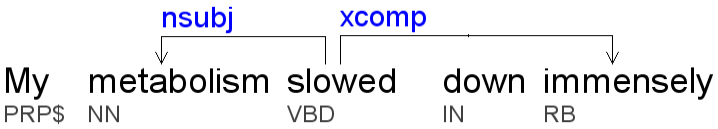
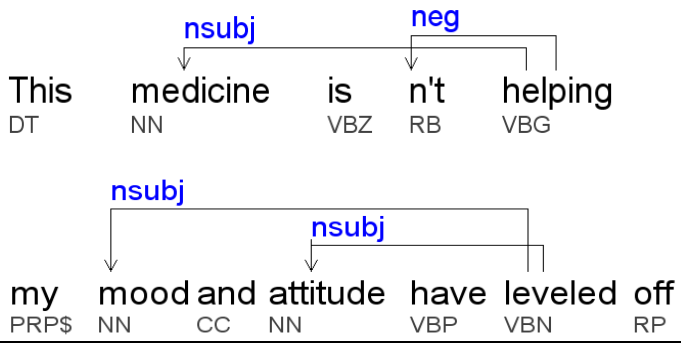
Caminho de dependência	Exemplos
amod NN_VB	
dobj NN_VB	
nsubj_xcomp NN_VB	
nsubj NN_VB	

Tabela 4.7: Adaptação do algoritmo *Aspectator* ao domínio médico: tratamento de verbo

Tendo em vista esta lacuna no algoritmo *Aspectator*, são propostos os caminhos de dependência apresentados na tabela 4.7. Os pares de opinião resultantes dos exemplos são demonstrados na tabela 4.8. Todos os pares apresentam um aspecto do tipo substantivo e um termo opinativo do tipo verbo. As sentenças que contêm termos ligados por conjunção também são extraídos, assim como também de extensões de termos conforme apresentado na tabela 4.3 são extraídos.

Pares de Opinião		**Caminhos de dependência**
Aspecto	**Termo Opinativo**	
aggression	increased	amod NN_VB
problems	sleeping	amod NN_VB
mood	improved	dobj NN_VB dobj NN_VB + conj dobj NN_VB + compound + conj
attention	improved	
energy disorders	improved	
seizures	developed	dobj NN_VB
metabolism	**immensely** slowed	nsubj_xcomp NN_VB + adverbial modifier
medicine	**n't** helping	nsubj NN_VB + simple negation
mood	leveled off	nsubj NN_VB nsubj NN_VB + conj
attitude	leveled off	

Tabela 4.8: Pares de opinião resultante dos exemplos da Tabela 4.7

Da mesma forma foi verificado que, para o domínio em questão, muitas opiniões são descritas em primeira pessoa. Por exemplo, observando a árvore sintática (Figura 4.4) das sentenças “I am feel very energetic” e “I became immensely depressed” é possível verificar que não há citação de termos substantivos em ambas as frases, mas representam sentenças opinativas e relevantes. Assim nessa tese, são propostos novos caminhos de dependência para extrair dados sobre sentenças em primeira pessoa onde foram considerados pares de opinião entre verbo e adjetivo ou verbo. Desta forma, para estas sentenças são resultados os pares de opinião {feel; very energetic} e {became; immensely depressed}.

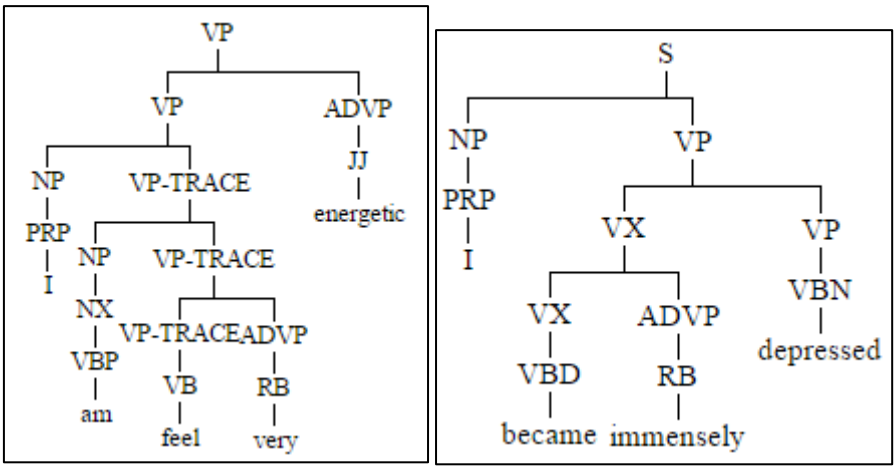


Figura 4.4: Exemplos de sentenças em primeira pessoa

Apresentamos na Tabela 4.9 os novos caminhos de dependência propostos para tratar o problema de extração par de opinião em sentenças em primeira pessoa. E, na Tabela 4.10 os pares de opinião resultantes dos exemplos apresentados.

Caminho de dependência	Exemplos
nsubj_xcomp VB_JJ	nsubj xcomp I felt very groggy PRP VBD RB JJ
nsubj_xcomp VB_VB	nsubj xcomp I mainly feel relaxed PRP RB VBP VBN

Tabela 4.9: Adaptação do algoritmo *Aspectator* ao domínio médico: tratamento de sentenças em primeira pessoa

Pares de Opinião		Caminhos de dependência
Aspecto	**Termo Opinativo**	
very groggy	felt	nsubj_xcomp VB_JJ + adverbial modifier
relaxed	feel	nsubj_xcomp VB_VB

Tabela 4.10: Pares de opinião resultante dos exemplos da Tabela 4.9

Ainda, propondo melhorias ao algoritmo *Aspectator* para extrair pares de opinião que não são atendidos no algoritmo original, é proposto o caminho apresentado

na tabela 4.11. Observando as sentenças exemplos na tabela, “The therapeutic effects are still evident” e “My sexual desire dropped sharply”, adotando o caminho de dependência AMOD seriam extraídos os pares <effects; therapeutic> e <desire; sexual> respectivamente.

No entanto, é visível, na sentença, que os pares extraídos <therapeutic, effects> e <sexual, desire> por AMOD são incompletos. Desta forma, com o novo caminho proposto, serão extraídos os pares <still evidente; therapeutic effects> e <sharply dropped; sexual desire>.

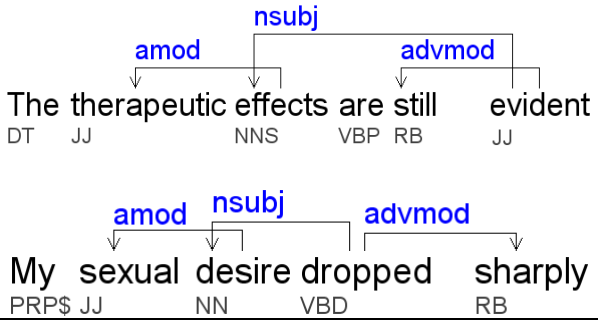
Caminho de dependência	Exemplos
amod_nsubj	

Tabela 4.11: Adaptação do algoritmo *Aspectator* ao domínio médico: tratamento do caminho AMOD composto

Pares de Opinião		Caminhos de dependência
Aspecto	**Termo Opinativo**	
therapeutic effects	**still** evident	amod_nsubj + adverbial modifier
sexual desire	**sharply** dropped	amod_nsubj + adverbial modifier

Tabela 4.12: Pares de opinião resultante dos exemplos da Tabela 4.11

Investigações foram realizadas a fim de adicionar adaptações ao algoritmo *Aspectator* quanto ao uso de verbos entre termos candidatos a aspectos e termos opinativos e citações em sentenças ligadas por conjunções e novas combinações de caminhos de dependência para extrair pares de opinião são levantados. Todos os novos caminhos de dependência consideraram as relações entre substantivos, adjetivos e verbos.

Assim, nesta fase de extração de dados, foi proposto adaptar o algoritmo *Aspectator* (BANCKEN et al., 2014) ao domínio médico para superar estas limitações. O objetivo do estudo de melhorias ao algoritmo é aumentar as possibilidades de extração de aspectos e termos opinativos.

A implementação do algoritmo foi realizada em linguagem *Java* e a combinação de caminhos de dependência foi computada usando o padrão *Universal Dependency Relations*[17] encapsulado na ferramenta *Stanford Parser*[18] para análise de árvore sintática. A implementação do algoritmo está disponível no endereço https://github.com/spdiana/OpinionPairExtraction.

A fim de averiguar a eficiência e qualidade de cada nova regra proposta para a extração de aspectos e termos de sentimentos, na seção 5.2, experimentos são realizados em um conjunto de comentários sobre medicamentos, e são apresentados os resultados quanto à precisão e cobertura obtidos no processo de mineração de aspectos referente a cada novo caminho de dependência sugerido. Também são avaliados os caminhos de dependência propostos no algoritmo original, e por fim o método proposto é comparado com dois outros trabalhos que desenvolveram regras para extração de aspecto baseado em pares de opinião.

4.3 Classificação de tipo de aspectos

A etapa de classificação de tipo de aspecto tem como entrada uma lista de pares de opinião descritos com um conjunto de características específicas de domínio e características linguísticas, e tem como saída cada par de opinião etiquetado em um dos quatro tipos de aspectos: Condição (C), Efeito Adverso (ADR), Dosagem (D) ou Eficácia (E). Um classificador construído a partir do uso de aprendizagem supervisionada foi adotado nesta solução proposta.

Uma questão importante no processo de classificação é o conjunto de características usadas para descrever os pares de opinião e, portanto, adotados como atributos preditores pelo classificador. Neste trabalho, um conjunto de sete diferentes recursos foi combinado para construção de um modelo preditivo de tipo de aspectos em

[17] http://universaldependencies.org/u/dep/
[18] https://nlp.stanford.edu/software/lex-parser.shtml

comentários de medicamentos. Inicialmente, o aspecto e seu termo opinativo, um par de opinião, são associados com suas respectivas classes gramaticais indicados por uma ferramenta de *POS* Tagging.

Também são atribuídas informações de lematização, radicalização, e presença de negação. Adicionalmente, ferramentas de *Named Entity Recognition* (NER) foram usadas para marcar cada aspecto e termo opinativo com um tipo semântico. Para isso, três ferramentas de NER baseados em conhecimento de domínio médico foram selecionadas: o *MetaMap, SIDER e MedTagger*. Estes recursos foram citados resumidamente na seção 3.2, e são descritos com mais detalhes a seguir.

4.3.1 Lematização (Lematization)

O processo de lematização transforma verbos flexionados para a sua forma no infinitivo e adjetivos e substantivos para a sua forma singular e, se existir, para sua forma no masculino. Este processo é também chamado de reduzir um termo a sua forma canônica. Por exemplo, a lematização para as palavras *livrinho e livros* é *livro* (ASGHAR, M. Z. et al., 2016; LIU, 2007). Para esta tarefa foi utilizado o recurso de lematização disponível na biblioteca *Stanford CoreNLP*.

4.3.2 Radicalização (Stemming)

Este processo reduz termos de diferentes classes gramaticais que possuem mesmo radical para forma denominada *stem*, sendo eliminados sufixos provenientes de flexão ou derivação, ou seja, todas as palavras flexionadas no texto são transformadas em sua forma base. Por exemplo, a radicalização para as palavras *books* para *book*, *laughing, laughed* and *laughs* para *laugh* (ASGHAR et al., 2016; LIU, 2007).

A principal diferença sobre o uso de radicalização e lematização é que a categoria morfológica é mantida no processo de lematização enquanto para *stemming* o *stem* será o mesmo. Citamos como exemplo, para as palavras *construiu* e *construções* o *stemming* será o mesmo *constru*. No entanto a lematização para a palavra *construiu* é

construir e para a palavra *construções* é *contrução* (ASGHAR et al., 2016; LIU, 2007). Esta tarefa foi realizada utilizando a ferramenta *Snowball*[19].

4.3.3 Negação

Palavras com a categoria *negação* são conhecidas como V*alence Shifters.* São termos que podem alterar a orientação de outro termo, como *never, no, not, nobody, nowhere, nothing, neither.* Negação de termos em textos médicos geralmente indicam a presença ou ausência de específicas conclusões médicas (ROKACH, et al., 2008; WU et al., 2014).

Por exemplo, considere as sentenças "Concentration hasn't improved much'' e "I did not experience any negative effects". Na primeira sentença os termos *concentration* e *improved* indicam positividade quanto à melhoria da concentração, mas o sentido é invertido ao ser ligado pelo termo *hasn't.* Na segunda sentença há citação negativa dos termos *experience negative effects*, mas ao ser citada como termo *not* a polaridade da sentença é invertida.

Neste trabalho, foi adotado o recurso *Negex*[20], um ferramenta de código aberto em linguagem *Java* para detecção de eventos de negação em documentos médicos. A ferramenta é aplicada para indicar se uma ocorrência de par de opinião é negado. *Negex* contém uma lista de termos de negação marcados com etiquetas padronizadas. Consideramos apenas termos que estão marcados com o rótulo *DEFINITE_NEGATED_EXISTENCE.*

4.3.4 MetaMap

MetaMap[21] é uma API em *Java* mantida pela Biblioteca Nacional de Medicina (*U.S. National Library of Medicine* - NLM) para mapear texto biomédico a partir do *Metathesaurus* existente no UMLS[22]. O UMLS corresponde a um conjunto de arquivos

[19] http://snowball.tartarus.org/
[20] https://code.google.com/archive/p/negex/
[21] https://metamap.nlm.nih.gov/
[22] https://www.nlm.nih.gov/research/umls/

e softwares que integra vários vocabulários de saúde e biomédicos (ver Figura 4.5). O UMLS é formado por três recursos de domínio (BODENREIDER, 2004):

(1) O *Metathesaurus*[23] é um extenso banco de dados de vocabulário, universal e multilíngüe que contém informações sobre conceitos biomédicos e relacionados à saúde, seus vários nomes e as relações entre eles. O *Metathesaurus* contém mais de um milhão de conceitos médicos e de termos gerais, a maioria em inglês, mas também abrange outras línguas, agrupados em mais de 130 tipos semânticos. Cada tipo semântico pertence a um dos 15 grupos semânticos.

(2) O *Semantic Network*[24] contém tipos semânticos e relações semânticas (relações entre tipos semânticos). Cada conceito do Metathesaurus é atribuído a pelo menos um tipo semântico.

(3) O *Specialist Lexicon e Lexical Tools*[25] correspondem a um componente central de sistema de processamento de linguagem natural. A entrada do léxico para cada palavra ou termo registra a informação sintática, morfológica e grafêmica.

O *MetaMap* usa abordagem baseada em processamento de linguagem natural (PNL) e técnicas linguísticas computacionais e fornece uma ligação entre o texto da literatura biomédica e os de conhecimento geral, incluindo relações de sinonímia incorporadas no *Metathesaurus* do UMLS. A Figura 4.6 apresenta exemplos da rede semântica UMLS[26]. O *MetaMap* está disponível através de acesso via web, através de versão instalável, e versão batch. Neste trabalho, foi adotado o acesso à ferramenta via web e coletado manualmente os rótulos retornado de cada aspecto e termo opinativo da base de experimento (ARONSON; LANG, 2010).

O *MetaMap* processa um texto dado como entrada separando-o inicialmente em tokens, identificando siglas e abreviaturas, POS tagging, pesquisa léxica de entradas no

[23] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/
[24] https://semanticnetwork.nlm.nih.gov/
[25] https://lexsrv3.nlm.nih.gov/Specialist/Home/index.html
[26] https://www.ncbi.nlm.nih.gov/books/NBK9679/

Specialist léxicon e análise sintática. As Figuras 4.7 e 4.8 apresentam exemplos de saída dado como entrada um par de opinião <obstructive; sleep apnea>, e um único termo "heart" respectivamente. O sistema retorna uma lista de conceitos semânticos agrupados em candidatos *Metathesaurus* (Meta Candidates), e também indica o melhor resultado (Meta Mapping) indicado pela pontuação perfeita (1000) para o termo consultado (ARONSON; LANG, 2010). Neste trabalho foi adotado para cada termo apresentando em um par de opinião a saída *Meta Mapping* do *MetaMap*.

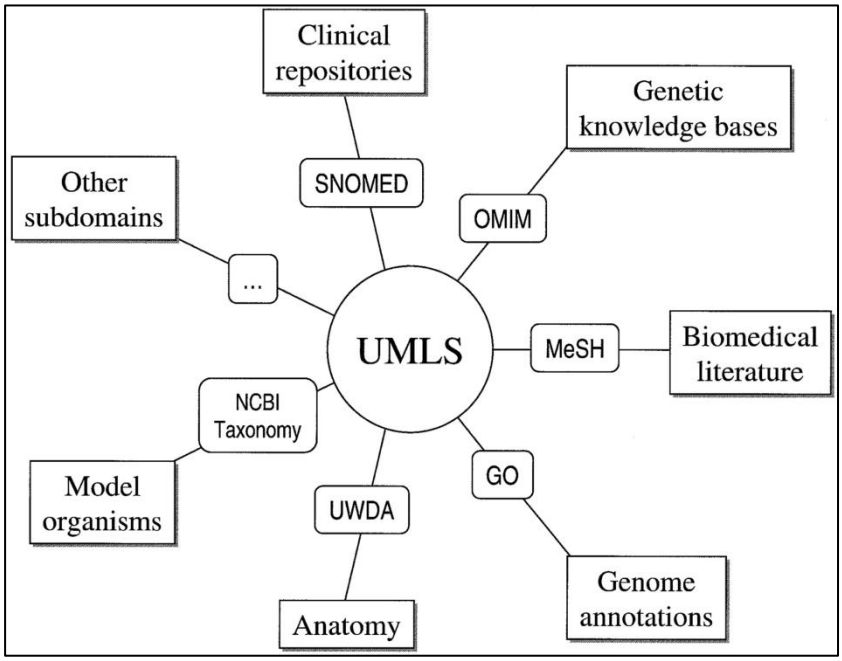


Figura 4.5: A representação de vários subdomínios integrados no UMLS (BODENREIDER, 2004)

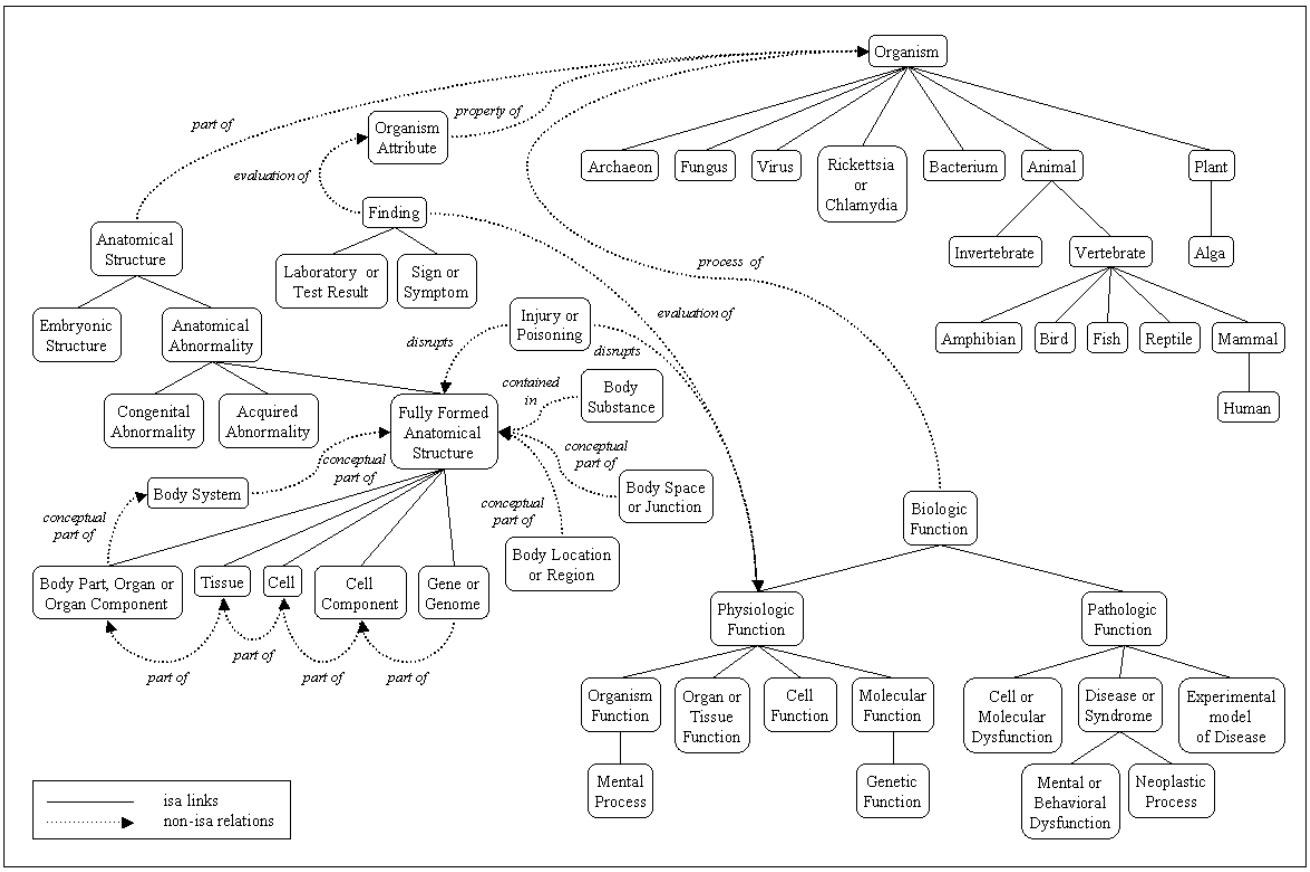


Figura 4.6: Uma porção da rede semântica UMLS: Relações

```
Phrase: "obstructive sleep apnea"
Meta Candidates (11):
  1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or
       Syndrome]
   901 Apnea, Sleep (Sleep Apnea Syndromes) [Disease or Syndrome]
   827 APNOEA (Apnea) [Pathologic Function]
   827 Sleep [Organism Function]
   827 Obstructive (Obstructed) [Functional Concept]
   827 Apnea (Apnea Adverse Event) [Finding]
   793 Sleeping (Asleep) [Finding]
   755 Obstruction [Individual Behavior,Pathologic Function]
   755 Sleepy [Finding]
   755 Sleeplessness [Sign or Symptom]
   755 Obstruction (Obstruction within Medical Device) [Phenomenon or
       Process]
Meta Mapping (1000):
  1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or
       Syndrome]
```

Figura 4.7: Exemplo de saída do *MetaMap* dado entrada uma frase (ARONSON; LANG, 2010)

```
    Phrase: "heart"
Meta Candidates (2):
    1000 Heart [Body Part, Organ, or Organ Component]
    1000 Heart (Entire heart) [Body Part, Organ, or Organ Component]
Meta Mapping (1000):
    1000 Heart (Entire heart) [Body Part, Organ, or Organ Component]
Meta Mapping (1000):
    1000 Heart [Body Part, Organ, or Organ Component]
```

Figura 4.8: Exemplo de saída do *MetaMap* dado entrada um termo (ARONSON; LANG, 2010)

4.3.5 SIDER

A base de dados *SIDER*[27] (Side Effect Resource) contém informações sobre medicamentos comercializados e as suas reações adversas extraídas de documentos públicos. A versão 4 do *SIDER* contém dados sobre 1430 drogas, 5880 ADRs e 140.064 pares drogas-ADR. Rótulos de medicamentos com informações para profissionais foram obtidos de registros nacionais e organizações de caridade. Estas informações heurísticas foram desenvolvidas para identificar automaticamente citações de ADRs nestes documentos coletados.

Foi utilizado do recurso SIDER os dados do arquivo "meddra_all_label_indications.tsv" disponível online. Foram adotados as tags "NLP_precondition" (Condições pré-existentes de pacientes), "NLP_indication" e "text_mention" (Detecção de menções de ADR). O arquivo foi persistido em tabela no banco de dados MySQL[28] e via algoritmo cada par de aspecto e termo opinativo foi consultado a sua categoria na tabela, caso retorne valor, é atribuído o valor retornado ao termo. As Figuras 4.9 e 4.10 apresentam exemplos de consulta ao recurso *SIDER* via web.

[27] http://sideeffects.embl.de/
[28] https://www.mysql.com/

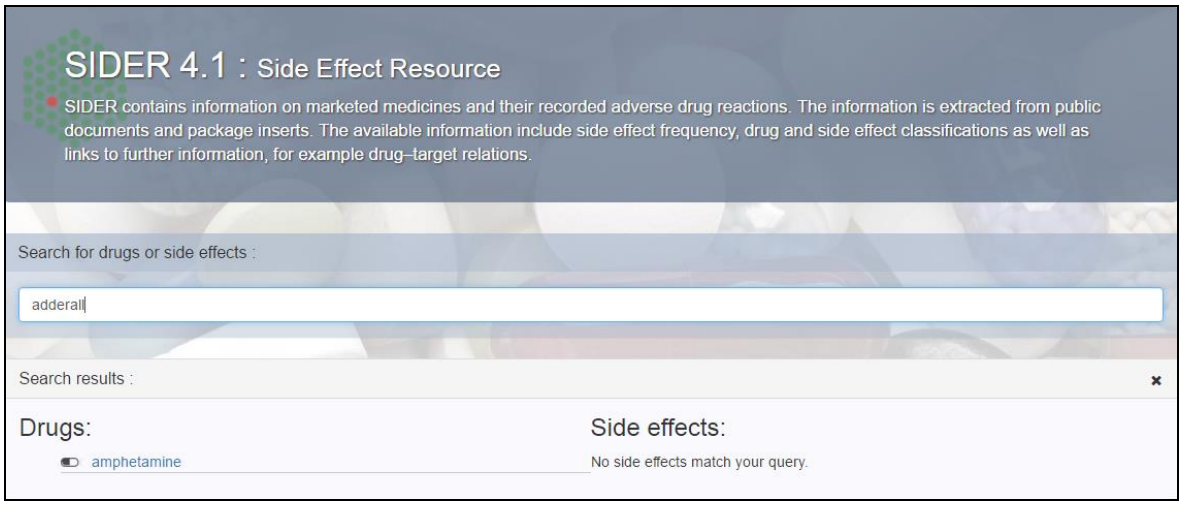


Figura 4.9: Exemplo de saída do recurso *SIDER* dado entrada "Adderall"

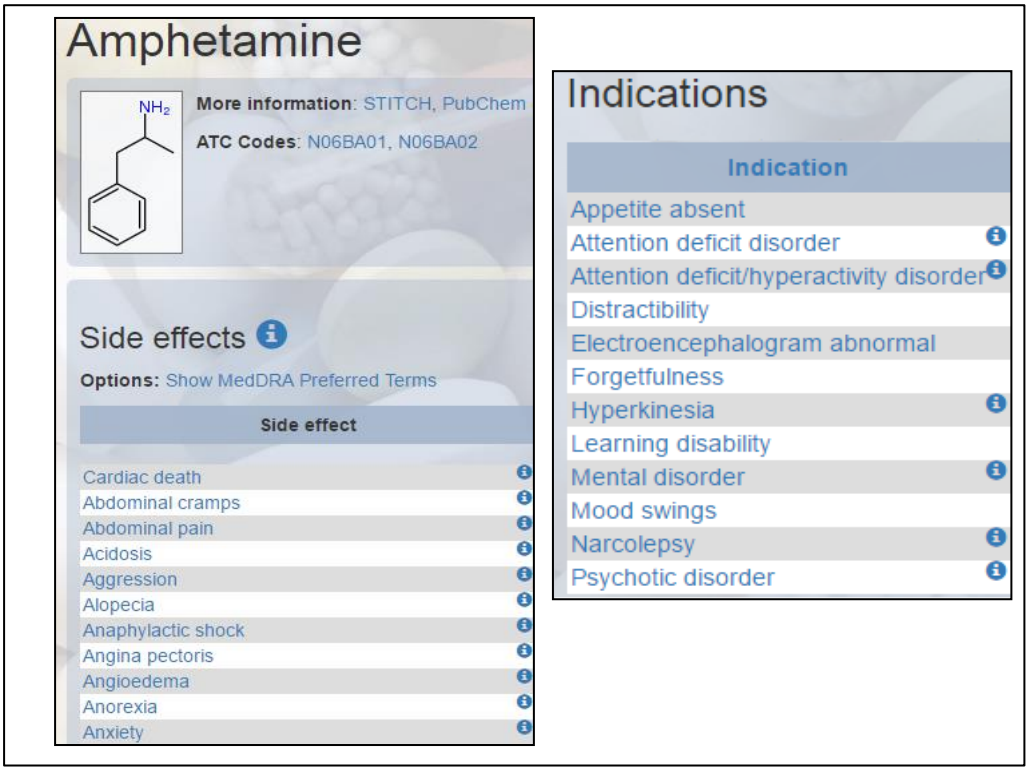


Figura 4.10: Exemplo de saída do recurso *SIDER* dado entrada "Amphetamine" (adaptado)

4.3.6 MedTagger

O *MedTagger*[29] é uma suíte de programas que utiliza dados estatísticos extraídos de grandes bases, aprendizagem de máquina, bases de conhecimento de domínio e análise sintática para detectar menções de conceito em texto clínico mantido pelo OHNLP Consortium (LIU; LI, 2012). Neste trabalho, foi adotado o dicionário de termos existente na ferramenta. O dicionário possui uma lista de termos clínicos anotados com um limitado conjunto de grupos semânticos, alguns exemplos: Anatomical Site (ANAT), Procedure (PROC), Diseases (DISO), Clinical Drug (DRUG), Chemical (CHEM). A lista de termos também foi persistida em tabela no banco de dados MySQL e via algoritmo, para cada aspecto e termo opinativo, foi consultado seu atributo semântico.

4.3.7 Descrição dos atributos do classificador

Para a etapa de extração de aspectos um conjunto de pares de opinião foi manualmente construído, identificando em cada base de experimento pares de opiniões válidos, o conjunto foi adotado para avaliar a eficácia do algoritmo de extração automatizado.

Par de Opinião		**Tipo de aspecto**
Aspecto	**Termo opinativo**	
mood swings	bad	C
hypertension	extreme	C
diarrhea	immediately had	ADR
muscle twitching	severe	ADR
anxiety	no obviously felt	E
depression	all reduced	E
dose	too low	D
dosage	regular	D

Tabela 4.13: Exemplos de pares de opinião rotulado

[29] http://ohnlp.org/index.php/MedTagger_Project_Page

Nesta etapa de classificação de tipo de aspecto, para cada par de opinião contido neste conjunto, foi manualmente atribuído um rótulo referente a um dos tipos de aspectos, Condição (C), Efeito Adverso (ADR), Eficácia (E) ou Dosagem (D), de acordo com o contexto do par de opinião citado no comentário do medicamento. Estes dados etiquetados são utilizados para treinamento e avaliação da performance do classificador. A tabela 4.13 apresenta exemplos de pares de opinião e seu respectivo rótulo de tipos de aspecto.

Característica	**Descrição das características**	**Tipo**
Tipo de Aspecto	Atributo alvo do classificador	Categórico {C, ADR, E, D }
POS Tagging	Etiqueta POS de cada palavra de acordo com a sua classe gramatical.	Categórico {NN, NNS, NNP, JJ, JJS, JJR,VBD, VBN, VBP, VBZ, VB, VBG }
Stemming	Redução do termo a sua forma canônica	Variável de acordo com o termo
Lematization	Redução do termo a sua forma base	Variável de acordo com o termo
Negex	Marcação de ocorrência de termos negativos, exemplo: never, dont, not, no, havent, wouldnt, non, wasnt, couldnt, coulnt, doesnt, cant, werent, didnt, never, isnt, hasn't.	Categórico {NA, DEFINITE_NEGATED_EXISTENCE}
MetaMap	Reconhecimento de entidade nomeada: conjuntos de dados baseados em conhecimento de domínio médico	Categórico (126 categorias) {Pharmacologic Substance, Qualitative Concept, Quantitative Concept, Disease or Syndrome, Sign or Symptom, Temporal Concept, NA, entre outros}
SIDER	Reconhecimento de entidade nomeada: conjuntos de dados baseados em conhecimento de domínio médico	Categórico {NLP_precondition, NLP_indication, txt_mention, NA}
MedTagger	Reconhecimento de entidade nomeada: conjuntos de dados baseados em conhecimento de domínio médico	Categórico {PROC, DISO, DRUG, ANAT, NA}

Tabela 4.14: Lista de características adotada para o classificador

Como já citamos anteriormente um conjunto de sete distintos recursos foi combinado para construção do modelo preditivo de tipo de aspectos. Logo, o conjunto citado acima, para cada aspecto e termo opinativo foi atribuído, se identificado no recurso, seu correspondente atributo de cada um dos sete recursos.

A Tabela 4.14 apresenta um resumo da lista de atributos preditores adotados para descrever os pares de opinião pelo classificador. NA é atribuído a termos que não são localizados no respectivo recurso. A classe alvo do classificador é o "tipo de aspecto", conforme apresentado na Tabela 4.14. A Tabela 4.15 apresenta um exemplo de par de opinião etiquetado com sua lista de atributos.

Característica	Aspecto	Negação	Termo Opinativo
Par de Opinião	panics	not	reduced
Tipo de aspecto	E		
POS Tagging	NNS		VBN
Stemming	panic		reduc
Lematization	panic		reduce
Negex		DEFINITE_NEGATED_EXISTENCE	
MetaMap	Mental or Behavioral Dysfunction		Qualitative Concept
SIDER	DISO		NA
MedTagger	NLP_precondition		NA

Tabela 4.15: Exemplo de par de opinião e sua lista de atributos

A fim de fazer uma previsão para uma determinada observação, normalmente é usado um modelo de treinamento para o conjunto de classe alvo. Neste trabalho, foi selecionado o algoritmo *Random Forest* (Floresta Aleatória), que é baseado em modelo de Árvore de Decisão (BREIMAN, 2001).

Desenvolvido em 2001, por Léo Breiman, o algoritmo *Random Forest* consiste em gerar várias árvores, cada qual construída a partir de uma reamostra aleatória do conjunto de treinamento original, no qual são combinadas para produzir uma previsão de consenso único, ou seja, dentro de um mesmo objeto, é gerado um conjunto de árvores de decisão. Cada conjunto de árvores é submetido a um mecanismo de votação (Figura 4.12), que elege a classe mais votada (BREIMAN, 2001; HAN et al, 2012; JAMES et al, 2013; TELOKEN et al, 2016).

Neste trabalho, foi utilizado, para treinar o classificador, o algoritmo *Random Forest* implementado na ferramenta *Weka*[30] de acordo com seus parâmetros padrão.

O algoritmo foi avaliado usando a metodologia de validação cruzada com 10 *folds*. De acordo com HAN et al. (2012), em *k-fold cross-validation* (validação cruzada), os dados iniciais são divididos aleatoriamente em *k* exclusivos subconjuntos ou "folds" mutuamente, D_1, D_2,..., D_k, cada um de tamanho aproximadamente igual. Treinamento e teste são realizados *k* vezes. Ou seja, para a tarefa de classificação, cada amostra é usada o mesmo número de vezes para treinamento e testes (LIU, 2007; HAN et al., 2012).

4.4 Considerações finais

Neste capítulo foi detalhada a arquitetura do modelo proposto nesta tese para extração e classificação de pares de opinião em citações de textos sobre o uso de medicamentos.

Neste trabalho assumimos um modelo de extração de aspectos baseado em regras de caminhos de dependência sintática. Para a classificação é assumido o algoritmo *Random Forest,* onde um conjunto de características linguísticas e especificas de domínio são adotadas para construção do modelo de dados a ser treinado no algoritmo de classificação.

Um conjunto de tarefas de pré-processamento, como a correção ortográfica, remoção de caracteres especiais, entre outros, são adotados com o objetivo de eliminar ruídos no conjunto de dados de experimentos. Esta fase é executada anteriormente as tarefas de extração e classificação.

No próximo capítulo abordaremos detalhes sobre os experimentos realizados a partir do modelo extração e classificação proposto. Assim como também, é detalhado sobre o conjunto de dados utilizado para experimentos. Várias questões foram identificadas e são discutidas.

[30] http://www.cs.waikato.ac.nz/ml/weka/

5 EXPERIMENTOS E RESULTADOS

Dentre os experimentos deste trabalho, foram realizados estudos em três bases de documentos referentes a comentários de usuários sobre o uso de medicamentos utilizando um modelo de extração e classificação de tipo de aspecto. Estes experimentos visaram avaliar o desempenho do modelo de extração e eficiência do classificador descritos nas seções 4.2 e 4.3 para o problema de mineração de aspecto.

Este capítulo está organizado da seguinte maneira: na Seção 5.1 é apresentado o conjunto de dados utilizado nos experimentos. Na seção 5.2, são apresentados os resultados de experimentos de extração de aspectos. Na seção 5.3, são apresentados os resultados de experimentos referentes à classificação de aspectos. Em seguida, na seção 5.4, são apresentadas as considerações finais.

5.1 Base de experimentos

Para este estudo, um conjunto de dados de textos opinativos sobre medicamentos foi coletado de um repositório público[31] em relação a três condições clínicas: “ADHD”, “AIDS” e “Anxiety”. O referido website possui uma boa variedade de comentários de medicamentos organizados por condições clínicas.

A fim de construir uma base de dados etiquetada para avaliação, inicialmente pares de opinião correspondente a um aspecto e seu respectivo termo de opinião presente em sentenças existentes nos comentários sobre medicamentos foram manualmente extraídos. Em seguida, cada par de opinião encontrado foi rotulado manualmente, por um único anotador, em um dos tipos de aspecto: Condição (C), Efeito Adverso (ADR), Eficácia (E), Dosagem (D).

Conforme citamos na seção 3.1, NA et al. (2012) e NA; KYAING (2015) discutiram que no contexto de opiniões sobre medicamentos existem seis tipos de aspectos mais comuns. Durante o desenvolvimento deste trabalho foi observado que o tipo de aspecto “Custo” tem baixa ocorrência nos comentários opinativos. Em relação ao conjunto de dados de experimento para “ADHD” ocorreram 16 pares de opinião,

[31] drugs.com

para “AIDS” 3 pares e “Anxiety” 8. Logo, o tipo de aspecto “Custo” foi desconsiderado na tarefa de classificação devido à baixa ocorrência no conjunto de experimento deste trabalho. Também foi verificado na fase de etiquetagem manual que muitos pares de opinião referente à “Eficácia” e “Opinião Geral” muitas vezes se enquadravam nos dois tipos de aspectos. Assim, foi que definido que aspectos relacionados seriam rotulados somente como tipo de aspecto “Eficácia”.

Inicialmente era previsto coletar um número de comentários distribuídos em várias condições clínicas para construção do modelo de base. Porém, coletar e etiquetar manualmente cada revisão é um processo demorado e custoso. As três bases foram selecionadas a partir da lista de condições clínicas mais comuns apresentadas no site (*Drugs by Condition*). As revisões coletadas correspondem a comentários cadastrados no sistema no período de abril de 2007 a janeiro de 2016.

	ADHD	**AIDS**	**Anxiety**
Número de medicamento	8	4	4
Total de comentários	802	201	500
Total de aspectos extraídos manualmente	6.528	1.841	3.955

Tabela 5.1: Conjunto de experimento

	Nome Remédio	**Nome Genérico**	**Comentário**	**Nota**
Anxiety	Ativan	Lorazepam	I was battling depression along with panic attacks. They first had me on Xanax. After a while wanted to get off that but was still having panic attacks. Not since I started Ativan. Also I find it much easier to focus and get things done now.	10
AIDS	Atripla	Efavirenz, emtricitabine, and tenofovir	I was diagnosed in April 2013 after getting really sick with fungal meningitis. My CD4 was 41 and the viral load was 260.000. After one month, the CD4 went to 71 (which wasn't supposed to increase as much), and the viral load was undetectable... I never had crazy dreams, though the first weeks were kind of hard because I used to get nauseous. Now it's all good. This pill rocks!	10
Anxiety	Buspar	Buspirone	Buspar was okay at first. I felt more relaxed and less tense. But I've had high blood pressure and headaches almost every day on this. Dr is taking me off and going to try something else.	5

Tabela 5.2: Base de experimento: exemplos de comentários

A Tabela 5.1 apresenta o total de documentos coletados para cada condição clínica, a quantidade de nomes de medicamentos relacionados a cada doença e o total de aspectos rotulados manualmente. Na tabela 5.2, apresentamos exemplos de comentários, onde foi coletado o nome da marca e nome genérico do medicamento, o texto do comentário e a nota dada pelo usuário ao uso do medicamento.

O conjunto de experimentos está disponível no endereço https://github.com/spdiana/DrugsReviewsCorpus.

5.2 Experimentos – Extração de Aspectos

Após a etapa de pré-processamento da base de dados, cada comentário foi separado em sentenças. Em seguida cada sentença foi verificada por um analisador sintático e, através de combinações de caminhos de dependência representados em árvore sintática, pares de opinião são extraídos conforme detalhado na seção 4.2. A fim de avaliar os resultados obtidos, foram computadas métricas clássicas para avaliar resultados de sistemas de recuperação, incluindo *Precisão* (P), *Cobertura* (C) e *F-Measure* (F) conforme descritas abaixo:

$$P = \frac{Opa}{Opi} \quad (5.2)$$

$$R = \frac{Opa}{Opn} \quad (5.3)$$

$$F = \frac{2 * P * R}{P + R} \quad (5.4)$$

- (Opi): Número de pares de opinião extraídos pelo método automatizado;
- (Op*a*): Número de pares de opinião extraídos automaticamente que ocorreram no conjunto de dados manualmente etiquetado, ou seja, pares relevantes extraídos;
- (Opn): Número de pares de opinião no conjunto de dados manualmente etiquetado;

Para analisar as medidas de precisão (equação 5.2) e cobertura (equação 5.3), a métrica *F-Measure* (equação 5.4) foi adotada representando a média harmônica entre a precisão e cobertura.

A fim de avaliar a sua eficácia, o método proposto também foi comparado a outros trabalhos anteriores que implementaram regras de caminhos de dependência sintática para extração de aspectos e seus respectivos termos opinativos, são eles:

- (BANCKEN et al., 2014), o algoritmo *Aspectator* original descrito na seção 4.2;
- ZHENG et al. (2014), descrito na seção 2.2.1.2. Os padrões de caminhos de dependência apresentados neste trabalho são comparados com as relações apresentadas no AEP;
- SAMHA (2016), descrito na seção 2.2.1.2.

Conjunto	Método	P	R	F
ADHD	ZHENG et al. (2014)	0,459	0,561	0,505
	SAMHA (2016)	0,580	0,540	0,560
	Aspectator	**0,816**	0,514	0,631
	Método Proposto	0,78	**0,665**	**0,718**
AIDS	ZHENG et al. (2014)	0,502	0,578	0,537
	SAMHA (2016)	0,643	0,604	0,623
	Aspectator	**0,772**	0,554	0,645
	Método Proposto	0,752	**0,678**	**0,713**
Anxiety	ZHENG et al. (2014)	0,515	0,543	0,529
	SAMHA (2016)	0,624	0,550	0,585
	Aspectator	**0,828**	0,582	0,683
	Método Proposto	0,787	**0,658**	**0,717**

Tabela 5.3: Precisão (P), Cobertura (C) e F-Measure (F)

A Tabela 5.3 apresenta os resultados obtidos de precisão, cobertura e F-Measure para os métodos avaliados nessa tese. Podemos afirmar que os métodos apresentados em ZHENG et al. (2014) e SAMHA (2016) de acordo com os resultados obtidos demonstraram não serem métodos competitivos em relação ao *Aspectator* original e o novo método proposto devido aos baixos valores de precisão em cada base.

Observe que o algoritmo *Aspectator* original obteve maior precisão para todos os conjuntos de dados, mas comparado ao método proposto em termos de cobertura obteve piores resultados. Isso se deve ao fato de no algoritmo proposto a adição de novos caminhos de dependência proporcionou atingir um número maior de aspectos, mas consequentemente estes novos caminhos de dependência também extraíram um número significativo de aspectos não relevantes, reduzindo então a precisão de pares de opinão válidos.

A partir dos resultados apresentados consideramos que o método proposto demostrou-se eficaz uma vez que comparado ao método original a precisão resultante obteve uma redução de apenas 4.41% para a base "ADHD", 2.59% para "AIDS" e 4.95% para "Anxiety". E, em relação a cobertura o método proposto atingiu um aumento de 29.37% para a base "ADHD", 22.38% para "AIDS" e 13.05% para "Anxiety". Desta forma, o método proposto obteve o melhor *trade-off* entre precisão e cobertura, ou seja, os melhores resultados em termos de *F-Measure*.

Caminhos	**ADHD**			**AIDS**			**Anxiety**		
Caminhos Originais	**Opa**	**Opi**	**P**	**Opa**	**Opi**	**P**	**Opa**	**Opi**	**P**
amod NN-JJ	1302	1627	**0,8**	591	787	**0,75**	1220	1493	**0,82**
nsubj-dobj NN-JJ	1	3	0,33	1	1	1	1	2	0,5
nsubj-xcomp NN-JJ	50	55	**0,91**	17	19	**0,89**	42	47	**0,89**
nsubj-cop NN-JJ	158	164	**0,96**	92	101	**0,91**	125	132	**0,95**
nsubjpass-advmod NN-VB	14	20	**0,7**	5	6	**0,83**	15	21	**0,71**
Novos Caminhos									
amod NN-VB	46	49	**0,94**	14	15	**0,93**	43	45	**0,96**
amod-conj NN-VB/JJ	46	73	0,63	24	50	0,48	48	74	0,65
amod-nsubj NNVB/JJ	113	161	**0,7**	82	110	**0,75**	106	149	**0,71**
dobj NN-VB	813	1127	**0,72**	285	394	**0,72**	750	1040	**0,72**
nsubj-xcomp NN-VB	50	108	0,46	15	29	0,52	42	84	0,5
nsubj-xcomp VB-JJ	237	247	**0,96**	51	53	**0,96**	209	215	**0,97**
nsubj-xcomp VB-VB	19	25	**0,76**	2	3	0,67	5	8	0,63
nsubj NN-VB	364	907	0,4	137	319	0,43	342	767	0,45
nsubj-conj NN-VB	13	18	**0,72**	8	11	**0,73**	15	21	**0,71**

Tabela 5.4: Precisão por caminho de dependência

Muitos estudos que trabalham com regras sintáticas geralmente apresentam os resultados de precisão geral, mas não apontam a precisão para cada regra sintática adotada para extração. Na Tabela 5.4, é demonstrada a precisão resultante para cada

regra sintática adotada neste trabalho, incluindo as regras do *Aspectator* Original e as novas combinações de caminhos sintáticos sugeridos.

Como pode ser visto, maioria das regras alcançaram bons valores de precisão. Observando as regras para os caminhos originais, o caminho "nsubj-dobj NN-JJ" teve baixa frequência nos três conjuntos de dados. Logo, consideramos esta opção de caminho de dependência não relevante para o domínio. Enquanto o caminho "amod NN-JJ" destaca-se por apresentar maior frequência de termos extraídos, portanto consideramos este caminho ser consideravelmente essencial por ter alta cobertura no domínio.

O caminho "nsubj NN-VB" apresenta alta frequência e baixa precisão igualmente nos três conjuntos. Observamos que este caminho se enquadra no comentário dado por LIU; GAO et al. (2016) de que determinadas regras sintáticas podem extrair muitos dados irrelevantes. Este caminho extrai muitos pares de opinião irrelevantes ao domínio e muitos sem sentido. Por exemplo, há ocorrência de verbos como "is", "are", "do", "went", "does", "been", entres outros, extraídos como termos opinativos.

O mesmo ocorre para o caminho "nsubj-xcomp NN-VB". Neste caminho são extraídos muitos pares de opinião que não são relevantes ao domínio deste estudo. Sugerimos para estes caminhos, o emprego de filtragem de termos relevantes ao domínio. Assim, consequentemente será possível alcançar melhorias na precisão.

O caminho "dobj NN-VB" atingiu bom desempenho de precisão e resolve o problema para a expressão "reduced my pain" discutido na seção 4.2. Os caminhos "nsubj-xcomp VB-VB" e "nsubj-xcomp VB-JJ" são capazes de extrair verbos como menção de aspecto. Conforme já discutimos anteriormente, comentários sobre medicamentos tem alta frequência de sentenças citadas em primeira pessoa. Por exemplo, "I feel extremely anxious", resultando a partir destes caminhos, o par de opinião < feel; extremely anxious>.

Realizamos experimento para averiguar os ganhos e perdas de precisão e cobertura a partir da seleção dos caminhos de dependência que resultaram precisão acima de 70% conforme apresentado na Tabela 5.4. Para isso, ordenamos em ordem decrescente a lista de caminhos de dependência de cada base de dados e simulamos a

junção de caminhos conforme listado na Tabela 5.5. Conforme apresentado para cada "conjunto", de 1 ao 9, é acrescido um caminho de dependência.

A Tabela 5.6 apresenta os resultados obtidos. Observe que o conjunto "1" resulta maior precisão, porém a cobertura é bastante irrelevante. A partir da adição do caminho AMOD NN-JJ ocorre um aumento expressivo na taxa de cobertura nas três bases de experimento.

Conjunto	ADHD	AIDS	Anxiety
1	nsubj-cop NN-JJ nsubj-xcomp VB-JJ	nsubj-xcomp VB-JJ amod NN-VB	nsubj-xcomp VB-JJ amod NN-VB
2	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ
3	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB nsubj-xcomp NN-JJ	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ
4	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB nsubj-xcomp NN-JJ **amod NN-JJ**	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ nsubjpass-advmod NN-VB	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ **amod NN-JJ**
5	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB nsubj-xcomp NN-JJ amod NN-JJ nsubj-xcomp VB-VB	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ nsubjpass-advmod NN-VB **amod NN-JJ**	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ amod NN-JJ dobj NN-VB
6	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB nsubj-xcomp NN-JJ amod NN-JJ nsubj-xcomp VB-VB nsubj-conj	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ nsubjpass-advmod NN-VB amod NN-JJ amod-nsubj NN-JJ/VB	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ amod NN-JJ **dobj NN-VB** nsubj-conj
7	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB nsubj-xcomp NN-JJ amod NN-JJ nsubj-xcomp VB-VB nsubj-conj **dobj NN-VB**	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ nsubjpass-advmod NN-VB amod NN-JJ amod-nsubj NN-JJ/VB nsubj-conj	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ amod NN-JJ dobj NN-VB nsubj-conj nsubjpass-advmod NN-VB
8	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB nsubj-xcomp NN-JJ amod NN-JJ nsubj-xcomp VB-VB nsubj-conj dobj NN-VB amod-nsubj NN-JJ/VB	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ nsubjpass-advmod NN-VB amod NN-JJ amod-nsubj NN-JJ/VB nsubj-conj **dobj NN-VB**	nsubj-xcomp VB-JJ amod NN-VB nsubj-cop NN-JJ nsubj-xcomp NN-JJ amod NN-JJ dobj NN-VB nsubj-conj nsubjpass-advmod NN-VB amod-nsubj NN-JJ/VB

9	nsubj-cop NN-JJ nsubj-xcomp VB-JJ amod NN-VB nsubj-xcomp NN-JJ amod NN-JJ nsubj-xcomp VB-VB nsubj-conj dobj NN-VB amod-nsubj NN-JJ/VB nsubjpass-advmod NN-VB		

Tabela 5.5: Ranking das precisões dos caminhos de dependência

	ADHD			**AIDS**			**Anxiety**		
Caminhos	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
1	**0,96**	0,08	0,15	**0,95**	0,03	0,06	**0,96**	0,06	0,11
2	0,95	0,09	0,17	0,92	0,08	0,15	0,96	0,09	0,17
3	0,95	0,10	0,18	0,92	0,09	0,16	0,95	0,10	0,18
4	0,83	**0,38**	0,52	0,92	0,09	0,17	0,84	**0,41**	0,55
5	0,83	0,38	0,52	0,78	**0,40**	0,53	0,80	0,59	0,68
6	0.83	0,39	0,53	0,78	0,45	0,57	0,80	0,60	0,68
7	0,79	0,56	0,66	0,78	0,45	0,57	0,80	0,60	0,69
8	0,79	0,58	0,67	0,76	**0,60**	**0,67**	0,79	**0,63**	**0,70**
9	0,79	**0,59**	**0,67**						

Tabela 5.6: Precisão, cobertura e F-Measure por ranking de precisão de caminhos de dependência

Conforme já citado anteriormente (ver Tabela 5.4), o caminho AMOD NN-JJ resulta o maior número de pares de opinião extraído, no entanto resulta queda na precisão devido ao número expressivo de taxa de acerto e erro. O mesmo ocorre a partir da inserção do caminho DOBJ NN-VB. Podemos concluir com esse experimento que com o incremento de caminhos de dependência ocorre redução na taxa de precisão, a taxa reduz gradativamente, mas mantém acima de 75%. No entanto, é possível atender uma maior cobertura de dados a partir da inclusão de caminhos de dependência e consequentemente alcançou uma taxa de *F-Measure* maior.

A Tabela 5.7 apresenta a cobertura relativa a cada tipo de aspecto, ou seja, tipos de aspectos relevantes (*Opa*) recuperados pelo método proposto que ocorrem no conjunto de dados rotulado dividido pelo número de aspectos presentes no conjunto de dados rotulado. É possível observar na tabela, que o tipo de aspecto Eficácia tem maior ocorrência, em seguida o aspecto Condição. Detectar ADRs em comentários é o tipo de aspecto de maior impacto sobre uso de medicamentos, porém sua ocorrência nas bases

de experimentos tem baixa incidência. E por fim, o aspecto Dosagem é o menos citado por usuários.

Conjunto	Eficácia	Condição	ADR	Dosagem
ADHD	0,534	0,231	0,169	0,065
AIDS	0,579	0,239	0,168	0,014
Anxiety	0,462	0,288	0,185	0,064

Tabela 5.7: Cobertura para cada tipo de aspecto

Conforme já citado na seção 4.2, confirmamos que a ocorrência do tipo de aspecto é bastante dependente ao tipo de condição clínica e tipo de medicamento citado. Isto pode ser explicado pela dificuldade da manipulação de alguns termos dependentes ao domínio que não foram corretamente analisados pela ferramenta de *parser* e possivelmente gerados classificações incorretas de *POS Tagging* e *parser* de caminho de dependência. Assim como também, abreviações de termos e correções de ortografia que não foram resolvidas na fase de pré-processamento também inferiram incorreta extração de aspecto e termo de sentimento.

Também observamos e confirmamos a premissa dada por YAZDAVAR et al., (2017) sobre a frequência de termos quantitativos em documentos relacionados a medicina. Observamos que há uma série de opiniões expressas sobre a "Eficácia" ou "Efeitos colaterais" que utilizam valores quantitativos. Isto ocorre principalmente na base de dados sobre a condição clínica "AIDS". Observe a sentença abaixo:

"I was diagnosed with HIV/AIDS in early October of 2015. My CD4 count was at 89 and my Viral Load up to 182,000/mL. I began taking Atripla on Nov. 10th 2015. I went in for my 6 week follow up on December 15th. After receiving my test results I was pleasantly surprised. My CD4 count had jumped to 394, and my Viral Load dropped to 2,185 - after only 6 weeks of treatment."

As sentenças "CD4 count was at 89" e "Viral Load up to 182,000/mL" podemos considerar citações sobre a "Condição" do paciente antes do uso do medicamento. Enquanto as sentenças "CD4 count had jumped to 394" e "Viral Load dropped to 2,185" podemos considerar como citação da "Eficácia" do medicamento

Observe que o usuário citou que, durante o tratamento, o número de células CD4 aumentou de 89 para 394. Enquanto, o remédio proporcionou uma redução na quantidade de carga de vírus (Viral Load) de 182,000 para 2,185 após semanas de uso do medicamento. Ou seja, através de citações numéricas/quantitativas o usuário cita sobre a eficiência do remédio no tratamento de AIDS e também sobre sua condição clínica antes de utilizar o medicamento.

Entretanto, verificamos que o algoritmo proposto neste trabalho não está habilitado para tratar relações com dados quantitativos. Analisamos que para o algoritmo atender esta deficiência é necessário investigar caminhos de dependência baseado em tipos numéricos. A lista de relações sintáticas Universais adotada neste trabalho, e implementado na biblioteca *Stanford CoreNLP*, possui a relação sintática *NumType*[32], o que poderia atender a esta lacuna.

No Apêndice A apresentamos a lista de pares de opinião mais frequentes e relevantes extraídos na tarefa de extração realizada pelo método proposto. Também é apresentada a lista de aspectos mais frequentes e a lista de termos opinativos mais frequentes.

Mediante os resultados apresentados consideramos os novos caminhos de dependência propostos relevantes ao domínio de opiniões sobre medicamentos. O algoritmo demonstrou resultados promissores para o problema de extração de aspecto e termos opinativos.

Ressaltamos que este algoritmo não requer rotulagem de dados para treinamento de modelos. Ou seja, este método não requer uma quantidade suficiente de dados anotados para funcionar corretamente, tem baixo custo de processamento e atende conjuntos de dados de grandes e pequenas proporções, como também extrair aspectos que possuem baixa frequência que podem ser relevantes ao domínio. Também destacamos que o algoritmo tem como base as relações Universais[33] de caminhos de dependências sendo facilmente possível adaptar o algoritmo para outras linguagens.

A fim de alcançar melhores resultados na precisão de cada caminho de dependência, estratégicas de filtragem e ranking de aspectos podem ser adotadas como

[32] http://universaldependencies.org/u/feat/NumType.html
[33] http://universaldependencies.org/

melhoria ao algoritmo, através do uso de léxicos de domínio, bases de conhecimento específico, léxicos de sentimento e relações de sinônimos e antônimos de termos com o objetivo de filtrar aspectos não-específicos ou irrelevantes ao domínio extraídos pelo algoritmo.

5.3 Experimentos – Classificação de Aspectos

Nesta seção, inicialmente foi avaliado o desempenho preditivo do algoritmo *Random Forest* e também foi medido o impacto dos conjuntos de recursos adotados. Para isso, o *Random Forest* foi avaliado com validação cruzada, *10-fold cross-validation*, usando diferentes conjuntos de características, como apresentado na Tabela 5.8 (primeira coluna). Em cada experimento, um novo conjunto de características é incluído para aumentar a descrição dos pares de opinião. O conjunto mais simples é o conjunto *Postag*, que é considerado como conjunto *baseline* para comparação.

Como pode ser visto na Tabela 5.8, para todos os conjuntos de dados, um ganho no desempenho foi observado quando o conjunto de características *MetaMap* foi incluído ao modelo de treinamento em comparação quando apenas utilizado o conjunto *Postag*. Embora melhores resultados possam ser alcançados incluindo outros conjuntos de recursos, resultados significantes que superam a combinação *Postag* + *Metamap* não foram verificados.

Conjunto de Característica	**ADHD**	**AIDS**	**Anxiety**
Postag	72,87	74,75	72,89
Postag + Metamap	76,08	77,34	75,9
Postag + Metamap + Medtagger	75,85	76,75	76
Postag + Metamap + Medtagger + Siders	76,31	77,4	76,25
Postag + Metamap + Medtagger + Siders + Negex	**76,86**	**78,16**	**76,78**
Postag + Metamap + Medtagger + Siders + Negex + Lemma	75,79	76,58	75,77
Postag + Metamap + Medtagger + Siders + Negex + Lemma + Stemming	72,69	73,38	72,97

Tabela 5.8: Pares de opinão corretamente classificados (em %)

Os melhores valores de precisão foram obtidos quando os conjuntos de *Postag, Metamap, Medtagger, Siders e Negex* foram considerados no modelo. No entanto, apenas um ganho muito pequeno no desempenho foi obtido sobre *Postag + MetaMap*.

Em alguns casos, incluir mais recursos pode até prejudicar a precisão, o que foi especialmente observado quando o conjunto de características *Stemming* foi adotado. O classificador não reagiu bem na predição com a remoção sufixos provenientes de flexão ou derivação aplicados pelo *Stemming*. Alguns termos resultam estrutura similar, o que pode ter confundido o classificador, como por exemplo, os termos "medicine" e "medication" tem *Stemming* resultante *"medic" "medicin".*

Uma razão para a alta precisão usando *MetaMap* é que este recurso, além de identificar termos médicos em grande cobertura, também abrange termos quantitativos, qualitativos e temporais como *30 mg, better, great, daily*, que são bastante frequentes no conjunto de dados da experimento.

Apresentamos na Tabela 5.9 os resultados de Precisão, Cobertura e F-Measure para cada tipo de aspecto observado no conjunto de características que resultou melhor desempenho na tabela 5.8. A seguir nas Tabelas 5.10, 5.11 e 5.12, destacamos a matriz de confusão resultante do melhor conjunto de características apresentado na tabela 5.8 para as bases ADHD, AIDS e Anxiety respectivamente para que possamos avaliar os resultados de desempenho citados na tabelas 5.9.

Observando cada matriz, o classificador falhou principalmente na classificação de "ADRs", "Condições" e "Eficácia". O erro de predição para aspectos "ADR" foi em maioria classificado como "Condição" ou "Eficácia", aspectos "Condição" em maioria classificado como "Eficácia" ou "ADR", e "Eficácia" foi classificada como "Condição" ou "ADR".

Aspecto	**ADHD**			**AIDS**			**Anxiety**		
	P	**C**	**F**	**P**	**C**	**F**	**P**	**C**	**F**
E	0,75	0,86	0,80	0,73	0,85	0,79	0,74	0,85	0,79
C	0,68	0,47	0,55	0,82	0,71	0,76	0,753	0,67	0,71
ADR	0,77	0,77	0,77	0,83	0,72	0,77	0,764	0,66	0,70
D	0,95	0,88	0,91	0,80	0,80	0,80	0,91	0,88	0,89

Tabela 5.9: Precisão, Cobertura e F-Measure

Observamos que uma das razões para indução do classificador ao erro é a ocorrência de termos opinativos de opinião repetidos para estes tipos de aspectos, como os termos, *low*, *lower*, *higher*, *declined*, *stopped*, *lost*, *less*, entre outros. Por exemplo, observe os pares de opinião: <*lost; focus*> rotulado como "Condição", < *lost; appetite*> como "ADR" e <*lost; pounds*> como aspecto de "Eficácia", note que nos três exemplos de pares de opinião o termo de opinião citado é o mesmo, *lost*, e é referenciado para diferentes situações, porém seus aspectos não são suficientes para que o classificador diferencie classes distintas.

ADHD				
	ADR	**C**	**D**	**E**
ADR	**1391**	121	3	290
C	185	**571**	4	455
D	4	3	**586**	72
E	211	142	20	**2470**

Tabela 5.10: Matriz de confusão do conjunto ADHD

AIDS				
	ADR	**C**	**D**	**E**
ADR	**313**	10	0	111
C	22	**375**	10	121
D	0	4	**66**	12
E	41	66	6	**684**

Tabela 5.11: Matriz de confusão do conjunto AIDS

Anxiety				
	ADR	**C**	**D**	**E**
ADR	**602**	106	2	202
C	85	**672**	13	225
D	0	6	**397**	48
E	101	109	21	**1366**

Tabela 5.12: Matriz de confusão do conjunto Anxiety

Também foi observado que um mesmo par de opinião pode ocorrer como "Condição" ou "ADR", e até mesmo em alguns casos como "Eficácia", como nas sentenças a seguir:

(1) "I had constant headache and horrible migraines which I am treated for by a neurologist."

(2) "The generic did nothing for me except to make me agitated/jittery for few hours, then crashed and had a headache."

Nas duas sentenças, (1) e (2), é citada a mesma referencia de par de opinião <had; headache>, porém na sentença (1) o par de opinião é descrito pelo usuário como uma "Condição", já na sentença (2) é descrito como uma "ADR". Para este par de opinião, e qualquer outro par de opinião que ocorra em diferentes tipos de aspectos, o conjunto de características será o mesmo. Logo, o classificador é induzido ao erro pois não há parâmetros que indiquem diferença entre estes aspectos. Desta forma, o classificador tende a classificar pela classe majoritária do par de opinião em questão.

Outro erro a mencionar é sobre pares de opiniões "falso negativo". Considere a seguinte sentença (3):

(3) "I was more satisfied with this one than any of the medicines. I never had headaches (with some I had terrible headaches!) the time release worked very well."

Nesta sentença (3), é extraído o par de opinião <never had; headaches> no qual o usuário indica uma "Eficácia" do medicamento, mas o modelo classificou este par de opinião como "Condição". Ressaltamos que o uso do recurso *Negex* não é suficiente para resolver o problema de citações de termos de negação que invertem o sentido de um par de opinião.

Todos esses exemplos observados estão induzindo o classificador a incorretas predições de tipos de aspectos. Alguns desses erros podem ser tratados incluindo outros tipos recursos para descrever os pares de opinião. Além disso, pode ser relevante analisar termos frequentes que ocorrem antes e depois de cada par de opinião para identificar possíveis padrões específicos.

No trabalho de SAMPATHKUMAR et al. (2014) foi definido uma lista de palavras chaves e frases que denotam a relação causal de um medicamento que causa um efeito colateral. São exemplos, *Caused by Caused by, Have been getting, Made me feel,* entre outros. Esta lista poderia ser uma possível entrada adicional ao modelo para distinguir tipos de aspectos "ADR".

No entanto, olhando para a frase anterior (2), a lista de termos descritos neste documento, SAMPATHKUMAR et al. (2014), não seria uma premissa válida para distinguir o aspecto <had; headache> ADR citado nesta sentença do tipo de aspecto "Condição", pois não há ocorrência de nenhuma palavra chave apresentada na lista que antecede este pair de opinião.

Outro detalhe observado é que pacientes usam frequentemente diferentes verbos para descrever diferentes tipos de aspectos. Por exemplo, os termos like *diagnosed, suffering* and *have* são geralmente utilizados para descrever o aspecto "Condição", frases como ocorrem são: *I have been diagnosed, I was suffering, I have ADHD, I have psychotic symptoms.*

Para o aspecto "ADR" termos como *felt* e *have been* ligado a outro verbo são frequentemente citados, por exemplo, *I have been losing my appetite, I felt very confused, I felt suicidal.* Já os verbos *work* e *help* são comumente usados para descrever a "Eficácia" do medicamento, são exemplos, *It helps me concentrate, Adderall really helps me focus, Didn't work well, Concerta worked wonders.*

Enquanto para o tipo de aspecto "Dosagem" os verbos mais comuns são: *prescribed, take, taken, up, upping* e *took.* Uma análise mais precisa poderia ser realizada para levantar quais os verbos que mais ocorrem com cada tipo de aspectos, a fim de levar uma lista de verbos comum a cada tipo de aspectos com o objetivo de incluir ao modelo de classificação como conjunto de características adicional.

Para além das limitações descritas acima, temos outras várias limitações deste estudo a citar que podem interferir no desempenho do classificador. Como já citamos anteriormente, erros de *parsing* podem ocorrer quando sentenças estão gramaticalmente escritas erradas ou incompletas. Erros ortográficos e gramaticais não corrigidos na etapa de pré-processamento induzem uma série de questões na análise gramatical.

O desempenho de método supervisionado é altamente dependente da qualidade e quantidade de dados de treinamento rotulados manualmente. E o *MetaMap* e outros recursos de domínios não retornam ou detectam a característica de certos termos corretamente, além de vários termos não terem sido detectados por esses recursos.

Ressaltamos que modelo de classificação apresentado neste trabalho apresentou resultados significantes para o problema de classificação de tipo de aspecto em comentários de medicamentos. E, outros recursos de domínio devem ser avaliados para melhorar a qualidade de desempenho do classificador a fim de amenizar o *trade-off* ocorrido na predição de tipos de aspectos "Condição", "Eficácia" e "ADR".

5.4 Considerações finais

Neste capítulo foram apresentados os resultados obtidos para as tarefas de extração e classificação de aspecto propostas nesta tese. Experimentos foram aplicados em três bases de citações sobre medicamentos referentes às condições clínicas "ADHD", "AIDS" e "Anxiety".

Para o problema de extração foi proposto adaptar o algoritmo *Aspectator* definido no trabalho de BANCKEN et al. (2014) para atender limitações observadas no algoritmo para o domínio da medicina. Os novos caminhos de dependência apresentados demonstraram relevantes e eficientes ao domínio. O modelo atingiu precisão acima de 70% para as três bases de dados e atingiu maior percentual de cobertura comparado aos modelos de *baseline*.

O algoritmo proposto para extração de aspectos e termos opinativos é simples de implementar, não exige rotulagem de dados para treinamento e atende bases de dados de qualquer tamanho. Além, de ter baixo custo de processamento e ser facilmente adaptável para outras línguas. Conforme discutido na seção 2.2.1.2, o algoritmo tende a extrair muitos pares de opinião que não são significativos ao domínio. Para atingir melhores resultados de precisão é necessário incorporar ao método estratégicas para filtragem ou ranking de aspectos relevantes ao domínio.

Para o problema de classificação foi proposto a adoção de um classificador supervisionado baseado nas características linguísticas *POS Tagging*, lematização,

radicalização e tratamento de negação, e o nos recursos específicos de domínio *MetaMap*, *SIDER* e *MedTagger*. O objetivo do classificador é categorizar pares de aspectos em um dos tipos de aspectos: Condição clínica, Reação Adversa, Dosagem e Eficácia.

O classificador é treinado utilizando o algoritmo *Random Forest* com validação cruzada. Para a classificação também foi atingido precisão acima de 70% para as três bases de experimentos.

Melhorias na precisão de predição de tipos ocorreu ao incorporar os dados do *MetaMap* ao conjunto de características. E, o modelo proposto atingiu maior precisão, 76,86% para ADHD, 78,16% para AIDS e 76,78% para Ansiedade, a partir do conjunto de características *Postag*, *Metamap*, *Medtagger*, *Siders* e *Negex*.

Os resultados demonstram que para atingir melhor resultado de precisão é necessário adoção de outros tipos de dados de características a fim de amenizar o problema o *trade-off* discutido sobre a predição dos tipos de aspectos “Condição”, “Eficácia” e “ADR”.

O modelo proposto para extração e classificação de aspectos em comentários de medicamentos demonstrou-se eficiente e atingiu resultados significativos ao problema de extração e classificação de múltiplos tipos de aspectos.

No próximo capítulo apresentamos as conclusões a respeito do trabalho de tese aqui desenvolvido, bem como melhorias e trabalhos futuros.

6 CONCLUSÕES

Nesta tese, foi proposto e desenvolvido um novo método para extrair e classificar aspectos em comentários opinativos sobre medicamentos. O método implementa as etapas de pré-processamento de textos e extração de pares de opinião, no qual um par de opinião corresponde um termo de aspecto e um termo opinativo sobre o aspecto. E, posteriormente pares de opinião são classificados a partir de um modelo supervisionado quanto aos tipos de aspectos "ADR", "Condição", "Eficácia" e "Dosagem".

Inicialmente, foi estendido o algoritmo *Aspectator* sugerindo novos caminhos de dependência para extrair pares de opinião relevantes ao domínio médico. Cada novo caminho foi testado em três conjuntos de citações de usuários sobre uso de medicamentos referente às condições clínicas "ADHD", "AIDS", "Anxiety".

A solução proposta obteve resultados melhores em comparação com os métodos de *baseline.* Os valores mais elevados de *F-Measure* foram observados em os conjuntos de dados de experimento.

Também foi apresentado um modelo de classificação de aspectos baseado em um conjunto de características semânticas e de domínio, e no classificador *Random Forest*. A avaliação do conjunto de dados mostra que são necessárias investigações mais aprofundadas e análise de outras formas de melhorar o desempenho do sistema.

Nas próximas seções, 6.1 e 6.2, serão apresentados as contribuições, limitações e trabalhos futuros deste trabalho respectivamente. Em seguida, as considerações finais.

6.1 Resumo das Contribuições

A seguir apresentamos um resumo das principais contribuições deste trabalho de doutorado:

- *Resumo das principais abordagens e tarefas associadas à Mineração de Opinião no domínio da Medicina:* Os documentos clínicos refletem o estado de saúde do paciente em termos de observações e informações, como descrição dos resultados de exames, diagnósticos, uso de medicamentos, entre outros. Avaliar adequadamente essas informações,

avaliar resultados clínicos positivos ou negativos ou julgar o impacto de uma condição médica para o bem-estar do paciente são essenciais. No capítulo 3, foi apresentado o estado da arte sobre Mineração de Opinião no contexto da Medicina. Embora métodos de Mineração de Opinião tenham sido desenvolvidos para abordar diversos contextos ainda não se encontraram em ampla aplicação no domínio médico. No capítulo 3, é fornecido também uma visão geral dos avanços recentes em farmacovigilância conduzida pela aplicação de Mineração de Opinião e são apresentados alguns trabalhos relacionados.

- *A apresentação de recursos de domínio aplicáveis a tarefas associadas à Mineração de Opinião no contexto da Medicina*: Na seção 3.2, descrevemos algumas ferramentas e recursos léxicos que dispõe informações sobre itens lexicais, como termos específicos de domínio da medicina e farmacêutica, que podem contribuir para a extração e classificação de opinião.

- *Conjunto de dados rotulados*: Produção de um conjunto de dados composto de comentários de usuários sobre o uso de medicamentos etiquetados por tipo de aspectos (Eficácia, Condição Clínica, ADR e Dosagem), dos quais podem ser adotadas para novos experimentos.

- *Um novo método não supervisionado para extração de opinião em comentários sobre medicamentos*: Nesta tese, foi implementado um modelo linguístico com intuito de extrair pares de opinião em textos opinativos sobre medicamentos. O algoritmo tem como base o algoritmo *Aspectator* onde caminhos de dependência representados em árvore sintática são adotados para extrair aspectos e termos opinativos a partir da relação de combinações de caminhos de dependência. Novas combinações de caminhos de dependência foram implementadas com o objetivo de adicionar adaptações para atender problemas específicos na extração de termos no domínio da medicina. O que não são tratados em trabalhos anteriores, assim como também trabalhos anteriores não focaram na extração de pares de opinião em comentários textuais sobre medicamentos. O modelo de extração, apresentado na seção 4.2, foi

executado em três diferentes bases de comentários de medicamentos sobre as condições clínicas “ADHD”, “AIDS” e “Anxiety”. O algoritmo de extração retornou resultados satisfatórios nas diferentes bases. Destacamos que a solução proposta pode ser facilmente adaptada a outras línguas, uma vez que não exige dados rotulados e possui baixo custo computacional.

- *Modelo de classificação de tipos de aspectos em textos sobre medicamentos*: Para a tarefa de classificação, um algoritmo de aprendizagem supervisionada é adotado para classificar pares de opinião entre quatro tipos de aspectos que são bastante frequentes em comentários textuais sobre medicamentos: “Condição”, “ADR”, “Dosagem” ou “Eficácia”. Durante o desenvolvimento desta pesquisa foi verificado que trabalhos anteriores apenas focaram na classificação de ADRs em comentários sobre remédios. Experimentos foram realizados em três conjuntos de dados relacionados a diferentes doenças e medicamentos comercializados.

6.2 Limitações e trabalhos Futuros

Inicialmente, é considerado como limitação nesta tese o fato do conjunto de experimentos adotado ter sido etiquetado manualmente apenas por um anotador. Uma vez que, a rotulação de dados manualmente é passível de erros.

Para certificar a qualidade da etiquetagem de rótulos a um conjunto de dados é necessário que seja realizada de forma separada e independente por mais de um anotador. E, em seguida as anotações de cada anotador devem ser comparadas e caso, para cada rótulo, houver a mesma concordância entre os anotadores, este rótulo é assumido para o conjunto. Caso haja discordância, os anotadores podem discutir e chegar a um denominador comum ou também é possível adotar a medida Cohen’s Kappa (COHEN, 1960), que corresponde a um método estatístico para avaliar o nível de concordância. Sendo assim, é possível gerar um conjunto de dados etiquetado confiável e certificado para experimentos.

O algoritmo de extração de aspecto proposto no capítulo 4, seção 4.2, foi utilizado para avaliar a extração de pares de opinião em documentos opinativos sobre comentários de uso de medicamentos utilizando regras linguísticas de caminho de dependência sintática, no qual atingiu bons resultados para a extração de termos. No entanto pontos de melhorias podem ser trabalhados no algoritmo para otimizar a sua precisão.

O algoritmo de extração proposto pode retornar muitos aspectos "não específicos" e "irrelevantes". Como possível tarefa de melhoria, estratégicas de ranking e filtragem de aspectos podem ser adotadas para minimizar este problema. Algoritmos específicos de ranking de características podem ser adotados, como também a filtragem de termos através do uso de léxicos de domínio, bases de conhecimento específico, léxicos de sentimento e relações de sinônimos e antônimos de termos, entre outros.

Conforme já citado, a gramática usada em frases de comentários pode ser bastante incoerente. Se as frases têm estruturas gramaticais inválidas ou fora do padrão, o analisador sintático não será capaz de extrair relações de dependência relevantes dessas sentenças, e pode até mesmo modificar ou indicar falsas dependências. Outras ferramentas de verificação ortográfica, assim como ferramentas alternativas de análise sintática podem ser adotadas para validar o processo. Assim como também métodos alternativos de extração de relação podem ser avaliados como melhoria ao processo de extração.

Propomos como trabalho futuro, como recurso adicional ao algoritmo de extração, habilitar o algoritmo a identificar aspecto que estão citados por termos "it" em uma sentença e que se refere a um aspecto citado na sentença anterior. Isto se dá através do tratamento de *Anaphora resolution* or *Coreference resolution* que descreve a tarefa de localizar todas as expressões que refere a uma entidade em um texto. *Anaphor* corresponde à repetição de uma palavra ou frase em sucessivas sentenças, linhas ou versos. Observe a sentença, "Atarax works great and it is not addicting", a segunda cláusula na sentença possui o termo "it" no qual referencia o termo "Atarax" na primeira cláusula. No algoritmo original, apenas o par "opinion pair" {works great, Atarax} será extraído, porém é possível inferir a extração do par {not addiction, Atarax}. O elemento linguístico a qual se refere ou a qual representa é o seu antecedente. Logo na sentença "Atarax" é o antecedente do termo "it".

Também como trabalho futuro, é proposto adicionar ao algoritmo tratamento de dados do tipo numérico para habilitar o algoritmo a tratar o problema de sentenças que apresentam citações opinativas utilizando dados quantitativos.

O algoritmo de classificação de aspecto proposto no capítulo 4, seção 4.3, foi desenvolvido para tratar o problema de predição de tipos de aspectos sobre pares de opinião extraídos em comentários sobre o uso de medicamento. O algoritmo atingiu taxa de acerto bastante significante e relevante ao domínio. A principal desvantagem deste método é que, ao contrário dos métodos não supervisionados, este método requer uma quantidade suficiente de dados anotados para funcionar corretamente. O modelo de classificação foi aplicado em três bases de experimentos onde foi levantado na seção 5.3 várias questões a serem investigadas para melhoria de performance do modelo.

Como investigação futura para o problema de classificação, é proposto analisar outros recursos de domínio específico que possam ser agregados ao conjunto de características do modelo. Assim como também analisar a estrutura de sentenças, bem como identificar padrões de sentenças e termos que precedem e antecedem termos aspectos e opinativos. O objetivo é localizar parâmetros adicionais ao modelo para representar aspectos do tipo “Condição”, “ADR” e “Eficácia” e reduzir a taxa de erros do classificador sobre estes tipos de aspectos.

Também consideramos como investigação futura, analisar a aplicabilidade de outros métodos de classificação para o problema de classificação de pares de opinião, sugerindo novos métodos de aprendizagem supervisionada ou métodos híbridos que integrem modelos supervisionados e não supervisionados e novos recursos léxicos para alcançar melhorias nos resultados.

Da mesma forma, como trabalho futuro, a construção de outros conjuntos de dados rotulados pode ser considerada, agregando outros rótulos de medicamentos e condições clínicas, com a finalidade de incorporá-los a novos experimentos e também disponibilizar para comunidade acadêmica.

Por fim, a análise de sentenças comparativas, onde são bastante frequentes em textos sobre medicamentos (exemplos, “Klonopin works longer than the Xanax”, “Found the brand name was by far more effective than generic alprazolam.”, “I'd say

that Adderall XR is better than Adderall IR") pode ser considerada para analisar a "Eficácia" sobre fármacos.

6.3 Considerações Finais

A automatização de sistema para extração e classificação de opinião em documentos textuais sobre comentários de medicamento é um tema importante, porém ainda há vários desafios a serem superados nesta área. Nesta tese foram resumidos algumas tarefas, métodos, aplicações e recursos léxicos associados à Mineração de Opinião no contexto da Medicina. E, também foi apresentado como contribuição um novo modelo de extração e classificação de pares de opinião que adota método não supervisionado com recursos linguísticos para o problema de extração e recursos de domínio e emprego de método supervisionado para o problema de classificação de múltiplos tipos de aspecto. Esperamos que esta tese possa ser referência como fonte de consulta e estudo a outras pessoas, e que o modelo implementado neste trabalho possa ser aproveitado por outros pesquisadores e estudantes da área.

Referências

ALGHUNAIM, A. **A Vector Space Approach for Aspect Based Sentiment Analysis**. *Ph.D. thesis*, Massachusetts Institute of Technology, 2015.

ALI, T. et al. **Can I Hear You? Sentiment Analysis on Medical Forums**. *Proceedings of the sixth international joint conference on natural language processing*, IJCNLP, p.667–673, 2013.

ARONSON, A. R.; LANG, F.-M. **An overview of MetaMap: Historical Perspective and recent Advances**. *Journal of the American Medical Informatics Association*, v.17, n.3, p.229, 2010.

ASGHAR, M. Z. et al. **Subjectivity Lexicon Construction for Mining Drug Reviews**. *Science International*, v.26, n.1, 2014a.

ASGHAR, M. Z. et al. **Medical Opinion Lexicon: An Incremental Model for Mining Health Reviews**. *International Journal of Academic Research*, v.6, n.1, p.295–302, 2014b.

ASGHAR, M. Z. et al. **SentiHealth: Creating Health-Related Sentiment Lexicon Using Hybrid Approach**. *SpringerPlus*, v.5, n.1, p.1139, 2016.

ATOLE, A. B. **Survey on Extractive Summary Generation from Opinion Targets and OpinionWords with Word Alignment Model**. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, v.4, n.11, 2015.

BANCKEN, W. et al. **Automatically Detecting and Rating Product Aspects from Textual Customer Reviews**. *Proceedings of the 1st International Workshop on Interactions between Data Mining and Natural Language Processing at ECML/PKDD*, p.1–16, 2014.

BECKER, K.; TUMITAN, D. **Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios**. *Simpósio Brasileiro de Banco de Dados*, p. 27-52, 2013.

BEHERA, P. N.; ELURI, S. **Analysis of Public Health Concerns using Two-step Sentiment Classification**. *International Journal of Engineering Research & Technology*, v.4, 2015.

BIYANI, P. et al. **Co-Training Over Domain-Independent and Domain-Dependent Features for Sentiment Analysis of an Online Cancer Support Community**. *International Conference On Advances In Social Networks Analysis And Mining*, IEEE/ACM, p.413–417, 2013.

BLEI, D. M. et al. **Latent dirichlet allocation**. *Journal of machine Learning research*, v.3, p.993–1022, 2003.

BODENREIDER, O. **The Unified Medical Language System (UMLS): Integrating Biomedical Terminology**. *Nucleic acids research*, v.32, n.suppl 1, p.D267–D270, 2004.

BORNEBUSCH, F. et al. **iTac: Aspect Based Sentiment Analysis Using Sentiment Trees and Dictionaries**. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval@COLING 2014)*, p.351, 2014.

BREIMAN, L. **Random Forests**. *Machine learning*, v.45, n.1, p.5–32, 2001.

BRODY, S.; ELHADAD, N. **An Unsupervised Aspect-sentiment Model for Online Reviews**. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p.804–812, 2010.

BRUN, C. et al. **XRCE: Hybrid Classification for Aspect-Based Sentiment Analysis**. *SemEval@COLING*, p.838–842, 2014.

CAMBRIA, E. et al. **SenticNet: A Publicly Available Semantic Resource for Opinion Mining**. *AAAI Fall Symposium: Commonsense Knowledge*, v.10, 2010.

CAMBRIA, E. et al. **Sentic PROMs: Application of Sentic Computing to The Development of a Novel Unified Framework for Measuring Health-Care Quality**. *Expert Systems with Applications*, v.39, n.12, p.10533–10543, 2012.

CASCAVEL, P.-C. de. **Mineração de Opinião Baseada em Extração de Aspectos**. *Monografia, Universidade Estadual do Oeste do Paraná*, 2016.

CASTAÑEDA, Y. et al. **UMCC_DLSI_Prob: A Probabilistic Automata for Aspect Based Sentiment Analysis**. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval@COLING 2014)*, p.722–726, 2014.

CAVALCANTI, D.; PRUDENCIO, R. **Aspect-based Opinion Mining in Drug Reviews**. *The 18th EPIA Conference on Artificial*, 2017.

CAVALCANTI, D.; PRUDENCIO, R. **Unsupervised Aspect Term Extraction in Online Drugs Reviews**. *The 30th International Florida Artificial Intelligence Research Society Conference*, 2017.

CHE, W. et al. **Sentence Compression for Aspect-Based Sentiment Analysis**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v.23, n.12, p.2111–2124, 2015.

CHEN, H.; ZIMBRA, D. **AI and Opinion Mining**. *IEEE Intelligent Systems*, v.25, n.3, p.74–80, 2010.

CHENG, V. et al. **Drug Review Mining with Regressional Probabilistic Principal Component Analysis**. *ACM SIGKDD Workshop on Health Informatics (HI-KDD)*, v.12, 2012.

CHENG, V. et al. **Probabilistic Aspect Mining Model for Drug Reviews**. *IEEE transactions on knowledge and data engineering*, v.26, n.8, p.2002–2013, 2014.

COHEN, Jacob. **A coefficient of agreement for nominal scales**. *Educational and psychological measurement*, v. 20, n. 1, p. 37-46, 1960.

DANIULAITYTE, R. et al. **"When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets**. *JMIR Public Health and Surveillance*, v.2, n.2, 2016.

DE CLERCQ, O. et al. **LT3: Applying Hybrid Terminology Extraction to Aspect-Based Sentiment Analysis**. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p.719–724, 2015.

DENECKE, K. **Sentiment Analysis from Medical Texts**. *Health Web Science*. Springer, p.83–98, 2015.

DENECKE, K. **Information Extraction from Medical Social Media**. *Health Web Science*. Springer, p.61–73, 2015.

DENECKE, K.; DENG, Y. **Sentiment Analysis in Medical Settings: New Opportunities and Challenges**. *Artificial intelligence in medicine*, v.64, n.1, p.17–27, 2015.

DENECKE, K.; NEJDL, W. **How Valuable is Medical Social Media Data? Content Analysis of The Medical Web**. *Information Sciences*, v.179, n.12, p.1870–1880, 2009.

DENG, Y. et al. **Retrieving Attitudes: Sentiment Analysis from Clinical Narratives**. *Workshop on Medical Information Retrieval (MEDIR@ SIGIR)*, p.12–15, 2014.

DING, X. et al. **Dynamic Topic Detection Model by Fusing Sentiment Polarity**. *Proceedings of the 38th Australasian Computer Science Conference* (ACSC 2015), v.27, p.30, 2015.

EBRAHIMI, M. et al. **Drug Side Effect Detection as Implicit Opinion from Medical Reviews**. *Research Journal of Applied Sciences, Engineering and Technology*, p. 1086-1094, 2015.

EGGER, D. et al. **Adverse Drug Reaction Detection using an Adapted Sentiment Classifier**. *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing (PSB)*, 2016.

ESULI, A. **Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications**. *ACM SIGIR Forum*, v.42, n.2, p.105–106, 2008.

ESULI, A.; SEBASTIANI, F. **SentiWordnet: A High-Coverage Lexical Resource for Opinion Mining.** *Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR)*, p.1–26, 2007.

FELDMAN, R. **Techniques and Applications for Sentiment Analysis**. *Communications of the ACM*, v.56, n.4, p.82–89, 2013.

FELDMAN, R. et al. **Utilizing Text Mining on Online Medical Forums to Predict Label Change due to Adverse Drug Reactions**. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.1779–1788, 2015.

GOEURIOT, L. et al. **Textual and Informational Characteristics of Health-Related Social Media Content: A Study of Drug Review Forums**. *Asia Pacific Conference Library & Information Education & Practice*, 2011.

GOEURIOT, L. et al. **Sentiment Lexicons for Health-Related Opinion Mining**. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI)*, p.219–226, 2012.

GOSAL, G. P. S. **Opinion Mining and Sentiment Analysis of Online Drug Reviews as a Pharmacovigilance Technique**. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*, v.3, p.4920–4925, 2015.

GREAVES, F. et al. **Use of Sentiment Analysis for Capturing Patient Experience from Free-Text Comments Posted Online**. *Journal of Medical Internet Research*, v.15, n.11, p.34, 2013.

GUHA, S. et al. **SIEL: Aspect Based Sentiment Analysis in Reviews**. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval@2015)*, p.759-766, 2015.

GUPTA, D. K. et al. **PSO-ASent: Feature Selection using Particle Swarm Optimization for Aspect Based Sentiment Analysis**. *International Conference on Applications of Natural Language to Information Systems*, p.220–233, 2015.

HAN, J. et al. **Data Mining: Concepts and Techniques**. *Morgan Kaufmann Publishers, Elsevier, 2012.*

HARPAZ, R. et al. **Text mining for Adverse Drug Events: The Promise, Challenges, and State of the Art**. *Drug safety*, v.37, n.10, p.777–790, 2014.

HTAY, S. S.; LYNN, K. T. **Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews**. The ScientificWorld Journal, v.3, p.1533–1541, 2013.

HU, M.; LIU, B. **Mining and Summarizing Customer Reviews**. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.168–177, 2004.

ISAH, H. et al. **Social Media Analysis for Product Safety Using Text Mining and Sentiment Analysis**. *14th UK Workshop on Computational Intelligence (UKCI)*, p.1–7, 2014.

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, p.1035–1045, 2010.

JAMES, G. et al. **An Introduction to Statistical Learning with Applications in R**. *Springer*, v.6, 2013.

JEYAPRIYA, A.; SELVI, C. K. **Extracting Aspects and Mining Opinions in Product Reviews Using Supervised Learning Algorithm**. *The 2nd International Conference on Electronics and Communication Systems (ICECS)*, p.548–552, 2015.

JI, X. et al. **Monitoring Public Health Concerns using Twitter Sentiment Classifications**. *IEEE International Conference on Healthcare Informatics (ICHI)*, p.335–344, 2013.

JIN, W. et al. **A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining**. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, p.465–472, 2009.

JIN, W. et al. **OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction**. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.1195–1204, 2009.

JO, Y.; OH, A. H. **Aspect and Sentiment Unification Model for Online Review Analysis**. *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*, p.815–824, 2011.

JONNAGADDALA, J. et al. **Binary Classification of Twitter Posts for Adverse Drug Reactions**. *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, p.4–8, 2016.

JUSOH, S.; ALFAWAREH, H. M. **Applying Information Extraction and Fuzzy Sets for Opinion Mining**. *Proceedings of the International Conference Image Processing, Computers and Industrial Engineering (ICICIE'14)*, 2014.

KHAN, K. et al. **Mining Opinion Components from Unstructured Reviews: A Review**. *Journal of King Saud University - Computer and Information Sciences*, v.26, n.3, p.258 – 275, 2014.

KIDO, G. S. et al. **Topic Modeling based on Louvain method in Online Social Networks**. *Proceedings in XII Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems: Information Systems in the Cloud Computing Era*, v.1, p.47, 2016.

KORKONTZELOS, I. et al. **Analysis of the Effect of Sentiment Analysis on Extracting Adverse Drug Reactions from Tweets and Forum Posts**. *Journal of Biomedical Informatics*, v.62, p.148–158, 2016.

KU, L.-W. et al. **Opinion Extraction, Summarization and Tracking in News and Blog Corpora**. *AAAI Symposium on Computational Approaches to AnalysingWeblogs* (AAAI-CAAW), p.100–107, 2006.

LAGUITAN, T. A. F. et al. **LikeOrNah?: Aspect Extraction from Movie Reviews using Rule-Based Approach**. *International Journal of Conceptions on Computing and Information Technology*, v.3, 2015.

LAL, M.; ASNANI, K. **Implicit Aspect Identification Techniques for Mining Opinions: a survey**. *International Journal of Computer Applications*, v.98, n.4, p.1–3, 2014.

LANCICHINETTI, A. et al. **A High-reproducibility and High-accuracy method for Automated Topic Classification**. *arXiv preprint arXiv:1402.0422*, 2014.

LI, F. et al. **Structure-aware Review Mining and Summarization**. *Proceedings of the 23rd International Conference on Computational Linguistics*, p.653–661, 2010.

LI, S. et al. **Improving Aspect Extraction by Augmenting A Frequency-based Method with Web-Based Similarity Measures**. *Information Processing & Management*, v.51, n.1, p.58–67, 2015.

LIN, C.; HE, Y. **Joint Sentiment/Topic Model for Sentiment Analysis**. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, p.375–384, 2009.

LIU, B. **Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data**. *Springer Science & Business Media*, 2007.

LIU, B. **Sentiment Analysis and Subjectivity**. *Handbook of Natural Language Processing*, Second Edition, 2010.

LIU, B. **Sentiment Analysis and Opinion Mining**. *Morgan & Claypool Publishers*, 2012.

LIU, H.; LI, D. et al. **Towards a Semantic Lexicon for Clinical Natural Language Processing**, *Annual Symposium of American Medical Informatics Association (AMIA)*, p.568-576, 2012.

LIU, Q.; GAO, Z. et al. **Automated Rule Selection for Opinion Target Extraction**. *Knowledge-Based Systems*, v.104, p.74–88, 2016.

MAATEN, L. et al. **Hidden-Unit Conditional Random Fields**. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (*AISTATS)*, v.15, p.479–488, 2011.

MAHADIK, A.; BHARAMBE, A. **Aspect Based Opinion Mining and Ranking: survey**. *International Journal of Current Engineering and Technology*, v.5, n.6, 2015.

MAJETHIA, R. et al. **PeopleSave: Recommending Effective Drugs Through Web Crowdsourcing**. *The 8th International Conference on Communication Systems and Networks (COMSNETS)*, p.1–6, 2016.

MANEK, A. S. et al. **Classification of Drugs Reviews using W-LRSVM Model**. *Proceedings of 12th IEEE India International Conference (INDICON 2015)*, p.1–6, 2015.

MANKAR, S. A.; INGLE, P. M. **Implicit Sentiment Identification using Aspect based Opinion Mining**. *International Journal of Science and Research (IJSR)*, v.4, n11, p.1731–1735, 2015.

MARRESE-TAYLOR, E. et al. **A Novel Deterministic Approach for Aspect-Based Opinion Mining in Tourism Products Reviews**. *Expert Systems with Applications*, v.41, n.17, p.7764–7775, 2014.

MCCOY, T. H. et al. **Sentiment Measured in Hospital Discharge Notes is Associated with Readmission and Mortality Risk: An Electronic Health Record Study**. *PloS one*, v.10, n.8, 2015.

MEDHAT, W. et al. **Sentiment Analysis Algorithms and Applications: A Survey**. Ain Shams Engineering Journal, v.5, n.4, p.1093–1113, 2014.

MEI, Q. et al. **Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs**. *Proceedings of the 16th International Conference on World Wide Web*, p.171–180, 2007.

MELZI, S. et al. **Patient's Rationale: Patient Knowledge Retrieval from Health Forums**. *The Sixth International Conference on eHealth, Telemedicine, and Social Medicine (TELEMED)*, 2014.

MERIS, S. R. et al. **Categorization of Drugs Based on Polarity Analysis of Twitter Data**. *Journal of Engineering and Applied Sciences (ARPN)*, v.11, n.13, 2016.

MILLER, G. A. **Wordnet: A Lexical Database for English**. *Communications of the ACM*, v.38, n.11, p.39–41, 1995.

MISHRA, A. et al. **Towards Automatic Pharmacovigilance: Analysing Patient Reviews and Sentiment on Oncological Drugs**. *IEEE International Conference On Data Mining Workshop (ICDMW)*, p.1402–1409, 2015.

MOGHADDAM, S.; ESTER, M. **Aspect-Based Opinion Mining from Product Reviews**. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p.1184–1184, 2012.

MONDAL, A. et al. **Lexical Resource for Medical Events: A Polarity Based Approach**. *IEEE International Conference on Data Mining Workshop (ICDMW)*, p.1302–1309, 2015.

MONDAL, A. et al. **A Hybrid Approach Based Sentiment Extraction from Medical Context**. *Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016*, p.35–40, 2016.

MONDAL, A. et al. **WME: Sense, Polarity and Affinity Based Concept Resource for Medical Events**. *Proceedings of the Eighth Global WordNet Conference*, p.242-246, 2016.

MOON, V.; BORKAR, P. **Extraction of Drug Reviews by Specific Aspects for Sentimental Analysis**. *International Journal of Computer Science and Information Technologies (IJCSIT)*, v.7, n.3, p.1186–1189, 2016.

MORE, P.; GHOTKAR, G. **A Study of Different Approaches to Aspect-based Opinion Mining**. *International Journal of Computer Applications*, v.145, n.6, 2016.

MUKHERJEE, S. et al. **People on Drugs: Credibility of User Statements in Health Communities**. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.65–74, 2014.

NA, J.-C. et al. **Sentiment Classification of Drug Reviews Using a Rule-Based Linguistic Approach**. *International Conference on Asian Digital Libraries (ICADL)*, p.189–198, 2012.

NA, J.C.; KYAING, W. Y. M. **Sentiment Analysis of User-Generated Content on Drug Review Websites**. *Journal of Information Science Theory and Practice*, v.3, n.1, p.6–23, 2015.

NEVES, B. V. A. **Detecção de Comunidades de Interesse em Microblogs por Meio de Modelagem de Tópicos**. *Dissertação de Mestrado, Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas*, 2016.

NIKFARJAM, A. et al. **Pharmacovigilance from Social Media: Mining Adverse Drug Reaction Mentions Using Sequence Labeling with Word Embedding Cluster Features**. *Journal of the American Medical Informatics Association*, v.22, n.3, p. 671-681, 2015.

NIU, Y. et al. **Analysis of Polarity Information in Medical Text**. *Proceedings of the American Medical Informatics Association (AMIA)*, 2005.

NIU, Y. et al. **Using Outcome Polarity in Sentence Extraction for Medical Question-Answering**. *Proceedings of the American Medical Informatics Association (AMIA)*, 2006.

NOFERESTI, S.; SHAMSFARD, M. **Using Linked Data for Polarity Classification of Patients' Experiences**. *Journal of biomedical informatics*, v.57, p.6–19, 2015.

NOFERESTI, S.; SHAMSFARD, M. **Resource Construction and Evaluation for Indirect Opinion Mining of Drug Reviews**. *PloS one*, v.10, n.5, 2015.

ONEATA, D. **Probabilistic Latent Semantic Analysis**. *Fifteenth Conference On Uncertainty*, p.1–7, 1999.

PATKI, A. et al. **Mining Adverse Drug Reaction Signals From Social Media: Going Beyond Extraction**. *Proceedings of BioLink-SIG*, v.2014, p.1–8, 2014.

PAUL, M. J. et al. **Social Media Mining for Public Health Monitoring and Surveillance**. *Pacific Symposium on Biocomputing (PSB)*, p.468–79, 2016.

PEKAR, V. et al. **UBham: Lexical Resources and Dependency Parsing for Aspect-Based Sentiment Analysis**. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval@COLING 2014)*, p.683, 2014.

PESTIAN, J. P. et al. **Sentiment Analysis of Suicide Notes: A Shared Task**. *Biomedical informatics insights*, v.5, n.Suppl. 1, p.3-16, 2012.

PONTIKI, M. et al. **Semeval-2014 Task 4: Aspect Based Sentiment Analysis**. *Proceedings of the 8th International Workshop on Semantic Evaluation*, p.27–35, 2014.

POPESCU, A.-M.; ETZIONI, O. **Extracting Product Features and Opinions from Reviews**. *Proceedings of the conference on human language technology and empirical methods in natural language processing*, p.339–346, 2005.

QIU, G. et al. **Opinion Word Expansion and Target Extraction Through Double Propagation**. *Association for Computational Linguistics*, v.37, n.1, p.9–27, 2011.

RANA, T. A.; CHEAH, Y.-N. **Hybrid Rule-Based Approach for Aspect Extraction and Categorization from Customer Reviews**. The 9th International Conference on IT in Asia (CITA), p.1–5, 2015.

RAVI, K.; RAVI, V. **A Survey on Opinion Mining and Sentiment Analysis: Tasks, approaches and applications**. *Knowledge-Based Systems*, v.89, p.14–46, 2015.

RILOFF, E.; WIEBE, J. **Learning Extraction Patterns for Subjective Expressions**. *Empirical Methods in Natural Language Processing (EMNLP)*, p.105–112, 2003.

ROKACH, L. et al. **Negation Recognition in Medical Narrative Reports**. *Information Retrieval*, v.11, n.6, p.499–538, 2008.

RUBTSOVA, Y.; KOSHELNIKOV, S. **Aspect Extraction Using Conditional Random Fields**. *SentiRuEval-2015*, 2015.

RUDER, S. et al. **A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis**. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

SAHANA, D. S.; GIRISH, L. **Automatic Drug Reaction Detection Using Sentimental Analysis**. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 2015.

SAMHA, A. K. **Aspect-Based Opinion Mining Using Dependency Relations**. *International Journal of Computer Science Trends and Technology (IJCST)*, v.4, n.1, 2016.

SAMPATHKUMAR, H. et al. **Mining Adverse Drug Reactions from Online Healthcare Forums using Hidden Markov Model**. *BMC Medical Informatics and Decision Making*, v.14, n.1, p.91, 2014.

SARKER, A. et al. **Outcome Polarity Identification of Medical Papers**. *Proceedings of Australasian Language Technology Association Workshop*, p.105–114, 2011.

SARKER, A. et al. **Utilizing Social Media Data for Pharmacovigilance: A Review**. Journal of Biomedical Informatics, v.54, p.202–212, 2015.

SARKER, A.; GONZALEZ, G. **Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-Corpus Training**. *Journal of Biomedical Informatics*, v.53, p.196–207, 2015.

SHARIF, H. et al. **Detecting Adverse Drug Reactions using a Sentiment Classification Framework**. *ASE International Conference on Social Computing (SOCIALCOM)*, 2014.

SHARIF, H. et al. **A Classification Based Framework to Predict Viral Threads**. *PACIS*, p.134, 2015.

SOKOLOVA, M.; BOBICEV, V. **What Sentiments Can Be Found in Medical Forums**? *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, p.633–639, 2013.

SOKOLOVA, M. et al. **How Joe and Jane Tweet about Their Health: Mining for Personal Health Information on Twitter**. *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, p.626–632, 2013.

SONDHI, P. et al. **Sympgraph: A Framework for Mining Clinical Notes Through Symptom Relation Graphs**. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.1167–1175, 2012.

STRAPPARAVA, C. et al. **Wordnet Affect: An Affective Extension of Wordnet**. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, v.4, p.1083–1086, 2004.

SU, Q. et al. **Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews**. *Conference: Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead, 21st International Conference on Computer Processing of Oriental Languages (ICCPOL)*, p.22–30, 2006.

SUTTON, C.; MCCALLUM. **An Introduction to Conditional Random Fields**. *Foundations and Trends® in Machine Learning*, v.4, n.4, p.267–373, 2012.

SWAMINATHAN, R. et al. **Opinion Mining for Biomedical Text Data: Feature Space Design and Feature Selection**. *The Nineth International Workshop on Data Mining in Bioinformatics (BIOKDD)*, 2010.

TANG, J. et al. **Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis**. *Proceedings of the 31st International Conference on Machine Learning*, v.32, p.190–198, 2014.

TELOKEN, A. et al. **Estudo Comparativo entre os Algoritmos de Mineração de Dados Random Forest e J48 na Tomada de Decisão**. *Simpósio de Pesquisa e Desenvolvimento em Computação*, v.2, n.1, 2016.

THOTA, G. N. S. P.; MEENAKUMARI, J. **Sentiment Analysis: Automation of Rating for Review Without Aspect Keyword Supervision**. *International Journal of Science and Research (IJSR)*, v.4, 2015.

TITOV, I.; MCDONALD, R. T. **A Joint Model of Text and Aspect Ratings for Sentiment Summarization**. *ACL*, v.8, p.308–316, 2008.

TOH, Z.; SU, J. NLANGP: **Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction**. *Proceedings of the 9th*

International Workshop on Semantic Evaluation (SemEval@2015), p. 496–501, 2015.

TOMAR, D.; AGARWAL, S. **A survey on Data Mining Approaches for Healthcare**. *International Journal of Bio-Science and Bio-Technology*, v.5, n.5, p.241–266, 2013.

VIVEKANANDAN, K.; ARAVINDAN, J. S. **Aspect-based Opinion Mining: A survey. International Journal of Computer Applications**, *International Journal of Computer Applications*, v.106, n.3, 2014.

WALLACH, H. **Efficient Training of Conditional Random Fields**. *Tese (Doutorado em Ciência da Computação) — Master's thesis, University of Edinburgh*, 2002.

WANG, H. **Sentiment-aligned Topic Models for Product Aspect Rating Prediction**. *Master of Science, Applied Sciences: School of Computing Science*, 2015.

WANG, H.; ESTER, M. **A Sentiment-aligned Topic Model for Product Aspect Rating Prediction**. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p.1192–1202, 2014.

WANG, W. et al. **Implicit Feature Identification via Hybrid Association Rule Mining**. *Expert Systems with Applications*, v.40, n.9, p.3518–3531, 2013.

WILEY, M. T. et al. **Pharmaceutical Drugs Chatter on Online Social Networks**. *Journal of Biomedical Informatics*, v.49, p.245–254, 2014.

WU, S. et al. **Negation's not Solved: Generalizability versus Optimizability in Clinical Natural Language Processing**. *PloS one*, v.9, n.11, p.e112774, 2014.

WU, Y. et al. **Phrase Dependency Parsing for Opinion Mining.** *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP '09), p.1533–1541, v.3, 2009.

XIA, L. et al. **Improving Patient Opinion Mining Through Multi-Step Classification**. *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, p.70–76, 2009.

XU, J. et al. **Citation Sentiment Analysis in Clinical Trial Papers**. *Proceedings of the American Medical Informatics Association (AMIA)*, v.2015, p.1334, 2015.

YANG, H.; YANG, C. C. **Mining a Weighted Heterogeneous Network Extracted from Healthcare-Specific Social Media for Identifying Interactions between Drugs**. *IEEE International Conference on Data Mining Workshop (ICDMW)*, p.196–203, 2015.

YAZDAVAR, A. H. et al. **Fuzzy Based Implicit Sentiment Analysis on Quantitative Sentences**, *Journal of Soft Computing and Decision Support Systems*, v.3, n.4, 2017.

YU, L. et al. **Topic Extraction Based on Product Reviews**. *Journal of Computational Information Systems*, v.9, n.2, p.773–780, 2013.

YUAN, B.; WU, G. **A Hybrid HDP-ME-LDA Model for Sentiment Analysis**. *2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)*, v.118, 2017.

ZENG, L.; LI, F. **A Classification-Based Approach for Implicit Feature Identification**. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, p.190–202, 2013.

ZHANG, L.; LIU, B. **Aspect and Entity Extraction for Opinion Mining**. Data Mining and Knowledge Discovery for Big Data. Springer, p.1–40, 2014.

ZHANG, Q. et al. **Opinion Mining with Sentiment Graph**. *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, v.1, p.249–252, 2011.

ZHANG, Y.; ZHU, W. **Extracting Implicit Features in Online Customer Reviews For Opinion Mining**. *WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web*, p.103–104, 2013.

ZHANG, Z. et al. **TJUdeM: A Combination Classifier for Aspect Category Detection and Sentiment Polarity Classification**. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p.772–777, 2015.

ZHAO, W. X. et al. **Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid**. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p.56–65, 2010.

ZHENG, X. et al. **Incorporating Appraisal Expression Patterns into Topic Modeling for Aspect and Sentiment Word Identification**. *Journal Knowledge-Based Systems*, v.61, n.1, p.29–47, 2014.

ZHOU, X. et al. **Representation Learning for Aspect Category Detection in Online Reviews**. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, p.417–424, 2015.

ZHUANG, L. et al. **Movie review mining and summarization**. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, p.43–50, 2006.

Apêndice A: Lista de Pares de Opiniões

	ADHD			AIDS			Anxiety		
	Aspecto	**Termo Opinativo**	**F**	**Aspecto**	**Termo Opinativo**	**F**	**Aspecto**	**Termo Opinativo**	**F**
1	mouth	dry	93	side effects	Never / had	41	anxiety	severe	62
2	life	changed	61	Cd4	climbed	17	anxiety	have	55
3	side effects	no / had	53	viral load	undetectable	13	side effects	No / had	41
4	ADHD	have	50	treatment	started	12	anxiety	social	40
5	medication	helped	47	hiv	positive	11	dose	hig /higher	34
6	effects	no / negative	28	side effects	no have	11	anxiety disorder	generalized	33
7	dose	too low	26	viral load	dropped	9	life	changed	31
8	feel	very energetic	23	issues	no had	8	dose	low	28
9	difference	noticed	21	nausea	had	7	life	saved	25
10	appetite	decreased	18	pill	once-a-day	7	panic attacks	severe	22
11	ADHD	severe	14	dreams	very / vivid	6	medicine	great	20
12	anxiety	extreme	12	life	saved	6	life	normal	19
13	headaches	bad	12	dreams	weird	5	anxiety	worse	19
14	difference	amazing	11	stomach	empty	5	nothing	did	17
15	depression	severe	9	headache	mild	5	feel	normal	16
16	improvement	made	9	medicine	effective	4	panic attacks	had	15
17	mouth	horrible	7	stool	loose	4	feel	calm	14
18	get	antsy	7	3glycerides	high	4	Stress disorder	traumatic	11
19	made	worse	5	fadigue	extreme	3	depression	severe	11
20	behavior	horrible	5	weight	ganeid	3	anxiety	increased	11

Tabela A. 1: Lista de pares de opiniões relevantes frequentes

	ADHD	F	AIDS	F	Anxiety	F
1	effects	471	effects	187	anxiety	772
2	medicine	212	viral load	162	effects	501
3	life	184	Cd4	47	medicine	302
4	medication	183	medicine	35	life	289
5	Vyvanse	175	dreams	32	dose	256
6	adderall	168	side effects	31	Klonopin	239
7	feel	142	Atripla	25	Xanax	216
8	dose	131	Complera	18	medication	214
9	anxiety	115	Stribild	18	feel	203
10	side effects	105	feel	17	time	161
11	ADHD	104	results	17	Panic attacks	138
12	difference	97	medication	15	day	122
13	Concerta	96	life	14	drug	108
14	mouth	84	treatment	13	depression	103
15	appetite	84	effect	13	Side effects	99
16	effect	74	pill	13	Ativan	98
17	drug	72	issues	11	buspar	94
18	medications	68	Hiv	11	anxiety disorder	78
19	feeling	65	dizziness	10	pill	77
20	Strattera	61	headache	9	attacks	74

Tabela A. 2: Lista de aspectos relevantes mais ocorridos

	ADHD	F	AIDS	F	Anxiety	F
1	had	327	had	97	had	424
2	have	204	undetectable	65	have	341
3	great	141	started	37	severe	199
4	high	118	have	31	prescribed	189
5	changed	92	good	16	few	154
6	took	90	great	14	great	149
7	made	88	positive	13	took	120
8	good	79	dropped	13	good	114
9	lost	68	slight	9	bad	106
10	increased	65	mild	9	made	91
11	better	63	normal	9	gave	72
12	bad	59	bad	9	helped	72
13	dry	57	diagnosed	9	worked	71
14	severe	53	took	9	normal	70
15	helped	49	high	9	changed	65
16	improved	47	vivid	8	saved	65
17	amazing	45	healthy	8	felt	64
18	worked	41	better	8	best	62
19	normal	40	horrible	7	daily	60
20	several	40	Severe	7	high	59

Tabela A. 3: Lista de termos opinativos relevanter mais ocorridos

Anexo A: Artigo (Cavalcanti; Prudêncio, 2017a)

Aspect-based Opinion Mining in Drug Reviews

Diana Cavalcanti[1] and Ricardo Prudêncio[1]

Centro de Informática, Universidade Federal de Pernambuco, Recife/PE, Brazil
dcc2,rbcp@cin.ufpe.br

Abstract. Aspect-based opinion mining can be applied to extract relevant information expressed by patients in drug reviews (e.g., adverse reactions, efficacy of a drug, symptoms and conditions of patients). This new domain of application presents challenges as well as opportunities for research in opinion mining. Nevertheless, the literature is still scarce of methods to extract multiple relevant aspects present in drug reviews. In this paper we propose a method to extract and classify aspects in drug reviews. The proposed solution has two main steps. In the aspect extraction, a method based on syntactic dependency paths is proposed to extract opinion pairs in drug reviews, composed by an aspect term associated to a sentiment modifier. In the aspect classification, a supervised classification is proposed based on domain and linguistics resources to classify the opinion pairs by aspect type (e.g., condition, adverse reaction, dosage and effectiveness). In order to evaluate the proposed method we conducted experiments with datasets related to three different diseases: ADHD, AIDS and Anxiety. Promising results were obtained in the experiments and various issues were identified and discussed.

Keywords: opinion mining; aspect extraction; aspect classification; natural language processing; machine learning; drugs reviews.

1 Introduction

Opinion mining is the process of detecting and analyzing subjective information, sentiments and opinionated aspects in large volumes of texts using computational methods [4]. Opinion mining has been applied in different domains, specially to evaluate the acceptance of products and services, as well as the general sentiment about people and brands. In our work, we are focused on opinion mining in drug reviews, in which patients express their experiences and opinions about treatments or medicines. This domain of application has received a great interest in recent years [1]. Specifically in pharmacovigilance, drug manufacturers can benefit from opinion mining since particular adverse reactions to a drug can be traced more quickly from public repositories or posts in social networks.

Previous works on opinion mining in drug reviews have focused on classifying the sentiment about a drug (e.g., positive or negative) expressed in the users' reviews [5, 6]. A more refined (and possibly useful) task is to identify the Adverse Drug Reactions (ADRs) mentioned in a drug review, which have been commonly reported by patients. ADRs are harmful reactions caused by medication intake

resulting in an intervention related to the use of the product. Due to various limitations in clinical trials, it is not possible to fully evaluate the consequences of using a particular drug before being released [1, 3, 2]. An ADR may not be detected before the product go to the market and can take some time after its sale to track new ADRs and relate them to the drug's label [9].

In this work, we focused on *aspect-based opinion mining* [4] in drug reviews. The aim is to identify in a drug review fragments of text that can be associated to specific aspects of interest like, adverse reactions, effectiveness, patient conditions, among others. The identification of ADRs can be seen as an example of task related to aspect-based opinion mining. Although ADR is in fact an important information to be identified, it is not the only aspect of interest mentioned in drug reviews. Patient conditions, for instance, once extracted, can be useful to reveal wrong uses of a medicine among patients. Other aspects can provide a better assessment of the reported patients' experience as well as a better summarization of the reviews associated to a drug. The extraction of multiple relevant aspects in drug reviews is a lack in the literature that we aim to focus on.

In the current work, a method for extracting and classifying aspects is proposed. In the aspect extraction, a linguistic method based on dependency paths in the syntactic tree of the reviews is adopted to extract *opinion pairs* (i.e., an aspect term and its opinion term). In the aspect classification, a supervised learning algorithm is adopted to classify opinion pairs in one of four aspects types: Condition, ADR, Dosage or Effectiveness. Previous work only focused on classifying ADRs. Experiments were performed in three datasets related to different diseases and drugs. The results revealed a gain in performance to extract relevant aspects (in terms of F-Measure) compared to previous work.

The current work filled in a gap in the literature by investigating and proposing effective methods for mining multiple aspects in drug reviews. We can mention the following contributions:

- Proposal of a method for extracting aspects in drug reviews based on syntactic dependency paths, with a good performance;
- Production of a corpus of drug reviews labeled by aspects, which can be adopted for new experiments;
- A supervised method based on domain resources and linguistic features for aspect type classification.

The remainder of this paper is organized as follows. Section 2 provides a brief introduction on opinion mining in drug reviews. In Section 3 we present the proposed approach. Section 4 presents the results of our experiments. Finally, Section 5 concludes the paper, along with recommendations for future work.

2 Opinion mining in drugs reviews

Recent studies have focused on mining reviews of drugs (available in social networks, forums, web,...) in order to provide useful information for pharmacovigilance. Two common tasks can be identified in the literature of opinion mining in drug reviews: (1) classification of reviews; and (2) extraction of opinion aspects.

Concerning the first task, a review can be automatically classified according to its sentiment about the drug (usually positive, negative or neutral) [5, 6]. Sentiment classification can be useful for filtering relevant reviews to be inspected (e.g., reviews classified as negative have a high chance of mentioning a negative side effect). A related task is to directly classify whether a review mentions an ADR or not [7, 10]. This is more specific and focused than simply classifying the sentiment of a review. The sentiment classification task can be dealt with both machine learning (e.g., [10, 7]) and knowledge-based approaches (e.g., [5, 6]).

The second task is to extract specific aspects in drug reviews. Hence, the aim is not only to classify whether a review mentions an ADRs, for instance, but also to extract the reaction itself or any other aspect considered as relevant [9]. In literature, this task is called *aspect-based opinion mining*, which aims to extract the main aspects mentioned about an item or entity and to provide the classification of the opinion given on every aspect [4]. According to [5, 6], six aspects (see examples in Table 1) are common in drug reviews:

- **Overall:** The general opinion of a medicinal product or when the clause is not mentioned any of the other five categories of aspects;
- **Effectiveness:** Changes noted after the use of the medicine, linked to the patient's condition or disease;
- **Side effects:** The reactions that are not related to the medicine. ADR is a negative side effect;
- **Dosage:** Reports the amount, frequency or the treatment period in which the medicine was used.
- **Condition:** Corresponds to a description of the patient's condition, e.g., a disease or health problems in general.
- **Cost:** Corresponds the cost/price of a medicine.

Table 1. Opinion sentence by aspects in drugs reviews

Aspect	**Opinion sentence**
Overall	Adderall is overall a **good** ADD **medicine**.
Effectiveness	It **helped** me stay **focused** on any tasks.
Side effects	I have only one side effect which is **dry mouth**.
Condition	I **had** clinical **depression/anxiety** for years.
Cost	I hate that the **price** is so **high**.
Dosage	I take **30mg twice a day**

As in the classification task, previous work on aspect-based opinion mining can be split in the two categories: (1) machine learning; and (2) knowledge-based approaches. In the machine learning approach, sequential learning algorithms like Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) have been adopted specially to extract ADRs, but also other aspects [1, 2]. The idea is to treat aspect extraction as a sequence labeling task: each word in a

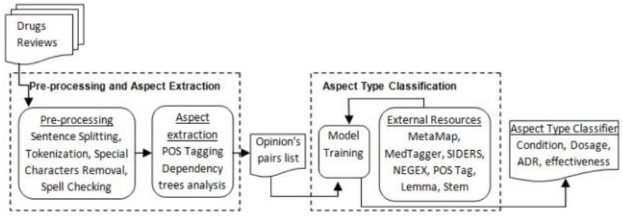


Fig. 1. An overview of the proposed approach

review is labeled with a tag associated to an aspect and sequences of words with the same tag are extracted. Sequential learning also requires labeled corpora for training the models. The need for a large training set can be even more critical for sequential learning, since whole sequences of inputs have to be classified.

Few attempts have been identified in the literature to build knowledge-based techniques for aspect extraction in drug reviews, although such techniques are very common in the general application of aspect-based opinion mining (e.g., for product reviews). Additionally, the existing works are not completely adequate for the task. For instance, in [8], a knowledge-based approach is proposed for extracting text fragments in a review that express an opinion about an entity. Although this approach can be used for filtering relevant opinions in the reviews, it does not directly extract specific aspects (like ADRs). In [5, 6], text fragments are extracted from reviews with the focus on classifying sentiments. Although some of the extracted information can coincide with aspects of interest, this approach is also not focused on aspect-based extraction.

3 Proposed Solution

In this paper, a method for extracting and classifying aspects in drug reviews is proposed. Our proposed method for mining aspects in drugs reviews is comprised of three main steps: Pre-processing, Aspect Extraction and Aspect Type Classification (see Figure 1). Initially, given an input review, the Pre-processing step normalizes the review's text. In the Aspect Extraction step, we extended an algorithm, in which dependency paths in the syntactic trees of the reviews are adopted to extract opinion pairs in the review's sentences (i.e., an aspect plus a sentiment modifier). This approach does not require to train an extraction model (e.g., by using a sequential learning algorithm), which is usually very expensive. Additionally, it will be useful to find low frequent aspects terms. Finally, in the Aspect Type Classification step, each opinion pair is classified into one of the four aspect types: Condition, Side Effects, Dosage or Effectiveness. For this a conventional supervised learning algorithm is adopted. To the best of our knowledge, there is no previous work which identifies multiple aspect classes in

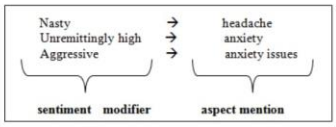


Fig. 2. Examples of opinion pairs in drug reviews.

drugs reviews. Previous work only focused on classifying ADRs, which is limited compared to our work. In the following sections we describe the implementation details for each step of the proposed approach.

3.1 Pre-processing

Each review in collected data were splited in sentences and tokenized by Stanford CoreNLP tool[1]. We performed spelling correction to correcting words that were misspelled frequently by users using the Google spell checking API[2] and special symbols were removed. Finally, the term *meds* and *med* frequently cited on corpus were replaced to *medicine*. Stopwords were not removed as it may infer wrong syntactic tree parse process.

3.2 Aspect extraction task

Knowledge-based approaches which adopted linguistic rules are an interesting alternative for aspect mining. Previous techniques successfully adopted in other domains (like product reviews) can be investigated in the domain of drug reviews, thus filling in a gap in the literature. Obviously, adaptation of previous methods has to be addressed to fit the specific characteristics of the drug review domain.

In the current work, we proposed an aspect extraction method for drug reviews, which is an extension of the *Aspectator* method [11]. This work was originally developed to automatically detect aspects of products on user feedback. It analyzes dependency paths in the syntactic tree of a review to find opinions expressed on candidates aspects. The selection and extraction is based on pairs of words or phrases called *opinion pairs* as illustrated in Figure 2. The first term called *sentiment modifier* is the word around the aspect that expresses an opinion and the second term called *aspect mention* is the mention of an aspect.

The original approach focuses on extracting opinion pairs based on nouns and adjectives, except to "nsubjpass-advmod" path. In our domain of interest, Sarker [2] comments that verbs have an important role when patients express their experiences in a review. Thus, important aspects cannot be extracted by the original *Aspectator*. For instance, in the expression *reduced my pain*, the term *reduced* is a verb (a verb not related by an adverb as suggested in"nsubjpass-advmod"

[1] https://stanfordnlp.github.io/CoreNLP/
[2] https://code.google.com/archive/p/google-api-spelling-java/

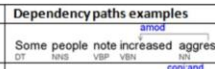
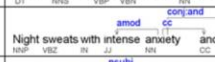
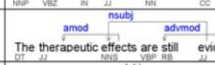
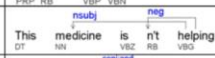
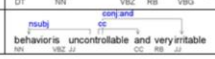
Dependency path	Dependency paths examples	Opinion pair
amod VB		<increased ; agression>
amod_conj		<intense; anxiety> <intense; nausea>
amod_nsubj		<still evident; therapeutic effects>
dobj NN_VB		<developed; seizures>
nsubj_xcomp VB		<immensely slowed; metabolism>
nsubj_xcomp VB_JJ		<very groggy; felt>.
nsubj_xcomp VB_VB		<mainly relaxed; fell>
nsubj VB		<n't helping; medicine>
nsubj_conj		<very irritable; behaviors> <uncontrollable; behaviors>

Fig. 3. New Dependency Paths: Aspectator algorithm adaptation to Medical domain

path) and thus the effectiveness aspect of the drug would not be identified. This paper proposes to extend the algorithm and adapt it to the medical domain in order to overcome this limitation as well as other gaps identified.

Figure 3 presents the new dependency paths proposed in our work. All paths considered relations between nouns and adjectives or verbs. For instance, in the dependency paths "nsubj-xcomp VB-JJ" and "nsubj-xcomp VB-VB", a verb is considered as aspect mention and an adjective or a verb respectively are sentiment modifiers. In our work, dependency trees were computed using the Universal Dependency Relations embedded in the Stanford CoreNLP tool.

3.3 Aspect Type Classification

In this step, each opinion pair is labeled as Condition, Side Effects, Dosage or Effectiveness. A model built by supervised learning is adopted in our solution. An important issue in this classification is the set of features used to describe the opinion pairs and thus adopted as predictor attributes by the classifier. We adopted in our work seven sets of features that were combined in the classification. Initially, the aspect and modifier in a pair are associated to their grammatical class indicated by the **POS tagging** tool. Additionally, named entity recognition tools were used to tag both the aspect and modifier with semantical types. Three knowledge based databases were adopted, containing lists of terms associated to different semantical types:

- **MetaMap:** The Unified Medical Language System (UMLS)[3] is a metathesaurus which contains over a million medical and general English concepts grouped in more than 130 semantic types. Each one belongs to one of the 15 semantic groups, e.g., the term *muscle tension* belongs to the semantic type *Sign* or *Symptom* from the *Disorders* semantic group. We used the MetaMap Java API[4] to tag the aspects and modifiers with UMLS semantic types.
- **SIDER[5]:** The SIDER side effect resource contains a list of terms tagged as precondition or indication of adverse reaction. The resource contains associations between drugs and adverse reactions terms extracted from pharmaceutical literature. For example, the term *Bipolar Disorder* is tagged as precondition and *Fever* is tagged as indication.
- **MedTagger[6]:** The MedTagger is a suite of programs including indexing based on dictionaries, syntactic parsing to detect concept mentions in clinical text, and named entity recognizers. Dictionary-based concept in MedTagger was used to tag aspects and modifiers matching a list of terms labeled with semantic groups like disorder (DISO), anatomy (ANAT), among others.

We also use **NegEX**[7], an open source Java tool for negation detection of events in medical documents. In our work, this tool was applied to indicate whether an opinion pair event is negated. NegEX contains a list of negation phrases tagged with default labels. We considered only terms tagged with the label *definiteNegatedExistence*. Finally it was included as attributes, the **Lemma** and **Stem** of the aspect and modifier in the opinion pair.

The Random Forest algorithm classifier implemented in Weka tool[8] was adopted to built the classifier. As it will be seen, the classifier was learned by using a manually labeled corpus of opinion pairs.

4 Experiments and results

In this section we describe the experiments to evaluate the proposed approach. Three drug review datasets[14] (see Table 2) were manually collected from the website *drugs.com* regarding to the clinical conditions: ADHD, AIDS and Anxiety.

In order to create a gold standard for evaluation, initially the opinion pairs present in each drug review were manually extracted. After extraction, each opinion pair was manually labeled (by one annotator) as Condition (C), Side Effects (ADR), Dosage (D) or Effectiveness (E), Overall aspect was considered as Effectiveness and Cost was not considered to the experiment due to the low frequency in the database. The aspect extraction step will be evaluated according

[3] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/
[4] https://metamap.nlm.nih.gov
[5] sideeffects.embl.de/
[6] https://sourceforge.net/projects/ohnlp/files/MedTagger/
[7] https://code.google.com/archive/p/negex/
[8] http://www.cs.waikato.ac.nz/ml/weka/

Table 2. Corpus of experiments related to three different clinical conditions

	ADHD	**AIDS**	**Anxiety**
Drugs Name	8	4	4
Total Reviews	802	201	500
Labeled Aspect	6,528	1,841	3,955

to its capability of retrieving the gold standard opinion pairs. The labeled pairs will be used to evaluate the predictive performance of the classification step. Following we present the experiments and obtained results respectively for the aspect extraction and the classification tasks.

4.1 Aspect Extraction

In order to evaluate the aspect extraction step, we computed classical metrics: precision, recall and F-measure. Precision (P) is defined by the number of automatically extracted opinion pairs that occur in the labeled corpus (i.e., relevant pairs extracted - Opa) divided by number of automatically opinion pairs extracted (Opi). Recall (R) is measured as the number of relevant opinion pairs extracted divided by the number opinion pairs present in the labeled dataset (Opn). The F-Measure (F) is the harmonic mean between precision and recall. A comparative analysis was performed with previous approaches:

- Zheng et al. [12]: developed an unsupervised dependency analysis-based approach to extract Appraisal Expression Patterns (AEPs). The AEP is applied to represent the syntactic relationship between aspect and sentiment words by using Shortest Dependency Path (SDP) that connects two words in the dependency graph. The AEP information is incorporated into the AEP-LDA model for mining aspect and sentiment words simultaneously. We compare our dependencies patterns with relations generated in the AEP.
- Samha [13]: proposed an approach undertakes Dependency Parsing, Lemma, and POS tagging in order to obtain the syntactic structure of sentences by means of a dependency relation rule. Dependency Relations is applied to map the dependencies between all words within the sentence in the form of relation (a grammatical relation holds between a head and a dependent).
- Original Aspectator.

Table 3 presents the obtained results to baselines, original Aspectator and new dependency paths. Original *Aspectator* achieved the highest precision for all datasets, but lower recall levels. The other baselines were not competitive due to the low values of precision in turn. The proposed method obtained the best trade-off between precision and recall, i.e., the best results in terms of F-Measure.

Table 4 presents the recall relative to each aspect type. Cost aspect retrieved the lowest frequency, only 27 pairs in all bases, users usually do not discuss about price in drug reviews. Dosage and Anxiety to ADHD had a higher occurrence than to AIDS, we noted, drugs belonging to ADHD and Anxiety are available

Table 3. Precision, Recall and F-Measure

	ADHD			AIDS			Anxiety		
Method	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Zheng et al. [12]	0,459	0,561	0,505	0,502	0,578	0,537	0,515	0,543	0,529
Samha [13]	0,580	0,540	0,560	0,643	0,604	0,623	0,624	0,550	0,585
Aspectator	0,816	0,514	0,631	0,772	0,554	0,645	0,828	0,582	0,683
Proposed Method	0,780	0,665	0,718	0,752	0,678	0,713	0,787	0,658	0,717

Table 4. Recall to each aspect type

Dataset	**Overall**	**Effectiveness**	**ADR**	**Dosage**	**Condition**	**Cost**
ADHD	0,796	0,696	0,629	0,708	0,713	0,615
AIDS	0,879	0,796	0,651	0,511	0,582	0,60
Anxiety	0,780	0,764	0,725	0,601	0,777	0,778

in different dosages (capsule, tablet or liquid), then users have more options to discuss about different dosages. Therefore, aspect type is quite dependent of disease type. For certain aspects like condition, the recall is lower, which could be explained by the difficulty in handling some domain dependent terms.

We also note Efficacy or Side Effects aspects are frequently expressed using quantitative values, mainly in the AIDS database, e.g., "*My **CD4** count was at **89** and **Viral Load** up to **182,000/mL**. After 6 weeks of treatment my **CD4** count had jumped to **394**, and **VL** dropped to **2,185***.", implies positive review about the medicine. The *Aspectator* is not prepared to handle with quantitative terms.

Table 5 presents the precision to original and new *Aspectator* paths. As it can be seen, most rules have good precision values. We consider "nsubj-dobj" path not relevant to our domain due to its lower frequency and low precision. The path "nsubj NN-VB" resulted high frequency and low precision, we observed that is necessary filtering the occurrence of some verbs types. For example, verbs "is", "do", "went", "does", "been" are frequent extracted as *sentiment modifier*.

The path "dobj NN-VB" reached high precision and solves the expression *reduced my pain* previously mentioned. The path "nsubj-xcomp VB-VB - VB-JJ" is able to extract verbs as *aspect mention*. Drugs reviews have high frequency of sentences in first person, as *I feel extremely anxious*, resulting opinion pair <extremely anxious; feel>. In addition, we considered the new dependency paths proposed relevant to drugs reviews domain.

4.2 Aspect Type Classification

In this section, we initially evaluated the predictive performance of the RF algorithm and measured the impact of the feature sets adopted. For this, the RF algorithm was evaluated with 10-fold cross-validation using different sets of features, as presented in Table 6 (first column). In each experiment, a new set is included to augment the description of the opinion pairs. The simplest set is the *Postag* set, which is considered as a baseline for comparison.

Table 5. Precision by Dependency Path

Path	ADHD			AIDS			Anxiety		
Original Paths	**Opa**	**Opi**	**P**	**Opa**	**Opi**	**P**	**Opa**	**Opi**	**P**
amod NN-JJ	1302	1627	**0,8**	591	787	**0,75**	1220	1493	**0,82**
nsubj-dobj NN-JJ	1	3	0,33	1	1	1	1	2	0,5
nsubj-xcomp NN-JJ	50	55	**0,91**	17	19	**0,89**	42	47	**0,89**
nsubj-cop NN-JJ	158	164	**0,96**	92	101	**0,91**	125	132	**0,95**
nsubjpass-advmod NN-VB	14	20	**0,7**	5	6	**0,83**	15	21	**0,71**
New Paths									
amod NN-VB	46	49	**0,94**	14	15	**0,93**	43	45	**0,96**
amod-conj NN-VB/JJ	46	73	0,63	24	50	0,48	48	74	0,65
amod-nsubj NN-VB/JJ	113	161	**0,7**	82	110	**0,75**	106	149	**0,71**
dobj NN-VB	813	1127	**0,72**	285	394	**0,72**	750	1040	**0,72**
nsubj-xcomp NN-VB	50	108	0,46	15	29	0,52	42	84	0,5
nsubj-xcomp VB-JJ	237	247	**0,96**	51	53	**0,96**	209	215	**0,97**
nsubj-xcomp VB-VB	19	25	**0,76**	2	3	0,67	5	8	0,63
nsubj NN-VB	364	907	0,4	137	319	0,43	342	767	0,45
nsubj-conj NN-VB	13	18	**0,72**	8	11	**0,73**	15	21	**0,71**

As it can be seen in Table 6, for all datasets, a gain in performance was observed when MetaMap was included in comparison to only using the Postag set. Although better results can be achieved by including other feature sets, the significance over the combination Postag+MetaMap was not verified. For instance, the best accuracy values were obtained when all domain feature sets were considered. However only a very small gain in performance was obtained over Postag+MetaMap. In some cases, including more features can even harm accuracy, which was specially observed when Stemming was adopted. One reason for high accuracy using MetaMap that this feature in addition to identifying medical terms also covers quantitative, qualitative, and temporal terms like 30 mg, better, great, daily in which are quite frequent in the experiment dataset.

Table 6. Correctly Classified Instances (in %).

Features Used	ADHD	AIDS	Anxiety
Postag	72.87	74.75	72.89
Postag + Metamap	76.08	77.34	75.9
Postag + Metamap + Medtagger	75.85	76.75	76
Postag + Metamap + Medtagger + Siders	76.31	77.4	76.25
Postag + Metamap + Medtagger + Siders + Negex	**76.86**	**78.16**	**76.78**
Postag + Metamap + Medtagger + Siders + Negex + Lemma	75.79	76.58	75.77
Postag+Metamap+Medtagger+Siders+Negex+Lemma+Stem	72.69	73.38	72.97

Table 7 presents the confusion matrix resulted from the best feature set presented in Table 6. The classifier has mainly failed in the ADR and Conditions

classes, which were commonly classified as Effectiveness. In the same way. Effectiveness was commonly assigned as Condition or ADR. We observed that the occurrence of modifiers terms like *low, lower, higher, declined, stopped, lost, less, etc,* are leading the classifier to errors. For example, observe the opinion pairs: <lost;focus> labeled as Condition, <lost;appetite> as ADR and <lost;pounds> as Effectiveness, the word *lost* was referenced in different ways as modifier, but their aspects are not sufficient for the classifier to distinguish different classes.

Table 7. Confusion matrix from using the optimal set of features

ADHD				
	ADR	C	D	E
ADR	1391	121	3	290
C	185	571	4	455
D	4	3	586	72
E	211	142	20	2470

AIDS				
	ADR	C	D	E
ADR	313	10	0	111
C	22	375	10	121
D	0	4	66	12
E	41	66	6	684

Anxiety				
	ADR	C	D	E
ADR	602	106	2	202
C	85	672	13	225
D	0	6	397	48
E	101	109	21	1366

We also observed that the same opinion pair may occur as Condition or ADR, as in the sentences: (1) "I **had** constant **headache** and horrible migraines which I am treated for by a neurologist" and (2) "The generic did nothing for me except to make me agitated/jittery for few hours, then crashed and **had** a **headache**". The opinion pair <had;headache> in sentence (1) is described as Condition and in sentence (2) is described as ADR.

Another error to mention is false negative opinion pair. For example, consider the sentence *I was more satisfied with this one than any of the medicines. I **never had headaches**(with some I had terrible headaches!) the time release worked very well.* The opinion pair <had;never;headaches> indicates Effectiveness, but the classifier predicted as Condition. We highlight that the use of the resource Negex is not enough to address problem .

All these examples are inducing the classifier to wrong predictions. Some of these errors can be dealt with by including other features to describe the opinion pairs. Additionally, it can be relevant to analyze frequent terms before and after each opinion pair in order to identify specific patterns. Sampathkumar et. al [9] defines a list of keywords and phrases that denote the causal relationship of a drug causing a side-effect. For example, expressions like *Caused by, Have been getting, Made me feel* could be adopted to distinguish ADR aspects.

We also observed that patients often use different verbs to describe different aspects. For example, terms like *diagnosed*, *suffering* and *have* are usually used to describe Condition (e.g., *I have been diagnosed, I was suffering, I have ADHD, I have psychotic symptoms*). For ADRs, terms like *felt* and *have been* are commonly used (e.g., *I have been losing my appetite, I felt very confused, I felt suicidal*). The verbs *work* and *help* are usually used to describe Effectiveness (e.g., *it helps me concentrate, Adderall really helps me focus, didn't work well, Concerta worked wonders.* For Dosage, common verbs are *prescribed*, *take*, *taken* and *took up.*

5 Conclusion and further work

In this work, we developed a method for extracting and classifying aspects in drug reviews. Initially, we extend the algorithm *Aspectator* suggesting new dependency paths to extract relevant opinion pairs to medical domain. We tested each path in three datasets of drugs reviews. The proposed solution achieved very competitive results compared to baseline methods (the highest values of F-Measure were observed for all datasets). We highlight that the solution can be easily adapted to other languages since it does not require labeled data. We also report an aspect term classification model based on Random Forest classifier. Evaluation on the dataset shows encouraging results that need further investigation and analysis of many ways to improve the performance of the system.

Other datasets can be considered in the future covering other medicines and diseases. We also aim to explore other supervised machine learning methods, hybrid approaches and new lexical resources in order to achieve improvements in results. Finally review analysis of comparative sentences can be performed in order to consider citations between drugs.

References

1. Denecke, K., Deng, Y.: Sentiment analysis in medical settings: New opportunities and challenges. Artificial Intelligence in Medicine 64(1), 17–27 (2015)
2. Sarker, A., et al.: Utilizing social media data for pharmacovigilance: A review. Journal of Biomedical Informatics 54, 202–212 (2015)
3. Gosal, G.P.S.: Opinion mining and sentiment analysis of online drug reviews as a pharmacovigilance technique. International Journal on Recent and Innovation Trends in Computing and Communication 3, 4920–4925 (2015)
4. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2012)
5. Na, J.C., et al.: Sentiment classification of drug reviews using a rule-based linguistic approach. Lecture Notes in Computer Science 7634, 189–198 (2012)
6. Na, J.C., Kyaing, W.Y.M.: Sentiment analysis of user-generated content on drug review websites. Journal of Information Science Theory and Practice 1(1) (2015)
7. Egger, D., et al.: Adverse drug reaction detection using an adapted sentiment classifier. Social Media Mining Shared Task Workshop at PSB (2015)
8. Noferesti, S., Shamsfard, M.: Resource construction and evaluation for indirect opinion mining of drug reviews. PLoS ONE 10 (2015)
9. Sampathkumar, H., et al.: Mining adverse drug reactions from online healthcare forums using hidden markov model. BMC Med. Inf. & Decision Making 14 (2014)
10. Jonnagaddala, J., et al.: Binary classification of twitter posts for adverse drug reactions. Social Media Mining Shared Task Workshop at PSB (2016)
11. Bancken, W., Alfarone, D., Davis, J.: Automatically detecting and rating product aspects from textual customer reviews. 1st international workshop, DMNLP (2014)
12. Zheng, X., et al.: Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification 61, 29–47 (2014)
13. Samha, A.K.: Aspect-based opinion mining using dependency relations. International Journal of Computer Science Trends and Technology (IJCST) 4 (2016)
14. Drugs Reviews Dataset, `https://github.com/spdiana/DrugsReviewsCorpus`

Anexo B: Artigo (Cavalcanti; Prudêncio, 2017b)

Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference

Unsupervised Aspect Term Extraction in Online Drugs Reviews

Diana C. Cavalcanti and Ricardo B. C. Prudêncio
Centro de Informática, Universidade Federal de Pernambuco, Cidade Universitária - 50740-560, Recife/PE, Brazil
{dcc2,rbcp}@cin.ufpe.br

Abstract

Aspect mining in drugs reviews has focused on extracting relevant information such as adverse reactions, efficacy of a drug, symptoms and conditions of patients. In our work, a new unsupervised and knowledge-based method is proposed for extracting aspects in drug reviews. The proposed solution is based on linguistic features, more specifically dependency paths in the syntactic tree of a review. The quality of the dependency path rules was investigated in a number of experiments in review corpora associated to three different diseases. Promising results were achieved compared to previous work.

Introduction

Sentiment and opinion about symptoms, treatments or medicines expressed in online media provide new opportunities for researches on opinion mining and sentiment analysis (Denecke and Deng 2015). Sentiment in medical context can be seen as a reflection of the health status of a patient, which can be good, bad or normal in some time interval. Expressed terms, like *severe pain*, can be a good indication of the health and quality of life of a patient. Descriptions of the use of a medicine may be related to symptoms presented before ingestion, such as *anxiety co-morbid*, *high blood pressure* or about adverse reactions, such as *extreme weight loss*.

Adverse Drug Reactions (ADR) have been commonly reported by patients in drug reviews. ADRs are harmful reactions caused by medication intake resulting in an intervention related to the use of the product, which provides risk in future use or specific treatment, change in dosage or even the withdrawal of product market (Yang and Yang 2015). The activities relating to the detection, assessment, understanding and prevention of adverse effects related to drugs is known as pharmacovigilance or drug safety monitoring. Pharmacovigilance starts during clinical trials of a drug and continued after released it for consumption (Gosal 2015).

Due to various limitations in clinical trials, it is not possible to fully evaluate the consequences of using a particular drug before being released (Gosal 2015; Cheng et al. 2014). An ADR may not be detected before the product go to the market and can take some time after its sale to track new ADRs and relate them to the drug's label (Sampathkumar, Chen, and Luo 2014).

Opinion mining is described as the process of detecting opinions and opinionated aspects in subjective texts from large volumes of structured or unstructured texts using computational methods (Veloso and Jr. 2007; Pang and Lee 2008; Harpaz et al. 2014). Many scientists have focused their research on the development of mining techniques in medical and pharmaceutical texts from publicly available data on the web. Specifically in pharmacovigilance, drug manufacturers can benefit from opinion mining since particular adverse reactions to a drug can be traced more quickly from public repositories or posts in social networks.

In this work, we focused on aspect-based opinion mining (Liu 2012) in drug reviews. The aim is to identify in a drug review fragments of text that can be associated to specific aspects of interest like, adverse reactions, effectiveness, patient conditions, among others. In general, we can distinguish in the literature two approaches for this task: (1) machine learning, in which sequential labeling algorithms (like Hidden Markov Models) are used to label the words in the review; and (2) knowledge-based systems, in which relevant parts of the review are extracted by the use of linguistic rules. There are advantages and limitations of each approach. We focused on the knowledge-based approach with linguistic techniques, since there is no need for a labeled training corpus, which can be very demanding in practice.

A new method for extracting aspects in drug reviews is proposed based on dependency paths identified in the syntactic tree of a review. Dependency path rules have been successfully adopted in other domains (Bancken, Alfarone, and Davis 2014). In our proposal, we derived new rules specific to the drug review domain and investigated their performance through experiments in three datasets related to different diseases, labeled with relevant aspects. The results revealed a gain in performance (in terms of F-Measure) compared to previous work.

The current work filled in a gap in the literature by investigating and proposing effective knowledge-based and unsupervised methods for aspect extraction in drug reviews. We can mention the following contributions:

- Proposal of a new method for extracting aspects in drug reviews based on syntactic dependency paths, which shown to have a good predictive performance;
- Investigation of previous knowledge-based systems that

use linguistic features for aspect extraction and their adequacy to the drug review domain;

- Production of a corpus of drug reviews labeled by aspects, which can be adopted for new experiments.

The remainder of this paper is organized as follows. Section II provides a brief introduction on opinion mining in drug reviews. In Section III we present the proposed approach. Section IV presents the results of our experiments. Finally, Section V concludes the paper with a discussion of our results, along with recommendations for future work.

Opinion mining in drugs reviews

Recent studies have focused on mining reviews of drugs (available in social networks, forums, web,...) in order to provide useful information for pharmacovigilance. Two common tasks can be identified in the literature of opinion mining in drug reviews: (1) classification of reviews; and (2) extraction of opinion aspects.

Concerning the first task, a review can be automatically classified according to its sentiment about the drug (usually positive, negative or neutral) (Na et al. 2012; Na and Kyaing 2015). Sentiment classification can be useful for filtering relevant reviews to be inspected (e.g., reviews classified as negative have a high chance of mentioning a negative side effect). A related task is to directly classify whether a review mentions an ADR or not (Egger, Uzdilli, and Cielebak 2015; Jonnagaddala, Jue, and Dai 2016). This is more specific and focused than simply classifying the sentiment of a review.

Previous work on review classification adopted in general two categories of techniques: (1) machine learning and (2) knowledge-based approaches. In the machine learning approach, a classification model is learned from a labeled set of reviews. These works can be distinguished by the learning algorithm and the features used for classification. For instance, in (Jonnagaddala, Jue, and Dai 2016; Egger, Uzdilli, and Cielebak 2015), Support Vector Machines (SVMs) were adopted as classifier and different features like n-grams, part-of-speech (POS) tags and lexicon words were considered for classifying tweets that mention an ADR. In (Sharif et al. 2014), feature ensemble was proposed to reduce the sparsity of the feature representation of reviews. Sarker and Gonzalez (2015) adopted different classifiers (Naive Bayes, Maximum Entropy (MNB) and SVMs) and lexicon features. In (Patki and Gonzalez 2014), MNB and SVMs were adopted and the *Wordnet* lexicon has been used to expand synonyms of verbs, adjectives and nouns. A sentiment lexicon is also employed in this work.

A known disadvantage of the machine learning approach is the need of a labeled corpus for training, which can be prohibitive in some applications. Alternatively, previous work has adopted knowledge-based approaches for review classification. For instance, in (Na et al. 2012; Na and Kyaing 2015), the authors proposed a set of rules based on linguistic features to identify patterns in the reviews and then to perform the sentiment classification. The adopted rules are based on semantic dependency analysis (extraction of grammatical dependencies in the texts) and a subjective lexicon. In (Noferesti and Shamsfard 2015), the authors proposed rules for classifying the sentiment polarity of opinions previously extracted from the reviews. The classification is based on POS tagger information and a domain knowledge base.

The second task of opinion mining in drug reviews in to extract specific aspects. Hence, the aim is not only to classify whether a review mentions an ADRs, for instance, but also to extract the reaction itself or any other aspect considered as relevant (Sampathkumar, Chen, and Luo 2014). In literature, this task is called *aspect-based opinion mining*, which aims to extract the main aspects mentioned about an item or entity and to provide the classification of the opinion given on every aspect (Liu 2012). According to (Na et al. 2012; Na and Kyaing 2015; Nikfarjam et al. 2015), six aspects (see examples in Table 1) are common in drug reviews:

- **Overall:** The general opinion of a medicinal product or when the clause is not mentioned any of the other five categories of aspects;
- **Effectiveness:** Changes noted after the use of the medicine, linked to the patient's condition or disease;
- **Side effects:** The reactions that are not related to the medicine. ADR is a negative side effect;
- **Dosage:** Reports the amount, frequency or the treatment period in which the medicine was used.
- **Condition:** Corresponds to a description of the patient's condition, e.g., a disease or health problems in general.
- **Cost:** Corresponds the cost/price of a medicine.

Table 1: Opinion sentence by aspects in drugs reviews

Aspect	**Opinion sentence**
Overall	Adderall is overall a **good** ADD **medicine**.
Effectiveness	It **helped** me stay **focused** on any tasks.
Side effects	I have only one side effect which is **dry mouth**.
Condition	I **had** clinical **depression/anxiety** for years.
Cost	I hate that the **price** is so **high**.
Dosage	I take **30mg twice a day**

As in the classification task, previous work on aspect-based opinion mining can be split in the two categories: (1) machine learning; and (2) knowledge-based approaches. In the machine learning approach, sequential learning algorithms like Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) have been adopted specially to extract ADRs, but also other aspects (Sampathkumar, Chen, and Luo 2014). The idea is to treat aspect extraction as a sequence labeling task: each word in a review is labeled with a tag associated to an aspect and sequences of words with the same tag are extracted. Sequential learning also requires labeled corpora for training the models. The need for a large training set can be even more critical for sequential learning, since whole sequences of inputs have to be classified.

Another task is the creation of medical opinion lexicon, usually created using information of a general lexicon. In (Asghar et al. 2013), a subjectivity lexicon is proposed based

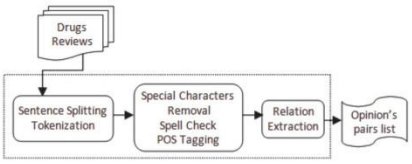


Figure 1: An overview of the proposed approach

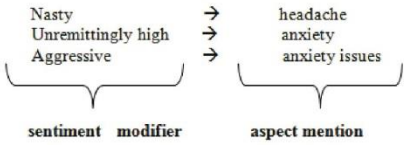


Figure 2: Opinion pair

on the corpus of drug reviews. The lexicon takes the initial medical seed list as input, expands it with SentiWordNet synonyms and antonyms, attaching polarity score with each word. In (Asghar et al. 2016), a health-related sentiment lexicon is proposed by a hybrid approach, which combines boot-strapping concepts and corpus-based strategies.

Few attempts have been identified in the literature to build knowledge-based techniques for aspect extraction in drug reviews, although such techniques are very common in the general application of aspect-based opinion mining (e.g., for product reviews). Additionally, the existing works are not completely adequate for the task. For instance, in (Noferesti and Shamsfard 2015), a knowledge-based approach is proposed for extracting text fragments in a review that express an opinion about an entity. Although this approach can be used for filtering relevant opinions in the reviews, it does not directly extract specific aspects (like ADRs). In (Na et al. 2012; Na and Kyaing 2015), text fragments are extracted from reviews with the focus on classifying sentiments. Although some of the extracted information can coincide with aspects of interest, this approach is also not focused on aspect-based extraction. So there is a gap in the literature that has to be addressed by new research.

Proposed Approach

Knowledge-based approaches which adopted linguistic rules are an interesting alternative for aspect mining. Previous techniques successfully adopted in other domains (like product reviews) can be investigated in the domain of drug reviews, thus filling in a gap identified in the literature. Obviously, adaptation of previous methods has to be addressed to fit the specific characteristics of the drug review domain.

In the current work, we proposed a new aspect extraction method for drug reviews, which is an extension of the *Aspectator* method (Bancken, Alfarone, and Davis 2014). This work was originally developed to automatically detect aspects of products on user feedback. It analyzes dependency paths in the syntactic tree of a review to find opinions expressed on candidates aspects.

The steps of this solution are shown in Figure 1. Initially, drug reviews are collected from a public repository. The reviews are divided into sentences, cleaned by removing special characters. Then tokenization and POS tagging is performed on each sentence (in our work Stanford CoreNLP tool was adopted). Spelling check is performed (in our work using the Google's spell checker). Finally, aspects are extracted by matching paths in dependency trees.

The selection and extraction is based on pairs of words or phrases called *opinion pairs* as illustrated in Figure 2. The first term called *sentiment modifier* is the word around the aspect that expresses an opinion and the second term called *aspect mention* is the mention of an aspect.

Aspectator is capable of extracting candidate aspect terms by combining certain syntactic dependency paths. Compared to other approaches, it is not necessary to have labeled data and the algorithm also does not require domain specific knowledge (Bancken, Alfarone, and Davis 2014).

Figure 3 presents the original dependency paths (given a dependency tree) proposed by *Aspectator*. The paths "amod", "nsubj-dobj", "nsubj-xcomp" and "nsubj-cop" are represented by a pair of substantive (NN) and adjective (JJ) and the path "nsubjpassa-advmod" is represented by a pair of substantive and verb (VB). For example, the sentence *I have been having pretty bad chest pains*, it is extracted the opinion pair <bad JJ ; pains NN>by dependency path "amod". Extensions dependency paths are dependent of main paths: "compound noun" path extracts compound noun (NN) to aspect mention, "adverbial modifier" extracts a modifier (RB) to sentiment term, "simple negation" and "negation through no determiner" extract negative term related to the respective opinion pair. Therefore, the sentence cited above results the opinion pair <pretty bad JJ; chest pains NN>by main and extension paths.

The original approach focuses on extracting opinion pairs based on nouns and adjectives, except to "nsubjpass-advmod" path. In our domain of interest, Sarker (2015) comments that verbs have an important role when patients express their experiences in a review. Thus, important aspects can not be extracted by the original *Aspectator*. For instance, in the expression *reduced my pain*, the term *reduced* is a verb (a verb not related by an adverb as suggested in"nsubjpass-advmod" path) and thus the effectiveness aspect of the drug would not be identified. This paper proposes to extend the algorithm and adapt it to the medical domain in order to overcome this limitation as well as other gaps identified.

Figure 4 presents new dependency paths proposed in our work. All new dependency paths considered relations between nouns and adjective or verbs. The dependency path "amod-nsubj" relations "amod" and "nsubj". In the dependency paths "nsubj-xcomp VB-JJ" and "nsubj-xcomp VB-VB", a verb is considered as aspect mention and an adjective or a verb respectively are sentiment modifiers. All new paths are matched with Extension paths suggested in original algorithm. Each new relationship was suggested by the observation that by considering only the original dependency paths,

relevant opinion pairs are ignored in the medical domain.

In our work, dependency trees were computed using the Universal Dependency Relations[1] embedded in the Stanford coreNLP(Mcdonald et al. 2013).

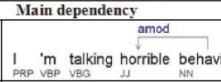
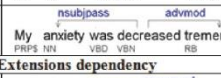
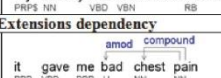
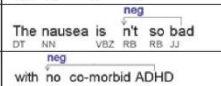
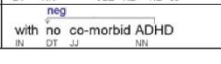
Dependency path	Example
Main dependency	
amod	
nsubj_dobj	
nsubj_xcomp	
nubj_cop	
nsubjpass_ advmod	
Extensions dependency	
compound noun	
adverbial modifier	
simple negation	
Negation through "no" determiner	

Figure 3: Main and Extensions Dependency paths used by by Aspectator Algorithm

Experiment and discussion

In this section we describe the experiments with the proposed approach. Drug review datasets were collected from a public repository[2] regarding three clinical conditions: ADHD, Aids and Anxiety. Table 3 presents an example of review from the Anxiety dataset. In order to create a labeled corpus for evaluation, we initially split each review into sentences and manually extracted opinion pairs found. Each opinion pair was classified in one of six aspect types described in Table 1. Table 2 summarizes the datasets.

Table 2: Corpus

	ADHD	Aids	Anxiety
Drugs Name	4	4	4
Total Reviews	502	201	500
Labeled Aspect	4.680	1.895	3.985

[1]Some relations have been updated, example, Acomp to Xcomp. Full list in http://universaldependencies.org/u/dep/

[2]drugs.com

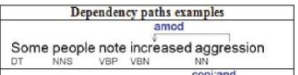
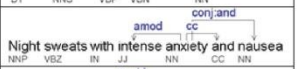
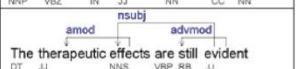
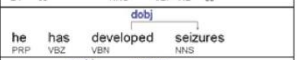
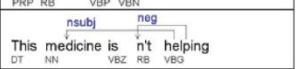
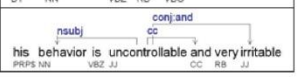
Dependency path	Dependency paths examples
amod VB	
amod_conj	
amod_nsubj	
dobj NN_VB	
nsubj_xcomp VB	
nsubj_xcomp VB_JJ	
nsubj_xcomp VB_VB	
nsubj VB	
nsubj_conj	

Figure 4: New Dependency Paths: Adaptation to Aspectator Algorithm to Medical Domain

Table 3: Anxiety Dataset: Review Example

Brand Name	Review	Rating
Ativan	I was battling depression along with panic attacks. They first had me on Xanax. After a while wanted to get off that but was still having panic attacks. Not since I started Ativan. Also I find it much easier to focus and get things done now.	10

All sentences were preprocessed to remove special characters and symbols, to convert capitalized terms as it may infer wrong syntactic tree parse process, to remove consecutive blank spaces between words, to perform spelling of terms and to standardize terms, as some characters can be used repeatedly in the same word. The term *meds* and *med* were replaced to *medicine*. Stopwords were not removed.

In order to evaluate the results, we computed classical metrics: precision, recall and F-measure. Precision (P) is defined by the number of automatically extracted opinion pairs that occur in the labeled dataset (i.e., relevant pairs extracted - Opa) divided by number of automatically opinion pairs extracted (Opi). Recall (R) is measured as the number of relevant opinion pairs extracted divided by the number opinion pairs present in the labeled dataset (Opn). The F-Measure (F) is the harmonic mean between precision and recall.

In order to do a comparative analysis, we also perform experiments with previous approaches:

- Zheng et al. (2014): developed an unsupervised depen-

Table 4: Precision, Recall and F-Measure to baselines, original Aspectator and new dependency paths

Dataset	Method	P	R	F
ADHD	Zheng et al. (2014)	0,459	0,561	0,505
	Samha (2016)	0,580	0,540	0,560
	Aspectator	0,816	0,514	0,631
	Method Proposed	0,78	0,665	**0,718**
Aids	Zheng et al. (2014)	0,502	0,578	0,537
	Samha (2016)	0,643	0,604	0,623
	Aspectator	0,772	0,554	0,645
	Method Proposed	0,752	0,678	**0,713**
Anxiety	Zheng et al. (2014)	0,515	0,543	0,529
	Samha (2016)	0,624	0,550	0,585
	Aspectator	0,828	0,582	0,683
	Method Proposed	0,787	0,658	**0,717**

dency analysis-based approach to extract Appraisal Expression Patterns (AEPs) from reviews regarded as a condensed representation of the syntactic relationship between aspect and sentiment words. In work, the AEP is applied to represent the syntactic relationship between aspect and sentiment words by using Shortest Dependency Path (SDP) that connects two words in the dependency graph, and it is an alternate sequence of POS tags and syntactic dependency relationships. The AEP information is incorporated into the AEP-LDA model for mining aspect and sentiment words simultaneously. We compare our dependencies patterns with relations generated in the AEP.

- Samha (2016): proposed propose a Natural Language Processing approach that undertakes Dependency Parsing, Pre-processing, Lemmatization, and part of speech tagging of natural texts in order to obtain the syntactic structure of sentences by means of a dependency relation rule. The Stanford Dependency Relations is applied to find the syntactic parsers that will allow us to map the dependencies between all words within the sentence in the form of relation (a grammatical relation holds between a head and a dependent). It's explored a set of syntactic rules and relations that were observed from the product dataset.
- Original Aspectator.

Table 4 presents the obtained results. The original *Aspectator* achieved the highest precision for all datasets, but lower recall levels. The other baselines were not competitive due to the low values of precision in turn. The proposed method obtained the best trade-off between precision and recall, i.e., the best results in terms of F-Measure.

Table 5: Recall to each aspect type

Dataset	Overall	Effectiveness	ADR	Dosage	Condition	Cost
ADHD	0,796	0,696	0,629	0,708	0,713	0,615
Aids	0,879	0,796	0,651	0,511	0,582	0,60
Anxiety	0,780	0,764	0,725	0,601	0,777	0,778

Table 5 presents the recall relative to each aspect type: aspects types retrieved by our method that occur in the labeled dataset divided by number of each type present in the labeled data set. The aspect Cost retrieved the lowest frequency where given the sum in all bases occurred only 27 pairs, users usually do not discuss about price in drug reviews. Regarding the aspect Dosage to ADHD and Anxiety had a higher occurrence than to AIDS, we observed that drugs belonging to the experiment-based ADHD and Anxiety are available in different dosages (capsule, tablet or liquid) which users have more options to discuss about different dosages.

We observed that aspect type occurrence is quite dependent of disease type. For certain aspects like condition, the recall is lower, which could be explained by the difficulty in handling some domain dependent terms. We also note that there is a number of opinions expressed about the efficacy or side effects of medications using quantitative values, which occurs mainly in the AIDS database. Note the review below:

- *My CD4 count was at 89 and my Viral Load up to 182,000/mL. I began taking Atripla on Nov. 10th 2015. After receiving my test results my CD4 count had jumped to 394, and my Viral Load dropped to 2,185 - after only 6 weeks of treatment. ...my CD4 count had jumped to 394, and my Viral Load dropped to 2,185*

implies positive review about the medicine while the *Aspectator* is not prepared to handle with quantitative terms.

Table 6: Precision by Dependency Path

Path	ADHD			Aids			Anxiety		
Original Paths	**Opa**	**Opi**	**P**	**Opa**	**Opi**	**P**	**Opa**	**Opi**	**P**
amod NN-JJ	1302	1627	**0,8**	591	787	**0,75**	1220	1493	**0,82**
nsubj-dobj NN-JJ	1	3	0,33	1	1	1	1	2	0,5
nsubj-xcomp NN-JJ	50	55	**0,91**	17	19	**0,89**	42	47	**0,89**
nsubj-cop NN-JJ	158	164	**0,96**	92	101	**0,91**	125	132	**0,95**
nsubjpass-advmod NN-VB	14	20	**0,7**	5	6	**0,83**	15	21	**0,71**
New Paths									
amod NN-VB	46	49	**0,94**	14	15	**0,93**	43	45	**0,96**
amod-conj NN-VB/JJ	46	73	0,63	24	50	0,48	48	74	0,65
amod-nsubj NN-VB/JJ	113	161	**0,7**	82	110	**0,75**	106	149	**0,71**
dobj NN-VB	813	1127	**0,72**	285	394	**0,72**	750	1040	**0,72**
nsubj-xcomp NN-VB	50	108	0,46	15	29	0,52	42	84	0,5
nsubj-xcomp VB-JJ	237	247	**0,96**	51	53	**0,96**	209	215	**0,97**
nsubj-xcomp VB-VB	19	25	**0,76**	2	3	0,67	5	8	0,63
nsubj NN-VB	364	907	0,4	137	319	0,43	342	767	0,45
nsubj-conj NN-VB	13	18	**0,72**	8	11	**0,73**	15	21	**0,71**

Table 6 presents the precision of each dependency path presented in Figures 2 and 3. As it can be seen, although most rules have good precision values, some of them are not so precise. For instance, as the dependency path "nsubj-doj" returned the lowest frequency and low precision, we consider this path not relevant to our domain. The dependency path "nsubj NN-VB" resulted high frequency and low precision, we observed that is necessary filtering the occurrence of some verbs types. For example, verbs "is", "do", "went", "does", "been" are frequent extracted as *sentiment modifier*.

The path "dobj NN-VB" reached high precision and solves the expression *reduced my pain* previously mentioned. The path "nsubj-xcomp VB-VB - VB-JJ" is able to extract verbs as *aspect mention*. Drugs reviews have high frequency of sentences in first person, as *I feel extremely anxious*, resulting opinion pair <extremely anxious; feel>. In addition, we considered the new dependency paths proposed relevant to drugs reviews domain.

Conclusion and further work

In this work, we extend the algorithm *Aspectator* suggesting new dependency paths to extract relevant *opinion pairs* to medical domain. We tested each new path in three datasets of drugs reviews. The proposed solution achieved very competitive results compared to baseline methods (the highest values of F-Measure were observed for all datasets). We highlight that the proposed solution can be easily adapted to other languages since it does not require labeled data.

As further work, we aim to investigate methods to automatically classify aspects types from unstructured text of drugs reviews. Other datasets can be considered in the future covering other medicines and diseases. We also aim to explore supervised machine learning, hybrid approaches and lexical resources in order to achieve improvements in results. Finally review analysis of comparative sentences can be performed in order to consider citations between drugs.

References

Asghar, M. Z.; Khan, A.; Ahmad, S.; and Ahmad, B. 2013. Subjectivity lexicon construction for mining drug reviews. *Science International* 26:145–149.

Asghar, M. Z.; Ahmad, S.; Qasim, M.; Zahra, S. R.; and Kundi, F. M. 2016. Sentihealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus* 5(1):1139.

Bancken, W.; Alfarone, D.; and Davis, J. 2014. Automatically detecting and rating product aspects from textual customer reviews. *International Workshop on Interactions between Data Mining and Natural Language Processing*.

Cheng, V. C.; Leung, C. H. C.; Liu, J.; and Milani, A. 2014. Probabilistic aspect mining model for drug reviews. *IEEE Trans. Knowl. Data Eng.* 26(8):2002–2013.

Denecke, K., and Deng, Y. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine* 64(1):17–27.

Egger, D.; Uzdilli, F.; and Cielebak, M. 2015. Adverse drug reaction detection using an adapted sentiment classifier. *Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*.

Gosal, G. P. S. 2015. Opinion mining and sentiment analysis of online drug reviews as a pharmacovigilance technique. *International Journal on Recent and Innovation Trends in Computing and Communication* 3:4920–4925.

Harpaz, R.; Callahan, A.; Tamang, S.; Low, Y.; Odgers, D.; Finlayson, S.; Jung, K.; LePendu, P.; and Shah, N. 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety: An Intern Jour. of medical Toxicology and Drug Experience* 37(10):777–790.

Jonnagaddala, J.; Jue, T. R.; and Dai, H.-J. 2016. Binary classification of twitter posts for adverse drug reactions. *Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Mcdonald, R.; Nivre, J.; Quirmbach-brundage, Y.; Goldberg, Y.; Das, D.; Ganchev, K.; Hall, K.; Petrov, S.; Zhang, H.; Tckstrm, O.; Bedini, C.; Bertomeu, N.; and Lee, C. J. 2013. Universal dependency annotation for multilingual parsing. In *In Proc. of ACL '13*.

Na, J.-C., and Kyaing, W. Y. M. 2015. Sentiment analysis of user-generated content on drug review websites. *Journal of Information Science Theory and Practice* 1(1).

Na, J.-C.; Kyaing, W. Y. M.; Khoo, C. S. G.; Foo, S.; Chang, Y.-K.; and Theng, Y.-L. 2012. Sentiment classification of drug reviews using a rule-based linguistic approach. *Lecture Notes in Computer Science* 7634:189–198.

Nikfarjam, A.; Sarker, A.; O'Connor, K.; Ginn, R.; and Gonzalez, G. 2015. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association* 22(3):671–681.

Noferesti, S., and Shamsfard, M. 2015. Resource construction and evaluation for indirect opinion mining of drug reviews. *PLoS ONE* 10.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* 2(1-2):1–135.

Patki, A., and Gonzalez, G. 2014. Mining adverse drug reaction signals from social media: Going beyond extraction. *22nd Annual International Conference on Intelligent Systems for Molecular Biology*.

Samha, A. K. 2016. Aspect-based opinion mining using dependency relations. *International Journal of Computer Science Trends and Technology (IJCST)* 4.

Sampathkumar, H.; Chen, X.-w.; and Luo, B. 2014. Mining adverse drug reactions from online healthcare forums using hidden markov model. *BMC Medical Informatics and Decision Making* 14(1):91.

Sarker, A., and Gonzalez, G. 2015. Portable automatic text classification for adverse drug reaction detection via multicorpus training. *J. of Biomedical Informatics* 53:196 – 207.

Sarker, A.; Ginn, R.; Nikfarjam, A.; OConnor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; and Gonzalez, G. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics* 54:202–212.

Sharif, H.; Zaffar, F.; Abbasi, A.; and Abbasi, A. 2014. Detecting adverse drug reactions using a sentiment classification framework. *The sixth ASE international conference on social computing (SocialCom)*.

Veloso, A., and Jr., W. M. 2007. Efficient on-demand opinion mining. In *XXII Simpósio Brasileiro de Banco de Dados*, 332–346.

Yang, H., and Yang, C. C. 2015. Mining a weighted heterogeneous network extracted from healthcare-specific social media for identifying interactions between drugs. In *ICDM Workshops*, 196–203. IEEE Computer Society.

Zheng, X.; Lin, Z.; Wang, X.; Lin, K.-J.; and Song, M. 2014. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. *Knowl.-Based Syst.* 61:29–47.