UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LEONARDO AUGUSTO SÁPIRAS

# Mineração de Opiniões Baseada em Aspectos em Fontes de Opiniões Fracamente Estruturadas

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof[a]. Dr[a]. Karin Becker

Porto Alegre
2015

**CIP — CATALOGAÇÃO NA PUBLICAÇÃO**

Sápiras, Leonardo Augusto

Mineração de Opiniões Baseada em Aspectos em Fontes de Opiniões Fracamente Estruturadas / Leonardo Augusto Sápiras. – Porto Alegre: PPGC da UFRGS, 2015.

75 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2015. Orientador: Karin Becker.

1. Mineração de opiniões. 2. Análise de sentimentos. 3. Identificação de aspectos. 4. Sumarização de opiniões. I. Becker, Karin. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Pós-Graduação: Prof. Vladimir Pinheiro do Nascimento
Diretor do Instituto de Informática: Prof. Luís da Cunha Lamb
Coordenador do PPGC: Prof. Luigi Carro
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

**AGRADECIMENTOS**

Agradeço à minha esposa, Fernanda, por me apoiar com carinho e dedicação durante a realização do mestrado. À minha família, por me auxiliar chegar onde estou.

A Deus, por proporcionar conhecimento e experiência por meio de desafios.

Faço um agradecimento especial à minha orientadora, Karin, pela orientação exemplar, conselhos, paciência e dedicação profissional.

Ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande do Sul, pela oportunidade de realizar o mestrado. Aos professores, por transmitir de forma especial seus ensinamentos. Aos meus colegas, por partilhar seus conhecimentos e experiências.

Todos vocês ajudaram a moldar meu futuro profissional. Os bons frutos que eu colher daqui para frente serão mérito de vocês. Humildemente digo, muito obrigado!

*“Acredite que você pode,*
*assim você já está no meio do caminho.”*
— THEODORE ROOSEVELT

**RESUMO**

Na WEB, são encontradas postagens sobre assuntos variados, notícias de celebridades, produtos e serviços. Tal conteúdo contém emoções positivas, negativas ou neutras. Minerar o sentimento da população sobre candidatos a eleições e seus aspectos em mídias virtuais pode ser realizado por meio de técnicas de Mineração de Opiniões. Existem soluções para fontes de opinião fortemente estruturadas, tais como revisões de produtos e serviços, no entanto o problema que se apresenta é realizar a mineração de opiniões em nível de aspecto em fontes de opiniões fracamente estruturadas. Além de avaliar conceitos relacionados à mineração de opiniões, o presente trabalho descreve a realização de um estudo de caso, o qual analisa fontes de opiniões fracamente estruturadas e propõe uma abordagem para minerar opiniões em nível de aspecto, utilizando como fontes de opinião comentários de leitores de jornais. O estudo de caso contribui (i) na concepção de uma abordagem para identificação da opinião em nível de aspecto sobre entidades eleitorais em comentários de notícias políticas, (ii) na aplicação de um método baseado em aprendizagem de máquina para classificar a opinião sobre entidades e seus aspectos em três classes (positivo, negativo e neutro), (iii) na representação da sumarização visual de opinião sobre entidades e seus aspectos. São descritos experimentos para identificar comentários que mencionam os aspectos saúde e educação, utilizando co-ocorrência, em que foram obtidos resultados satisfatórios utilizando as técnicas Expected Mutual Information Measure e phi-squared. Já para a polarização de sentenças, são realizados experimentos com duas abordagens de classificação: uma que classifica sentenças em três classes e outra que realiza classificações binárias em duas etapas.

**Palavras-chave:** Mineração de opiniões. Análise de sentimentos. Identificação de aspectos. Sumarização de opiniões.

**Aspect-Based Opinion Mining in Weakly Structured Opinion Sources**


## ABSTRACT

In the WEB are found posts about various subjects like celebrity news, products and services. Such content has positive, negative or neutral emotions. Mining the population's sentiments about elections candidates and their aspects in virtual media can be performed using Opinion Mining techniques. There are solutions for highly structured opinion sources, such as reviews of products and services, however the problem is how to perform aspect-based opinion mining in less structured opinions sources. Besides evaluating concepts related to opinion mining, this work describes a case study which analyzes weakly structured sources and proposes an approach to mine aspect-based opinions using as sources of sentiment reviews of newspaper readers. The case study contributes (i) designing an approach to identify the aspect-based opinion about electoral candidates in news political comments, (ii) to the application of a machine learning-based method to classify the opinion about entities and their aspects in three classes (positive, negative and neutral) (iii) to the representation of a visual summarization review of entities and their aspects. It describes experiments to identify comments about health and education aspects using co-occurrence where satisfactory results were obtained using the techniques Expected Mutual Information Measure and phi-squared. In which regards sentences polarization, experiments are performed with two classification approaches, one that classifies sentences in three classes and another that performs binary classifications in two stages.



**Keywords:** Opinion Mining, Sentiment Analysis, Aspect Identification, Opinion Summarization.

## LISTA DE FIGURAS

Figura 1.1 Revisão sobre produto no site Best Buy ..... 13
Figura 1.2 Trecho de notícia sobre educação extraído da Folha de São Paulo ..... 15
Figura 1.3 Comentário sobre educação existente na notícia da Figura 1.2. ..... 15
Figura 1.4 Comentário cuja opinião tem como alvo educação e candidatos ..... 15
Figura 1.5 Comentário cuja opinião tem como alvo um dos candidatos ..... 16
Figura 1.6 Comentário sobre outros assuntos que não são relacionados ao conteúdo da notícia..... 16

Figura 2.1 Processo de descoberta de conhecimento em bases de dados ..... 18
Figura 2.2 Abordagem utilizada para o desenvolvimento de modelos de classificação ..... 20
Figura 2.3 Etapas da mineração de opinões ..... 22

Figura 3.1 Modelo de sumarização de opiniões sobre aspectos de um produto ..... 29
Figura 3.2 Sumarização do sentimento em relação a candidatos alemães ..... 30
Figura 3.3 Modelo de sumarização apresentado no Observatório da WEB. ..... 31
Figura 3.4 Modelo de sumarização do Manchetômetro ..... 31
Figura 3.5 Dashboard de monitoramento do LCI com destaque para o gráfico de opiniões e nuvem de termos ..... 32
Figura 3.6 Nuvem de palavras de um filme monitorado pelo LCI. ..... 33
Figura 3.7 Framework de previsão de variações em intenções de voto baseada em sentimentos. ..... 34

Figura 4.1 *Profiling* de notícias e comentários. ..... 42
Figura 4.2 Abordagem proposta para minerar opiniões em nível de aspecto. ..... 46
Figura 4.3 Extração de comentários ..... 49
Figura 4.4 Identificação de comentários que mencionam aspectos ..... 49
Figura 4.5 Mineração de opiniões em nível de sentença ..... 50
Figura 4.6 Mineração de opiniões em nível de aspecto ..... 51

Figura 5.1 Processo de identificação de aspectos. ..... 53
Figura 5.2 Avaliação de resultados para o aspecto saúde em comentários. ..... 57
Figura 5.3 Avaliação de resultados para o aspecto educação em comentários. ..... 58
Figura 5.4 Avaliação de resultados para o aspecto saúde em notícias. ..... 59
Figura 5.5 Avaliação de resultados para o aspecto educação em notícias. ..... 60

Figura 7.1 Modelo visual de sumarização de opiniões. ..... 69
Figura 7.2 Nuvem de palavras para termos representativos de saúde. ..... 70
Figura 7.3 Nuvem de palavras para termos representativos de educação. ..... 70

## LISTA DE TABELAS

Tabela 2.1 Comparação entre dicionários de opinião ..... 25

Tabela 3.1 Experimentos para classificação de sentenças..... 37

Tabela 4.1 Conjuntos de notícias, comentários e sentenças anotadas..... 41
Tabela 4.2 Concordância entre os anotadores na identificação de menções aos aspectos. ..... 42
Tabela 4.3 Quantificação de comentários e sentenças associados a notícias que mencionam os aspectos analisados..... 43
Tabela 4.4 Quantificação de comentários e sentenças associados a notícias que não mencionam os aspectos analisados..... 44

Tabela 5.1 Palavras semente escolhidas..... 54

Tabela 6.1 Datasets utilizados nos experimentos de polarização..... 62
Tabela 6.2 Resultados da polarização das sentenças com a abordagem ternária ..... 63
Tabela 6.3 Resultados da avaliação de sentenças com e sem sentimento ..... 65
Tabela 6.4 Resultados da avaliação de sentenças com sentimento ..... 65
Tabela 6.5 Síntese comparativa dos resultados obtidos com os experimentos realizados. ..... 67

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| EMIM | Expected Mutual Information Measure |
| KDD | Knowlegde Discovery in Databases |
| LCI | Live Customer Intelligence |
| LSI | Indexação Semântica Latente |
| POS | Part-of-Speech |
| SMO | Sequential Minimal Optimization |
| SVM | Support Vector Machine |

SUMÁRIO

**1 INTRODUÇÃO** ..... **12**
**1.1 Problema** ..... **14**
**1.2 Objetivo** ..... **16**
**1.3 Contribuições** ..... **17**
**1.4 Organização do trabalho** ..... **17**
**2 FUNDAMENTAÇÃO TEÓRICA** ..... **18**
**2.1 Mineração de dados** ..... **18**
2.1.1 Processo de descoberta de conhecimento em bases de dados ..... 18
2.1.2 Tarefas da mineração de dados ..... 19
2.1.3 Técnicas de classificação de dados ..... 19
**2.2 Mineração de opiniões** ..... **21**
2.2.1 Níveis de análise textual ..... 23
2.2.2 Tipos de opiniões ..... 24
2.2.3 Técnicas para classificação de opiniões ..... 24
**3 TRABALHOS RELACIONADOS** ..... **26**
**3.1 Técnicas para identificação de aspectos na mineração de opiniões** ..... **26**
3.1.1 Frequência de substantivos ou locuções substantivas ..... 26
3.1.2 Relação entre a opinião e seu alvo ..... 27
3.1.3 Aprendizado supervisionado ..... 28
3.1.4 Modelagem de tópicos ..... 28
3.1.5 Conclusões sobre identificação de aspectos ..... 28
**3.2 Modelos de sumarização de resultados** ..... **29**
**3.3 Mineração de opiniões em comentários sobre notícias políticas** ..... **33**
3.3.1 Descrição do processo de mineração de opiniões ..... 33
3.3.1.1 Extração ..... 33
3.3.1.2 Identificação ..... 34
3.3.1.3 Classificação da polaridade ..... 35
3.3.1.4 Sumarização ..... 37
3.3.2 Considerações ..... 37
**4 ESTUDO DE CASO SOBRE MINERAÇÃO DE OPINIÕES EM NíVEL DE ASPECTOS EM COMENTÁRIOS DE NOTíCIAS** ..... **39**
**4.1 Objetivos** ..... **39**
**4.2 Fonte de dados e análise do corpus** ..... **40**
4.2.1 Anotação manual e Profiling ..... 41
**4.3 Abordagem proposta** ..... **45**
4.3.1 Visão geral ..... 45
4.3.2 Formalização ..... 45
4.3.3 Etapas ..... 47
4.3.4 Ilustração da abordagem ..... 48
**5 IDENTIFICAÇÃO DE ASPECTOS** ..... **53**
**5.1 Abordagem proposta** ..... **53**
**5.2 Experimentos** ..... **55**
5.2.1 Palavras Semente ..... 55
5.2.2 Expected Mutual Information Measure ..... 55
5.2.3 Phi-squared ..... 56
5.2.4 Indexação Semântica Latente ..... 56
**5.3 Resultados** ..... **57**

**6 MINERAÇÃO DE OPINIÕES EM NíVEL DE SENTENÇA ..... 61**
**6.1 Abordagens para polarizar sentenças ..... 61**
**6.2 Corpus e Gold Standard ..... 61**
**6.3 Classificação ternária ..... 62**
6.3.1 Descrição do experimento ..... 62
6.3.2 Discussão dos resultados ..... 63
**6.4 Classificações binárias em duas etapas ..... 64**
6.4.1 Descrição do experimento ..... 64
6.4.2 Discussão dos resultados ..... 64
**6.5 Comparação das abordagens para polarização de sentenças ..... 66**
**7 MINERAÇÃO DE OPINIÕES EM NíVEL DE ASPECTO E SUMARIZAÇÃO ..... 68**
**7.1 Mineração em Nível de Aspecto ..... 68**
**7.2 Sumarização visual de opiniões ..... 69**
**8 CONCLUSÕES E TRABALHOS FUTUROS ..... 71**
**REFERÊNCIAS ..... 73**

# 1 INTRODUÇÃO

O advento de redes sociais e de outros meios de comunicação tornou o conteúdo disponível na WEB mais dinâmico e massivo. As pessoas passaram a compartilhar suas opiniões, ideias, sentimentos e críticas ao invés de apenas obter informações em documentos virtuais. Postagens sobre assuntos variados, notícias de celebridades, produtos e movimentos sociais realizados em mídias virtuais contém emoções como alegria, felicidade, decepções, tristeza e raiva. As opiniões podem estar relacionadas à qualidade do atendimento médico em um determinado município, ou ao tempo de duração da bateria de um celular. Esses pensamentos possuem um grande valor estratégico para organizações que têm interesse em obter informações sobre seus produtos, serviços, marketing, propaganda, etc. Descobrir a opinião pública é um fator importante para o processo de tomada de decisões. Devido à grande quantidade de conteúdo subjetivo existente em documentos textuais na WEB, a atividade de recuperação e de interpretação desse tipo de informação é complexa. Esse problema torna-se ainda maior quando é necessário extrair e classificar a subjetividade em diferentes fontes de dados. Por tais razões, a WEB vem sendo minerada com uma combinação de técnicas de descoberta de conhecimento em bases de dados e processamento de linguagem natural, em uma área denominada Mineração de Opiniões (PANG; LEE, 2008; TSYTSARAU; PALPANAS, 2012; LIU, 2012; BECKER, 2014).

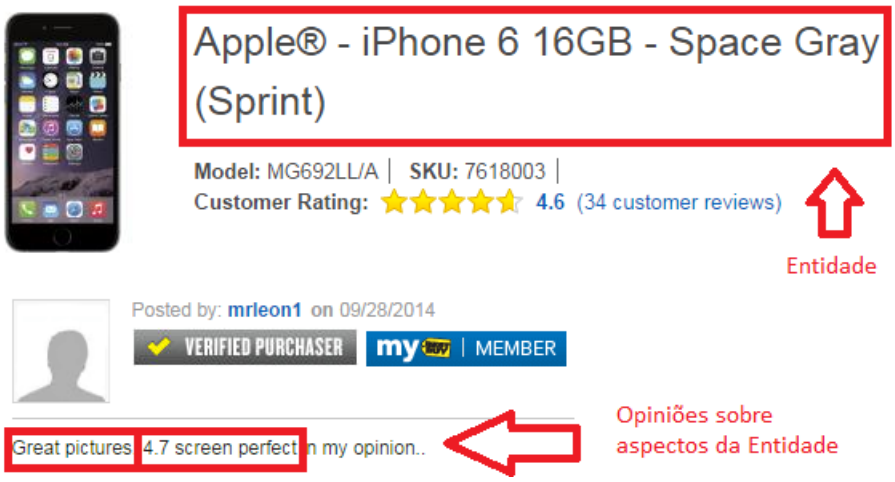
A Mineração de Opiniões tem como objetivo extrair, processar e classificar opiniões descritas em mídias virtuais sobre os mais diversos assuntos. Ela permite classificar a opinião sobre aspectos e características de entidades através de classes, sendo as mais usuais as polaridades positivas, negativas e neutras (PANG; LEE, 2008; QIU et al., 2011; CASTELLANOS et al., 2011; TSYTSARAU; PALPANAS, 2012; LIU, 2012).

Os trabalhos pioneiros na área de Mineração de Opiniões buscaram identificar a opinião global de usuários contida em revisões sobre produtos e serviços (TURNEY, 2002; PANG; LEE; VAITHYANATHAN, 2002). Com o passar dos anos, a opinião passou a ser detalhada em termos dos diferentes aspectos de cada entidade. Exemplos de aspetos de uma entidade computador seriam o preço, peso ou desempenho. Essas informações mais detalhadas quando classificadas e sumarizadas são extremamente úteis e podem influenciar potenciais consumidores (HU; LIU, 2004; GUO et al., 2009; QIU et al., 2011; LIU et al., 2013).

Revisões de produto são mais propícias para a realização da mineração de opiniões pois o conteúdo textual, neste contexto, é altamente focado e rico em opiniões, o que permite realizar a mineração de opiniões mais facilmente do que em outras formas de conteúdo (LIU,

2012). Documentos textuais neste âmbito permitem identificar de maneira relativamente simples opiniões sobre uma entidade ou sobre seus aspectos. No exemplo de revisão apresentado na Figura 1.1, é ilustrada uma revisão de produto extraída do site Best Buy. Percebe-se que a entidade-alvo da opinião é clara, um iPhone 6. É possível também identificar opiniões sobre os aspectos imagem (*picture*) e tamanho de tela (*4.7 screen*) da entidade, que são detalhados no texto da revisão. Fontes de opiniões assim são denominadas fortemente estruturadas pois (i) possuem metadados que permitem identificar qual a entidade alvo das opiniões, a qual não precisa portanto ser explicitamente citada no texto da revisão, (ii) pressupõe-se que os comentários terão opiniões e (iii) os comentários possuem menções explícitas à entidade alvo e/ou seus aspectos.

Figura 1.1: Revisão sobre produto no site Best Buy.




Fonte: Adaptado de Best Buy (2014).

Revisões de produto são, portanto, fontes de opiniões fortemente estruturadas e permitem minerar opiniões em nível de aspecto mais facilmente. Já em fontes de opiniões fracamente estruturadas, como publicações em redes sociais e comentários de notícias, minerar a opinião sobre aspectos de entidades é uma atividade bem mais complexa e difícil, já que estas fontes podem conter opiniões sobre múltiplas entidades, ou não conter nenhuma opinião. O presente trabalho considera como fontes de opiniões fracamente estruturadas aquelas cujos documentos textuais não possuam necessariamente opiniões e não é simples identificar o alvo das opiniões, seja uma entidade ou aspecto dessa entidade.

## 1.1 Problema

Neste trabalho, busca-se minerar opiniões em nível de aspecto sobre candidatos a eleições, usando como fonte de dados comentários de leitores de jornais. O desafio deste trabalho é identificar e classificar a opinião baseada em aspectos a partir de fontes de opiniões fracamente estruturadas. A mineração de opiniões será aplicada em fontes de dados relacionadas com eleições, já que a plataforma eleitoral de candidatos inclui propostas relevantes para a população em diversas áreas, tais como saúde, educação, segurança, as quais são exploradas durante uma campanha com o intuito de obter votos. Essas áreas são tratadas neste trabalho como os aspectos de um candidato sobre os quais os leitores podem opinar. Assim, parte-se da premissa que, além da percepção global de um candidato, é possível identificar a opinião pública em relação a aspectos específicos deste. Por exemplo, deseja-se identificar que a percepção sobre um candidato X em relação à saúde é mais positiva que a do candidato Y. Contudo, no que se refere à educação, a percepção é negativa.

O problema tratado nesta pesquisa é relevante pois existem poucos trabalhos que se propõem a minerar a opinião sobre aspectos em fontes de opinião fracamente estruturadas. Opiniões expressas em fontes de opiniões menos estruturadas, tais como redes sociais (TUMASJAN et al., 2010; CASTELLANOS et al., 2011), blogs e notícias (BALAHUR et al., 2010), são mais difíceis de serem mineradas, pois as tarefas de encontrar o conteúdo de opinião e o seu alvo são bem mais complexas. Como mencionado, fontes de opinião assim podem conter sentimentos sobre múltiplas entidades ou mesmo podem não conter nenhuma opinião. Além disso, existem situações nas quais o alvo da opinião é implícito ou cujo conteúdo textual expressa sentimentos sobre vários domínios (e.g. informática, cinema), diferente de revisões de produto nas quais o domínio abordado pelos comentários é único, e a entidade sendo revisada também é única. Considerando a notícia da Folha de São Paulo, ilustrada na Figura 1.2, as figuras 1.3 a 1.6 apresentam comentários de leitores relacionados a esta notícia, que ilustram o problema abordado.

A notícia relata desafios relacionados à educação que os candidatos à prefeitura de São Paulo teriam caso eleitos. Espera-se que os comentários dos leitores da notícia sejam sobre esse mesmo assunto, ou seja, educação.

Figura 1.2: Trecho de notícia sobre educação extraído da Folha de São Paulo.

**Entenda os desafios e as propostas dos candidatos de SP para a educação**

FÁBIO MAZZITELLI
DIÓGENES CAMPANHA
DE SÃO PAULO

*15/09/2012 05h52*

Recomendar 143 | Tweetar 2 | g+1 3 | OUVIR O TEXTO | Mais opções

Criado como escola-modelo para revitalizar áreas da periferia da capital, o CEU (Centro Educacional Unificado) se firmou como política pública nas três últimas gestões, tornou-se uma rede de 45 escolões, e o rumo desse modelo deve determinar o desenho da educação municipal nos próximos anos.

O futuro prefeito vai definir se a cidade investirá na expansão de CEUs, como propõem Celso Russomanno e Fernando Haddad, ou se o melhor é aproveitar ao máximo a estrutura já em funcionamento, como sugerem José Serra e Gabriel Chalita.

Fonte: Adaptado de Grupo Folha (2014).

Figura 1.3: Comentário sobre educação existente na notícia da Figura 1.2.

Assunto da notícia

Deixando de lado a política,a educação e a saúde deveriam ficar por conta do Estado. O prefeito deveria cuidar sómente da cidade, com seus diversos problemas ( limpeza, áreas verdes, asfalto, buracos, ambulantes, zoneamento, obras viárias, moradores de rua, etc,etc) Na verdade o prefeito de uma cidade como S. Paulo, precisa ser quase um "feitor", sem seu lado pejorativo, e zelar 24 hs. pela cidade.

Fonte: Adaptado de Grupo Folha (2014).

Figura 1.4: Comentário cuja opinião tem como alvo educação e candidatos.

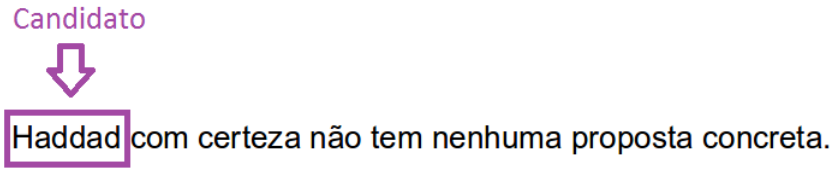
Assunto da notícia

Projetos para a educação não devem ser que nem uma lista de supermercado ou ideias mirabolantes e impossiveis que vão prejudicar enormemente a rotina dos estudantes ja matriculados (viu Serra e Russomano?. O programa de governo de Haddad é super bem escrito e esclarecedor, não joga promessas no ventilador sem nos permitir vislumbrar como de fato as por em pratica. Fica a dica para os criticos-mala , Leiam!!! No site do PenseNovo!

Candidato

Fonte: Adaptado de Grupo Folha (2014).

Figura 1.5: Comentário cuja opinião tem como alvo um dos candidatos.




Fonte: Adaptado de Grupo Folha (2014).

Figura 1.6: Comentário sobre outros assuntos que não são relacionados ao conteúdo da notícia.

Para que pesquisar no Google, temos em todas as manchetes "mensalão". Muito melhor, muitos assuntos e matérias.SP não corre perigo nenhum, em breve não teremos mais nenhum solto. Vão certamente descobrir milhares de outras falcatruas e todos serão presos.

Fonte: Adaptado de Grupo Folha (2014).

A Figura 1.3 apresenta um exemplo em que a opinião contida no comentário apresenta como alvo o mesmo aspecto apresentado na notícia. O comentário da Figura 1.4 opina sobre o aspecto educação e sobre alguns dos candidatos a prefeito. O comentário da Figura 1.5 apresenta uma opinião sobre a proposta de um dos candidatos, mas não descreve de forma explícita que a proposta é sobre educação. A Figura 1.6 ilustra um exemplo de um comentário sobre assuntos não abordados na notícia apresentada acima e que não menciona nenhum candidato.

## 1.2 Objetivo

O objetivo deste trabalho é apresentar um estudo de caso que se propõe minerar a opinião pública sobre entidades eleitorais em nível de aspecto, usando fontes de opiniões fracamente estruturadas, mais especificadamente de comentários de leitores de jornais. Para tanto, busca-se:

- desenvolver um conjunto de experimentos para avaliação de técnicas de identificação de aspectos em fontes de opinião fracamente estruturadas;
- desenvolver um método para distinguir a opinião global sobre um candidato, e a opinião sobre aspectos;
- representar as opiniões usando um modelo de sumarização.

## 1.3 Contribuições

Esta pesquisa contribui para a mineração de opiniões em nível de aspecto em um contexto diferente de revisão de produtos. De acordo com a revisão da literatura, esse é um dos trabalhos pioneiros de mineração de opiniões em nível de aspecto em fontes de opinião fracamente estruturadas. As três principais contribuições são apresentadas abaixo:

- na concepção de uma abordagem para identificação da opinião em nível de aspecto sobre entidades eleitorais em comentários de notícias políticas;
- na aplicação de um método baseado em aprendizagem de máquina para classificar a opinião sobre entidades e seus aspectos em três classes (positivo, negativo e neutro);
- na representação da sumarização de opinião sobre entidades e seus aspectos.

Este trabalho integra os resultados desenvolvidos pelo grupo de pesquisa em Mineração de Opiniões do Instituto de Informática da UFRGS. O problema de extração e de classificação de opiniões a partir de comentários de notícias de jornais já foi tratado nos trabalhos de Tumitan and Becker (2013) e Tumitan and Becker (2014), que trataram da extração de notícias e de comentários, bem como da mineração de opiniões em nível de sentença sobre candidatos a eleições. O presente trabalho complementa essas pesquisas, realizando a mineração de opiniões em nível de aspecto. Resultados parciais foram apresentados à comunidade científica em Sapiras and Becker (2014a) e Sapiras and Becker (2014b).

## 1.4 Organização do trabalho

O trabalho está estruturado da seguinte forma. O Capítulo 2 apresenta a fundamentação teórica. O Capítulo 3 realiza uma discussão a respeito dos trabalhos relacionados à Mineração de Opiniões. O Capítulo 4 apresenta a abordagem proposta para minerar opiniões em nível de aspecto. O processo de identificação de aspectos é discutido no Capítulo 5. As técnicas utilizadas para classificar as opiniões são apresentadas no Capítulo 6. O Capítulo 7 expõe a sumarização desenvolvida e os resultados obtidos com a abordagem implementada. Por fim, o Capítulo 8 apresenta as conclusões e as perspectivas de trabalhos futuros.

# 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo apresentar os conceitos necessários para o entendimento do trabalho. São apresentadas definições e conceitos relativos ao processo de KDD, à mineração de opiniões em documentos textuais, à identificação de aspectos, à classificação de subjetividade e à sumarização de resultados.

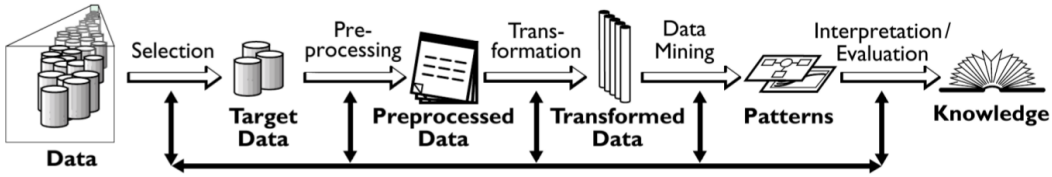
## 2.1 Mineração de dados

A mineração de dados apresenta-se como um processo cujo significado e cujas etapas são descritos a seguir.

### 2.1.1 Processo de descoberta de conhecimento em bases de dados

O trabalho utiliza técnicas de mineração de dados, que, segundo Fayyad, Piatetsky-Shapiro and Smyth (1996), é uma das etapas do processo de descoberta de conhecimento em bases de dados (KDD - Knowlegde Discovery in Databases). O processo de KDD é composto por uma série de etapas que são apresentadas na Figura 2.1 e descritas abaixo:

Figura 2.1: Processo de descoberta de conhecimento em bases de dados.




Fonte: Fayyad, Piatetsky-Shapiro and Smyth (1996).

1. **Seleção de dados**: nessa etapa, é criado um conjunto de dados-alvo, selecionados de acordo com os objetivos do processo de descoberta de conhecimento que será realizado.
2. **Pré-processamento**: essa fase tem como objetivo tratar os dados previamente selecionados, realizando a limpeza dos dados e remoção de ruídos.
3. **Transformação**: transforma os dados de acordo com a tarefa de mineração de dados. Fazem parte dessa etapa a redução de dimensionalidade do conjunto de dados (quantidade de atributos), a redução de registros existentes, entre outros.

4. **Mineração de dados**: nessa etapa, conforme o objetivo de negócio, é escolhida uma tarefa de mineração. O algoritmo é escolhido e parametrizado de acordo com as características do conjunto de dados-alvo.
5. **Interpretação e avaliação dos resultados**: os resultados obtidos são interpretados e avaliados. Medições estatísticas ou métodos de teste de hipóteses podem ser aplicados com o objetivo de eliminar resultados não legítimos.

A mineração de opiniões utiliza o processo de descoberta de conhecimento em bases de dados descrito acima.

### 2.1.2 Tarefas da mineração de dados

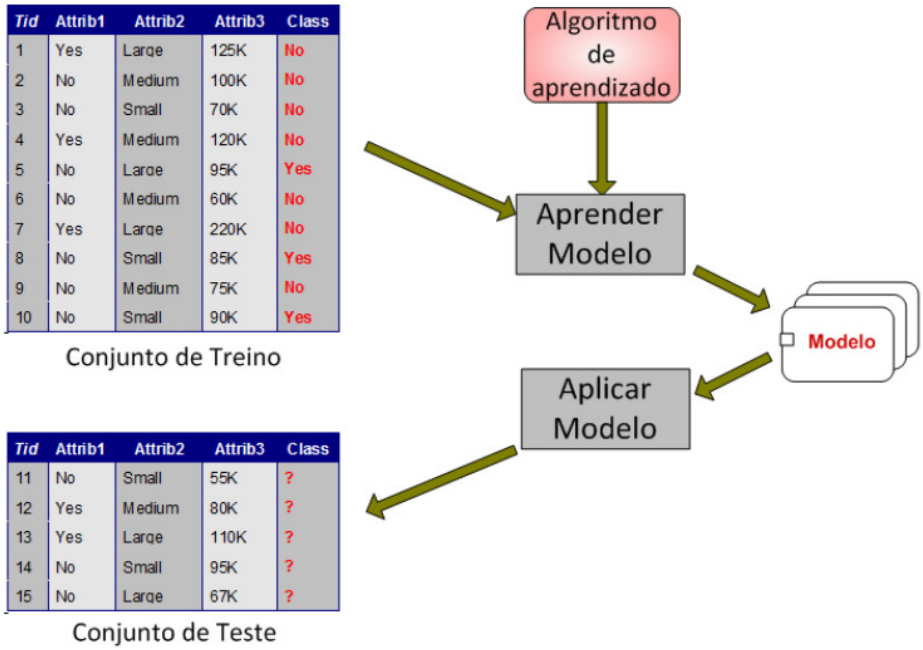
As tarefas de mineração de dados podem ser divididas em tarefas de previsão e em tarefas descritivas. As tarefas de previsão têm como objetivo prever o valor de um determinado atributo com base nos valores de outros atributos. Já nas tarefas descritivas, o objetivo é derivar padrões. As tarefas descritivas são geralmente explanatórias e frequentemente necessitam de técnicas de pós-processamento para avaliar os resultados obtidos. Essas tarefas também podem ser classificadas como (i) Análise de Agrupamentos, (ii) Modelagem Preditiva, (iii) Análise de Associação e (iv) Detecção de Anomalias (TAN; STEINBACH; KUMAR, 2005). A modelagem preditiva é a tarefa de construir um modelo para uma determinada variável alvo como uma função das variáveis explicativas ou discriminatórias. Existem dois tipos de modelagem preditiva: a classificação, em que o alvo da previsão é uma variável discreta, e a regressão, que tem como alvo variáveis contínuas (TAN; STEINBACH; KUMAR, 2005).

O presente trabalho faz uso de modelos de classificação para identificar se uma determinada sentença é positiva, negativa ou neutra. Para isso, são utilizadas técnicas de classificação, descritas na Seção 2.1.3.

### 2.1.3 Técnicas de classificação de dados

A classificação é a tarefa de categorizar objetos, tais como a identificação de e-mails spams, categorização de células malignas ou benignas ou a classificação de galáxias com base em seus formatos (TAN; STEINBACH; KUMAR, 2005). Uma técnica de classificação recebe como entrada um conjunto de registros (também denominados exemplos ou instâncias). Cada registro é um par $(x, y)$, em que $x$ é o conjunto de atributos explicativos e $y$ o atributo especial,

Figura 2.2: Abordagem utilizada para o desenvolvimento de modelos de classificação.




Fonte: (TAN; STEINBACH; KUMAR, 2006 apud BECKER, 2014).

denominado rótulo da classe, atributo alvo ou de categorização. O conjunto de atributos $x$ pode possuir valores discretos ou contínuos, enquanto que $y$ deve ser um atributo discreto. Essa é a característica-chave que distingue modelos de classificação de modelos de regressão. Os modelos de classificação são utilizados para prever o rótulo da classe de registros cujo atributo $y$ seja desconhecido. O modelo de classificação pode ser considerado um mecanismo que atribui automaticamente um rótulo de classe quando recebe o conjunto de atributos de um registro desconhecido (TAN; STEINBACH; KUMAR, 2005).

Uma técnica de classificação é uma abordagem sistemática para a criação de modelos de classificação a partir de um conjunto de dados de entrada. São exemplos de técnicas de classificação árvores de decisão, classificadores baseados em regras, redes neurais, algoritmos Support Vector Machine (SVM) e classificadores Bayesianos. Essas técnicas utilizam um algoritmo de aprendizagem para identificar o modelo mais apropriado para o relacionamento entre o conjunto de atributos e o rótulo da classe dos dados de entrada. A abordagem para a construção de um modelo de classificação é apresentada na Figura 2.2.

Um conjunto de registros cujos rótulos sejam conhecidos são fornecidos como entrada. Esse conjunto também é conhecido como Conjunto de Treinamento (ou *Training Set*) e é utilizado para construir um modelo de classificação. O modelo criado é então aplicado em um Conjunto de Testes (*Test Set*), que possui registros com rótulos de classes desconhecidos.

A avaliação de um classificador pode ser realizada por meio de métricas de desempenho, como Precisão, Revocação, F-score e acurácia (AGGARWAL; ZHAI, 2012). Quanto maior a precisão, menor é a quantidade de erros falso positivos cometidos pelo classificador. A revocação mede a fração de exemplos positivos classificados corretamente pelo classificador. A precisão e a revocação podem ser resumidas em uma métrica denominada F-score, que representa a média harmônica entre a precisão e a revocação. A acurácia é a proporção de resultados classificados corretamente sobre o total de registros analisados pelo classificador (TAN; STEINBACH; KUMAR, 2005).

## 2.2 Mineração de opiniões

A mineração de opiniões é uma área de estudos que abrange pesquisas de mineração de dados, linguística computacional, recuperação de informações e inteligência artificial. Ela pode ser definida como um estudo realizado computacionalmente para processar opiniões, sentimentos, emoções e subjetividade expressos em forma textual (PANG; LEE, 2008; TSYTSARAU; PALPANAS, 2012; LIU, 2012; BECKER; TUMITAN, 2013). A mineração de opiniões é dividida nas etapas listadas abaixo e apresentadas na Figura 2.3:

- identificação de documentos textuais com textos subjetivos sobre um assunto ou uma entidade;
- classificação da opinião em diferentes classes, tais como positivo, negativo ou neutro;
- sumarização dos resultados obtidos, a partir de modelos de visualização. Os resultados sumarizados podem ser a entrada para outras aplicações (e.g. predição baseada em sentimento).

Uma opinião é definida formalmente por Liu (2012) como uma quíntupla ($e_i$, $a_{ij}$, $s_{ijkl}$, $h_k$, $t_l$), onde:

$\mathbf{e_i}$ : é nome de uma entidade;

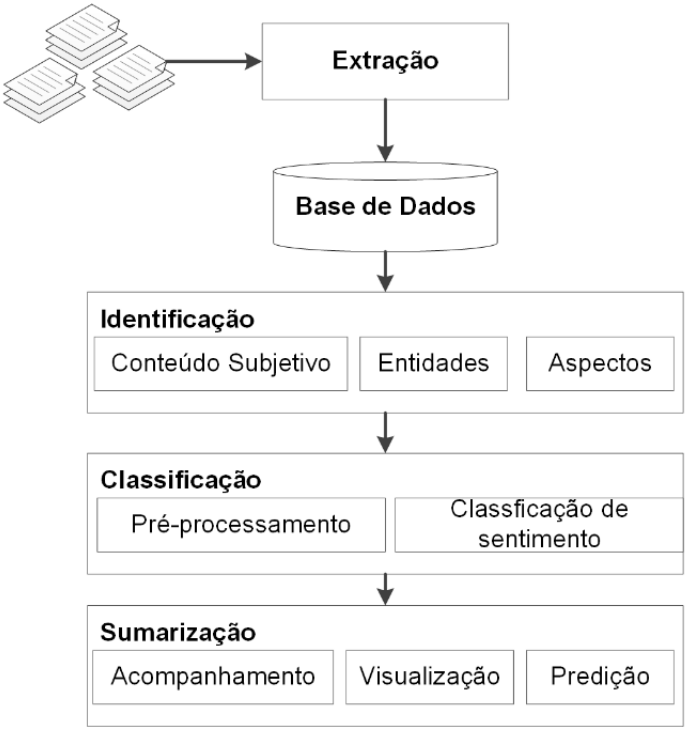
$\mathbf{a_{ij}}$ : é o aspecto da entidade $e_i$. Um aspecto também é denominado tópico;

$\mathbf{s_{ijkl}}$ : é a opinião sobre aspecto $a_{ij}$ da entidade $e_i$;

$\mathbf{h_k}$ : é a entidade que expressa a opinão, também chamado de fonte de opinião;

$\mathbf{t_l}$ : é o tempo no qual a opinião foi expressa por $h_k$.

Figura 2.3: Etapas da mineração de opiniões.




Fonte: Becker and Tumitan (2013).

A opinião $s_{ijkl}$, expressa sobre uma entidade ou aspecto, é medida em termos de uma polaridade, podendo ser categorizada em classes, tais como: positiva, negativa ou neutra (TSYTSARAU; PALPANAS, 2012).

Em fontes de opiniões fortemente estruturadas, a entidade-alvo de uma opinião está fora do conteúdo de um comentário, enquanto que a opinião sobre a entidade e seus aspectos é descrita no conteúdo dos comentários, como já foi ilustrado na Figura 1.1, a partir da qual é possível visualizar a descrição da entidade e as opiniões sobre os aspectos.

Já em fontes de opinião fracamente estruturadas, tais como Twitter e comentários de leitores de jornais, a identificação das entidades é realizada a partir de menções no próprio *tweet* ou comentário. Para identificar os aspectos, geralmente são considerados substantivos ou locuções substantivas (LIU, 2012).

Este trabalho irá contribuir para essas três etapas. Quanto à etapa de identificação, propõe um estudo para identificar aspectos, o que complementa trabalhos anteriores de identificação de entidades (TUMITAN; BECKER, 2013). Quanto à etapa de classificação, será abordada utilizando o processo proposto por Tumitan and Becker (2014), mas agregando contribuições quanto à classificação de polaridade em três classes. Finalmente, será abordada a etapa de sumarização de resultados, agregando as opiniões por entidade e aspecto usando por meio de

modelos visuais de sumarização.

### 2.2.1 Níveis de análise textual

A mineração de opiniões pode ser realizada em diferentes níveis de análise, sendo os mais comuns (LIU, 2012; BECKER, 2014):

**Documento:** nessa granularidade, a tarefa é classificar se um documento é, como um todo, positivo ou negativo. A análise em nível de documento é adequada quando o documento menciona apenas uma entidade.

**Sentença:** determina o sentimento de uma sentença em um documento. Esse tipo de análise é útil quando um documento apresenta opiniões sobre mais de uma entidade. Ela permite também identificar e distinguir sentenças objetivas (fatos) de sentenças subjetivas (opiniões).

**Entidade e aspecto:** tem como foco a opinião expressa. O objetivo é identificar qual o alvo da opinião, ou seja, a entidade ou algum de seus aspectos.

Em determinadas situações, classificar a opinião sobre entidades nos níveis de documento e sentença não é suficiente. Mesmo que um documento apresente uma opinião positiva geral sobre uma única entidade, não significa que tal documento não possa apresentar alguma opinião negativa sobre a mesma entidade, variando o aspecto de opinião. Uma análise mais detalhada das opiniões descritas em um documento pode ser realizada por meio da Mineração de Opiniões Baseada em Aspectos (HU; LIU, 2004; LIU, 2012).

Este trabalho irá abordar a mineração de opiniões em nível de aspecto, cujo processo é dividido nas mesmas etapas descritas na Figura 2.3. A tarefa de extração de aspectos identifica todos os atributos (ou *features*) de uma entidade que estão descritos em um documento ou sentença. Para o exemplo “A duração da bateria e imagem deste tablet são excelentes, mas o som precisa melhorar”, essa etapa irá extrair os aspectos “duração da bateria”, “imagem” e “som”. A etapa de classificação de sentimento irá utilizar técnicas estatísticas, baseadas em aprendizagem de máquina ou dicionários para identificar a polarização de opinião sobre cada aspecto dentro da sentença analisada. Já a etapa de sumarização irá apresentar, por meio de gráficos e outras técnicas de visualização, a opinião geral sobre uma entidade e seus aspectos.

### 2.2.2 Tipos de opiniões

As opiniões podem ser regulares ou comparativas. Nas opiniões regulares, o autor expressa uma opinião sobre um alvo (e.g. "A cerveja belga é realmente boa"). Em opiniões comparativas, uma sentença ou documento apresenta opiniões sobre mais de uma entidade ou aspecto, de forma a realizar uma comparação entre eles uma vez estabelecido um referencial (e.g. "A cerveja belga é melhor que a cerveja americana") (LIU, 2012).

As opiniões regulares dividem-se em opiniões diretas e indiretas. Nas opiniões diretas, o sentimento é expresso diretamente sobre uma entidade ou aspecto (e.g. "A qualidade de som deste televisor é ótima"). Já nas opiniões indiretas, o sentimento é expresso indiretamente sobre uma entidade (e.g. "Após assistir ao filme X, fiquei contente por ter comprado as entradas de cinema").

As opiniões podem ainda ser caracterizadas como explícitas e implícitas (BECKER, 2014). No exemplo "Carreguei a bateria há duas horas e já acabou", é expressa uma opinião implícita negativa para o aspecto "bateria".

Tal como na maioria dos trabalhos em Mineração de Opiniões (TURNEY, 2002; HU; LIU, 2004; CASTELLANOS et al., 2011; QIU et al., 2011), este trabalho pressupõe que as opiniões são regulares, diretas e explícitas.

### 2.2.3 Técnicas para classificação de opiniões

São utilizadas abordagens baseadas em dicionário, aprendizagem de máquina, estatísticas ou semânticas para realizar a classificação de uma determinada opinião (TSYTSARAU; PALPANAS, 2012).

As abordagens baseadas em dicionário, também conhecidas como abordagens léxicas ou linguísticas, utilizam dicionários de sentimentos, que contêm conjuntos de palavras ou expressões associados a cada classe de sentimento (e.g. palavra "bom" como positiva ou "ruim" como negativa). A composição básica de um léxico de sentimento é a palavra de sentimento e sua respectiva polaridade, expressa como uma categoria, valor ou escala (SILVA; CARVALHO; SARMENTO, 2012). Estas técnicas são bastante dependentes da qualidade dos dicionários existentes.

Os léxicos possuem diferenças entre si, tais como números de termos, e informações adicionais, como flexões, lema, *part-of-speech tags*, etc. Na Tabela 2.1, são apresentados alguns léxicos e suas propriedades. Existem dicionários de opinião para diversos idiomas, e destaca-se

o SentiLex-PT para o idioma português de Portugal, e o OpLexicon, que possui expressões de opinião para o português do Brasil (SOUZA M.; VIEIRA, 2011).

Tabela 2.1: Comparação entre dicionários de opinião.

| **Dicionário** | **Positivos** | **Negativos** | **POS** | **Radical** | **Lema** | **Idioma** |
|---|---|---|---|---|---|---|
| General Inquirer | 1.915 | 2.291 | Sim | Não | Não | Inglês |
| OpinionFinder | 2.718 | 4.912 | Sim | Sim | Não | Inglês |
| OpLexicon | 8.675 | 14.469 | Sim | Não | Não | Português |
| SentiLex-PT | 82.347 entradas | | Sim | Não | Sim | Português |
| SentiWordNet | 117.659 entradas | | Sim | Não | Sim | Inglês |

Fonte: Tumitan and Becker (2013).

Na mineração de opiniões, observa-se o maior uso de métodos de aprendizagem supervisionada, tais como de classificação e regressão. Os algoritmos de classificação mais utilizados são Support Vector Machine, Naïve Bayes, Maximum Entropy e baseados em redes neurais (DAVE; LAWRENCE; PENNOCK, 2003; PANG; LEE, 2008; TSYTSARAU; PALPANAS, 2012; BECKER, 2014). Embora apresentem bons resultados, são altamente dependentes de um corpus anotado para aprendizado de tamanho significativo.

As técnicas estatísticas baseiam-se na premissa de que termos que traduzem opiniões frequentemente são encontrados juntos no corpus de textos. Ou seja, se uma dada palavra ocorrer mais frequentemente com palavras definidas como positivas no mesmo contexto, então é provável que tal palavra também seja positiva. Essas técnicas só podem ser aplicadas quando houver um corpus de tamanho significativo para validação das frequências dos termos. Os algoritmos estatísticos mais utilizados são Pointwise Mutual Information (PMI) (TURNEY, 2002), Semantic-orientation Latent Semantic Analysis (SO-LSA) e Latent Dirichlet Allocation (LDA) (MAOWEN et al., 2012).

Abordagens Semânticas são semelhantes às estatísticas, com a diferença que a polaridade é calculada em termos de alguma medida de distância entre termos. O princípio das técnicas nessa categoria é que termos semanticamente próximos devem possuir a mesma polaridade. A abordagem semântica pode ser utilizada como complemento às abordagens descritas anteriormente, como forma de expansão ou descoberta de um vocabulário específico, caso não existam bons léxicos de opinião (BECKER, 2014).

Este trabalho irá utilizar uma técnica de aprendizado de máquina para classificar a polaridade das opiniões, por ter apresentado resultados significativamente superiores aos de técnicas baseadas em dicionário (TUMITAN; BECKER, 2014).

# 3 TRABALHOS RELACIONADOS

Este capítulo irá apresentar os trabalhos relacionados ao processo de mineração de opiniões, com maior foco para aqueles relacionados às etapas de identificação de aspectos e de sumarização de resultados.

## 3.1 Técnicas para identificação de aspectos na mineração de opiniões

Uma opinião pode possuir como alvo uma entidade ou um aspecto dessa entidade. Esta seção irá descrever conceitos e técnicas para identificar aspectos na mineração de opiniões.

As atividades de identificação de aspectos na mineração de opiniões dividem-se em identificação de aspectos explícitos ou implícitos. A sentença "A duração de bateria do iPhone é excelente" apresenta uma opinião positiva explícita para o aspecto "bateria" da entidade "iPhone". Liu (2012) descreve que quando uma sentença apresentar uma opinião direta sobre uma entidade, sem que seja descrito um aspecto dela, existe a possibilidade de tal opinião na verdade ter como alvo aspectos implícitos. Por exemplo, para a sentença "O iPhone é muito caro.", existe uma opinião negativa implícita para o aspecto "preço" da entidade "iPhone". Por ser mais fácil de ser efetuada, a maioria dos trabalhos que realizam mineração de opiniões em nível de aspecto tem como foco a identificação de aspectos explícitos.

Liu (2012) descreve quatro abordagens principais para identificação de aspectos:

1. frequência de substantivos ou locuções substantivas;
2. relação entre a opinião e seu alvo;
3. aprendizado supervisionado;
4. modelagem de tópicos.

Cada abordagem é discutida com mais detalhes no restante desta seção.

### 3.1.1 Frequência de substantivos ou locuções substantivas

A abordagem para identificação de aspectos baseada na frequência de substantivos, como o próprio nome já diz, procura por expressões que são substantivos ou locuções substantivas em documentos ou sentenças de um determinado domínio. Para cada expressão encontrada, é registrada também a frequência na qual ela ocorre. Dessa forma é possível identificar quais

são os aspectos mais comentados em um corpus (LIU, 2012).

O trabalho pioneiro na mineração de opiniões em nível de aspecto é apresentado por Hu and Liu (2004). A fonte de opiniões utilizada por eles são revisões de produtos, ou seja, uma fonte de opiniões bem estruturada. Com o intuito de realizar a identificação dos aspectos a partir de substantivos, eles propõem o uso de regras de associação para encontrar co-ocorrências frequentes de substantivos usados em revisões de um mesmo tipo de produto. Essa abordagem funciona em revisões de produtos pois, nessas fontes, os termos utilizados para representar um aspecto são comuns em um mesmo domínio. No entanto, o mesmo não ocorre em comentários de leitores de jornais, pois os comentários são sobre qualquer domínio, assim como os termos utilizados.

### 3.1.2 Relação entre a opinião e seu alvo

Técnicas de identificação de aspectos baseadas na relação entre a opinião e seu alvo conseguem extrair aspectos a partir do conhecimento prévio de uma palavra de opinião e de padrões léxico-sintáticos ou relação de dependências. Por exemplo, para a sentença "O atendimento do hospital X é ruim", se soubermos que "ruim" é uma opinião negativa, e que "hospital X" é uma entidade, com o uso de uma relação de dependência, é possível inferir que a palavra "atendimento" é um aspecto da entidade "hospital X". Quando não for encontrada uma relação de dependência compatível com uma dada sentença, podem ser utilizadas janelas de palavras para identificar qual substantivo ou locução substantiva está mais próxima de uma opinião ou de uma entidade previamente conhecida. Atualmente o uso de relações de dependência e padrões léxico sintáticos é custoso, pois a variedade de padrões que precisam ser mapeados em fontes de opinião menos estruturadas é muito grande (CASTELLANOS et al., 2011).

Qiu et al. (2011) desenvolvem um método para identificar opiniões e seus alvos em sentenças utilizando o relacionamento entre opiniões e seus alvos. Caso uma sentença possua uma palavra de opinião que exista em um léxico de opinião, os autores identificam o seu alvo a partir do relacionamento entre tal opinião e seu alvo com o auxílio de dependências sintáticas. Caso a sentença possua um termo de alvo conhecido, é possível também fazer o contrário, ou seja, encontrar a palavra de opinião correspondente. A abordagem apresentada por Qiu et al. (2011) não pode ser utilizada no presente trabalho, pois os recursos para tratamento da linguagem que são necessários para identificar padrões sintáticos ou relações de dependências na língua portuguesa são escassos. Tumitan and Becker (2014), em uma experiência com o *parser* "Palavras", o melhor para a língua portuguesa, relatam resultados ruins para o corpus

considerado no presente trabalho. Esse mau desempenho é atribuído ao excesso de erros feitos na língua portuguesa, e à informalidade do vocabulário.

Castellanos et al. (2011) desenvolveram uma ferramenta denominada Live Customer Intelligence (LCI) para monitorar a opinião sobre produtos e serviços no Twitter. O LCI quebra os *tweets* em sentenças, utilizando o OpenNLP[1]. Cada sentença é então rotulada como declarativa, imperativa ou interrogativa a partir de padrões léxico-sintáticos. A extração dos aspectos de cada produto é realizada a partir dos substantivos existentes em cada sentença. Para isso, eles assumem que as palavras de opinião são associadas aos aspectos de entidades verificando a relação entre a opinião e o seu alvo. Todos os substantivos que são considerados alvo de opiniões são extraídos e depois agrupados em categorias, utilizando técnicas de aprendizado de máquina. Castellanos et al. (2011) utilizam substantivos para caracterizar os termos de aspectos.

### 3.1.3 Aprendizado supervisionado

A abordagem baseada em aprendizagem de máquina supervisionada faz uso de corpus anotados para conseguir identificar aspectos em um documento ou sentença. Algoritmos de identificação incluem SVM, cadeias de Markov (HMM) e Conditional Random Fields (CRF).

### 3.1.4 Modelagem de tópicos

Os algoritmos baseados em modelagem de tópicos são os principais métodos para identificação de aspectos em grandes coleções de texto. A modelagem de tópicos é um método de aprendizado não supervisionado que utiliza estatística para identificar a probabilidade de um conjunto de palavras estar relacionada a um dado tópico (ou aspecto), gerando como saída um agrupamento de palavras. Alguns algoritmos para modelagem de tópicos são o pLSA (Probabilistic Latent Semantic Analysis) e o LDA (LIU, 2012).

### 3.1.5 Conclusões sobre identificação de aspectos

O problema tratado neste trabalho não permite o uso das abordagens para extração de aspectos descritas acima em comentários de notícias eleitorais, devido ao fato de essa fonte de opiniões ser fracamente estruturada. Então, este trabalho apresenta uma técnica diferente para

[1]OpenNLP: https://opennlp.apache.org/

identificação de aspectos, que será descrita no Capítulo 5.

## 3.2 Modelos de sumarização de resultados

A etapa de sumarização no processo de mineração de opiniões tem como objetivo criar métricas e sumários para quantificar as opiniões sobre entidades em um conjunto de dados. Um sumário em nível de aspecto pode ajudar a identificar os pontos fortes e fracos sobre os aspectos de entidades. Existem diversas formas de representar a sumarização da opinião, e os modelos podem ser textuais, gráficos, associar as opiniões a informações geográficas ou temporais (BECKER, 2014). São apresentados alguns modelos de sumarização visual desenvolvidos em trabalhos relacionados à mineração de opiniões.

Uma proposta de sumarização baseada em aspectos é apresentado por Hu and Liu (2004), a qual mostra a quantidade de opiniões positivas/negativas de cada aspecto de um produto, sendo selecionados os mais frequentemente citados. Um exemplo de modelo é mostrado na Figura 3.1, no qual a sumarização das opiniões é realizada em nível de aspecto.

Figura 3.1: Modelo de sumarização de opiniões sobre aspectos de um produto.

*Digital_camera_*1:
Feature: **picture quality**
Positive: 253
<individual review sentences>
Negative: 6
<individual review sentences>
Feature: **size**
Positive: 134
<individual review sentences>
Negative: 10
<individual review sentences>
...

Fonte: Hu and Liu (2004).

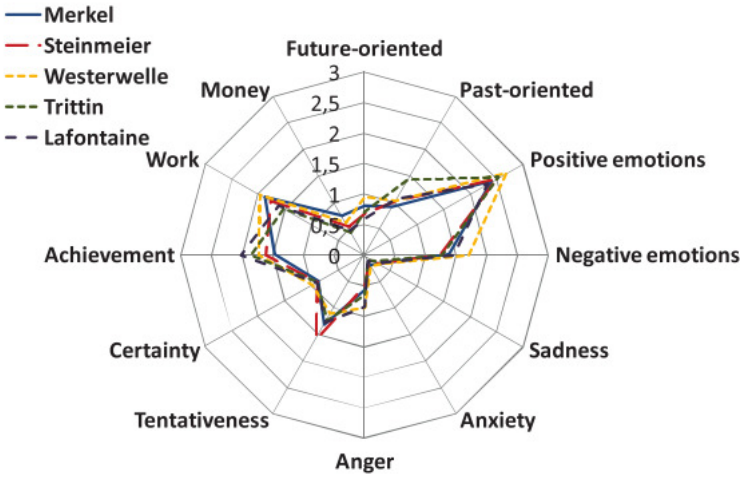
Um dos inconvenientes dessa proposta é a exibição de todas as sentenças individualmente. Dependendo da quantidade de sentenças, a visualização do modelo pode ser comprometida.

A pesquisa apresentada em Tumasjan et al. (2010) tem como objetivo verificar se redes sociais, como o Twitter, podem refletir o sentimento político da população, provendo meios para prever a popularidade de partidos, coalizões ou o resultado de eleições. O estudo de caso

aborda as eleições alemãs. Ao invés de verificar se um *tweet* é positivo, negativo ou neutro, eles classificam as mensagens em termos de emoções (e.g. ansiedade, raiva, certeza, etc...).

O modelo de sumarização, mostrado na Figura 3.2, apresenta as dimensões sobre cinco dos principais candidatos a Chanceler da Alemanha. Ele mostra, por meio de uma escala, qual a relação de cada um dos candidatos com as emoções analisadas. Esse modelo de sumarização não é o ideal para o estudo de caso desenvolvido no presente trabalho, pois não seria possível representar por meio dele a opinião da população sobre os aspectos de cada candidato.

Figura 3.2: Sumarização do sentimento em relação a candidatos alemães.




Fonte: Tumasjan et al. (2010).

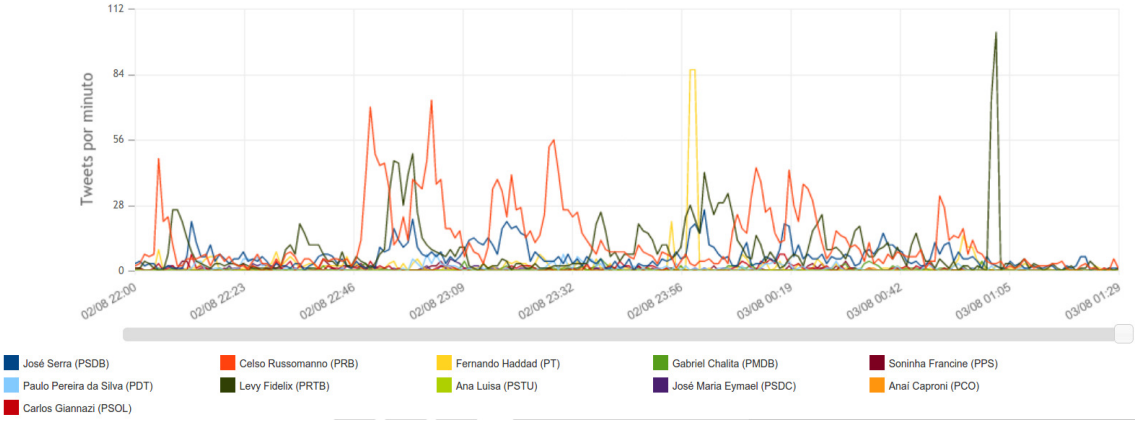
O Observatório da WEB[2] é um portal de monitoramento de *tweets* sobre vários assuntos, entre eles o processo eleitoral. O observatório permite monitorar em *tweets* as menções sobre partidos, candidatos relacionados a algum evento (e.g. Debate entre os candidatos a prefeito de São Paulo). O modelo de sumarização, ilustrado na Figura 3.3, apresenta a evolução de quantidade de menções sobre as entidades eleitorais no tempo. O Observatório da WEB também exibe nuvens de palavras com as menções e *hashtags* mais frequentes.

No entanto, o observatório não realiza mineração de opiniões, preocupando-se apenas em contabilizar as menções sobre partidos e candidatos. Ele também não se propõe a sumarizar opiniões sobre aspectos específicos das entidades analisadas.

O Manchetômetro[3] é um portal para acompanhamento de notícias, as quais tem como alvo uma entidade política. Ele verifica em diversas mídias se a chamada de notícias remete a uma opinião neutra ou negativa sobre um candidato. O modelo de sumarização utilizado pelo

[2]http://www.observatorio.inweb.org.br/eleicoes2012/destaques/
[3]http://www.manchetometro.com.br/

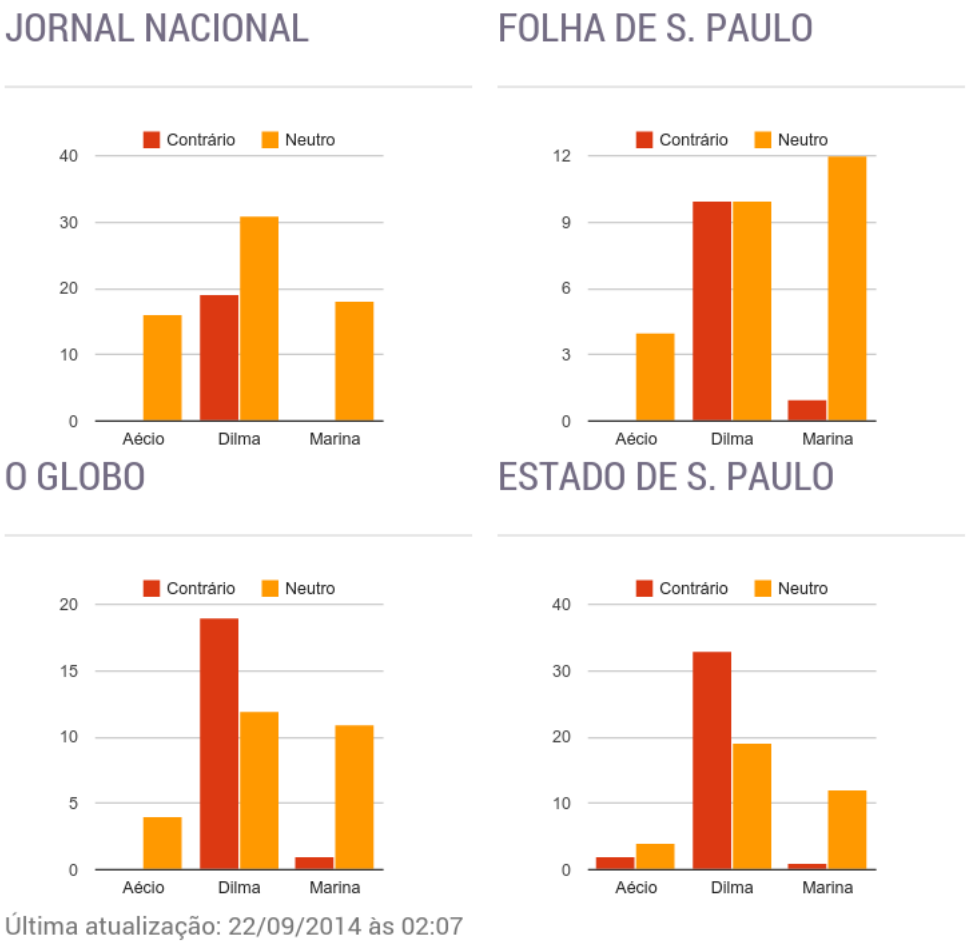
Figura 3.3: Modelo de sumarização apresentado no Observatório da WEB.




Fonte: Instituto Nacional de Ciência e Tecnologia para a Web (2014).

Manchetômetro é ilustrado na Figura 3.4. Nesse modelo, é apresentada por meio de um gráfico de barras, a quantidade de menções neutras e contrárias sobre os candidatos.

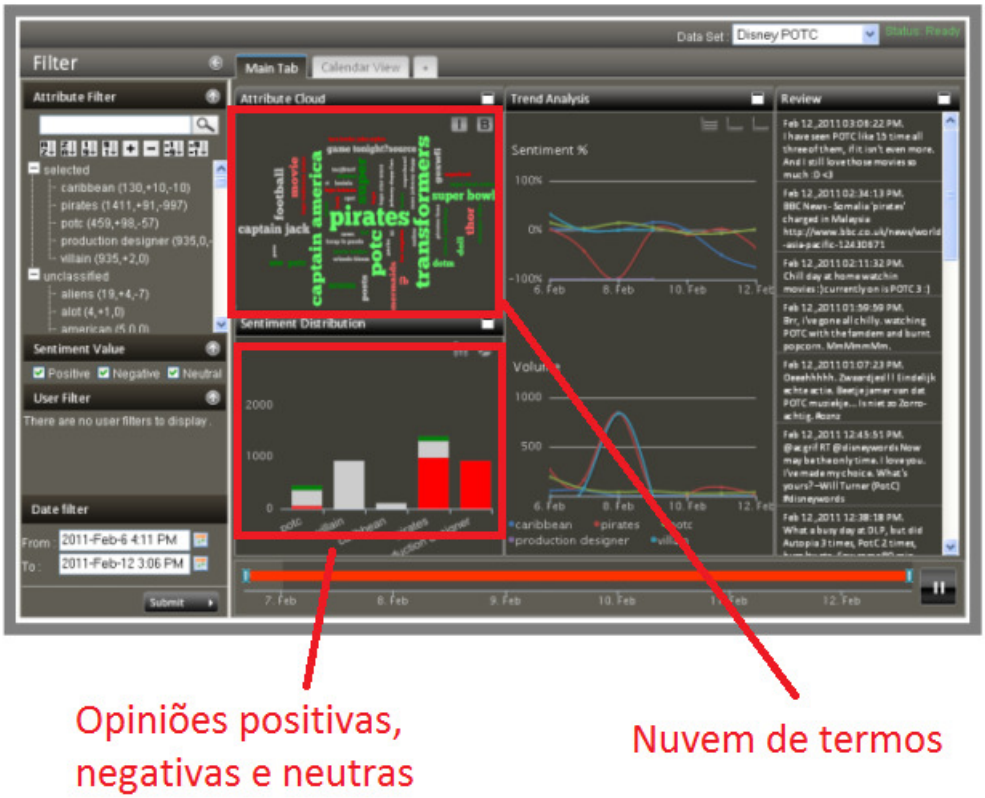
Figura 3.4: Modelo de sumarização do Manchetômetro.




Fonte: Laboratório de Estudos de Mídia e Esfera Pública (2014).

A abordagem de sumarização adotada neste trabalho foi baseada no LCI (Live Customer Intelligence). A Figura 3.5 ilustra o modelo que apresenta a quantidade de opiniões positivas, negativas e neutras sobre os aspectos de um determinado produto ou serviço. Além disso, o LCI também apresenta uma nuvem de palavras (Figura 3.6) com os termos mencionados nos *tweets*.

Figura 3.5: Dashboard de monitoramento do LCI com destaque para o gráfico de opiniões e nuvem de termos.




Fonte: Castellanos et al. (2011).

Dos modelos de sumarização apresentados, a sumarização no formato de gráfico de barras foi escolhido como sendo o mais adequado pelo autor do trabalho, pois permite uma fácil comparação das opiniões sobre as entidades eleitorais e seus aspectos. Outro modelo interessante é o que utiliza nuvens de palavras, pois apresenta visualmente os termos que aparecem com maior frequência.

Figura 3.6: Nuvem de palavras de um filme monitorado pelo LCI.

Fonte: Castellanos et al. (2011).

## 3.3 Mineração de opiniões em comentários sobre notícias políticas

O estudo de caso apresentado no presente trabalho complementa trabalhos realizados pelo grupo de pesquisa em Mineração de Opiniões, os quais identificam, em comentários de leitores do jornal A Folha de São Paulo, a opinião expressa em relação a candidatos a cargos eletivos (TUMITAN; BECKER, 2013; TUMITAN; BECKER, 2014). A Figura 3.7 ilustra o processo proposto por Tumitan and Becker (2014) para prever variações de intenção de votos baseado em atributos de sentimento. As etapas relativas à mineração de opiniões sobre candidatos políticos, identificadas na área com borda tracejada da Figura 3.7, serão complementadas pelas contribuições do presente trabalho.

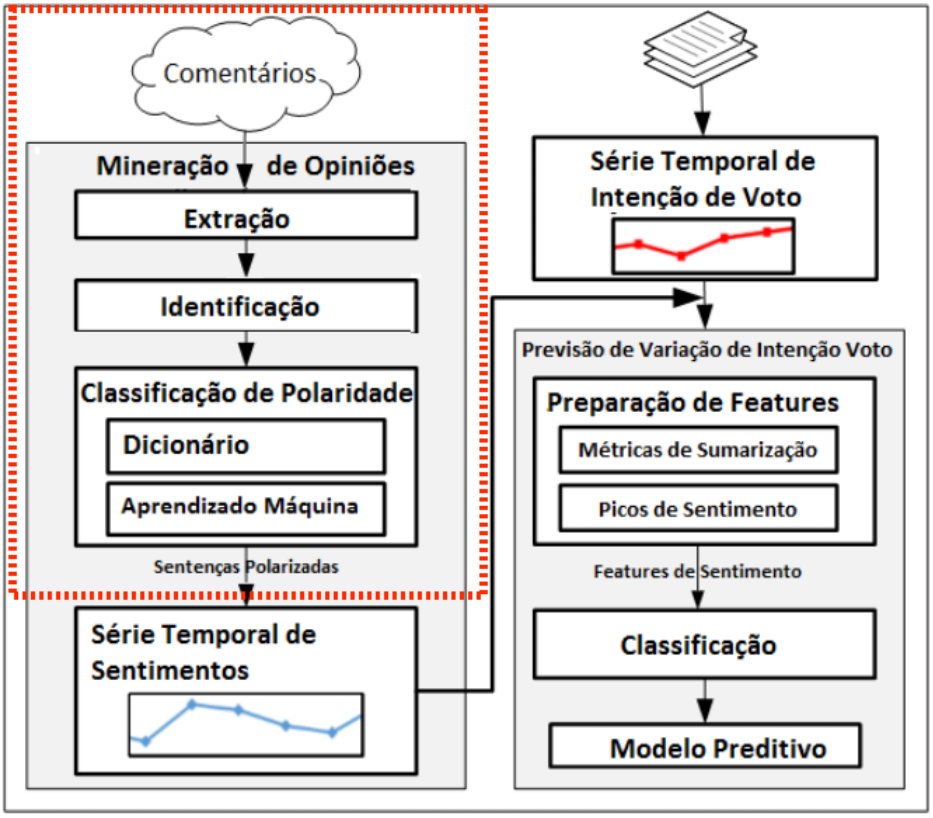
O restante desta seção descreve os aspectos relevantes deste processo necessários ao entendimento do presente trabalho, e destaca as diferenças com base nas extensões propostas na presente dissertação.

### 3.3.1 Descrição do processo de mineração de opiniões

#### *3.3.1.1 Extração*

Esta etapa obtém notícias e comentários de leitores da versão digital da Folha de São Paulo. A extração implementada faz uso do Google Reader como indexador da seção de notícias

Figura 3.7: Framework de Previsão de Variações em Intenções de Voto baseada em Sentimentos.




Fonte: Tumitan and Becker (2014).

políticas da Folha sob o rótulo *Poder*. Foram extraídas notícias e comentários sobre as eleições presidenciais de 2010, eleições governamentais de São Paulo de 2010, bem como das eleições municipais de São Paulo de 2012.

#### *3.3.1.2 Identificação*

Após a extração, são realizadas atividades de remoção de ruídos e pré-processamento. Notou-se, neste corpus, a presença de um elevado número de comentários (quase) duplicados, os quais foram identificados usando técnicas de similaridade. Removeram-se todos comentários com no mínimo 85% de similaridade em relação a outros comentários usando a técnica de Cosseno. Comentários muito curtos (menos de quatro palavras) também foram removidos, pois não continham texto algum, possivelmente devido a um erro de manipulação do sistema, ou as poucas palavras não caracterizavam nenhuma opinião.

Os comentários remanescentes são quebrados em sentenças, utilizando a ferramenta NLTK[4], mais precisamente, um módulo para o idioma português chamado *punkt*. Outras operações também são realizadas, tais como transformação de caracteres para minúsculo e a conver-

[4]Natural Language Toolkit - http://www.nltk.org/

são de caracteres acentuados em não acentuados (e.g. educação -> educacao). A transformação de caracteres visa minimizar o efeito negativo de erros de ortografia frequentemente encontrados nos comentários.

Finalmente, esta etapa busca sentenças que possuam menções às entidades monitoradas. Os candidatos com maior intenção de voto nas respectivas eleições e, portanto, mais comentados são selecionados para monitoração. A abordagem assume que, se um candidato é mencionado em uma sentença, essa sentença tem tal candidato como alvo da opinião contida nela.

Em razão de muitos nomes de candidatos serem escritos com erros de ortografia ou na forma de apelidos (e.g. vampiserra, pastorusso, malddad), buscou-se identificar as variações de escrita nas menções dos candidatos. Para isso foi criado um dicionário de nomes, usando expressões regulares envolvendo (parte) do nome dos candidatos, os quais foram validados manualmente.

#### *3.3.1.3 Classificação da polaridade*

Nesta etapa, as sentenças são classificadas quanto a sua polaridade, na forma de um problema binário (i.e. classes positivo e negativo). Foram experimentadas duas abordagens: baseada em dicionário (TUMITAN; BECKER, 2013) e baseada em aprendizado de máquina (TUMITAN; BECKER, 2014). Um conjunto de sentenças anotadas foi utilizado para validar e/ou treinar classificadores. Tumitan and Becker (2014) utilizaram 1000 sentenças anotadas quanto a sua polaridade para as eleições de 2010 e 600 sentenças anotadas para as eleições de 2012. A anotação do conjunto de sentenças de cada eleição foi realizada por um grupo de três anotadores, observando-se que apenas um anotador era comum aos dois conjuntos. Foram consideradas apenas as sentenças nas quais houve concordância entre, no mínimo, dois anotadores.

Para a abordagem baseada em dicionário, foi utilizado o dicionário de sentimentos SentiLex-PT (SILVA; CARVALHO; SARMENTO, 2012), junto com um dicionário de domínio extraído diretamente a partir dos comentários. Este último contém termos de sentimento encontrados pelos anotadores, bem como as palavras de sentimento dentre as 1000 palavras mais frequentemente encontradas nos comentários e que não estivessem contidas no Sentilex-PT. Foram feitos vários experimentos variando o uso de caracteres acentuados, negação, uso de dicionário de domínio e *stemming*[5]. Os resultados são apresentados detalhadamente em (TUMITAN; BECKER, 2013), onde observa-se que todas as variações apresentaram resultados insatisfatórios devido à falta de um dicionário apropriado para a língua portuguesa, em particular

[5]Stemming: método para a redução de um termo ao seu radical, por meio da remoção de desinências, afixos e vogais temáticas (HULL, 1996).

a do Brasil. O melhor resultado obtido por meio dessa abordagem é apresentado na Tabela 3.1.

Também foi experimentada a abordagem de polarização baseada em aprendizagem de máquina, usando como conjunto de treino/teste as sentenças anotadas. Foram testados diferentes tipos de extração e preparações de *features*, tais como n-gramas, representação binária de pesos, remoção de *stopwords*, além de várias técnicas de seleção de *features*. Foram removidas menções diretas aos nomes dos candidatos, pois essas eram as *features* mais determinantes na polaridade de uma sentença, tornando o classificador muito específico a cada eleição. Por outro lado, foram mantidas *features* representando os apelidos, pois muitos desses são pejorativos e implicam opinião quando utilizados (e.g. motoserra, malddad, tarusso). Por terem influenciado negativamente o desempenho do classificador baseado em dicionário, não foram desenvolvidos experimentos envolvendo *stemming* ou lematização [6].

Os melhores resultados foram obtidos com as seguintes preparações de *features*:

- unigramas;
- representação por pesos com TF-IDF;
- uso de *Information Gain* como critério para seleção de *features*.

Tumitan and Becker (2014) testaram diferentes algoritmos de classificação disponíveis no WEKA (HALL et al., 2009). Os melhores resultados foram obtidos com o algoritmo SMO (Sequential Minimal Optimization), que pertence à família SVM (Support Vector Machine) (PLATT, 1998). A Tabela 3.1 apresenta os resultados encontrados para as abordagens de classificação usando dicionário e aprendizagem de máquina. Para as sentenças relativas a 2012, foram realizados dois experimentos: com todas sentenças anotadas usando *10-fold cross validation*, e com conjunto das sentenças anotadas de 2010 como conjunto de treino e as de 2012 como conjunto de teste.

Considerando a acurácia, os melhores resultados foram obtidos para a abordagem baseada em aprendizagem de máquina. Em experimentos que utilizam validação cruzada, a acurácia de polarização foi de 81,37% para as sentenças das eleições de 2010 e de 83,24% para as da eleição de 2012. Nos experimentos com dicionário, os melhores resultados tiveram acurácia de 50,39% e 52,14% para as eleições de 2010 e 2012 respectivamente. No experimento usando a eleição de 2012 como conjunto de teste, o desempenho foi inferior ao da validação cruzada, devido a problemas de overfiting, pois termos distintos foram usados em cada eleição. Por exemplo, nas eleições de 2010 existiam várias referências às primeiras candidatas à presidência

[6]Lematização: processo para agrupar as diferentes inflexões e variantes de um termo para que possam ser analisadas como um único item, o lema (KANIS; SKORKOVSKá, 2010).

Tabela 3.1: Experimentos para classificação de sentenças, onde (A) acurácia, (P) precisão, (R) revocação e (F) f-measure.

| Eleição | Abordagem | A (%) | Polaridade | P (%) | R (%) | F (%) |
|---|---|---|---|---|---|---|
| 2010 | Baseado em dicionário | 50,39 | Positivo | 29,39 | 62,99 | 40,08 |
| | | | Negativo | 54,79 | 44,94 | 49,38 |
| | SVM | 81,37 | Positivo | 70,63 | 65,58 | 68,01 |
| | | | Negativo | 85,56 | 88,2 | 86,86 |
| 2012 | Baseado em dicionário | 52,14 | Positivo | 26,99 | 56,41 | 36,51 |
| | | | Negativo | 90,18 | 51,45 | 65,52 |
| | SVM | 83,24 | Positivo | 27,59 | 10,53 | 15,24 |
| | | | Negativo | 86,45 | 95,38 | 90,7 |
| | SVM Conjunto de teste | 77,40 | Positivo | 25,56 | 30,26 | 27,71 |
| | | | Negativo | 87,98 | 85,27 | 86,61 |

Fonte: Tumitan and Becker (2014).

(e.g. "guerreira", "presidenta"), enquanto que, em 2012, foram evocados feitos passados dos candidatos e os escândalos de seus partidos.

#### *3.3.1.4 Sumarização*

A partir das sentenças polarizadas, são derivadas séries temporais diárias de sentimento por candidato, a partir das quais são derivadas diversas métricas que contabilizam o sentimento baseado em sentenças. A sumarização desenvolvida permite visualizar a evolução diária do sentimento da população por candidato, bem como prever a variação de intenções de voto. Uma das limitações é que a sumarização é baseada nas sentenças. Isso introduz um viés nos resultados, uma vez que um comentário longo, com muitas opiniões emitidas por uma única pessoa, pode ter o mesmo peso que muitos comentários curtos, emitidos por pessoas distintas.

### 3.3.2 Considerações

O presente estudo de caso reutiliza todo o conhecimento e procedimentos propostos nos trabalhos anteriores da pesquisa de Tumitan and Becker (2014), acrescentando contribuições nos seguintes pontos:

- A mineração de opiniões é realizada em nível de aspecto, baseada na classificação em nível de sentença.
- Estende a proposta de polarização baseada em aprendizagem de máquina. Além de tratar como um problema ternário (i.e. classes positiva, negativa, neutra), desenvolve experimentos que comparam essa classificação ternária com a polarização em duas etapas

(identificação de sentenças com sentimento, seguida da classificação das sentenças de sentimento em positivas/negativas). Os experimentos de classificação, detalhados no Capítulo 6, também apresentam resultados superiores.

- O modelo de sumarização agrega as opiniões por candidato e seus aspectos, baseado nos comentários e não nas sentenças, representando, assim, opiniões emitidas por leitores, independente do tamanho de seus comentários.

# 4 ESTUDO DE CASO SOBRE MINERAÇÃO DE OPINIÕES EM NÍVEL DE ASPECTOS EM COMENTÁRIOS DE NOTÍCIAS

Este capítulo apresenta uma visão geral do estudo de caso desenvolvido, a saber, uma experiência de mineração de opiniões em nível de aspecto fora do contexto de revisão de produtos. São detalhados os objetivos, as características do corpus, bem como a abordagem proposta para o desenvolvimento da Mineração de Opiniões em nível de aspecto em comentários de jornais.

## 4.1 Objetivos

O objetivo geral deste estudo de caso é realizar uma abordagem que combine técnicas para minerar opiniões em nível de aspecto em fontes de opiniões pouco estruturadas. Foram consideradas as eleições municipais de 2012 da cidade de São Paulo e os três candidatos a prefeito mais comentados. Para isso, são utilizados comentários escritos em português extraídos da Folha de São Paulo que contenham menções explícitas aos candidatos.

O estudo de caso trata da identificação de aspectos dos candidatos, os quais podem ser alvo de opiniões. Foram considerados neste trabalho dois aspectos frequentemente associados a um plano eleitoral, os quais são relevantes à população: saúde e educação. A partir da identificação dos aspectos, este trabalho também propõe avaliar a percepção global da população em relação a um candidato e a percepção em relação a aspectos específicos (e.g. a percepção de X no tocante à saúde é mais positiva do que a do candidato Y no mesmo aspecto).

Os objetivos específicos do estudo de caso são:

- identificar a fonte de dados que deve ser usada para encontrar aspectos, ou seja, as notícias, os comentários, ou ainda, uma combinação destes;
- uma vez identificada a fonte de dados, selecionar, a partir de experimentos, as técnicas mais adequadas de identificação de aspectos;
- adaptar a Mineração de Opiniões em nível de sentença discutida na Seção 3.3.1 ao nível de aspecto;
- selecionar modelos de sumarização adequados à apresentação dos resultados.

## 4.2 Fonte de dados e análise do corpus

Para identificar os aspectos, foram utilizados, como fonte de opiniões fracamente estruturadas, os comentários de leitores do portal de notícias Folha de São Paulo[1], mais precisamente, os comentários de notícias sobre as eleições municipais de 2012 da cidade de São Paulo. Foram considerados, no estudo de caso, os três candidatos à prefeitura mais comentados, sendo eles: José Serra (PSDB), Fernando Haddad (PT) e Celso Russomanno (PRB). O período analisado corresponde aos meses que antecedem o primeiro turno das eleições (setembro e outubro de 2012). O corpus utilizado é o mesmo descrito na Seção 3.3 e consiste de 407 notícias, associadas a 14.848 comentários. Tais comentários são divididos em 79.752 sentenças.

A pesquisa considera como fonte de opiniões os comentários dos leitores como reação a uma notícia. Cada comentário pode incluir opiniões sobre um ou mais candidatos, e utiliza-se o nível de sentença para buscar frases que contenham opiniões e menções a candidatos, os quais são considerados como entidades-alvo da opinião. No nível de aspecto, surge o problema de como identificar um aspecto e o que deve ser considerado como documento para sua identificação: a notícia, o comentário ou ambos.

Assim, busca-se saber se os comentários associados a uma notícia sobre o aspecto saúde (e.g. "candidato X promete duplicar os postos de saúde") são representativos da opinião que os leitores têm sobre o candidato no tocante a esse aspecto, pois as seguintes situações são possíveis:

- o comentário enfatiza o aspecto da notícia (e.g. "confio que X acabará com as filas nos postos");
- o comentário revela uma opinião sobre uma entidade não descrita na notícia (e.g. "Y também prometeu acabar com as filas de ônibus, mas não cumpriu");
- o comentário revela uma opinião sobre um aspecto não mencionado na notícia (e.g. "X é corrupto", "X e Y parecem estar mais preocupados com problemas pessoais de familiares").

Nota-se que muitos comentários podem também expressar alguma opinião sobre um candidato de forma implícita ou indireta, ou seja, sem mencionar o nome do candidato nem o aspecto (e.g. "ponho fé", considerando a notícia exemplificada acima). Há também comentários que expressam uma opinião explícita sobre um aspecto de um candidato que nem tem relação com a notícia (e.g. "segurança nunca foi seu forte quando X estava no poder", em uma

[1] www.folha.uol.com.br

notícia sobre postos de saúde). As diferentes situações discutidas estão ilustradas na Figura 4.1. Sem uma análise profunda da semântica da sentença, não é possível tratar opiniões indiretas e implícitas, razão pela qual esses comentários e respectivas opiniões não serão abordados neste trabalho.

### 4.2.1 Anotação manual e Profiling

Para descobrir qual fonte de dados é a mais adequada para a identificação de aspectos, foi buscada uma resposta para a seguinte questão: *deve-se realizar a identificação dos aspectos levando em conta apenas as notícias, os comentários ou a combinação de ambos?*

Como passo inicial, foi realizado um processo de anotação manual do corpus de comentários, sentenças e notícias. Os conjuntos anotados são apresentados na Tabela 4.1. Todas as 407 notícias eleitorais do corpus foram anotadas por um único anotador, que avaliou se as notícias evocavam um dos tópicos analisados (i.e. saúde ou educação). Já os comentários foram anotados em nível de sentença por 3 anotadores. Foi anotada uma amostra de 2072 sentenças selecionadas aleatoriamente. Cada anotador avaliou se a sentença evocava um dos tópicos, quais candidatos mencionava explicitamente, se expressava uma opinião um candidato. Os anotadores foram orientados a informar a polaridade apenas para as sentenças que mencionassem um dos candidatos analisados. Eles também foram orientados a basear sua avaliação apenas no conteúdo explicitamente escrito, sem usar julgamento próprio ou conhecimento do domínio político para inferir entendimento. Assumiu-se que todo comentário que contivesse pelo menos uma sentença anotada como mencionando um dado aspecto, por transitividade, também mencionava o aspecto. Essa relação entre sentenças e comentários resultou em uma amostra de 487 comentários anotados. O nível de concordância entre os anotadores das sentenças é apresentado na Tabela 4.2.

Tabela 4.1: Conjuntos de notícias, comentários e sentenças anotadas.

| **Anotações** | **Quantidade** |
|---|---|
| Notícias anotadas | 407 |
| Comentários anotados | 487 |
| Sentenças anotadas | 2072 |

Fonte: Autoria própria.

Apenas foram considerados os comentários e sentenças em que houve concordância entre os três anotadores, definindo os casos de comentários/sentenças relevantes indicados na

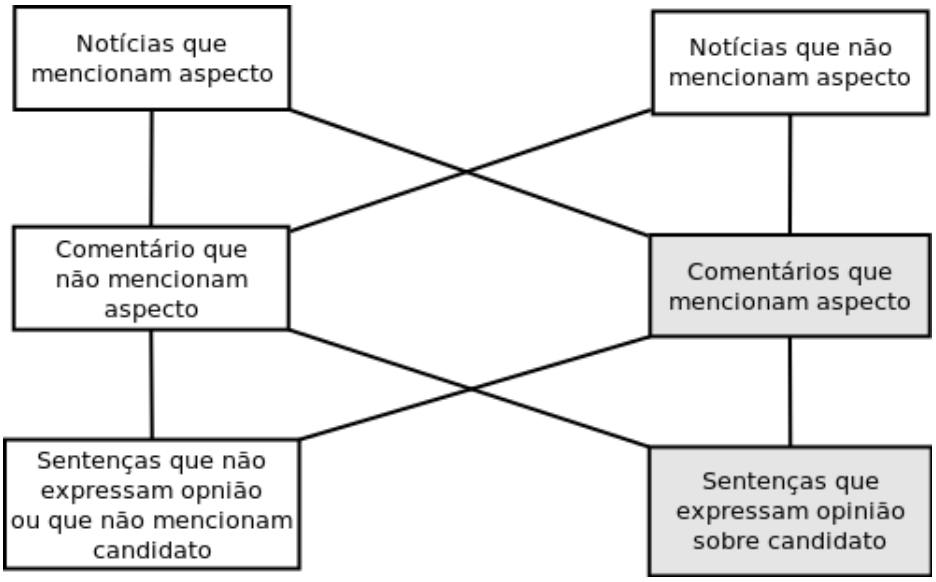
Tabela 4.2: Concordância entre os anotadores na identificação de menções aos aspectos.

| **Anotadores** | **Aspecto** | **Concordância** |
|---|---|---|
| Anotadores 1 e 2 | Saúde | 90,7% |
| | Educação | 84,6% |
| Anotadores 1 e 3 | Saúde | 86,6% |
| | Educação | 80% |
| Anotadores 2 e 3 | Saúde | 89,6% |
| | Educação | 83% |
| Entre todos anotadores | Saúde | 77,2% |
| | Educação | 65,6% |

Fonte: Autoria própria.

Figura 4.1 com a cor cinza. Com base na anotação realizada, foram quantificadas todas as situações da Figura 4.1. As quantificações são apresentadas nas tabelas 4.3 e 4.4.

Figura 4.1: *Profiling* de notícias e comentários.




Fonte: Autoria própria.

- Notícias que fazem menção a um aspecto;
- notícias que não fazem menção a um aspecto;
- comentários que fazem menção ao aspecto e que estão relacionados a sentenças com opiniões com menções a, pelo menos, um candidato;
- comentários que não fazem menção ao aspecto e/ou que não estão relacionados a sentenças com opiniões com menções a, pelo menos, um candidato.

Tabela 4.3: Quantificação de comentários e sentenças associados a notícias que mencionam os aspectos analisados

| | **Notícias que mencionam aspecto** | | | | |
|---|---|---|---|---|---|
| | Educação | 35 / 407 (8,59%) | | | |
| | Saúde | 32 / 407 (7,86%) | | | |
| Educação | 11 comentários e 50 sentenças | Comentários que mencionam aspecto | 4 / 11 (36,36%)<br>4 / 487 (0,82%) | Sentenças que expressam opinião sobre candidato | 4 / 50 (8%)<br>4 / 2072 (0,19%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 7 / 50 (14%)<br>7 / 2072 (0,34%) |
| | | Comentários que não mencionam aspecto | 7 / 11 (63,63%)<br>7 / 487 (1,44%) | Sentenças que expressam opinião sobre candidato | 8 / 50 (16%)<br>8 / 2072 (0,39%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 31 / 50 (62%)<br>31 / 2072 (1,5%) |
| Saúde | 25 comentários e 56 sentenças | Comentários que mencionam aspecto | 17 / 25 (68%)<br>17 / 487 (3,49%) | Sentenças que expressam opinião sobre candidato | 5 / 56 (8,92%)<br>5 / 2072 (0,24%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 23 / 56 (41,07%)<br>23 / 2072 (1,11%) |
| | | Comentários que não mencionam aspecto | 8 / 25 (32%)<br>8 / 487 (1,64%) | Sentenças que expressam opinião sobre candidato | 5 / 56 (8,92%)<br>5 / 2072 (0,24%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 23 / 56 (41,07%)<br>23 / 2072 (1,11%) |

Fonte: Autoria própria.

Como pode ser visto na Tabela 4.3, no processo de anotação, 35 notícias foram classificadas como relevantes para o aspecto Educação. Os anotadores entenderam que, dos 11 comentários anotados relativos a essas notícias, apenas 4 comentários explicitamente expressavam opiniões sobre esse aspecto. Isso representa 36,6% dos comentários anotados relativos a essas notícias e apenas 0,82% de todos comentários anotados. Das 32 notícias anotadas como Saúde, apenas 17 comentários foram anotados como relacionados ao mesmo assunto da notícia, em um conjunto de 25 comentários anotados (63,63%). Em relação ao total de comentários anotados, isso representa apenas 3,49%.

Quando consideradas as notícias que não evocam os tópicos monitorados (Tabela 4.4), um número significativamente maior de comentários relevantes foi encontrado. Para o aspecto educação, foram encontrados 93 comentários relevantes, representando 19,10% de todos comentários anotados. Já para a área de saúde, foram encontrados 70 comentários relevantes, representando 14,37% de todos os comentários anotados.

Tabela 4.4: Quantificação de comentários e sentenças associados a notícias que não mencionam os aspectos analisados

| | **Notícias que não mencionam aspecto** | | | | |
|---|---|---|---|---|---|
| | Não são sobre educação | 372 / 407 (91,40%) | | | |
| | Não são sobre saúde | 375 / 407 (92,13%) | | | |
| Não são sobre educacão | 476 comentários e 2022 sentenças | Comentários que mencionam aspecto | 93 / 476 (19,54%)<br>93 / 487 (19,10%) | Sentenças que expressam opinião sobre candidato | 113 / 2022 (5,59%)<br>113 / 2072 (5,45%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 326 / 2022 (16,12%)<br>326 / 2072 (15,73%) |
| | | Comentários que não mencionam aspecto | 383 / 476 (80,46%)<br>383 / 487 (78,64%) | Sentenças que expressam opinião sobre candidato | 303 / 2022 (14,99%)<br>303 / 2072 (14,62%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 1280 / 2022 (63,30%)<br>1280 / 2072 (61,78%) |
| Não são sobre saúde | 462 comentários e 2016 sentenças | Comentários que mencionam aspecto | 70 / 462 (14,71%)<br>70 / 487 (14,37%) | Sentenças que expressam opinião sobre candidato | 101 / 2016 (5,01%)<br>101 / 2072 (4,87%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 289 / 2016 (14,29%)<br>289 / 2072 (13,95%) |
| | | Comentários que não mencionam aspecto | 392 / 462 (82,35%)<br>392 / 487 (80,49%) | Sentenças que expressam opinião sobre candidato | 328 / 2016 (16,22%)<br>328 / 2072 (14,83%) |
| | | | | Sentenças que não expressam opinião ou não mencionam candidato | 1298 / 2016 (64,19%)<br>1298 / 2072 (62,64%) |

Fonte: Autoria própria.

Ao analisar a proporção entre os comentários e as notícias com ou sem o tópico em questão, é possível observar algumas situações. No tocante à educação, há uma proporção maior de comentários sobre esse aspecto em notícias sobre o tema (36,36%) do que em notícias específicas sobre outros assuntos (19,54%). Para o aspecto saúde, essa proporção é de 68% e 14,71% para notícias que mencionam e que não mencionam saúde, respectivamente. Além disso, é observado um número absoluto muito pequeno de comentários sobre os aspectos específicos em notícias sobre o mesmo aspecto, o qual é uma consequência do número igualmente baixo dessas notícias. Assim, considerar somente esses comentários poderia comprometer os resultados do estudo de caso.

Por outro lado, há um número não negligenciável de comentários sobre os aspectos saúde e educação relacionados a notícias totalmente diversas. Sob o universo de comentários anotados, 0,82% dos comentários mencionam o aspecto Educação em notícias sobre o mesmo tema. Já para notícias sobre outros temas, o percentual sobe para 19,10%. Ao realizar a mesma análise para o aspecto Saúde, 3,49% dos comentários são sobre o mesmo aspecto mencionado nas notícias, e para notícias sobre outros assuntos o percentual sobe para 14,37%. Assim, optou-

se por extrair aspectos de comentários sem levar em consideração o assunto abordado na notícia à qual eles estão associados.

De posse dessas informações, é apresentada no Capítulo 5 uma abordagem para realizar a identificação de aspectos em fontes de opinião fracamente estruturadas. Essa abordagem utilizará o conteúdo dos comentários como fonte para encontrar os aspectos alvos de uma opinião.

## 4.3 Abordagem proposta

### 4.3.1 Visão geral

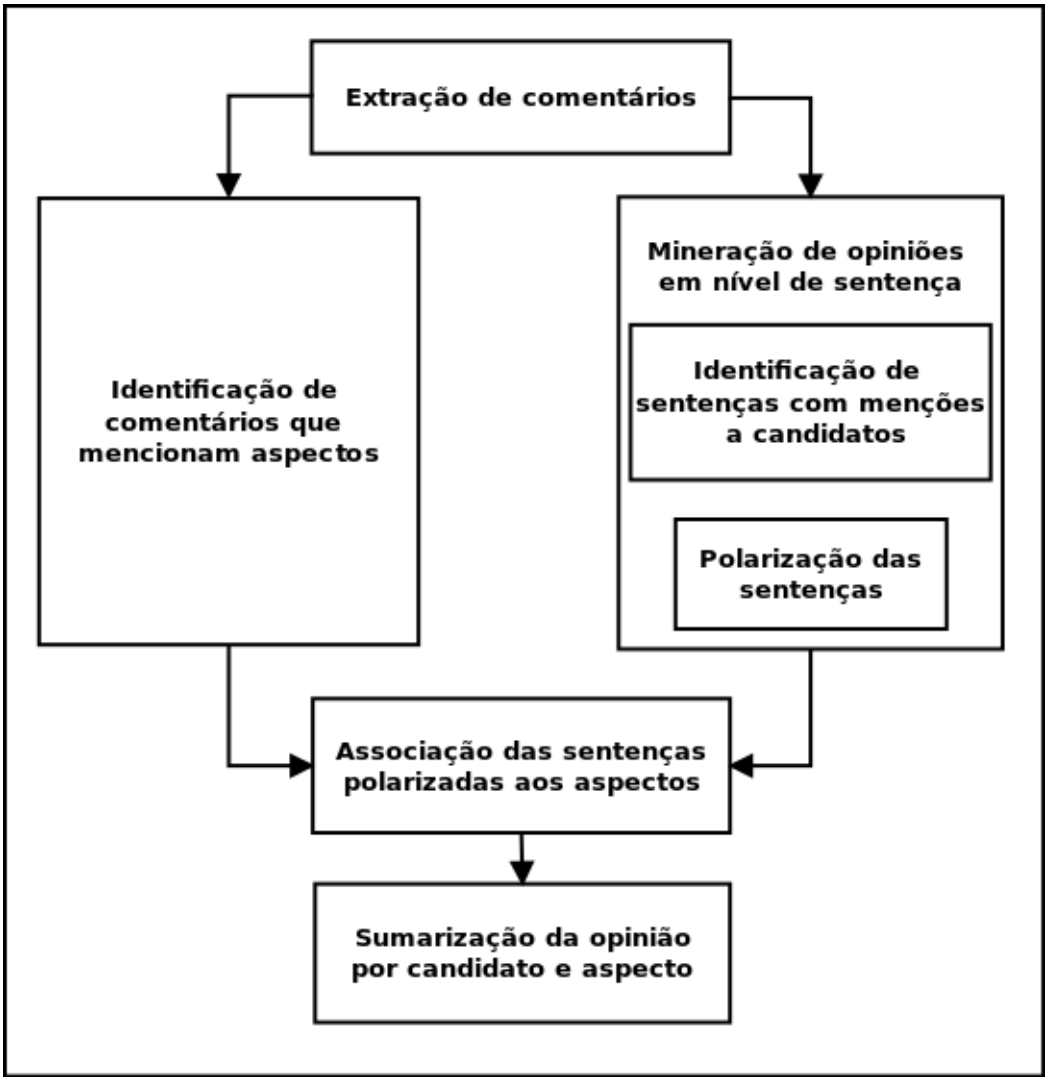
A visão geral da abordagem proposta é esboçada na Figura 4.2. Uma vez extraídos os comentários relevantes ao estudo de caso, dois procedimentos são realizados: a) a identificação dos comentários envolvendo os aspectos-alvo e b) a mineração de opiniões em nível de sentença para sentenças mencionando os candidatos observados. O nível de sentença foi mantido para a polarização da opinião porque cada comentário pode envolver opiniões sobre mais de um candidato. Para ajustar a polarização ao nível de aspecto, é analisada a relação entre as sentenças polarizadas e os respectivos comentários. Finalmente, a opinião dos leitores é sumarizada por candidato e por seus aspectos. No restante da seção, são apresentados os conceitos formais utilizados no estudo de caso e a descrição das etapas da abordagem proposta.

### 4.3.2 Formalização

A discussão detalhada das etapas propostas para mineração de opiniões em nível de aspecto é baseada nos seguintes conceitos:

- $A = \{a_i\}$: conjunto de aspectos monitorados;
- $E = \{e_j\}$: conjunto de candidatos monitorados;
- $C = \{c_k\}$: conjunto de comentários extraídos;
- $C_a = \{c_a : a \in A \wedge c_a \in C\}$: conjunto de comentários que mencionam um dado aspecto $a_i \in A$, onde $C_a \subseteq C$;
- $S = \{< s_i, c_k >: c_k \in C\}$: conjunto de sentenças relativas a cada comentário $c_k \in C$;
- $SM = \{< s_i, e_j, c_k >: e_j \in E \wedge < s_i, c_k > \in S\}$: subconjunto de sentenças $S$ que mencionam cada uma das entidades eleitorais $e_j \in E$;

Figura 4.2: Abordagem proposta para minerar opiniões em nível de aspecto.




Fonte: Autoria própria.

- $SP = \{< s_i, e_j, pol, c_k >: pol \in \{pos, neut, neg\} \ \wedge$ $< s_i, e_j, c_k > \in SM\}$: conjunto de sentenças polarizadas que mencionam cada uma das entidades $e_j \in E$. Três classes são consideradas: positivo (*pos*), negativo (*neg*) e neutro (*neut*);
- $SP_a = \{< s_i, e_j, pol, c_k >: pol \in \{pos, neut, neg\} \ \wedge \ c_k \in C_a\}$: subconjunto de sentenças com menções polarizadas $SP_a \subseteq SP$ associadas aos comentários $c_k$ que mencionam $a_i$;
- $CP = \{< c_k, e_j, pol >: pol \in \{pos, neut, neg\} \wedge \ < c_k, e_j > \in SP\}$: subconjunto de comentários $CP \subseteq C$, que agregam por comentário as sentenças polarizadas $SP$ sobre uma entidade $e_j \in E$;
- $CP_a = \{< c_k, e_j, a_i, pol >: pol \in \{pos, neut, neg\} \wedge \ < c_k, e_j > \in CP \ \wedge \ a_i \in A\}$: subconjunto de comentários $CP_a \subseteq CP$, que agregam as sentenças polarizadas $SP_a SP$

por comentário sobre uma entidade $e_j \in E$ e um aspecto $a_i \in A$.

### 4.3.3 Etapas

São descritos abaixo os detalhes de cada etapa do processo de mineração de opiniões em nível de aspecto proposto no estudo de caso.

**Extração de comentários:** os comentários são extraídos, e os ruídos são removidos, como descrito na Seção 3.3.1.1, junto com o pré-processamento realizado para minimizar o efeito dos problemas de português presentes nos comentários, tal como descrito na Seção 3.3.1.2. O resultado dessa etapa é o conjunto $C$ de comentários pré-processados.

**Identificação de comentários que mencionam aspectos:** é realizada a identificação dos comentários $C_a$. O Capítulo 5 apresenta experimentos nos quais diferentes métodos de extração foram utilizados, incluindo LSI, phi-squared e EMIM.

**Identificação de sentenças com menções a candidatos:** nessa etapa, são buscadas todas as sentenças contendo menções aos candidatos observados. Para isso, primeiramente os comentários são divididos em sentenças utilizando o módulo *punkt* do NLTK. A relação entre sentença e comentário de origem é armazenada. O resultado é o conjunto $S$. Em seguida, são filtradas as sentenças que fazem menções explícitas aos candidatos, usando seus nomes (e.g. Serra, Russomano) ou variações (e.g. vampissera, tarussomano), conforme descrito na Seção 3.3.1.2. O resultado é o conjunto das sentenças com menções $SM$.

**Polarização das sentenças:** as sentenças com menções $SM$ são polarizadas em três classes, gerando como resultado o conjunto de sentenças polarizadas $SP$. A classificação das sentenças foi realizada por meio de algoritmos de aprendizagem de máquina. Estendendo os experimentos descritos na Seção 3.3.1.3, foram experimentadas duas abordagens, uma que classifica as sentenças em três classes e outra que realiza duas classificações binárias seguidas. Nesta última, primeiro são identificadas as sentenças que possuem sentimento, as quais são, então, classificadas como positivas e negativas. Uma descrição completa dos experimentos de polarização é apresentada no Capítulo 6.

**Associação das sentenças polarizadas aos aspectos:** o objetivo dessa etapa é associar as sentenças polarizadas aos aspectos mencionados nos respectivos comentários. Para isso, as sentenças com opiniões sobre aspectos são separadas, usando o seu relacionamento com os respectivos comentários. Para cada aspecto $a_i \in A$, buscam-se as sentenças $SP_a$,

considerando-as como opiniões sobre o aspecto $a_i \in A$ do candidato $e_j \in E$. O resultado é o conjunto de sentenças associadas aos aspectos $SP_a$.

**Sumarização da opinião sobre as entidades e seus aspectos:** A partir do conjunto $SP_a$, as sentenças são agregadas por comentários que mencionam uma entidade $e_j \in E$ e um aspecto $a_i \in A$. O resultado dessa etapa é o conjunto $CP_a$. Além disso, também são obtidos os comentários $CP$, que agregam as sentenças polarizadas $SP$ sobre uma entidade $e_j \in E$. As sentenças são agregadas por comentário e candidato, para representar o número de pessoas emitindo opiniões sobre cada candidato. Dessa forma consegue-se chegar à mineração de opiniões em nível de aspecto em um domínio não estruturado. Essa etapa também apresenta sumarizações gráficas das opiniões sobre os candidatos em geral, utilizando o conjunto $CP$, e sobre seus aspectos, utilizando o conjunto $CP_a$. No Capítulo 7, são apresentados dois modelos: um que exibe as opiniões sobre os candidatos e aspectos e uma nuvem de palavras que ilustra os termos representativos mencionados de cada aspecto $a \in A$. Tais modelos permitem acompanhar as opiniões sobre os candidatos e seus aspectos, bem como quais são os termos representativos mais mencionados sobre cada aspecto.

### 4.3.4 Ilustração da abordagem

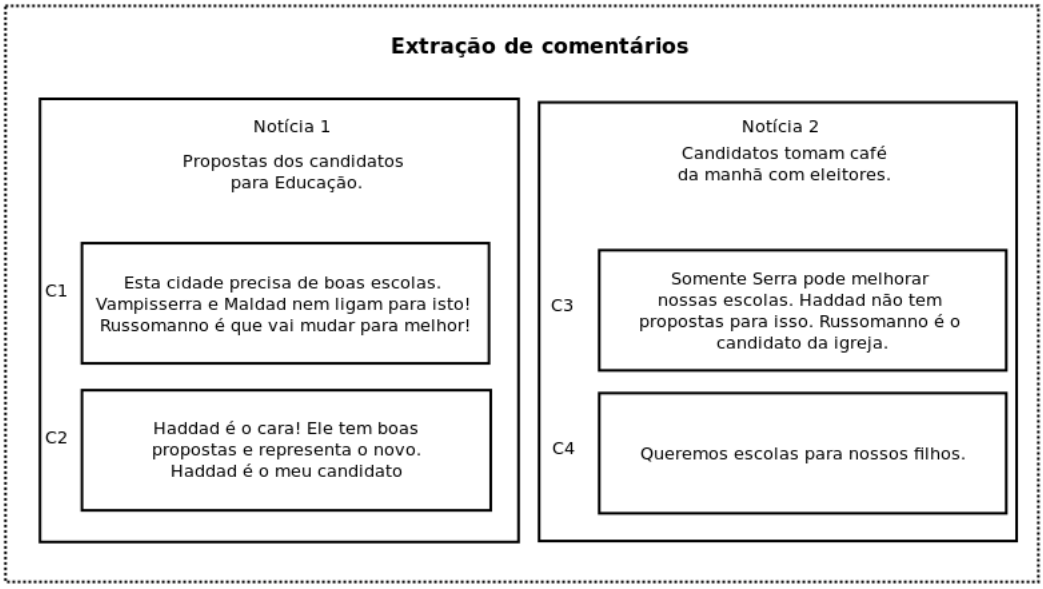
Essa seção irá demonstrar como a abordagem proposta irá minerar opiniões em nível de aspecto. Para tanto, serão utilizados quatro comentários relativos a duas notícias, sendo uma sobre educação e outra sobre outro assunto

legendFonte: Autoria própria.Para a demonstração prática da abordagem optou-se por utilizar notícias e comentários fictícios. Esses comentários não devem de forma alguma serem interpretados como o posicionamento político desse trabalho.. No exemplo, são utilizados o conjunto de entidades $E = \{Serra, Haddad, Russomanno\}$ e o conjunto de aspectos $A = \{educacao\}$.

O resultado da etapa *Extração de comentários* é apresentado na Figura 4.3. Como as notícias relacionadas aos comentários não são relevantes para a abordagem, elas serão desconsideradas.

Em seguida, a etapa *Identificação de comentários que mencionam aspecto*, ilustrada na Figura 4.4, cria o conjunto $C_a$, com os comentários que mencionam termos representativos do aspecto *educação*. Observa-se que nesse exemplo o termo “escolas” caracteriza o aspecto

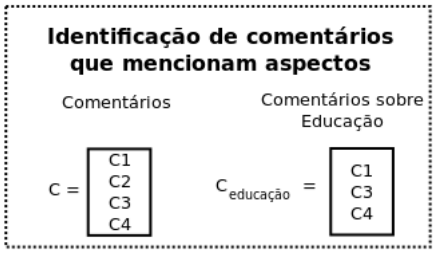
Figura 4.3: Extração de comentários.




Fonte: Autoria própria.

*educação*. Nota-se também que o comentário C2 não é atribuído ao conjunto $C_a$, pois não menciona o aspecto analisado.

Figura 4.4: Identificação de comentários que mencionam aspectos.

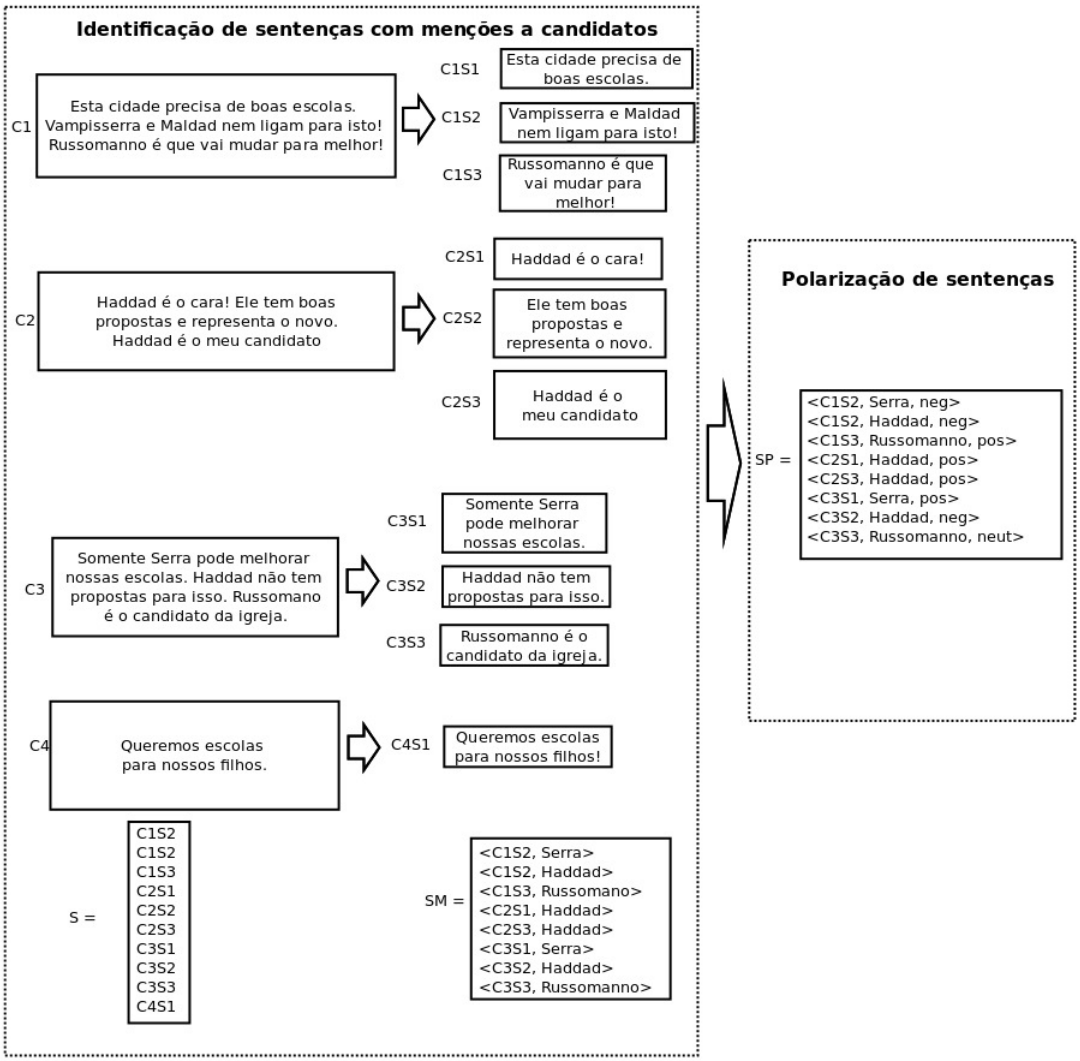



Fonte: Autoria própria.

Paralelamente, a abordagem realiza a *Mineração de opiniões em nível de sentença*, ilustrada na Figura 4.5. Primeiro, a etapa *Identificação de sentenças com menções a candidatos* quebra os comentários em sentenças, gerando como saída o conjunto de sentenças $S$. Além disso, ela analisa as sentenças procurando por aquelas que possuem menção aos candidatos analisados, gerando como saída o conjunto $SM$. Ao analisar as sentenças extraídas, observa-se que a sentença C2S2 não faz parte do conjunto $SM$, pois ela não possui menção direta ao nome do candidato Haddad ou a algum apelido dele. Nota-se também que existem sentenças que são associadas a mais de um candidato (e.g. C1S2).

Em seguida é realizada a *Polarização de sentenças*. O resultado dessa etapa é o conjunto $SP$, que possui as sentenças polarizadas que mencionam os candidatos analisados.

Figura 4.5: Mineração de opiniões em nível de sentença.

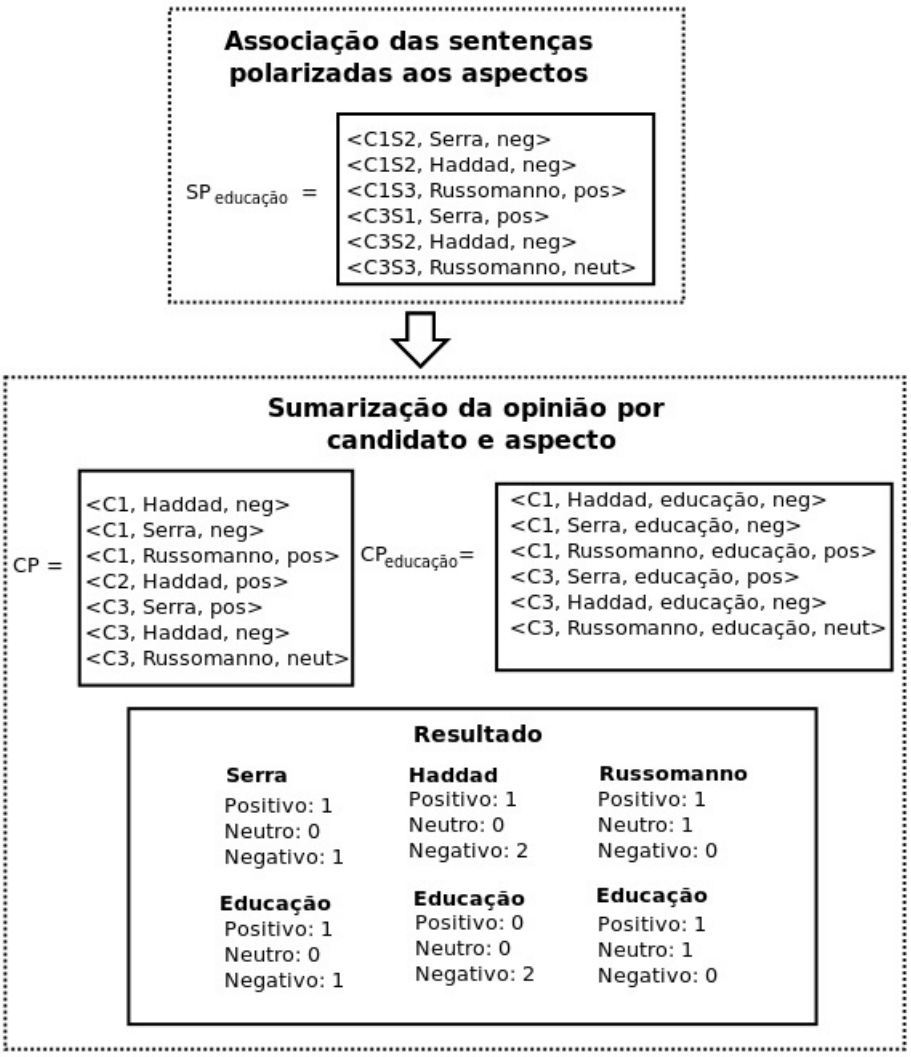



Fonte: Autoria própria.

Após minerar as opiniões em nível de sentença e identificar os comentários com menções aos aspectos, é realizada a mineração de opiniões em nível de aspecto, apresentada na Figura 4.6, através da *Associação das sentenças polarizadas aos aspectos*. Para isto, as sentenças polarizadas são associadas aos aspectos mencionados em seus comentários de origem. Isso gera como saída o conjunto de sentenças $SP_{educacao}$. Ao analisar as sentenças do exemplo, observa-se que fazem parte desse conjunto apenas as sentenças dos comentários C1 e C3, pois apenas esses comentários possuem sentenças polarizadas que mencionam candidatos.

Finalmente, na etapa *Sumarização da opinião por candidato e aspecto*, a partir dos conjuntos $SP$ e $SP_{educacao}$, a abordagem contabiliza os comentários por candidato e por aspecto

Figura 4.6: Mineração de opiniões em nível de aspecto.




Fonte: Autoria própria.

de cada candidato. O presente trabalho não propõe contabilizar sentenças, pois existem comentários (e.g. C2) que possuem diversas sentenças com a mesma opinião sobre um determinado candidato. Esse tipo de situação poderia prejudicar o resultado final caso um único comentário tivesse centenas de sentenças com uma determinada opinião sobre um candidato. Para contabilizar os comentários, são somadas as polarizações das sentenças de cada comentário sobre um candidato (ou seu aspecto). Tomando como exemplo o comentário C1, a soma das polaridades das sentenças resultou em um valor menor que zero para os candidatos José Serra e Fernando Haddad. Logo, tal comentário é contabilizado como sendo negativo para eles, e também como negativo para o aspecto *educação* desses candidatos. Já para o candidato Celso Russomano, o comentário C1 é contabilizado como positivo para o aspecto *educação* e também para o próprio candidato. Ao analisar o comentário C2, observa-se que ele é é contabilizado como positivo

apenas uma vez para o candidato Fernando Haddad, pois existem duas sentenças positivas nesse comentário para esse candidato.

Quanto ao comentário C3, ele é contabilizado como positivo para Serra e para seu aspecto *educação*. Para Haddad, esse comentário é contabilizado como negativo tanto para o candidato, quanto para seu aspecto. Observa-se que a sentença C3S3, polarizada como neutra para o candidato Russomanno, contabiliza o comentário C3 como neutro para o aspecto Educação desse candidato, mesmo que a opinião contida nessa sentença não tenha relação alguma com o aspecto *educação*. Já o comentário C4 não é contabilizado para nenhum aspecto dos candidatos, pois tal comentário não possui sentenças com menções aos candidatos analisados.

Após a contabilização de comentários ser realizada, os resultados são apresentados por meio de um modelo de sumarização. A Seção 7.2, apresenta um modelo de sumarização gráfico com os resultados obtidos no estudo de caso do presente trabalho.

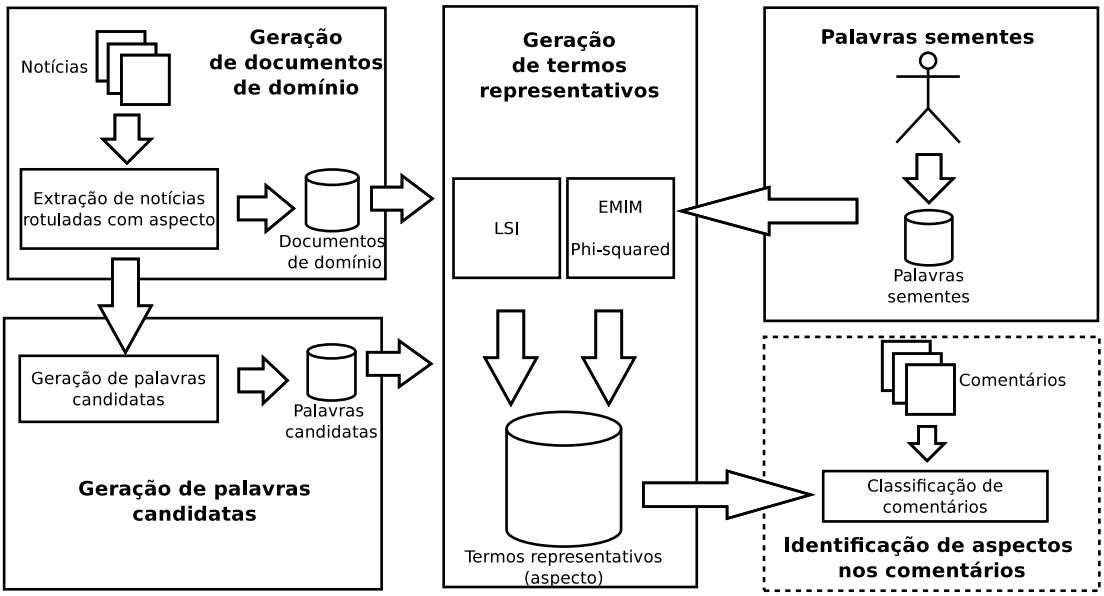
# 5 IDENTIFICAÇÃO DE ASPECTOS

Este capítulo apresenta uma abordagem para identificar aspectos em fontes de opiniões fracamente estruturadas. Foram experimentadas diferentes técnicas para identificar quais palavras podem representar os aspectos analisados (saúde e educação), e os resultados obtidos são avaliados.

## 5.1 Abordagem proposta

A abordagem proposta para identificar aspectos encontra os comentários que mencionam os aspectos analisados a partir de um conjunto de termos representativos de cada aspecto. Os *Termos Representativos* são encontrados com base no processo esboçado na Figura 5.1. Esses termos são selecionados dentre um conjunto de *Palavras Candidatas*, extraídas de documentos de domínio. Para algumas técnicas experimentadas, é também utilizado um conjunto de *Palavras Sementes*, definidas a partir do conhecimento de domínio. Deve-se observar que o conjunto de *Palavras Semente* é independente do conjunto de *Palavras Candidatas*. As etapas desse processo são melhor descritas no restante desta seção.

Figura 5.1: Processo de identificação de aspectos.




Fonte: Autoria própria.

**Geração de documentos de domínio:** para encontrar termos que representassem cada um dos aspectos, buscaram-se notícias classificadas pelo jornal como sendo sobre Saúde e Educação. Foi realizada a extração de um corpus diferente de notícias, contendo 1000 notícias classificadas pela Folha de São Paulo com o rótulo Educação e 1000 notícias com o rótulo Saúde.

**Geração de palavras candidatas:** de cada um desses *corpora*, foram extraídos todas as palavras existentes, junto com suas respectivas frequências e classes gramaticais (*part-of-speech tags*). Para isso, foi utilizada a ferramenta NLTK [1]. *Stopwords* foram eliminadas. Para a criação do conjunto de palavras candidatas de cada aspecto, utilizaram-se apenas substantivos e termos que apareciam no respectivo corpus (e.g. notícias sobre saúde) e que não apareciam no outro (e.g. notícias sobre educação). Uma proposta de melhoria futura para essa etapa é utilizar também as palavras que co-ocorram até um limite de vezes em ambos conjuntos.

**Geração dos termos representativos:** utilizando uma combinação entre técnicas baseadas em frequência, dependência ou independência de termos, foram selecionadas as palavras mais representativas para um determinado aspecto. As técnicas de co-ocorrência usadas foram Phi-squared (CHURCH; GALE, 1991) e Expected Mutual Information Measure (CHURCH; HANKS, 1990). Para as duas técnicas citadas, foram utilizadas as palavras semente listadas na Tabela 5.1. Além disso, foram realizados experimetnos utilizando Indexação Semântica Latente (DEERWESTER et al., 1990).

Tabela 5.1: Palavras semente escolhidas.

| | **Aspecto** | |
|---|---|---|
| | **Saúde** | **Educação** |
| **Palavras Sementes** | Saúde | Educação |
| | Hospital | Aula |
| | Médico | Aulas |
| | Médicos | Professores |
| | Vacina | Escola |
| | Vacinação | Escolas |
| | Hospitalar | - |
| | Doença | - |
| | Doenças | - |

Fonte: Autoria própria.

Com posse dos termos representativos dos aspectos Saúde e Educação, são filtrados todos os comentários das notícias eleitorais que contêm tais termos em seu conteúdo.

[1] Natural Language Toolkit - http://www.nltk.org/

## 5.2 Experimentos

Esta seção descreve as técnicas experimentadas para identificar os aspectos nos comentários sobre notícias políticas, bem como a avaliação dos resultados obtidos. Para isso, foram realizados experimentos que tinham como objetivo descobrir quais das palavras candidatas poderiam ser consideradas palavras representativas para um dado aspecto. Também é apresentada a avaliação de resultados para o conjunto de palavras semente.

Durante a realização dos experimentos, todas as palavras passaram por um processo de limpeza, no qual foram transformadas para minúsculo, e os acentos foram removidos (e.g. "Vacinação" foi transformada em "vacinacao"). A razão é que muitos comentários foram escritos com erros de ortografia. Essa medida não resolve todos os problemas de ortografia, mas minimiza o efeito negativo deles para essa abordagem.

### 5.2.1 Palavras Semente

O primeiro experimento para determinar qual deveria ser o conjunto de termos representativos considerou como tal conjunto apenas as palavras semente. Logo, se um comentário possuir em seu conteúdo uma ou mais palavras semente, ele passa a ser classificado como um comentário que menciona aspecto. Nas abordagens descritas a seguir, verificou-se se os comentários possuíam em seu conteúdo um ou mais termos representativos de um aspecto.

Observa-se que não foi utilizado o termo "professor" como uma palavra semente. O motivo é que o processo de *profiling* revelou muitos comentários que não mencionavam o aspecto Educação, mas mencionavam alguns candidatos como "Professor Fulano". Como o uso desta palavra poderia comprometer a identificação de comentários específicos para este aspecto, optou-se por utilizar apenas o plural deste termo como palavra semente.

### 5.2.2 Expected Mutual Information Measure

Este experimento é baseado na utilização da técnica Expected Mutual Information Measure (EMIM) (CHURCH; HANKS, 1990), que compara a probabilidade de observar duas palavras, $x$ e $y$, junto com a probabilidade de observá-las independentemente. EMIM é definida por meio da Equação 5.1, em que $a$ é número de vezes que as palavras $x$ e $y$ co-ocorrem em um documento; $b$ é o número de vezes em que $x$ ocorre em um documento e $y$ não ocorre; $c$ é o

número de vezes que *y* ocorre é *x* não e, *d* é o número de vezes que nem *x* nem *y* ocorrem um documento, em que *x* representa uma palavra candidata e *y* representa uma palavra semente.

$$EMIM_{(x,y)} = \log_2 \frac{a(a+b+c+d)}{(a+b)(a+c)} \tag{5.1}$$

Caso o resultado de EMIM seja maior que zero, *x* passa a ser considerado um termo representativo. Experimentos preliminares demonstraram resultados insatisfatórios utilizando apenas esse critério. Buscando melhorar os resultados obtidos, mesclou-se o resultado do EMIM com o resultado da quantidade de vezes que *x* e *y* co-ocorrem em todos os documentos de domínio. O experimento que apresentou a melhor avaliação de resultados para essa abordagem considerou a união entre o conjunto de palavras semente com o conjunto de palavras candidatas em que *x* e *y* co-ocorrem no mínimo 10 vezes em todos os documentos de um dado aspecto.

### 5.2.3 Phi-squared

A técnica phi-squared ($\phi^2$) é uma medida estatística que favorece uma alta ocorrência de eventos (CHURCH; GALE, 1991). Phi-squared é definida por meio da Equação 5.2, cujas variáveis têm o mesmo significado das utilizadas na Equação 5.1.

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}, \text{onde } 0 \leq \phi^2 \leq 1 \tag{5.2}$$

Foram realizados diversos experimentos variando o resultado phi-squared com o *threshold*. Da mesma forma como realizado na abordagem EMIM, foi considerada também a quantidade de vezes que *x* e *y* co-ocorrem em todos os documentos. No final, o experimento que apresentou a melhor avaliação de resultados considerou um phi-squared mínimo de 0,02 e uma co-ocorrência mínima entre *x* e *y* de 20 vezes. Por fim, o conjunto de termos representativos foi a união entre o conjunto resultante dessa abordagem com o conjunto de palavras semente do aspecto.

### 5.2.4 Indexação Semântica Latente

Este experimento utilizou a Indexação Semântica Latente (LSI) para identificar quais palavras candidatas poderiam ser consideradas termos representativos. O LSI implementa o

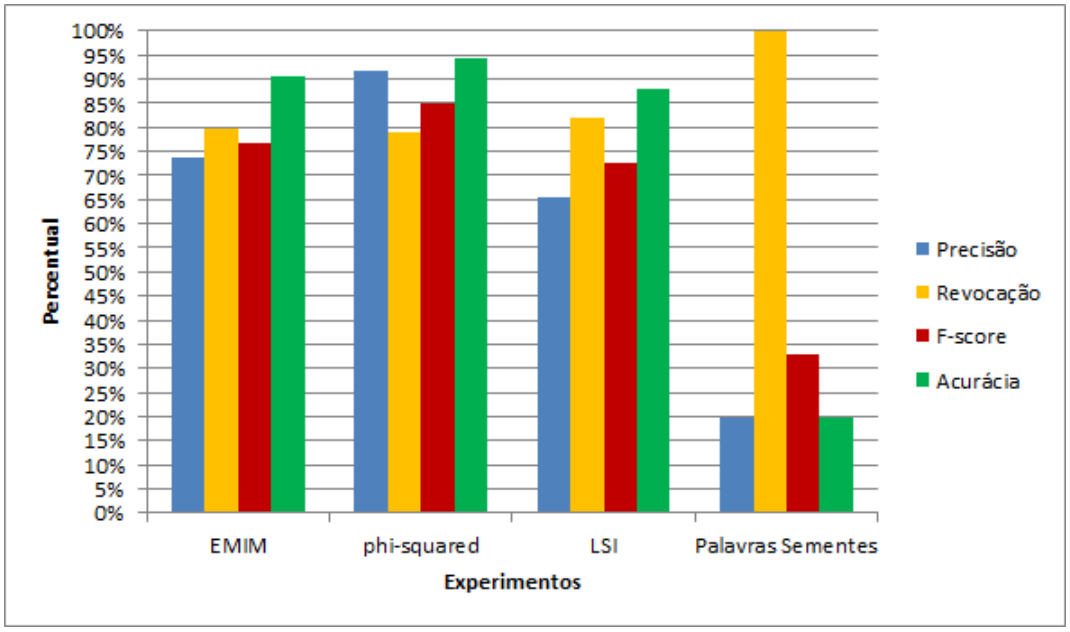
modelo vetorial utilizando técnicas estatísticas para recuperar informações de um corpus com base no conteúdo semântico das consultas (DEERWESTER et al., 1990).

Assim, foi calculado o LSI das palavras candidatas em relação aos documentos de domínio, utilizando a biblioteca Python gensim (ŘEHŮŘEK; SOJKA, 2010). Foi utilizada a função LsiModel parametrizado para encontrar 100 tópicos[2]. Depois foram realizados diversos experimentos, variando a similaridade resultante do LSI. Diferentemente das abordagens que utilizam EMIM e phi-squared, com o LSI não se verificou a co-ocorrência das palavras candidatas com as palavras semente.

## 5.3 Resultados

Para realizar a avaliação dos resultados, foram utilizadas as métricas Precisão, Revocação, F-score e Acurácia (AGGARWAL; ZHAI, 2012). A Figura 5.2 apresenta um gráfico com a avaliação dos resultados das abordagens descritas acima em comentários sobre Saúde; a Figura 5.3, os resultados para os comentários sobre Educação. O experimento LSI considera como ponto de corte resultados LSI iguais a ou maiores a 0,75.

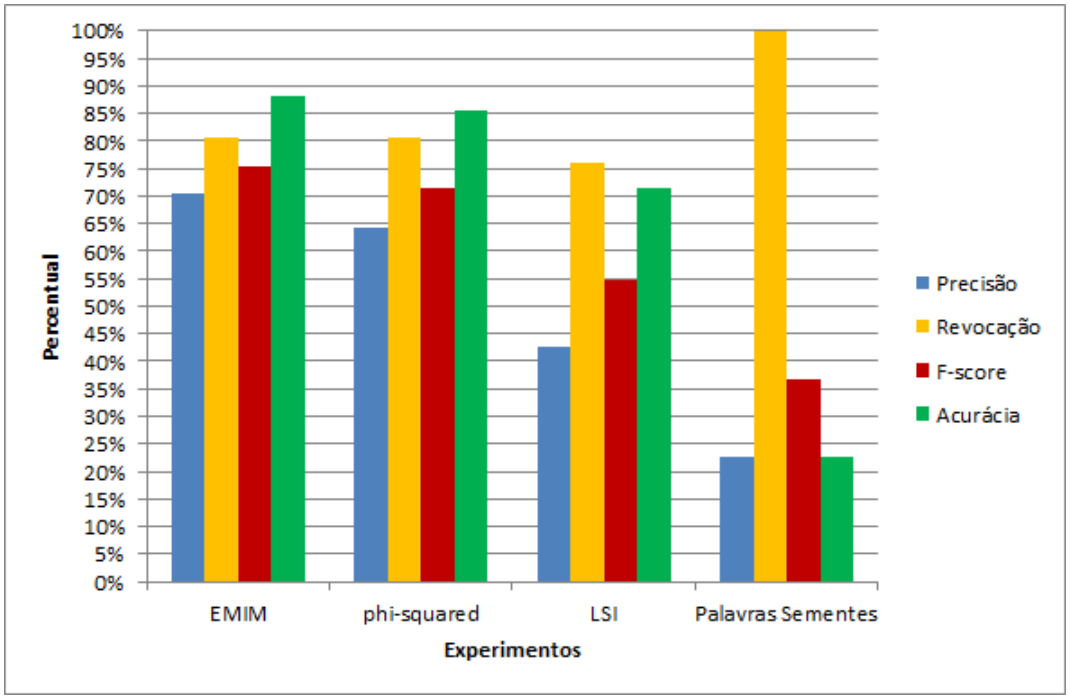
Figura 5.2: Avaliação de resultados para o aspecto saúde em comentários.




Fonte: Autoria própria.

[2]Não houve variação desse parâmetro nos experimentos realizados.

Figura 5.3: Avaliação de resultados para o aspecto educação em comentários.




Fonte: Autoria própria.

Ao analisar os resultados para comentários sobre saúde, observa-se que a técnica EMIM obteve 73,83% de precisão, 79,8% de revocação, 76,7% de F-score e 90,42% de acurácia. Em comparação, a técnica phi-squared obteve 91,8% de precisão, 78,8% de revocação, 84,8% de F-score e 94,4% de acurácia. Paralelamente, a técnica com LSI apresentou resultados inferiores, sendo 65,32% de precisão, 81,82% de revocação, 72,65% de F-score e 87,82% de acurácia. Utilizando apenas as palavras sementes escolhidas, obteve-se uma alta revocação 100%, mas uma precisão, F-score e acurácia muito baixos (19,8%, 33% e 19,8%, respectivamente). Observa-se então, que para saúde, a técnica com phi-squared obteve melhores resultados para precisão, F-score e acurácia.

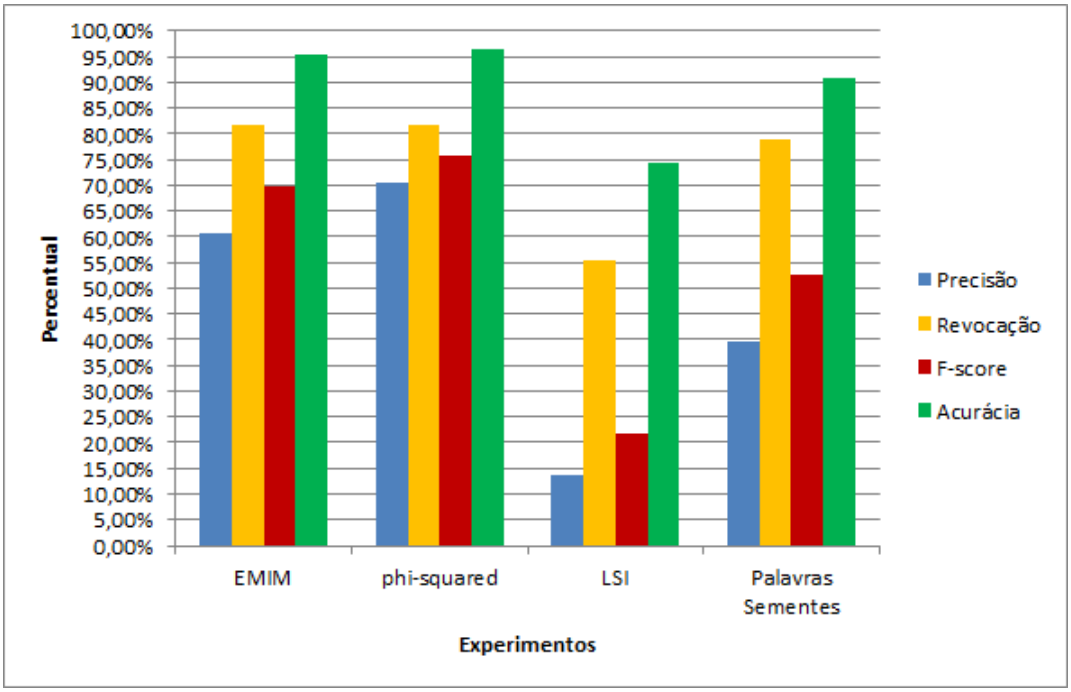
Já para comentários sobre o aspecto educação, a técnica EMIM apresentou um resultado melhor se comparado com as outras técnicas, com 70,54% de precisão, 80,53% de revocação, 75,21% de F-score e 88,02% de acurácia, enquanto que a phi-squared resultou em 64,08% de precisão, 80,53% de revocação, 71,37% de F-score e 85,43% de acurácia. Da mesma forma como nos comentários para saúde, os experimentos com LSI e palavras sementes apresentaram resultados insatisfatórios para precisão e F-score, no entanto, a revocação foi relativamente alta com LSI (76,11%) e de 100% utilizando palavras sementes.

Com base no F-score, o experimento que apresentou a melhor avaliação de resultados para o aspecto Saúde foi o utilizando a técnica Phi-squared. Já para o aspecto Educação, o melhor resultado foi obtido com a técnica EMIM. Foram realizados também experimentos

mesclando as técnicas EMIM e phi-squared, mas não houve melhoria nos resultados finais.

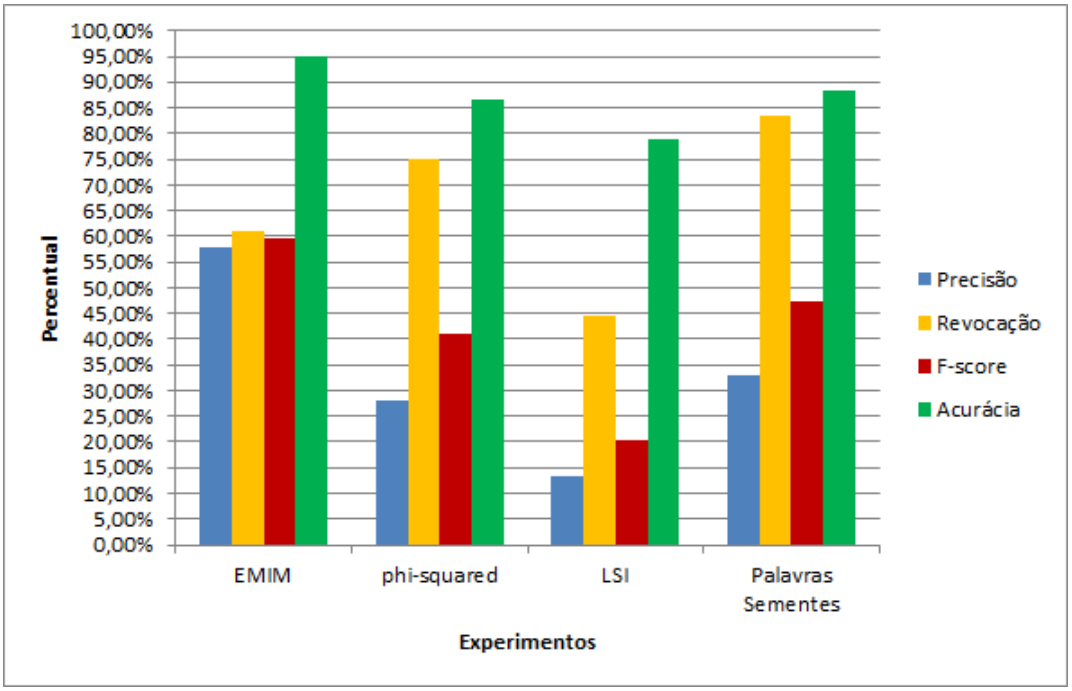
Como existe um corpus de notícias políticas rotuladas como relevantes para os aspectos analisados, foi realizada também a validação de resultados usando tais notícias ao invés dos comentários. Os resultados são apresentados nas tabelas 5.4 e 5.5. Apesar de ter sido descartada a utilização das notícias nessa abordagem, o objetivo foi verificar o comportamento das técnicas em documentos que possuem uma qualidade melhor de escrita. Os resultados obtidos foram semelhantes aos realizados com os comentários, favorecendo as técnicas phi-squared para saúde e EMIM para educação. Assim como nos comentários, os experimentos com LSI e palavras sementes não apresentaram bons resultados.

Figura 5.4: Avaliação de resultados para o aspecto saúde em notícias.




Fonte: Autoria própria.

Figura 5.5: Avaliação de resultados para o aspecto educação em notícias.




Fonte: Autoria própria.

# 6 MINERAÇÃO DE OPINIÕES EM NÍVEL DE SENTENÇA

Este capítulo tem como objetivo apresentar as abordagens utilizadas para classificar as opiniões em nível de sentença. Serão apresentados os experimentos e a avaliação dos resultados obtidos.

## 6.1 Abordagens para polarizar sentenças

As abordagens para polarizar sentenças utilizadas referem-se ao trabalho de Tumitan and Becker (2014), descrito na Seção 3.3. As abordagens são baseadas em aprendizagem de máquina, considerando os resultados significativamente superiores sobre os classificadores baseados em dicionário. O presente trabalho complementa o trabalho apresentado na Seção 3.3 com duas abordagens para realizar a classificação:

- classificação de sentenças em três classes (positivo, negativo e neutro);
- realização de classificações binárias em duas etapas: na primeira, as sentenças são classificadas *com* ou *sem* sentimento; na segunda, as sentenças com sentimento são classificadas em positivas ou negativas.

## 6.2 Corpus e Gold Standard

Para treinar e validar os classificadores de opinião em nível de sentença, foram adotados os mesmos *corpora* utilizados por Tumitan and Becker (2014), que correspondem aos comentários sobre notícias das seguintes eleições:

- eleição municipal de São Paulo em 2012;
- eleição governamental para o estado de São Paulo em 2010;
- eleição presidencial brasileira em 2010.

Desse conjunto, foi selecionado aleatoriamente um conjunto de sentenças, criando assim 2 *datasets*, em que cada *dataset* foi anotado por três anotadores. Apenas sentenças com, no mínimo duas concordâncias foram consideradas. A Tabela 6.1 apresenta os resultados de cada *dataset* de acordo com a respectiva anotação realizada.

Tabela 6.1: Datasets utilizados nos experimentos de polarização

| Eleição | Dataset | Total | Negativo | Positivo | Neutro | Descartado | Concordância |
|---|---|---|---|---|---|---|---|
| 2012 | Dataset1 | 471 | 208 | 86 | 136 | 41 | 91,29% |
| 2012 | Dataset2 | 600 | 482 | 72 | 28 | 18 | 97% |
| 2010 | Dataset3 | 1000 | 356 | 154 | 417 | 73 | 92,7% |
| Total | | 2071 | 1046 | 312 | 581 | 132 | |

Fonte: Autoria própria.

Os *datasets* 2 e 3 são os mesmos utilizados nos experimentos descritos na Seção 3.3.1.3. Para criar o *dataset1* buscou-se aproveitar as 2072 sentenças anotadas durante o processo descrito na Seção 4.2.1. No entanto, na anotação realizada no presente trabalho os anotadores informaram apenas a polaridade para as sentenças que opinassem sobre algum dos candidatos analisados. Das 2072 sentenças, apenas 471 possuem menções aos candidatos. Excluindo as sentenças com total discordância entre os anotadores, os classificadores construídos no presente trabalho contam com um conjunto maior de sentenças (430). Isso representa um acréscimo de 24,82% nas sentenças negativas, 38,05% nas positivas e 30,56% nas sentenças neutras. Em particular os dois últimos tipos de sentença eram bastante escassos no conjunto de treino dos classificadores anteriormente desenvolvidos.

## 6.3 Classificação ternária

### 6.3.1 Descrição do experimento

Basicamente, o primeiro experimento consistiu em repetir a extração e preparação de *features* descrita na Seção 3.3.1.3. Em termos da preparação de dados, a única diferença foi que a seleção de atributos não foi realizada com o algoritmo Information Gain, mas sim com o CfsSubsetEval (HALL, 1998), utilizando como método de busca o algoritmo BestFirst. A razão para essa alteração foi que essa forma de seleção de atributos apresentou melhores resultados. A classificação também foi realizada com o algoritmo SMO. As demais diferenças em relação àquela experiência são:

- um conjunto maior de dados, com a inclusão das 430 instâncias do *dataset1*;
- o tratamento do problema da polarização com três classes.

### 6.3.2 Discussão dos resultados

A Tabela 6.2 apresenta os resultados usando Precisão (P), Revocação (R), F-Score (F) e as respectivas médias ponderadas (WP, WR e WF), bem como também a Acurácia (A). Foram feitos dois experimentos: com validação cruzada (10-folds) e com conjuntos de treino e teste distintos. No experimento SVM (Cross-validation), os *datasets* 1, 2 e 3 foram mesclados.

Já no segundo experimento, utilizou-se o *Dataset1* como conjunto de testes e os *Dataset2* e *Dataset3* mesclados como conjunto de treino. Dessa forma, o conjunto de treino é composto por 445 sentenças positivas, 226 neutras e 838 negativas. Já o conjunto de testes é composto por 136 sentenças positivas, 86 neutras e 208 negativas.

Tabela 6.2: Resultados da polarização das sentenças com a abordagem ternária

| **Experimento** | **Polaridade** | **P %** | **R %** | **F %** | **WP %** | **WR %** | **WF %** | **A %** |
|---|---|---|---|---|---|---|---|---|
| SVM Cross-validation | Positivo | 82,2 | 26,3 | 39,8 | 68,7 | 62,2 | 56,8 | 62,17 |
| | Neutro | 77,6 | 26,7 | 39,7 | | | | |
| | Negativo | 58,9 | 95,7 | 72,9 | | | | |
| SVM Conjunto de testes | Positivo | 45,5 | 6 | 10,6 | 48,2 | 53,1 | 45,5 | 53,07 |
| | Neutro | 33,3 | 17,3 | 22,8 | | | | |
| | Negativo | 56,8 | 87,1 | 68,7 | | | | |

Fonte: Autoria própria.

Se comparado com os resultados de validação cruzada relatados na Tabela 3.1 como SVM (para Eleição de 2012), verifica-se:

- Precisão: acréscimo de 54,61 pontos percentuais para a classe positiva, e uma redução para a classe negativa (redução de 27,55%);
- Revocação: acréscimo de 15,77 pontos percentuais para a classe positiva, e praticamente nenhuma alteração para a classe negativa (0,32%);
- F-Score: acréscimo de 24,56 pontos percentuais para a classe positiva, e redução de 17,8% na classe negativa.

Comparando com os resultados do segundo experimento da Tabela 3.1, ou seja, conjuntos de treino e teste distintos (também para 2012), verifica-se:

- Precisão: acréscimo de 19,94 pontos percentuais para a classe positiva, mas uma piora para a classe negativa (redução de 31,18%);
- Revocação: houve uma redução de 24,26 pontos percentuais para a classe positiva e uma leve melhora para para a classe negativa (acréscimo de 1,83%);

- F-Score: para sentenças positivas obteve-se uma retração de 17,11 pontos percentuais, assim como também houve uma redução de 17,91 pontos percentuais para a classe negativa.

Nessa comparação, observa-se que os resultados melhoraram apenas para as sentenças positivas, o que pode ter sido resultado do acréscimo de sentenças anotadas positivas. Contudo, a introdução de uma nova classe degradou o desempenho geral, com impacto nos resultados para a classe negativa. No futuro, os resultados podem ser reavaliados com uma quantidade de sentenças anotadas ainda mais equilibradas.

## 6.4 Classificações binárias em duas etapas

### 6.4.1 Descrição do experimento

Em busca de resultados melhores, foram realizados experimentos usando classificações binárias em duas etapas. Classificações binárias tendem a ser melhor que aquelas que utilizem mais classes pois assim, espera-se que os erros não sejam propagados para a quantidade de classes analisadas. Experimentos similares são descritos em Alves et al. (2014), trabalho que polariza *tweets* em três classes, usando os seguintes passos:

1. São identificadas as sentenças que possuem ou não sentimento. Dessa forma, as sentenças neutras são classificadas *sem* sentimento, enquanto que as positivas e negativas são classificadas *com* sentimento.
2. A polarização positiva e negativa é realizada apenas nas sentenças *com* sentimento. Logo, o resultado desse processo são sentenças classificadas neutras (sem sentimento), positivas e negativas.

Assim, são realizadas duas classificações binárias ao invés de uma ternária. Foram realizados experimentos iguais aos descritos na Seção 6.3, com a mesma preparação de *features* e os mesmos *datasets*.

### 6.4.2 Discussão dos resultados

As tabelas 6.3 e 6.4 apresentam os melhores resultados obtidos. Foram feitos duas variações dos experimentos, com validação cruzada e conjuntos de treino e teste distintos, tal

como na seção anterior.

Tabela 6.3: Resultados da avaliação de sentenças com e sem sentimento

| **Experimento** | **Possui sentimento** | **P %** | **R %** | **F %** | **WP %** | **WR %** | **WF %** | **A %** |
|---|---|---|---|---|---|---|---|---|
| SVM | Sim | 74 | 96,5 | 83,8 | 75,6 | 74,6 | 70,2 | 74,57 |
| Cross-validation | Não | 79,1 | 28 | 41,4 | | | | |
| SVM | Sim | 73,2 | 86,4 | 79,2 | 62,7 | 67,5 | 64 | 67,51 |
| Conjunto de teste | Não | 36,1 | 19,5 | 25,4 | | | | |

Fonte: Autoria própria.

Tabela 6.4: Resultados da avaliação de sentenças com sentimento

| **Experimento** | **Polarização** | **P %** | **R %** | **F %** | **WP %** | **WR %** | **WF %** | **A %** |
|---|---|---|---|---|---|---|---|---|
| SVM | Positivo | 74,5 | 23,9 | 36,2 | 66 | 65,1 | 55,7 | 65,09 |
| Cross-validation | Negativo | 64,5 | 97,6 | 77,7 | | | | |
| SVM | Positivo | 30,4 | 8,3 | 13,1 | 59,2 | 67,8 | 60,7 | 67,82 |
| Conjunto de testes | Negativo | 71,1 | 92,2 | 80,3 | | | | |

Fonte: Autoria própria.

Comparando os resultados da Tabela 6.4 com os apresentados na Tabela 3.1 para validação cruzada, observa-se novamente piora nos resultados das sentenças negativas e melhorias para a classe positiva. Mais especificamente:

- Precisão: acréscimo de 46,91 pontos percentuais para a classe positiva, e redução de 21,95 pontos para a classe negativa;
- Revocação: acréscimo de 13,37 pontos percentuais para a classe positiva, e pequena melhora para a classe negativa (acréscimo de 2,22%);
- F-Score: acréscimo de 20,96 pontos percentuais para a classe positiva, e piora para a classe negativa negativa (redução de 13%).

Considerando a validação cruzada, ambas as abordagens de classificação ternária e a classificação binária em duas etapas apresentaram uma melhora significativa de precisão para a classe positiva, à custa de um desempenho pior para a classe negativa. Contudo, ressalta-se que a revocação para ambas as classes foi superior aos resultados descritos na Seção 3.3.1.3.

Ao comparar os resultados do experimento com conjuntos de treino e teste distintos com os descritos na Tabela 3.1, observa-se uma melhora para a precisão de sentenças positivas e na revocação de sentenças negativas, mas resultados inferiores nas demais métricas:

- Precisão: acréscimo de 4,84 pontos percentuais para a classe positiva mas uma piora para a classe negativa (redução de 16,88%);

- Revocação: redução de 21,86 pontos para a classe positiva, mas acréscimo para a classe negativa de 6,93 pontos;
- F-Score: ambas as classes apresentaram redução, sendo que para a positiva houve redução de 14,61 pontos percentuais, e para negativa, de 6,31%.

## 6.5 Comparação das abordagens para polarização de sentenças

Nos experimentos realizados, nota-se que a classificação de sentenças positivas apresentou resultados superiores àqueles anteriormente atingidos pelo grupo de pesquisa, relatados na Seção 3.3.1.3, ao custo de uma piora para as classes negativas. Contudo, uma das contribuições é considerar a classe neutra nos modelos desenvolvidos. Este capítulo conclui com uma comparação entre as duas abordagens utilizadas nos experimentos desenvolvidos no presente trabalho. Uma síntese dos resultados é apresentada na Tabela 6.5.

Para os experimentos com validação cruzada, observa-se a que abordagem de classificações binárias em duas etapas em relação à classificação ternária apresentou:

- Precisão: redução de 7,7 pontos percentuais para a classe positiva, e acréscimos de 5,6 pontos para a negativa, e de 1,5% para classe neutra;
- Revocação: redução de 2,4 pontos para a classe positiva, e acréscimos de 1,9 pontos para a negativa e de 1,3% classe neutra;
- F-Score: a classe positiva apresentou redução de 3,6 pontos, diferente das classes negativa e neutra, que apresentaram acréscimos de 4,8 e 1,7 pontos, respectivamente.

Já para o experimento com conjuntos de treino e testes, a abordagem de classificações binárias em duas etapas em relação à classificação ternária apresentou:

- Precisão: redução de 15,1 pontos para a classe positiva e acréscimos de 14,3 pontos percentuais para a classe negativa, e de 2,8 pontos a classe neutra;
- Revocação: melhora para todas as classes, sendo acréscimos de 2,3 pontos para a positiva, 5,1 para negativa e 2,2 para a classe neutra;
- F-Score: melhora para todas as classes também, sendo acréscimos de 2,5 pontos para a positiva, 11,6 pontos para negativa e 2,6 pontos para a classe neutra.

Tabela 6.5: Síntese comparativa dos resultados obtidos com os experimentos realizados.

| | **Experimento** | **Abordagem de Classificação** | **Positivo (%)** | **Neutro ou sem sentimento (%)** | **Negativo (%)** |
|---|---|---|---|---|---|
| **Precisão** | SVM Cross-validation | Ternária | 82,2 | 77,6 | 58,9 |
| | | Binária em duas etapas | 74,5 | 79,1 | 64,5 |
| | SVM Conjuntos de treino e teste distintos | Ternária | 45,5 | 33,3 | 56,8 |
| | | Binária em duas etapas | 30,4 | 36,1 | 71,1 |
| **Revocação** | SVM Cross-validation | Ternária | 26,3 | 26,7 | 95,7 |
| | | Binária em duas etapas | 23,9 | 28 | 97,6 |
| | SVM Conjuntos de treino e teste distintos | Ternária | 6 | 17,3 | 87,1 |
| | | Binária em duas etapas | 8,3 | 19,5 | 92,2 |
| **F-score** | SVM Cross-validation | Ternária | 39,8 | 39,7 | 72,9 |
| | | Binária em duas etapas | 36,2 | 41,4 | 77,7 |
| | SVM Conjuntos de treino e teste distintos | Ternária | 10,6 | 22,8 | 68,7 |
| | | Binária em duas etapas | 13,1 | 25,4 | 80,3 |
| **Acurácia** | SVM Cross-validation | Ternária | 62,17 | | |
| | | Binária em duas etapas | 65,09 | | |
| | SVM Conjuntos de treino e teste distintos | Ternária | 53,07 | | |
| | | Binária em duas etapas | 67,82 | | |

Fonte: Autoria própria.

Ao comparar a acurácia das duas abordagens desenvolvidas, verifica-se que a abordagem com classificações binárias em duas etapas obteve melhores resultados, sendo que nos experimentos com validação cruzada houve um acréscimo de 2,92 pontos percentuais, enquanto que nos experimentos com conjuntos de treino e teste disjuntos o acréscimo foi de 14,75 pontos percentuais.

# 7 MINERAÇÃO DE OPINIÕES EM NÍVEL DE ASPECTO E SUMARIZAÇÃO

Este capítulo tem como objetivo descrever como a abordagem proposta minera opiniões em nível de aspecto e sumariza as opiniões sobre candidatos e seus aspectos. São apresentados modelos visuais e uma discussão dos resultados obtidos.

## 7.1 Mineração em Nível de Aspecto

A estratégia para minerar opiniões em nível de aspecto consiste em associar a polarização das sentenças que mencionam os candidatos aos aspectos mencionados nos comentários. Por exemplo, se um determinado comentário mencionar um termo representativo do aspecto *Educação*, e esse comentário possuir sentenças polarizadas que mencionam candidatos, a abordagem assume que a opinião sobre o aspecto *Educação* dos candidatos mencionados é a opinião média das sentenças polarizadas sobre esses candidatos.

A opinião média pode ser entendida da seguinte forma: caso a soma da polaridade das sentenças em um comentário for superior a zero, a opinião média do comentário é positiva, caso inferior a zero então é negativa, e quando ela for igual a zero a opinião média no comentário é neutra. Assim, as opiniões são agregadas por comentários para representar *pessoas* emitindo opiniões. O uso de sentenças, como em Tumitan and Becker (2014), provocaria distorções nos resultados, por exemplo, um comentário com diversas sentenças negativas teria mais peso que um comentário com uma única sentença positiva.

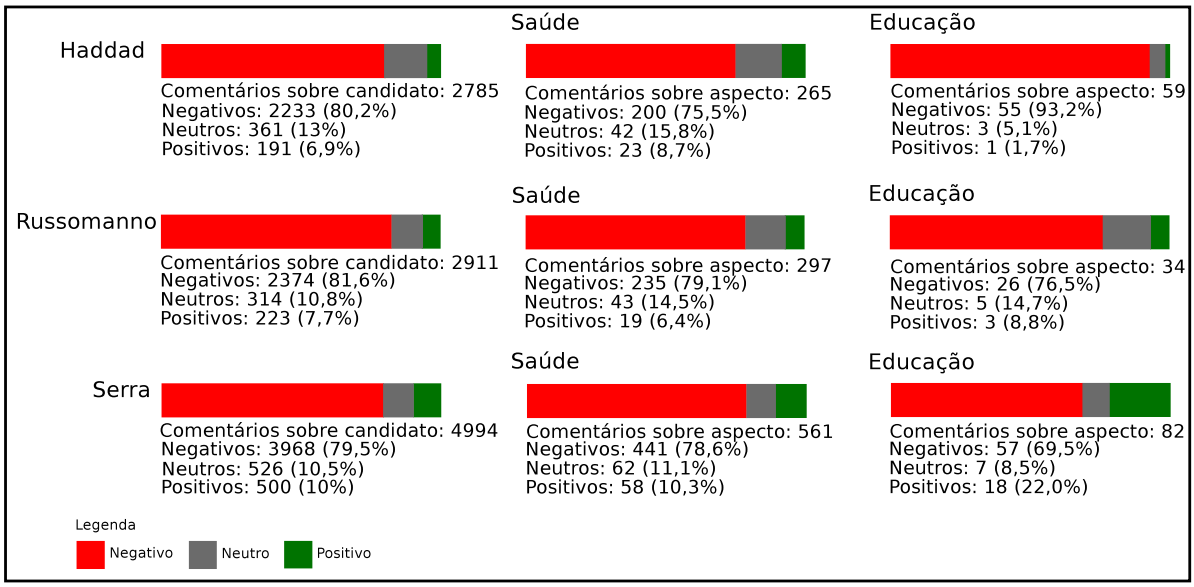
Observa-se que o mesmo comentário pode ser contabilizado para mais de um candidato, caso expresse opiniões sobre múltiplos candidatos. Nesse caso, são contabilizadas as sentenças referentes a cada candidato.

A mesma lógica foi aplicada na agregação dos comentários que mencionam candidatos e aspectos, ou seja, caso um comentário faça menção a uma entidade e aspecto, e possua mais sentenças positivas que mencionem o mesmo candidato, o comentário é classificado como positivo para o aspecto do candidato. Se existirem mais sentenças negativas ele é classificado como negativo, e caso o número de sentenças positivas e negativas sejam iguais, o comentário é classificado como neutro.

## 7.2 Sumarização visual de opiniões

Para sumarizar visualmente os resultados agregados, foram utilizados modelos visuais, que são ilustrados nas figuras 7.1, 7.2 e 7.3. Tais modelos são semelhantes ao apresentado em (CASTELLANOS et al., 2011), como descrito na Seção 3.2. Por meio do modelo da Figura 7.1, é possível visualizar a percepção da população sobre cada candidato em geral, os aspectos analisados em particular, bem como os termos mais comentados em cada tópico.

Figura 7.1: Modelo visual de sumarização de opiniões.




Fonte: Autoria própria.

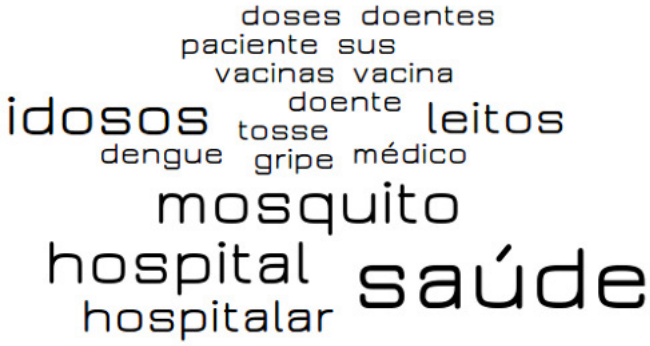
A opinião sobre cada candidato é sumarizada usando o número absoluto e percentual de comentários positivos, negativos ou neutros ($CP$). O mesmo é feito especificamente para cada aspecto ($CP_a$). Os resultados mostrados correspondem à aplicação da abordagem proposta no *corpus* de comentários das eleições de 2012, que totalizam 14.848 comentários.

Com base na sumarização apresentada na Figura 7.1, conclui-se que os leitores expressam um número substancialmente maior de opiniões negativas tanto para os candidatos em geral, quanto para seus aspectos. Observa-se também que, apesar de existir uma grande quantidade de comentários com opinião sobre os candidatos, a proporção em relação ao total de comentários foi relativamente baixa (51%). Uma explicação para esse baixo número é que muitos comentários fazem referência aos candidatos por meio de pronomes ou mencionam seus partidos. Além disso, observa-se uma quantidade de opinião da população sobre os aspectos analisados (saúde e educação) também muito baixa. Isso está relacionado ao fato de que o

conteúdo dos comentários apresenta opiniões sobre outros aspectos (e.g. corrupção) ou opiniões diretas aos candidatos. Considerando que dois candidatos haviam sido ex-ministros de educação e saúde, é interessante interpretar os comentários por aspecto. Os leitores têm uma percepção mais positiva sobre educação em relação ao ex-ministro da saúde do que no tocante ao ex-ministro da educação sobre educação.

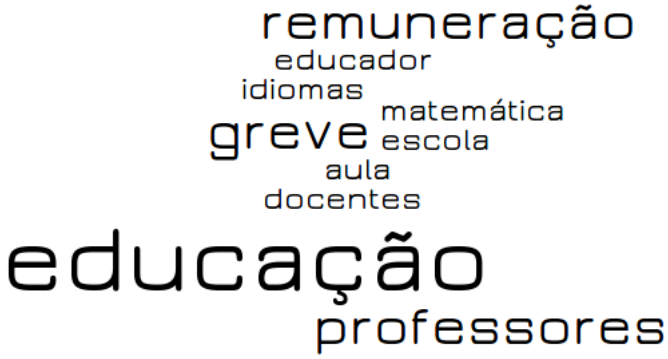
As nuvens de palavras (figuras 7.2 e 7.3) mostram os termos representativos para cada aspecto, em que o tamanho das palavras está relacionado à frequência de ocorrência de tais termos nos comentários. Percebe-se que os termos saúde e educação, que representam os aspectos de mesmo nome, são os mais mencionados nos comentários.

Figura 7.2: Nuvem de palavras para termos representativos de saúde.




Fonte: Autoria própria.

Figura 7.3: Nuvem de palavras para termos representativos de educação.




Fonte: Autoria própria.

Algo que poderia ser realizado em trabalhos futuros seria utilizar cores para representar se um termo é mencionado de forma positiva ou negativa mais vezes.

# 8 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma abordagem para minerar opiniões em nível de aspecto, usando fontes de opinião fracamente estruturadas escritas em português. Foi desenvolvido um estudo de caso completo, o qual identificou, classificou e sumarizou a percepção dos leitores de um jornal (Folha de São Paulo) sobre os aspectos saúde e educação de candidatos, baseado em seus comentários.

A revisão do estado da arte descreveu o processo de Mineração de Opiniões, enquanto os trabalhos relacionados tiveram como foco as etapas de identificação de aspectos e de sumarização de resultados. A análise dos trabalhos relacionados demonstrou que o presente trabalho é pioneiro na mineração de opiniões em nível de aspecto fora do contexto de revisões de produtos.

No que tange à identificação de aspectos, foi possível verificar, por meio do estudo de caso realizado, que é válido considerar comentários sobre notícias de qualquer assunto. Assim, foi apresentada uma abordagem para classificar comentários em relação aos aspectos Saúde e Educação. Experimentos realizados com diferentes técnicas demonstraram resultados satisfatórios para as técnicas de co-ocorrência EMIM e phi-squared, e resultados ruins para LSI.

Para polarizar as sentenças, foram utilizadas técnicas de classificação baseadas em aprendizagem de máquina, com o algoritmo SMO. Foram realizados dois experimentos para classificação de sentenças: um com classificação em três classes e outro com classificações binárias em duas etapas. Ao comparar a polarização de sentenças positivas, observa-se que ambas abordagens obtiveram bons resultados que trabalhos relacionados. No entanto, o acréscimo de uma classe prejudicou a polarização de sentenças negativas.

Após a polarização das sentenças ser realizada, o presente trabalho classifica as opiniões em nível de aspecto. Para tanto, as polarizações das sentenças são associadas aos aspectos dos candidatos mencionados nos comentários.

A sumarização de resultados desenvolvida, baseia-se em modelos visuais. Ela apresenta a percepção da população sobre cada candidato e sobre os aspectos abordados. Para realizar a sumarização, as opiniões são agregadas por meio de comentários. Dessa forma, permite-se representar as pessoas emitindo suas opiniões. Além disso, a abordagem também apresenta, por meio de nuvens de palavras, os termos representativos de cada aspecto que mais foram mencionados nos comentários.

Por meio da sumarização, foi possível observar que, apesar de existir uma grande quantidade de opiniões em relação aos candidatos, a quantidade em relação aos aspectos analisados é pequena. Isso se deve ao fato de a população apresentar opiniões sobre outras dimensões destes

candidatos.

Os resultados obtidos durante a pesquisa foram publicados em conferências:

1. SAPIRAS, L.; BECKER, K. Identificação de aspectos de candidatos eleitorais em comentários de notícias. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING - BRASNAM, 2014, Brasilia, Brasil. **Anais...** [S.l.], 2014.
2. SAPIRAS, L.; BECKER, K. Mineração da opinião sobre aspectos de candidatos a eleições em comentários de notícias. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2014, Curitiba, Brasil. **Anais...** [S.l.], 2014. p. 117–126.

Como trabalhos futuros, faz-se necessário realizar algumas melhorias na identificação dos aspectos, tais como a utilização de expressões idiomáticas, expressões multi-palavras e padrões léxico-sintáticos. É necessária a resolução de algumas limitações também, tais como a extensão do estudo de caso para outros aspectos, o uso de corpus balanceado para classificação de opiniões, o tratamento de opiniões irregulares e implícitas e a identificação automática das entidades eleitorais nos comentários. Também é importante a identificação automática de aspectos. Para resolver o problema de uma sentença expressar opiniões sobre vários candidatos (e.g. "X é horrível, votarei em Y"), a solução seria polarizar a opinião em nível de cláusulas. Pretende-se também utilizar outros meios além de comentários de notícias, tais como *posts* em redes sociais, como Twitter e Facebook.

# REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. X. **Mining Text Data**. Boston, MA: Springer US, 2012. ISBN 978-1-4614-3222-7.

ALVES, A. L. F. et al. Temporal analysis of sentiment in tweets: A case study with fifa confederations cup in brazil. In: DECKER, H. et al. (Ed.). **Database and Expert Systems Applications**. [S.l.]: Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8644). p. 81–88. ISBN 978-3-319-10072-2.

BALAHUR, A. et al. Sentiment analysis in the news. In: CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 2010, Valletta, Malta. **Procedings...** [S.l.]: ELRA, 2010. v. 10, p. 2216–2220.

BECKER, K. Introdução à Mineração de Opiniões. In: JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA / CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 2014, Brasilia, Brasil. **Anais...** 1. ed. Porto Alegre: SBC, 2014. chp. 2, p. 125–175.

BECKER, K.; TUMITAN, D. Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2013, Recife, Brasil. **Anais...** 1. ed. Pernambuco, 2013. p. 27–52.

Best Buy. **Best Buy**. 2014. Disponível em <http://www.bestbuy.com/>. Acesso em: 14 set. 2014.

CASTELLANOS, M. et al. Lci: a social channel analysis platform for live customer intelligence. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2011, Athens, Greece. **Proceedings...** New York, USA: ACM, 2011. p. 1049–1058.

CHURCH, K. W.; GALE, W. A. Concordances for parallel text. In: CONFERENCE OF THE UW CENTRE FOR THE NEW OED AND TEXT RESEARCH, 1991. **Proceedings...** [S.l.], 1991. p. 40–62.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 16, n. 1, p. 22–29, mar. 1990. ISSN 0891-2017.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 2003. **Proceedings...** [S.l.], 2003. ISBN 1-58113-680-3.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 391–407, 1990.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Commun. ACM**, ACM, New York, NY, USA, v. 39, n. 11, p. 27–34, nov. 1996. ISSN 0001-0782.

Grupo Folha. **Folha de São Paulo**. 2014. Disponível em <http://www.folha.uol.com.br>. Acesso em: 15 set. 2014.

GUO, H. et al. Product feature categorization with multilevel latent semantic association. In: ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 2009, Hong Kong, China. **Proceedings...** New York, NY, USA: ACM, 2009. (CIKM '09), p. 1087–1096. ISBN 978-1-60558-512-3.

HALL, M. et al. The weka data mining software: an update. **ACM SIGKDD Explorations Newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HALL, M. A. **Correlation-based Feature Subset Selection for Machine Learning**. Thesis (PhD) — University of Waikato, Hamilton, New Zealand, 1998.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2004, Seattle, USA. **Proceedings...** New York, USA: ACM, 2004. (KDD '04), p. 168–177.

HULL, D. A. Stemming Algorithms: A Case Study for Detailed Evaluation. **Journal of the American Society for Information Science**, v. 47, n. 1, p. 70–84, 1996.

Instituto Nacional de Ciência e Tecnologia para a Web. **Observatório da WEB**. 2014. Disponível em <http://observatorio.inweb.org.br/>. Acesso em: 10 out. 2014.

KANIS, J.; SKORKOVSKá, L. Comparison of different lemmatization approaches through the means of information retrieval performance. In: SOJKA, P. et al. (Ed.). **Text, Speech and Dialogue**. [S.l.]: Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6231). p. 93–100. ISBN 978-3-642-15759-2.

Laboratório de Estudos de Mídia e Esfera Pública. **Manchetômetro**. 2014. Disponível em <http://www.manchetometro.com.br/>. Acesso em: 1 set. 2014.

LIU, B. **Sentiment Analysis and Opinion Mining**. [S.l.]: Morgan & Claypool Publishers, 2012. (Synthesis Lectures on Human Language Technologies).

LIU, Q. et al. A logic programming approach to aspect extraction in opinion mining. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCES ON WEB INTELLIGENCE, 2013. **Proceedings...** [S.l.], 2013. p. 276–283.

MAOWEN, W. et al. Text topic mining based on LDA and co-occurrence theory. In: INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE & EDUCATION (ICCSE), 2012. **Proceedings...** [S.l.]: IEEE, 2012. p. 525–528. ISBN 978-1-4673-0242-5.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Found. Trends Inf. Retr.**, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1-2, p. 1–135, jan. 2008. ISSN 1554-0669.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: Sentiment classification using machine learning techniques. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2002. **Proceedings...** Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. p. 79–86.

PLATT, J. C. Fast training of support vector machines using sequential minimal optimization. In: **Advances in Kernel Methods - Support Vector Learning**. [S.l.]: MIT Press, 1998.

QIU, G. et al. Opinion word expansion and target extraction through double propagation. **Computational Linguistics**, v. 37, n. 1, p. 9–27, 2011.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: **Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**. Valletta, Malta: ELRA, 2010. p. 45–50. <http://is.muni.cz/publication/884893/en>.

SAPIRAS, L.; BECKER, K. Identificação de aspectos de candidatos eleitorais em comentários de notícias. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING - BRASNAM, 2014, Brasilia, Brasil. **Anais...** [S.l.], 2014.

SAPIRAS, L.; BECKER, K. Mineração da opinião sobre aspectos de candidatos a eleições em comentários de notícias. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2014, Curitiba, Brasil. **Anais...** [S.l.], 2014. p. 117–126.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: CASELI, H. de M. et al. (Ed.). **PROPOR**. [S.l.]: Springer, 2012. (Lecture Notes in Computer Science, v. 7243), p. 218–228.

SOUZA M.; VIEIRA, R. B. D. R. A. I. M. Construction of a portuguese opinion lexicon from multiple resources. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY - STIL, 2011, Cuiabá, Brasil. **Anais...** [S.l.], 2011.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Pearson Addison Wesley, 2006. (Pearson International Edition). ISBN 9780321420527.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining, (First Edition)**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. **Data Mining Knowledge Discovery**, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 3, p. 478–514, may 2012. ISSN 1384-5810.

TUMASJAN, A. et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. In: INTERNATIONAL CONFERENCE ON WEB AND SOCIAL MEDIA - ICWSM, 2010. **Proceedings...** [S.l.], 2010.

TUMITAN, D.; BECKER, K. Tracking Sentiment Evolution on User-Generated Content: A Case Study on the Brazilian Political Scene. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2013, Recife, Brasil. **Anais...** [S.l.], 2013. p. 135–144.

TUMITAN, D.; BECKER, K. Sentiment-based features for predicting election polls: a case study on the brazilian scenario. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCES ON WEB INTELLIGENCE, 2014, Warsaw, Poland. **Proceedings...** [S.l.]: IEEE Computer Society, 2014. p. 8p.

TURNEY, P. D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2002, Philadelphia, USA. **Proceedings...** Stroudsburg, USA: Association for Computational Linguistics, 2002. p. 417–424.