Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo

Lucas Lo Ami Alvino Silva

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Flávio de Barros Vidal

Brasília
2013

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Coordenador: Prof.ª Dr.ª Maristela Terto de Holanda

Banca examinadora composta por:

Prof. Dr. Flávio de Barros Vidal (Orientador) — CIC/UnB
Prof.ª Dr.ª Fernanda Lima — CIC/UnB
Prof.ª Dr.ª Carla Silva Rocha Aguiar — Egenharia de Software/UnB-FGA

CIP — Catalogação Internacional na Publicação

Silva, Lucas Lo Ami Alvino.

Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo / Lucas Lo Ami Alvino Silva. Brasília : UnB, 2013. 129 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2013.

1. análise de sentimentos em contexto, 2. naive-bayes, 3. modelos probabilísticos, 4. startups

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo

Lucas Lo Ami Alvino Silva

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Flávio de Barros Vidal (Orientador)
CIC/UnB

Prof.ª Dr.ª Fernanda Lima
CIC/UnB

Prof.ª Dr.ª Carla Silva Rocha Aguiar
Egenharia de Software/UnB-FGA

Prof.ª Dr.ª Maristela Terto de Holanda
Coordenador do Bacharelado em Ciência da Computação

Brasília, 24 de julho de 2013

Dedicatória

Dedico este trabalho à minha família, que sempre me apoiou em todos os momentos da minha vida e me permitiu ter a grande experiência de estudar na Universidade de Brasilia - UnB. Também dedico a todas as pessoas que me ajudam e me apóiam a cada dia que passa.

Agradecimentos

Agradeço a Deus, a minha família e a minha namorada, Ana Patricia por sempre me acompanharem e me abençoarem nos caminhos que desejo seguir. Aos meus amigos que sempre me deram forças nos momentos difícies e mais desafiadores desse projeto, em especial minha amiga Amanda Karolinne, gostaria de dizer muito obrigado!

Agradeço à minha orientadora, Carla Rocha, que sempre esteve ao meu lado neste trabalho desafiador. Agradeço a toda paciência, auxílio e dedicação a mim como orientando. Você sem dúvidas é um exemplo não apenas de professora, mas de pessoa e com certeza foi um exemplo pra mim nesses últimos nove meses nos quais dividimos essa maravilhosa tarefa que foi essa monografia.

Agradeço a Empresa Júnior de Computação - CJR, por ter iniciado a minha carreira profissional e ter sido a base de muitos conhecimentos necessários para o mercado de trabalho. Com certeza foi uma das melhores experiências que tive durante minha vida universitária. Carregarei todos vocês, membros da CJR, no meu coração para sempre!

Agradeço a Federação de Empresas Juniores do Distrito Federal - Concentro, por ter me proporcionado o maior de todos os desafios que já passei na vida até hoje. Consegui superar e hoje sou mais maduro por causa de vocês, "concentrados"!

Agradeço a Confederação Brasileira de Empresas Juniores- Brasil Júnior, por me dar acesso a uma das maiores redes de transformadores do Brasil! Se hoje conheço pessoas e profissionais excelentes, devo isso a todos vocês, membros e ex-membros da confederação!

Por fim, agradeço ao Vitor Fujimoto e ao Rafael Mezzomo por abrirem as portas para mim do novo rumo profissional que estou ingressando: o mundo de *Startups*. Muito obrigado pela oportunidade que me deram, foi engrandecedor para mim trabalhar com vocês no Ugla!

Resumo

Esta pesquisa foca na melhoria dos resultados da análise de sentimentos aplicada em um contexto específico, a partir da utilização de uma base de dados composta por elementos desse contexto. A análise de sentimentos é a área de estudos que foca na identificação automática dos estados privados, como opiniões, emoções, sentimentos, avaliações, crenças e especulações na linguagem natural. Esta identificação pode ser realizada em dois níveis diferentes: frases e documentos. Este trabalho apresenta uma abordagem de análise de sentimentos que começa na coleta de insumos da web até a melhoria dos classificadores contextualizados a partir do trabalho realizado sobre os atributos definidos. Durante esse processo, os dois níveis de análise de sentimentos são executados. Também são utilizados anotadores neste trabalho, que contribuem para a criação de todos os novos arquivos de treino de classificadores e na criação dos textos utilizados como teste de validação dos classificadores produzidos. É apresentada uma abordagem de avaliação da congruência entre as anotações realizadas dentro da plataforma que permite garantir a assertividade entre as anotações e evita que os textos anotados estejam enviesados com a opinião de apenas um anotador. Foi desenvolvida também, uma plataforma código aberto, o Analisador de Sentimentos, que permite realizar todos os processos necessários para realização deste estudo. Ela se caracteriza como uma contribuição para a comunidade científica para que esta possa avançar os estudos nesta área.

Palavras-chave: análise de sentimentos em contexto, naive-bayes, modelos probabilísticos, startups

Abstract

This work shows that contextual sentiment analysis is more accurate than generic sentiment analysis. To achieve this goal, a supervised learning algorithm - Naive Bayes - was used with different training bases, contextualized and not contextualized. Sentiment analysis is the field of study that focuses on automatic identification of private states, such as beliefs, emotions, feelings, evaluations, beliefs and speculations in natural language. This assessment can be carried out at two different levels: phrases and documents. This paper presents an approach to sentiment analysis that starts in collecting inputs from web to improve classifiers contextualized from the work done on the defined features. During this process, the two levels of sentiment analysis are performed. Annotators are used in this work to contribute to the creation of all new training files of classifiers and the creation of texts used as a validation test for classifiers produced. It is presented an approach for assessing the congruence between the notes made within the platform which ensures assertiveness among annotations and prevents annotated texts are skewed to the opinion of one annotator. An open source platform was developed, the Analisador de Sentimentos, to perform all procedures required for this study. It is characterized as a contribution to the scientific community so that it can advance study in this area.

Keywords: context sentiment analysis, naive-bayes, probabilistic models, startups

Sumário

1 Introdução **1**

2 Análise de Sentimentos - Conceitos Base **4**
2.1 Análise de Sentimentos 4
2.2 Aplicações 5
2.3 Definição do Problema 6
2.4 Desafios 7

3 Fases da Análise de Sentimentos **11**
3.1 Macro visão 11
3.2 Micro visão 14
3.2.1 Coleta de dados 14
3.2.2 Extração de atributos 16
3.2.3 Filtro de subjetividade 21
3.2.4 Identificação da orientação semântica 22
3.3 Anotação 33

4 O projeto **35**
4.1 Concepção 35
4.1.1 Processo de classificação 36
4.1.2 Processo de anotação e retreino 40
4.2 Desenvolvimento 41

5 Estudo de caso **46**
5.1 O experimento 46
5.2 Análise dos resultados 51

6 Conclusão **52**
6.1 Principais contribuições 52
6.2 Trabalhos futuros 52

Referências **54**

Lista de Figuras

1.1 Metodologia científica aplicada neste trabalho 2

3.1 Modelo esquemático da análise de sentimentos. Adaptado de Kumar[16] . . 11
3.2 Modelo esquemático - entradas e saídas do Web *Crawler* 12
3.3 Modelo esquemático - entradas e saídas da extração de atributos 12
3.4 Modelo esquemático - entradas e saídas do Filtro de subjetividade 13
3.5 Modelo esquemático - entradas e saídas da Orientação Semântica 13
3.6 Funcionamento básico de um *crawler*. Adaptado de Liu[17] 15
3.7 Processo realizado no aprendizado supervisionado. Adpatado de Liu [17] . 23
3.8 Cluster em plano bidimensional. Adpatado de Liu [17] 24
3.9 Exemplo de *Support Vector Machine*. Adpatado de Liu [17] 27
3.10 Distribuição de elementos em um plano. Adpatado de Statsoft 28
3.11 Distribuição de elementos em um plano. Adpatado de Statsoft 29

4.1 Modelo esquemático do primeiro processo da plataforma Analisador de Sentimentos 35
4.2 Modelo esquemático do segundo processo da plataforma Analisador de Sentimentos 36
4.3 Modelo esquemático do *crawler* do Analisador de Sentimentos 37
4.4 Modelo esquemático do *scraper* do Analisador de Sentimentos 38
4.5 Tela que permite ao usuário executar o *crawler* e o *web scraper* 43
4.6 Tela de anotação de um texto após a finalização da atividade do anotador 43
4.7 Tela de associação entre usuários anotadores e os textos que irão anotar . 44
4.8 Tela de tutorial, passo 1: explicação de conceitos 44
4.9 Tela de tutorial, passo 2: explicação de conceitos 45
4.10 Tela de comparação entre classificadores e anotadores 45

5.1 Gráfico com percentual entre categorias positiva e negativa após a classificação das frases com o classificaador genérico 47
5.2 Apresentação do sistema de frases classificadas como positivas e negativas. Frases em verde são positivas e frases em vermelho negativas 47
5.3 Tela de validação de classificadores do Analisador de Sentimentos 50
5.4 Gráfico com polaridade e tendência dos textos classificados com o classificador não balanceado 50
5.5 Gráfico com polaridade e tendência dos textos classificados com o classificador balanceado por retirada 50

Lista de Tabelas

3.1 Tags da técnica de Part-of-Speech (POS) 19
3.2 Padrões de extração de frases 33

4.1 Estudo das frases do blog *Techcrunch* 38
4.2 Funcionalidades da plataforma Analisador de Sentimentos 42

5.1 Congruência de anotação no grupo *Startup* 1 48
5.2 Congruência de anotação no grupo *Startup* 2 48
5.3 Congruência entre classificadores e anotadores para frases não ambíguas . . 49

Capítulo 1

Introdução

Word of Mouth (WOM) consiste no processo de passagem de informações de pessoa para pessoa e tem um papel importante nas decisões de compra dos consumidores. Subhabrata [22] explica que em situações comerciais, WOM engloba questões como atitudes de compartilhamento de consumidores, opiniões ou reações sobre produtos ou serviços. WOM é baseada na confiança e nos ciclos sociais das pessoas. Em outras palavras, as pessoas confiam em sua família, amigos, e outras pessoas que estejam presentes em seu ciclo social.

Pesquisas mostram que as pessoas aparentemente confiam na opinião de outras que não estejam diretamente em seu ciclo social, como aquelas presentes em avaliações online de produtos [10]. Aliado a isso, também há uma crescente disponibilidade de recursos e canais de opinião, como blogs e redes sociais. De acordo com a Nielsen Company [5], as redes sociais, que são um tipo de mídia social, foram o fenômeno consumista global de 2008. Dois terços da população mundial que utiliza internet visita redes sociais ou blogs, setores estes que agora representam quase 10% de todo o tempo gasto na internet.

Para suportar esses dois pontos, temos a explosão da web 2.0, que deu um poder sem precedentes aos consumidores de compartilhar sua opinião. Dentro desse contexto cresce a importância da análise de sentimentos como elemento capaz de identificar de forma eficiente a opinião das pessoas sobre determinado assunto ou empresa. Esta atividade tem potencial para auxiliar as empresas e startups, uma instituição humana modelada para criar um novo produto ou serviço sob condições de extrema incerteza [30], em seu processo de tomada de decisão a partir da melhor compreensão do seu mercado consumidor.

Inserida a análise de sentimentos no contexto acima, é válido ressaltar o problema identificado e avaliado neste trabalho e que será discutido nos capítulos posteriores é: **classificadores genéricos de análise de sentimentos possuem eficiência baixa em frases contextualizadas**. Entende-se por eficiência do classificador como o nível de acerto em suas inferências se comparado a polaridade real das frases.

A fim de solucionar a problemática, esta pesquisa trabalhará com a hipótese de que **é possivel aumentar a eficiência de um classificador de análise de sentimentos a partir de dados fornecidos por anotadores que produzem insumos contextualizados**.

O presente trabalho tem por finalidade **comprovar que a análise de sentimentos possui resultados mais precisos quando aplicada a determinado contexto através da utilização de um algoritmo de aprendizado supervisionado**. Para cumprir

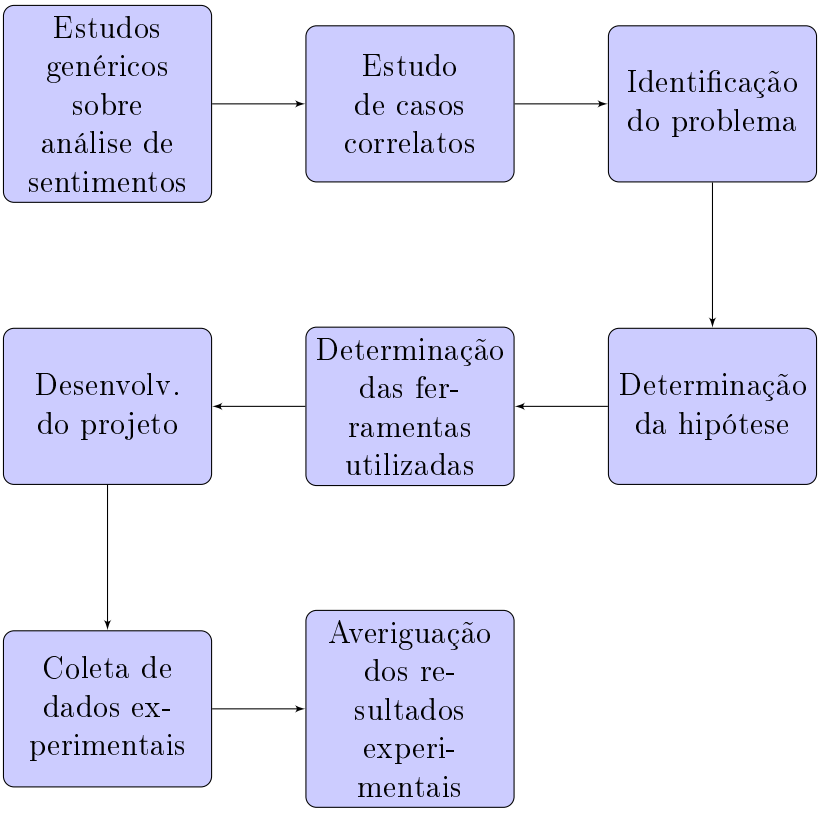

Figura 1.1: Metodologia científica aplicada neste trabalho

tal finalidade, este trabalho apresentará uma plataforma, o Analisador de Sentimentos, capaz de fazer inferências a respeito dos sentimentos envolvidos nos artigos de um blog de empreendedorismo e negócios, cuja temática principal são *startups* e que também é capaz de simular a anotação manual das frases dos textos do sistema. A priori, a plataforma será capaz de analisar as frases categorizando-as como "positivas" ou "negativas". Espera-se que, após retroalimentação do Analisador de sentimentos a partir de dados fornecidos pelos anotadores, os resultados de suas inferências sejam mais robustos e adequados ao domínio no qual está sendo aplicado, ou seja, o de *startups*, além de conseguir utilizar mais uma categoria de classificação: "neutra".

Para atingir seu objetivo, será aplicada uma metodologia científica cujas etapas estão apresentadas na figura 1.1 e serão descritas nos próximos parágrafos.

As duas primeiras etapas da metodologia apresentadas na figura 1.1 consistem na ambientação à temática da análise de sentimentos através da aquisição dos conceitos-base, aprofundamento sobre os principais estudos já exibidos na área e estudos de casos correlatos ao explicitado neste trabalho. O resultado é apresentado ao leitor nos capítulos 2 e 3.

O problema e a hipótese deste trabalho foram citados acima. As etapas seguintes da metodologia, que dizem respeito ao experimento realizado, serão descritas nas seções 4.1, 4.2 e capítulo 5.

O presente trabalho está dividido em nos seguintes capítulos: conceitos base de análise de sentimentos, que dará ao leitor os pontos mais básicos para melhor leitura deste

trabalho; fases da análise de sentimentos, que explica o processo de realização da análise de sentimentos de forma macro e de forma micro; projeto, que apresenta as principais estruturas utilizadas para realização deste trabalho; estudo de caso, que contém a descrição do experimento e a análise dos resultados obtidos; e conclusões.

Capítulo 2

Análise de Sentimentos - Conceitos Base

Para melhor compreensão do estudo que foi realizado neste trabalho, é interessante que sejam expostos ao leitor alguns conceitos base sobre análise de sentimentos e também apresentados alguns trabalhos anteriormente realizados nesta área.

Tais tópicos serão tratados neste capítulo e no capítulo 3, respectivamente.

2.1 Análise de Sentimentos

Segundo Mukherjee [22] , análise de sentimentos consiste em uma atividade que envolve Processamento de Linguagem Natural (PLN) e Extração de Informações e que visa obter o sentimento das pessoas que são expressados em comentários positivos ou negativos, perguntas e pedidos em documentos escritos através da análise de uma grande quantidade destes elementos. Em outras palavras, a análise de sentimentos tem por finalidade determinar a atitude de uma pessoa diante de um tópico específico ou diante da totalidade de um documento que aborda uma temática específica. Já de acordo com Morency e colaboradores [21] a análise de sentimentos é uma área de estudos que foca na identificação automática de estados privados, como opiniões, emoções, sentimentos, avaliações, crenças e especulações na linguagem natural. Esta área atua ainda na classificação de dados subjetivos como positivo, negativo ou neutro.

Para Liu [17], análise de sentimentos (também conhecida como mineração de opinião) realiza a avaliação da opinião das pessoas, apreciações, atitudes e emoções frente entidades, indivíduos, tópicos específicos, eventos e seus atributos e que indicam sentimentos positivos ou negativos. Este trabalho utiliza como base esta definição de análise de sentimentos.

Cientes da definição desta área de estudos, é importante que nos atentemos a algumas questões referentes a esta temática, a saber: aplicações, definição do problema da análise de sentimentos e aplicações desta.

2.2 Aplicações

A análise de sentimentos encontra utilização em diversos setores do mercado consumidor, tais como avaliação de produtos, descoberta de atitudes e suas tendências de consumidores para o fortalecimento de campanhas de marketing, encontrar opiniões a cerca de tópicos em alta ou também avaliar filmes. Pang e Lee [28] apresentam alguns empregos da análise de sentimentos, como:

1. **Sites de avaliação**:

   Sites com mecanismo de busca orientado a avaliações e opiniões agregadas a produtos, serviços, candidatos em eleições, questões políticas e assim por diante. Dois exemplos interessantes são os sites IMDb [1] e Kantar [2].

2. **Subcomponente de sistemas**:

   Neste tópico podemos abordar as seguintes utilidades:

   (a) Incremento para sistemas de recomendação na medida em que o estes podem não recomendar itens que receberam muitos feedbacks negativos.

   (b) Detecção de "flames" em e-mails, que consistem de declarações depreciativas contidas nestes canais de comunicação e que são vistas apenas pelo receptor da mensagem.

   (c) Detecção de páginas que contém conteúdo sensível à colocação de anúncios em sistemas que mostram propagandas como barras laterais. Em sistemas mais robustos é possível ainda apresentar propaganda dos produtos quando estes possuem feedbacks positivos dos consumidores e escondê-los quando os feedbacks são negativos.

   (d) Melhora nas respostas a questões orientadas a opiniões.

3. **Inteligência de mercado e inteligência governamental**:

   Neste tópico é possível explorar sistemas capazes de identificar as respostas de importantes questões de mercado, como por exemplo, os motivos pela diminuição nas vendas de determinado produto ou compilar avaliações de clientes ou pontos consensuais destes em relação a determinado produto. É importante frisar que existem diversas fontes na internet de onde essas informações podem ser extraídas e cada uma delas tem suas especificidades, tais como forma, teor das opiniões, língua e hábitos de escrita. Este fato implica na necessidade de técnicas mais robustas de identificação das opiniões.

   Para inteligência governamental, é interessante citar plataformas que podem monitorar as fontes de aumento nas comunicações hostis e negativas. Mukund e Avik [23] discutem alguns exemplos de aplicações nesta área em seu trabalho.

[1] http://www.imdb.com/
[2] http://www.kantar.com/

2.3 Definição do Problema

A definição do problema da análise de sentimentos dada por Liu [17] nos permite determinar estruturas em um texto complexo de forma didática, o que viabiliza uma compreensão mais clara da área de estudos. Em sua obra, Liu [17] explora o seguinte exemplo (os números entre parênteses identificam as frases contidas no texto):

> "(1)Eu comprei um iPhone alguns dias atrás. (2)Ele parecia ser um ótimo celular. (3) O touch screen era realmente bom. (4)A qualidade de voz era clara também. (5)Porém, minha mãe ficou furiosa comigo por não ter avisado a ela antes de ter feito a compra dele. (6)Ela também achava que o celular era muito caro e queria que eu o devolvesse para a loja."

A questão-chave neste ponto é: o que nós queremos minerar ou extrair dessa passagem? Liu [17] mostra que na passagem existem diversas opiniões, sejam elas positivas, como as expressas nas sentenças (2), (3) e (4), ou negativas , como nas sentenças (5) e (6). Ele percebe que todas as opiniões possuem alvos. O alvo da opinião na sentença (4), por exemplo, é a qualidade da voz no iPhone. E por último, vale ressaltar que todas as opiniões possuem um dono. O dono das opiniões nas sentenças (2), (3) e (4) são o próprio autor da passagem (o "eu").

Baseado nesse exemplo, o pesquisador citado formaliza os seguintes conceitos:

1. **Entidade**: uma entidade e é um produto, serviço, pessoa, evento, organização ou tópico. Ela é associada com um par: $e(T, W)$, onde T é a hierarquia de componentes e subcomponentes e W é o conjunto de atributos de e. Aqui vale a ressalva de que cada componente ou subcomponente tem seus próprios atributos.

   Exemplo: um carro de uma marca em particular é considerado uma entidade, como por exemplo, um Palio. No conjunto de seus componentes podemos encontrar elementos como pneus, para-brisas e teto solar. Já o conjunto de atributos é composto por elementos como peso, tamanho e cor.

2. **Aspecto**: os aspectos de uma entidade e são os componentes e atributos de e.

3. **Nome de aspecto e Expressão de aspecto**: o nome de aspecto faz jus ao nome de um aspecto dado por um usuário, enquanto uma expressão de aspecto é uma palavra real ou frase que apareceu no texto indicando um aspecto.

   Exemplo: no domínio de carros, um aspecto pode receber o nome de pneus. Existem algumas expressões que representam esse aspecto, tais como "jogo de rodas", "rodas" e o próprio "pneu".

4. **Nome de entidade e Expressão de entidade**: um nome de entidade é o nome de uma entidade dado por um usuário. Já uma expressão de entidade é uma palavra real ou fase que apareceu no texto indicando uma entidade.

5. **Dono da opinião**: o dono da opinião é a pessoa ou organização que expressou uma opinião em determinado contexto.

6. **Opinião**: uma opinião é uma quíntupla:

$$(e_i, a_{ij}, oo_{ijkl}, h_k, t_l) \tag{2.1}$$

onde e_i é o nome de uma entidade, a_{ij} é um aspecto de e_i, h_k é o dono da opinião e t_l é o tempo no qual h_k expressou sua opinião. Já o componente oo_{ijkl} pode ser positivo, negativo ou neutro ou ainda pode ser expresso com diferentes níveis de força/intensidade.

Dados os conceitos acima, Liu [17] define o problema da análise de sentimentos em dois níveis: classificação de sentimentos em nível de documento e classificação de sentimentos em nível de sentenças. Ele o faz da seguinte forma:

1. **Definição do problema em nível de documento**: dado um documento opinado d que avalia uma entidade e, determine a orientação da opinião oo_{ijkl} em um aspecto GERAL na quíntupla *(e, GERAL, oo, h, t)*. Assume-se que os valores de e, h e t são conhecidos ou irrelevantes.
2. **Definição do problema em nível de sentenças**: dada uma sentença s, duas subatividades devem ser realizadas:
   (a) **Classificação de subjetividade**: determine se s é uma sentença subjetiva ou objetiva.
   (b) **Classificação de sentimento em nível de sentença**: se s é subjetiva, determine se ela expressa uma opinião positiva, negativa ou neutra.

Perceba que o segundo problema não leva em consideração a quíntupla de opinião citada no início desta seção. Liu[17] explica que isso acontece porque a classificação de sentimentos em nível de sentenças é, em geral, uma etapa intermediária, trabalhando como filtro de subjetividade dentro da classificação de sentimentos em nível de sentenças.

2.4 Desafios

Subhabrata [22] aborda em seu trabalho os seguintes desafios para a análise de sentimentos:

1. **Sentimentos implícitos e sarcamos**:
   As sentenças de um texto podem apresentar sentimentos implicitamente mesmo que não haja presença clara destes através das palavras. Considere o exemplo abaixo:
   > "Como alguém consegue passar uma tarde inteira na fila do SUS?"
   Percebe-se que essa sentença não carrega explicitamente um teor negativo através das palavras, apesar de ela possuir essa polaridade.
2. **Dependência de domínio**:
   Aqui é preciso compreender que algumas palavras tem sua polaridade modificada de acordo com o domínio no qual se encontram. Considere o seguinte exemplo:
   > "Leia o livro O Caçador de Pipas"

Essa frase possui um sentimento positivo se considerarmos o domínio "leitura de livros", tendo em vista que é uma indicação do produto, porém, se considerarmos o domínio de "filmes que reproduzem livros" pode haver um teor negativo, tendo em vista que o diretor do filme pode receber o feedback negativo de que precisa ler o livro para reproduzi-lo melhor no filme.

3. **Expectativas frustradas**:

   Este caso está relacionado a situações em que o autor insere o contexto em sua mensagem apenas para refutá-lo ao final desta. Considere o seguinte exemplo:

   > "Nossa, essa viagem deveria ser maravilhosa! Encontramos passagens áreas baratas, conseguimos um excelente hotel e receberíamos um serviço turístico renomado na região. Porém, tudo acabou indo por água abaixo!"

   Apesar de termos palavras agradáveis durante a maior parte do texto e que indicam um sentimento positivo, o texto na realidade expressa um sentimento negativo em relação a situação, posto que a última sentença é crucial na definição da polaridade da mensagem.

4. **Pragmática**:

   A pragmática, ramo da linguística que estuda a linguagem dentro do domínio de sua aplicação – a comunicação, é um elemento importante a ser detectado, posto que ela é capaz de alterar completamente o sentimento do usuário. Veja os seguintes exemplos:

   > "Caramba, meu time DESTRUIU o seu no jogo de ontem, hein?!"

   > "Estou completamente destruído após um dia inteiro de trabalho!"

   Percebe-se que o uso de caixa alta no verbo destruir na primeira frase denota um sentimento (positivo). Por outro lado, este mesmo verbo denota um sentimento negativo na segunda frase, significando o mesmo que exausto.

5. **Conhecimento de mundo**:

   É importante que algum conhecimento de mundo seja adicionado aos sistemas que analisam sentimentos. Veja o seguinte exemplo:

   > "Ela é uma verdadeira bruxa!"

   Esta frase retrata um sentimento negativo. Porém, para identifica-lo, é preciso ter um conhecimento de mundo sobre o que representa o termo bruxa na frase.

6. **Detecção de subjetividade**:

   Consiste na diferenciação de textos com opinião e textos sem opinião. É uma técnica que auxilia na detecção de sentimentos na medida em que pode compor filtros que retiram as mensagens objetivas do conjunto de dados a serem analisados. Veja os seguintes exemplos:

   > "Comprei o novo smartphone da Motorola ontem, o Milestone 3."

"Gostaria de comprar um celular que fosse leve."

"Odiei o novo celular que ganhei!."

O primeiro exemplo representa uma frase objetiva. O segundo exemplo apresenta uma frase subjetiva, porém o autor da mesma não expressa uma opinião positiva ou negativa em relação a algo. Por fim, a terceira frase é subjetiva e expressa uma opinião negativa em relação ao celular.

7. **Identificação de Entidade**:

   Por vezes em uma mesma sentença temos a presença de mais de uma entidade. Frases comparativas são exemplos clássicos desse caso. É importante, portanto, identificar para qual das entidades do contexto a opinião dada é direcionada. Veja os exemplos abaixo:

   "O Motorola Razr é melhor que o iPhone 5."

   "O Vasco foi muito superior ao Flamengo no jogo do Campeonato Brasileiro de ontem."

   Os exemplos acima representam sentimento positivo em relação a Motorola Razr e Vasco e sentimento negativo em relação a iPhone 5 e Flamengo.

8. **Negação**:

   A manipulação de negações dentro do contexto de uma frase é um dos grandes desafios da análise de sentimentos. Isso se deve ao fato de que, além de uma negação poder ser expressa de várias maneiras (algumas vezes explícita e outras implicitamente), é preciso que se considere o escopo da negação para saber se ela realmente se caracteriza como uma. Vamos considerar os seguintes exemplos:

   "Eu não gosto do Fantástico."

   "Eu não gostei do elenco do filme, mas adorei a direção do Steven Spielberg."

   "Eu não apenas gostei do elenco do filme, bem como adorei a direção do Steven Spielberg."

   O primeiro exemplo representa o método mais simples para identificação de negação, que consiste em encontrar o operador de negação (no exemplo a palavra "não") na frase e reverter a polaridade desta.

   Este método, porém, é muito simples para identificar que a frase não está completamente negativada no segundo exemplo, pois ele não é capaz de identificar o escopo da negação. Neste caso é preciso perceber que o operador de negação "mas" está modificando a polaridade da segunda sentença. O método, portanto, consiste em modificar a polaridade de todas as palavras seguintes a negação até se encontrar outra negação.

   O terceiro exemplo representa um caso que os métodos anteriores não conseguem solucionar. A polaridade da primeira sentença não é alterada em vista da presença

do operador "não" em vista do termo "apenas" que vem em seguida. Portanto, um método mais robusto deve ser capaz de identificar expressões como "não apenas" e afins.

Capítulo 3

Fases da Análise de Sentimentos

Neste capítulo será construída junto ao leitor uma compreensão das etapas de realização da análise de sentimentos. Esta atividade será feita de forma macro a fim de se obter uma visão genérica do processo e também de forma micro, com uma breve revisão bibliográfica em torno de cada uma das etapas da análise de sentimentos. Ademais, neste capítulo será feita uma explanação sobre o processo de anotação da análise de sentimentos.

Vale a ressalva de que as etapas do processo descrito acima são semelhantes àquelas presentes na mineração de dados, podendo haver ou não etapas que incluem mecanismos de aprendizado de máquina.

Este capítulo será dividido em três seções, à saber: macro visão (3.1), micro visão (3.2) e anotação (3.3).

3.1 Macro visão

A figura 3.1 abaixo mostra um modelo esquemático da análise de sentimentos. Cada etapa mostrada funciona como um filtro de dados para a etapa posterior. Os modelos esquemáticos apresentados nas figuras (3.2),(3.4),(3.3) e (3.5) mostram as entradas e as saídas desses filtros. Os próximos parágrafos discursarão brevemente sobre cada etapa.

O primeiro passo consiste na coleta de dados para montar uma base de dados a partir da qual possam se realizar os experimentos científicos. Por ser utilizado neste trabalho, esta etapa trabalhará bastante com o conceito de *crawlers*, máquinas capazes de capturar páginas web a partir de uma regra de negócio pré-estabelecida. Este recebe um conjunto de documentos web como entrada e retira deste um subconjunto de documentos web que atendem à regra de negócio estabelecida (vide figura 3.2). A literatura sugere também

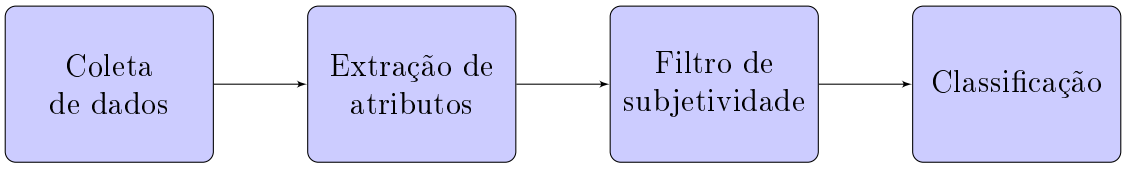

Figura 3.1: Modelo esquemático da análise de sentimentos. Adaptado de Kumar[16]

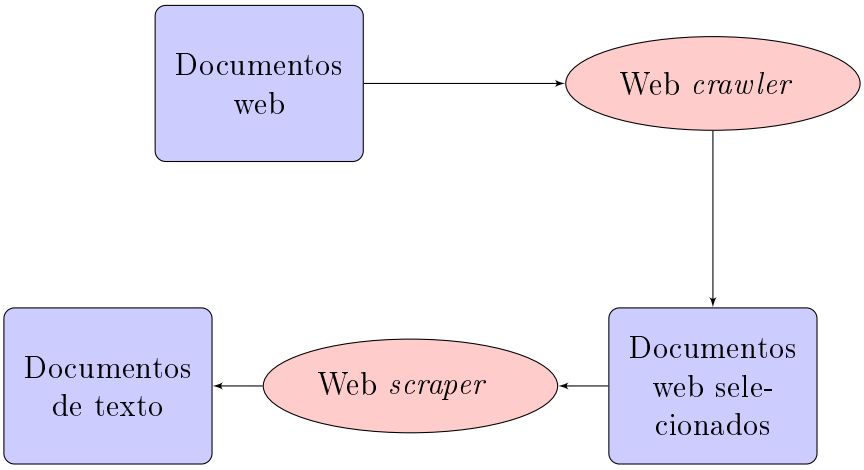

Figura 3.2: Modelo esquemático - entradas e saídas do Web *Crawler*

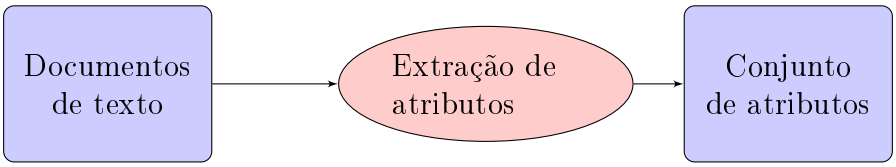

Figura 3.3: Modelo esquemático - entradas e saídas da extração de atributos

para a execução desta etapa a utilização de bases de dados previamente montadas, tais como Blog TREC[27] , MPQA [1], entre outras que serão exploradas na seção 3.2.

A próxima etapa é determinar e extrair, a partir da base de dados construída, os elementos-chave que serão utilizados para realizar a inferência da orientação semântica dos textos que estão sendo avaliados, que são chamados de atributos. Produz-se, então, um conjunto de elementos-chave (vide figura 3.3). Estes elementos serão utilizados para inferir a orientação de opinião (o elemento "oo" da quíntupla de opinião) e a subjetividade em frases ou textos.

A etapa de filtro de subjetividade consiste em identificar e selecionar em documentos de texto os trechos destes nos quais há subjetividade no discurso do autor. Esta fase pressupõe a ideia de que os sentimentos são expressos apenas em textos subjetivos [40] [29]. A saída produzida consiste em um subconjunto de documentos de textos que contém apenas os trechos subjetivos dos textos originais, vide figura 3.4.

O último estágio é a classificação. Ela recebe os atributos da fase anterior e produz um resultado que informa a polaridade sentença, frase ou documento, vide figura 3.5.

O modelo proposto nessa seção foi produzido com o intuito de facilitar a compreensão do leitor sobre a análise de sentimentos. Ele não é seguido à risca na literatura utilizada como referência deste trabalho, como é percebido em alguns exemplos da seção 3.2.

[1] http://mpqa.cs.pitt.edu/

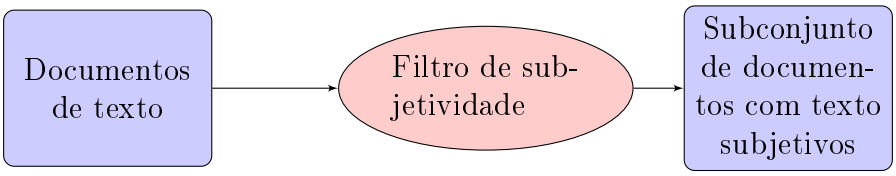

Figura 3.4: Modelo esquemático - entradas e saídas do Filtro de subjetividade

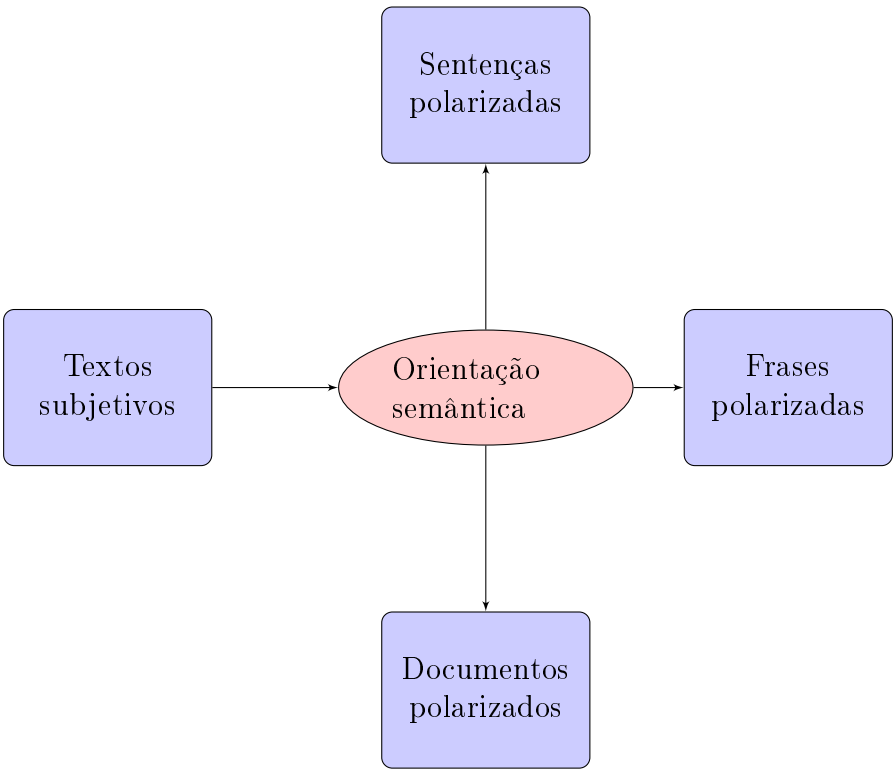

Figura 3.5: Modelo esquemático - entradas e saídas da Orientação Semântica

3.2 Micro visão

A seguir cada uma das etapas mostrada na figura 3.1 será explorada nas subseções. Conforme descrito anteriormente, esta seção visa informar ao leitor detalhes sobre cada uma das etapas do processo de análise de sentimentos.

3.2.1 Coleta de dados

As próximas duas subseções abordam os dois tipos de atividades realizadas na coleta de dados, conforme descrito na seção 3.1.

Base de dados previamente montadas

Pang [28] explica que uma fonte de opiniões e sentimentos é a anotação manual de textos. Por ser uma atividade com alto custo de tempo, pesquisadores buscam formas alternativas de obter textos com sentimentos através de recursos pré-existentes, como sites de análise, como Rotten Tomatoes [2] , Epinions [3] e Amazon [4]. Estes, por sua vez, fornecem uma base de dados que pode ser utilizada para treinar classificadores de aprendizado de máquina – como será desenvolvido neste trabalho, vide capítulos 4 e 5 - bem como elementos de teste de robustez em experimentos.

Pang [28] ainda apresenta em seu trabalho alguns conjuntos de dados anotados já criados e que podem ser utilizados. Abaixo, os principais são apresentados:

1. **Blog 06**: a universidade de Glasgow distribui essa coleção de dados de teste de 25GB, que consiste em postagens de blogs sobre vários tópicos. Esta base inclui informações sobre blogs de maior referência, fornecidas pela Nielsen BuzzMetrics [5].
2. **Base de dados de análise de filmes da Universidade de Cornell** [6] esta base consiste dos seguintes conjuntos de dados:
   (a) Conjunto de dados de polaridade de sentimentos
   (b) Conjunto de dados com escala de sentimentos
   (c) Conjuunto de dados de subjetividade
3. **MPQA *corpus***: esta base contém mais de 500 notícias de artigos de diversas fontes. Tais notícias foram anotadas de forma manual em nível de sentenças e sub-sentenças para opiniões e outros estados privados.

O presente trabalho faz uso do conjunto de dados de polaridade de sentimentos para treinar um classificador genérico.

[2] http://www.rottentomatoes.com/
[3] http://www.epinions.com/
[4] http://www.amazon.com/
[5] http://www.nielsen-online.com/buzzmetrics/
[6] http://www.cs.cornell.edu/people/pabo/movie-review-data/

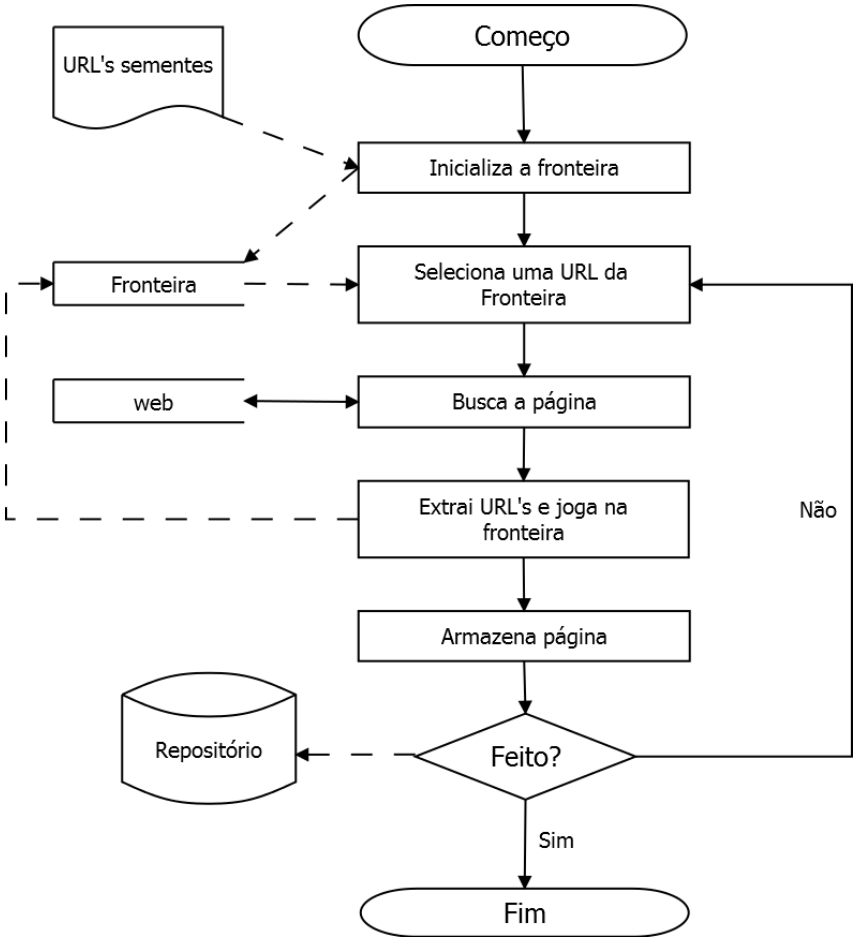

Figura 3.6: Funcionamento básico de um *crawler*. Adaptado de Liu[17]

Web Crawler

Web crawlers, também chamados de *spiders*, *bots*, *ants* ou *web scutters* são programas que automaticamente recuperam e fazem o *download* de páginas Web [17] [4].

Castillo [2] cita em seu trabalho que a World Wide Web se tornou em poucos anos uma verdadeira Biblioteca de Alexandria da atualidade, funcionando como a mais larga difusora de cultura de todos os tempos. Com essa grande difusão de informações através de bilhões de páginas na internet, estas que são servidas por e para milhares de servidores ao redor do mundo, os usuários de computador tem acesso rápido a vários *hyperlinks* que lhes permitem avançar mais e mais em diversas páginas novas. Um *web crawler* pode acessar milhares de sites e coletar informações específicas dos mesmos para que sejam analisadas e mineradas tanto online quanto off-line.

A figura 3.6 mostra um *crawler* sequencial. Ele se inicia com um as URL's sementes inicializando e alimentando a fronteira, que consiste de um conjunto de URL's não visitadas. A cada iteração do loop principal do crawler uma URL nova é lançada, a página a qual ela se refere é buscada na web e então desta são extraídas novas URL's que alimentarão o sistema. Paralelamente a essa atividade, a página buscada pelo *crawler* é armazenada em uma base de dados. Esse processo pode ser finalizado após um número

pré-determinado de páginas for capturado. O *crawler* também pode parar devido ao esvaziamento da fronteira, mas esse caso é pouco provável de acontecer, devido a média de *hyperlinks* por página web, que de acordo com Liu [17] é dez por página.

É possível dividir os *crawlers* em três grandes tipos: universais, focados e em tópicos [17]. Aqueles do tipo universal foram desenvolvidos para atuar em larga escala e busca cobrir o máximo de páginas importantes possível, ou seja, tornando a busca por novas páginas mais próxima do desejo de quem está manipulando a ferramenta. Já o do tipo focado objetiva influenciar a busca do *crawler* por páginas em determinada categoria selecionada pelo usuário. Este tipo de *crawler* funciona através de um mecanismo de aprendizado das categorias de página desejadas. Por último vem o os *crawlers* em tópicos, que funcionam forma semelhante aos focados, porém ele é utilizado em casos nos quais não há páginas suficientes na fronteiras ou nas URL's sementes para treinar ou o sistema possui apenas uma pequena *query* [20] [17].

A área de estudos sobre *Web Crawlers* possui diversos desafios que envolvem tópicos como: aumentar a eficiência da busca por páginas novas (URL's), realizar o *parsing* das páginas HTML de forma mais efetiva (tendo em vista que muitos sites são construídos com formatação equivocada do código HTML), extração de links e canonicalização das URL's, escalabilidade dos algoritmos e limitações de hardware [4] [17].

É válido citar o trabalho realizado por Chakrabarti e colaboradores [3], no qual desenvolveu-se uma plataforma denomiada *Focused Crawler* que é capaz de buscar, adquirir, indexa e mantém páginas em um conjunto específico de tópicos que representam um segmento estreito da web. O diferencial deste trabalho jaz no fato de que o software desenvolvido contém dois módulos de mineração de hipertexto chamados *classifier* e *distiller*. O *classifier* tem a função de avaliar se um documento de hipertexto está relacionado ao tópico em foco. Já o *distiller* possui a responsabilidade de identificar nos documentos de hipertexto os nós que dão acesso a páginas muito relevantes através de links.

3.2.2 Extração de atributos

Sufian e Ranjith [35], os quais explicam em seu trabalho que o termo atributo (*"feature"*) se refere aos principais elementos sobre os usuários querem mirar para gerar uma opinião sobre os eles. Matematicamente falando, os atributos correspondem a elementos da entrada que são selecionados para reduzir o tamanho desta e torna-la administrável para processamento e análise dos dados.

A engenharia de atributos é uma atividade-base da análise de sentimentos. A conversão de um pedaço de texto em um vetor de atributos ou qualquer outra representação capaz de tornar os atributos mais importantes e salientes disponíveis é o primeiro passo de qualquer abordagem dirigida a dados em análise de sentimentos. Nesta seção veremos alguns atributos que são usados comumente na análise de sentimentos.

Presença de termos *versus* frequência de termos

A frequência de termos é considerada essencial em sistemas de Recuperação de Informações e de Classificação Textual tradicionais [28]. A técnica do TF-IDF (*Term Frequency – Inverse Document Frequency*) é uma medida bem utilizada na modelagem de documentos. A ideia da abordagem da frequência de termos é de que os termos que aparecem frequentemente em um documento são mais informativos no que tange a determinação

da ideia sobre a qual o documento fala do que aqueles termos que apenas aparecem uma única vez ou poucas vezes.

Essa realidade muda quando se está trabalhando no universo da Análise de Sentimentos. Pang e colaboradores [29] obtiveram uma performance melhor usando a presença de termos. Isto é, vetores de atributos com valores binários cujas entradas representavam meramente se um termo ocorre (valor 1) ou não (valor 0) formaram uma base mais eficaz para a averiguar a classificação da polaridade do que utilizar vetores de atributo com valores reais em que estes valores aumentam com a ocorrência do termo correspondente.

Wiebe [38] averiguou em seu trabalho também que as *hapax legomena*, que são palavras que aparecem uma única vez dado um corpo de avaliação, são indicadores de alta precisão de subjetividade em textos.

Atributos baseados em termos e n-gramas

Informações sobre posição espacial no texto ganham cada vez mais espaço quando se fala de atributos. Mas o que isso significa? Mukherjee [22] explica que palavras que aparecem em certas posições no texto carregam mais sentimentos ou peso sobre estes do que palavras que aparecerem em outras posições. Isto é semelhante ao que acontece na Recuperação de Informação onde as palavras que aparecem em tópicos, tais como título, subtítulos, resumos, etc. tem um maior peso do que aquelas que aparecem no corpo do texto.

Um caso interessante que ilustra essa situação é a mostrada no exemplo da seção 2.4 no tópico de "Expectativa Frustrada", no qual apesar de o texto conter muitas palavras positivas , a presença de um sentimento negativo na sentença final tem uma participação crucial do sentimento final relacionado a passagem. De forma geral, as palavras pertencentes as n primeiras sentenças e as m últimas em um texto tem um peso crucial na determinação do sentimento.

A informação de posição espacial é utilizada por Pang [29] e pelos pesquisadores Kim e Hovy [14] na codificação de vetores de atributos.

N-gramas também são utilizados como atributos. Um n-grama consiste em uma sequência contígua de n itens de um dado trecho de texto ou discurso. Um n-grama pode ser qualquer combinação de letras, fonemas, sílabas, palavras, etc. Eles são capazes de captar o contexto até certo ponto e são amplamente utilizados em Processamento de Linguagem Natural.

Dave [8] utiliza considera em seu trabalho cada palavra dos textos da base de dados como n-gramas. Ou seja, palavras como *"this"* e *"pretty"* são consideradas unigramas, trechos de texto com duas palavras, como *"this is"* são considerados bigramas e assim por diante. Estes n-gramas recebem pesos a partir da sua frequência de aparecimento dentro dos textos da base de dados. O pesquisador usa também o sistema de ponderação de Gauss em seu trabalho.

Uma questão a ser debatida é se a utilização de n-gramas de ordem elevada são atributos utilizáveis. Pang [29] reportaram que unigramas superam em performance os bigramas na análise de sentimentos de críticas de filmes, apresentando como resultados em sua pesquisa uma performance de 82,9% de precisão com utilização de unigramas com a técnica *Support Vector Machines* contra 77,1% de precisão com a utilização de bigramas com a mesma técnica. Entretanto, Dave e colaboradores [8] encontraram em seu trabalho que

sob algumas configurações, bigramas e trigramas produzem uma melhor classificação da polaridade na análise de produtos. Apresentam um resultado de 87,2% de precisão com a utilização de bigramas com a técnica *Support Vector Machines* contra 81,1% de precisão com unigramas.

A distância contrastiva entre termos foi utilizada como um atributo automaticamente computado por Snyder e Barzilay [33] como parte de um sistema de inferência de avaliação. Para melhor visualizar o conceito trabalhado nesse projeto é possível usar o exemplo do par de palavras **delicioso** e **sujo**, que possuem um alto contraste em termos de avaliação implícita da polaridade das mesmas. A distância contrastiva para um par de palavras é computada considerando a diferença de peso relativo associado às palavras nos modelos de *Ranking* individualmente treinados. Além desses atributos, Snyder e Barzilay [33] utilizaram os mesmos n-gramas de Pang e Lee [29].

Part-of-speech tagging

Part-of-speech (POS) de uma palavra é a categoria linguística que é definida através do seu comportamento sintático e morfológico. A informação de *part-of-speech* é comumente explorada em análise de sentimentos. Uma razão básica para isso acontecer é o fato de que essa categoria de atributo pode ser considerada uma forma bruta de desambiguação do sentido da palavra.

POS *tagging* consiste na atividade de rotular cada palavra na sentença com sua apropriada part-of-speech. A tabela 3.1 mostra as tags utilizadas por Santorini [31] em seu trabalho.

Como citado anteriormente, foi determinado que adjetivos são bons indicadores de sentimento dentro de um texto. Os primeiros trabalhos realizados para predizer a orientação semântica das palavras foram desenvolvidos com o viés para adjetivos. O pesquisador Hatzivassiloglou [12] [13] possui trabalhos interessantes na área.

O fato de os adjetivos serem um bom indicador de que uma sentença é subjetiva ou não implica que outras partes do discurso não contribuem na expressão de sentimentos ou opiniões. Porém, em seu estudo de classificação da polaridade da análise de filmes, Pang [29] encontrou em seu trabalho que usando atributos compostos apenas por adjetivos possuem uma performance muito pior do que aquela utilizando unigramas. Em suma, eles mostraram que nomes e verbos podem ser também forte indicadores de sentimento.

A combinação de advérbios com adjetivos também se mostra um indicador interessante de subjetividade, como citado anteriormente. A maioria dos advérbios não possui uma polaridade *a priori*. Porém, quando eles aparecem modificando um adjetivo, são peças significativas para a determinação do sentimento de uma sentença. Benamara [1] mostrou como os advérbios alteram o valor do sentimento do adjetivo com o qual são utilizados.

Negação

Manipular negações dentro de uma sentença é uma ideia interessante na análise relacionada à opinião e a sentimento. Enquanto a abordagem de conjunto de palavras (conhecida como *bag of words*) entende sentenças como "Eu gosto desse livro" e "Eu não gosto desse livro" similares nos mecanismos de mensuração mais comuns, uma única palavra, o termo negativo, faz com que as duas sentenças estejam em classes opostas. Pang

Tabela 3.1: Tags da técnica de Part-of-Speech (POS)

Tag	Descrição
CC	conjunção coordenativa
CD	número cardinal
DT	determinador
EX	verbo "haver"com sentido de existir
FW	palavra estrangeira
IN	preposição ou conjunção subordinada
JJ	adjetivo
JJR	adjetivo comparativo
JJS	adjetivo no superlativo
LS	marcador de item de lista
MD	verbo modal
NN	substantivo, singular ou coletivo
NNS	substantivo no plural
NNP	nome próprio no singular
NNPS	nome próprio no plural
PDT	pré-determinador
POS	final possessivo
PRP	pronome pessoal
PRP$	pronome possessivo
RB	advérvio
RBR	advérbio comparativo
RBS	advérbio superlativo
RP	partícula
SYM	símbolo
TO	para
UH	interjeição
VB	verbo na forma básica
VBD	verbo conjugado no passado
VBG	verbo, gerúndio ou particípio do presente
VBN	verbo no particípio do passado
VBP	verbo no presente sem estar na 3º pessoa do singular
VBZ	verbo no presente na 3º pessoa do singular
WDT	determinador com Wh
WP	pronome com Wh
WP$	pronome possessivo com Wh
WRB	advérbio com Wh

e Lee [28] afirmam que não há situação paralela a essa na Recuperação de Informação, na qual um simples termo pode ser aplicado em uma regra de classificação.

Mejova [19] reporta que as negações são geralmente consideradas nos resultados pós-processamento, enquanto a representação original do texto as ignora. Pang e Lee [28] compreendem isso afirmando que negações podem ser trabalhadas como elementos secundários do atributo de um segmento de texto, onde a representação inicial deste como um vetor ignora a negação, mas posteriormente a representação é alterada de forma a se preocupar com a negação.

Outra solução interessante é a proposta por Das e Chen [7] na qual é anexado o termo "NOT" a palavras que ocorrem próximos a termos negativos, tais como *"no"* ou *"don't"*. Pela ilustrar essa ideia considere o seguinte exemplo:

> *"I don't like deadlines!"*

O *token "like"* é convertido em um novo *token "like*-NOT".

Porém, um ponto a ser analisado é que nem todas as aparições explícitas de negação invertem a polaridade da sentença na qual aparecem. Para ilustrar essa ideia considere o seguinte exemplo fornecido por Pang e Lee [28]:

> *"No wonder this is considered one of the best."*

É incorreto associar o termo "NOT" ao *token "best"*. Para manipular esses casos Na e colaboradores [24] usaram padrões específicos de *part-of-speech tags* para identificar as negações relevantes para polaridade do sentimento em um texto. Em seu conjunto de dados de análises eletrônicas, observaram um aumento de 3% na precisão como resultado da modelagem que fizeram em relação as abordagens que não consideravam negações no trabalho, tais como unigramas e POS *tags*.

Outra dificuldade para modelar a negação diz respeito ao fato de que ela pode ser expressa de muitas formas sutis (vide seção 2.4). Sarcasmo e ironia podem ser elementos muito difíceis de serem identificados. Wilson e colaboradores [40] discutem outros efeitos complexos da negação.

Atributos orientados a tópicos

Interações entre tópico e sentimento tem um importante papel na análise de sentimentos. Pang e Lee [28] utilizam os seguintes exemplos para ilustrar essa questão:

> "Wal-mart informa que os lucros aumentaram."

> "Concorrentes informam que lucros aumentaram."

Essas duas sentenças podem indicar tipos de notícias completamente diferentes (uma boa e outra ruim) considerando o sujeito do documento, no caso, Wal-mart. Dessa forma, informações sobre tópicos devem ser incorporadas aos atributos. Por exemplo, poderiamos ter os tópicos: Wal-Mart e concorrentes que contem informações sobre cada uma das entidades.

Para a análise de opiniões preditivas (isto é, se uma mensagem M sobre um partido político P prediz se P irá vencer nas eleições), Kim e Hovy [15] propuseram utilizar generalização de atributos. Especificamente, para cada sentença em M, cada nome de candidato

e de partido político é substituído pelo termo "PARTY" (ou seja, P) ou "OTHER" (não-P). A partir daí, padrões como "PARTY ganhará" ou "OTHER irá ganhar" são extraídos como atributos de n-gramas. Este modelo supera a utilização de n-gramas simples como atributos em cerca de 10% de precisão na classificação de qual partido político uma mensagem prediz que ganhará.

3.2.3 Filtro de subjetividade

Esta segunda etapa do processo de análise de sentimento diz respeito a detecção de sentimentos ou opiniões em frases. Ela pode ser vista como uma classificação de um texto como subjetivo ou objetivo. Aqui é perceptível que esta etapa está relacionada com a resolução do problema da classificação de sentimentos em nível de sentenças, descrito em seção 2.3.

Para melhor compreender essa etapa, é interessante que fique claro o conceito de linguagem subjetiva. De acordo com Wiebe [38] uma linguagem subjetiva consiste naquela utilizada para expressar estados privados no contexto de um texto ou uma conversação. Estados privados é um termo genérico que engloba opiniões, avaliações, emoções e especulações. Os exemplos a seguir, extraídos do trabalho de Wiebe [38] e traduzidos para a língua portuguesa, mostram alguns casos de subjetividade:

> "Em diversas camadas diferentes, este conto é fascinante." (George Melloan, "Whose Spying on Our Computers?" Wall Street Journal, November 1,1989).
>
> "Os custos com os cuidados de saúde estão corroendo o nosso padrão de vida e minando nossa força industrial." (Kenneth H. Bacon, "Business and Labor Reach a Consensus on Need to Overhaul Health-Care System," Wall Street Journal, November 1, 1989).

Wiebe [38] também aborda em seu trabalho o conceito de elementos subjetivos, que consistem em expressões linguísticas dos estados privados dentro de um contexto, tal como é o termo "corroendo" no exemplo acima.

Por se tratar de um problema de classificação, a etapa de filtragem acopla bem soluções tradicionais de aprendizado supervisionado, tais como classificação de *Naive-Bayes* ou *Support Vector Machines* [29]. Wiebe e colaboradores [39] utilizaram o método de classificação de Naive-Bayes em seu projeto, que tinha por finalidade demonstrar um procedimento capaz de criar, de forma automatizada, as melhores *tags* únicas (para serem utilizadas no classificador) quando existem julgadores de subjetividade que discordam.

Kim e Hovy [14] descrevem em seu trabalho um método efetivo para obter dentro de um texto palavras que são suporte para opinião e palavras que não são suporte para opinião. Este trabalho foi testado em três conjuntos de dados diferentes: MPQA, o novo modelo do TREC 2003 [34] e a própria base de dados montada por eles. Com base nisso, mostraram que o método citado pode ser utilizado eficientemente para identificação de sentenças que dão suporte a opinião. Trabalhos semelhantes a estes são citados por Pang [28].

É interessante citar o TREC-2006, projeto desenvolvido por Ounis e colaboradores [27], no qual uma coleção de teste em larga escala foi desenvolvida. Sobre esta coleção foi rodado um experimento no qual um grupo de participantes avaliaria o conteúdo dos

elementos dessa coleção e avaliaria se os textos sobre determinado tópico presentes na mesma possuíam algum tipo de opinião sobre a temática do tópico. Os autores abordam as técnicas utilizadas pelos participantes para completar suas atividades, tais como abordagens baseadas em dicionários, soluções apoiadas em classificação textual e uma abordagem voltada a linguística simples.

Outro trabalho relevante é o elaborado por Godbole e Skiena [11], no qual os autores apresentam um sistema que atua na fase de identificação de sentimento em textos, que associa as opiniões expressas com cada entidade relevante. Além disso, a plataforma possui etapas de agregação de sentimentos e pontuação de sentenças. Neste trabalho é utilizado o seguinte conceito de subjetividade:

> "A subjetividade reflete a quantidade de sentimento que é associada a uma entidade, sem considerar se as sentenças relacionadas a tal entidade são positivas ou negativas."

3.2.4 Identificação da orientação semântica

O último passo do processo de análise de sentimentos é dedicada a classificação dos atributos de forma a associá-los a um conjunto definido. O termo "conjunto" pode ser considerado como categorias que classificam um determinado elemento. Por exemplo, se considerarmos uma aplicação que identifica frases subjetivas e objetivas dentro de um texto, então os atributos estão vinculados ao conjunto C = {subjetivo, objetivo}.

Os algoritmos classificadores podem ser de aprendizado supervisionado ou de aprendizado não supervisionado.

O aprendizado supervisionado, também chamado de aprendizado induzido, possui o mesmo princípio de aprendizagem baseado em experiências passadas que os humanos utilizam, por exemplo, para adquirir novos conhecimentos com a finalidade de alavancar sua expertise para realizar determinada atividade do mundo real [17]. Emanuel [9] afirma que os algoritmos de aprendizado supervisionado visam construir um modelo de distribuição de classes (categorias) em função das características dos dados em questão. Este modelo é definido (ou "aprendido") a partir de uma base de treinamento, onde as instâncias e suas classes são previamente conhecidas (funciona como as experiências passadas citadas por Liu [17]).

Liu[17] explica alguns conceitos básicos sobre esse tipo de classificação. O conjunto de dados usado na atividade de aprendizado consiste em um conjunto de registros de dados que pode ser descrito da seguinte maneira $A = \{A_1, A_2, ..., A_{|A|}\}$, onde $A_n (1 <= n <= |A|)$ é um atributo de um registro e |A| é o tamanho do conjunto A ou o número de atributos. Este conjunto de dados também possui um atributo especial chamado de atributo de classe (este será chamado de C). C é considerado separadamente dos demais atributos de A, ou seja, é assumido que C não está em A. O atributo de classe pode ser representado por um conjunto discreto $C = \{c_1, c_2, ..., c_{|c|}\}$, onde |C| é o número de classes e $|C| > 2$.

Dado um conjunto de dados D, a meta do aprendizado é produzir uma função de classificação capaz de relacionar os valores de atributo em A e classes em C. Esta função encontrada é usada na predição de dados futuros.

Os passos para realização da atividade de aprendizado supervisionado são descritos na figura 3.7. Durante o passo 1, o algoritmo de aprendizado usa os dados de treinamentos para produzir um modelo de classificação. O passo 2 consiste em utilizar os dados de teste no modelo aprendido a fim de obter maior precisão na classificação. Vale a ressalva de que

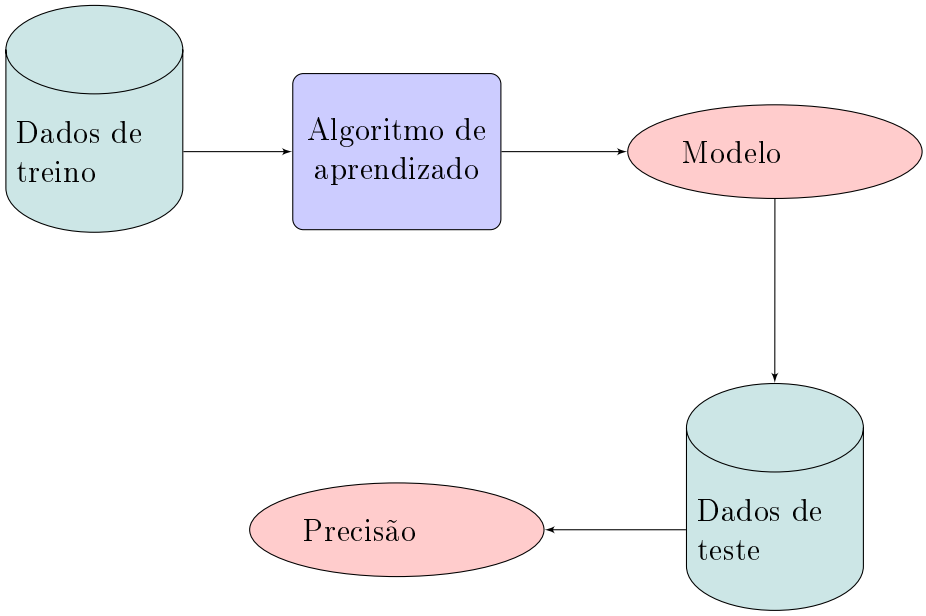

Figura 3.7: Processo realizado no aprendizado supervisionado. Adpatado de Liu [17]

uma premissa fundamental do aprendizado de máquina é tida nesse tipo de classificação: a distribuição dos exemplos de treinamento é análoga à distribuição dos exemplos de teste. Essa premissa assegura a acurácia dos resultados obtidos. Porém, no mundo real essa premissa é violada diversas vezes. Caso esta situação gere uma precisão insatisfatória é preciso repetir o processo de aprendizado utilizando outro algoritmo de aprendizado.

Serão apresentados ainda nessa seção três algoritmos de aprendizado supervisionado: Naive-Bayes Text Classification, Support Vector Machines e Entropia Máxima, que foram utilizados por Pang e Lee[29].

Por outro lado, aprendizado não supervisionado parte do princípio de que os atributos de classe não são conhecidos, portanto é preciso que sejam descobertos padrões que relacionem os atributos dos dados aos atributos de classe[17]. Portanto, os algoritmos de aprendizagem não supervisionada devem fazer agrupamento das instâncias analisadas tendo em vista algumas medidas de similaridade entre elas [9].

Liu [17] também explica alguns conceitos básicos sobre esse tipo de classificação. A organização das instâncias de dados em grupos de similaridade é o processo chamado de clusterização. Esta é uma das técnicas de análise de dados mais utilizada. Uma instância de dados é chamada de ponto de dados e pode ser visualizada como um ponto dentro de um espaço n-dimensional, onde n é o número de atributos nos dados.

Na figura 3.8 é possível visualizar um exemplo de *clusters* (como são chamados os grupos de similaridade) em um espaço bidimensional. Cada grupo de dados é um cluster diferente, o que significa que os seus atributos são afins dentro de si próprios, mas diferentes em relação a outro cluster.

O desafio da clusterização é, portanto, encontrar esses três clusters dentro do conjunto de dados trabalhado. Essa demanda parece simples a olho nu quando se trabalha com espaços bi ou tridimensionais, entretanto ela se torna complexa quando o número de dimensões do espaço aumenta. Outro ponto a ser levantado aqui é o fato de que nem

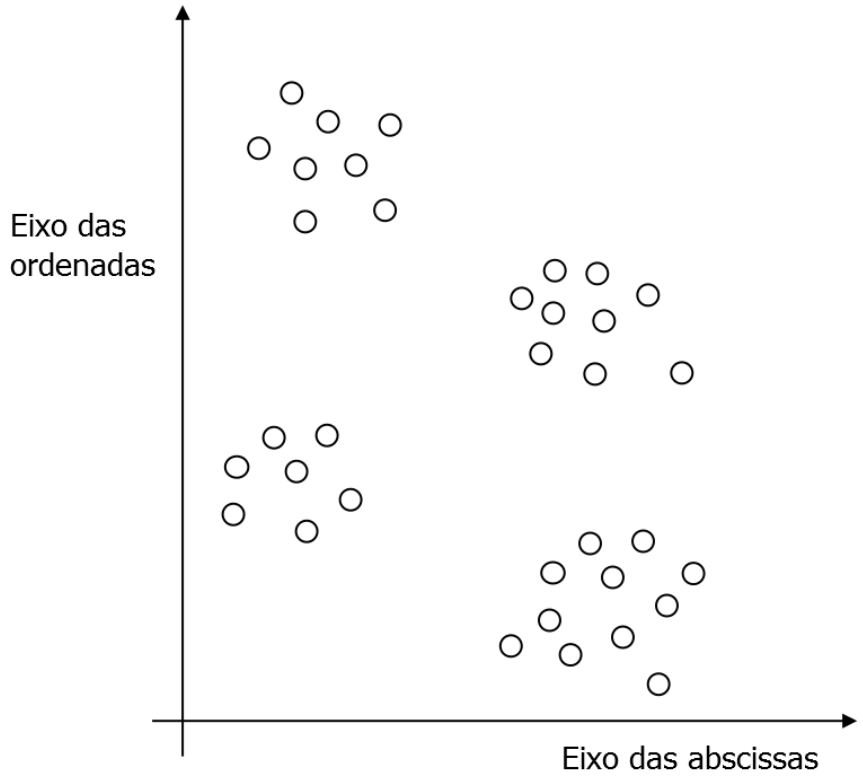

Figura 3.8: Cluster em plano bidimensional. Adpatado de Liu [17]

sempre os *clusters* estão bem delineados como no exemplo da figura 3.8.

A fim de obter melhores resultados, a clusterização precisa de uma função de similaridade que mensura o quão parecidos são dois pontos de dados ou uma função de distância para mensurar a distância entre dois pontos de dados. Elas servem de base para responder a seguinte pergunta: este ponto de dados pertence a um *cluster* X ou não?

Existem dois tipos principais de clusterização: a particional e a hierárquica. O primeiro trabalha com a premissa de que cada ponto de dado pertence a um único *cluster*. O algoritmo *K-means* pertence a esse tipo. O segundo trabalha aninhando sequências de clusters como uma árvore. Em outras palavras, um cluster filho também pertence ao *cluster* pai.

Considere os seguintes exemplos, retirados de Liu [17], para melhor compreensão da clusterização com enfoque na sua utilização no dia-a-dia:

> **Exemplo 1:** uma empresa deseja conduzir uma campanha de marketing para promover os seus produtos. A estratégia mais efetiva é criar modelos personalizados de materiais de *marketing* para cada cliente levando em consideração o seu perfil e situação financeira. Porém, essa estratégia é muito cara para um número de clientes elevado. Outra estratégia que pode ser usada é a de desenvolver um modelo genérico de *marketing* que seria utilizado para todos os tipos de clientes. Entretanto, essa estratégia pode não ser efetiva, na medida em que cada tipo de cliente pode não se sentir valorizado, o que diminui seu potencial de compra. Dessa forma, o melhor custo benefício neste caso é usar segmentação de clientes a fim de agrupá-los em

grupos pequenos com similaridades. Este agrupamento é feito através de algoritmos de clusterização.

Exemplo2: diariamente, agências de notícias em todo o mundo produzem um número de novos artigos gigantesco. Caso um site deseje coletar esses novos artigos com o objetivo de fornecer um serviço de notícias integrado, ele terá que organizar todos os artigos coletados de acordo com algum tipo de hierarquia. Dado o contexto o desafio a ser respondido, portanto, é: quais devem ser tais tópicos e como devem ser organizados? Uma estratégia para solucionar essa questão é contratar pessoas que realizem essa atividade. Contudo, essa estratégia é custosa, tanto financeiramente (a empresa deve pagar todos os editores) quanto em questão de tempo. Outra possível saída é utilizar métodos de classificação de aprendizado supervisionado para classificar os artigos de acordo com tópicos pré-determinados. Todavia, os tópicos de notícia mudam constante e rapidamente, o que torna os dados de treinamento voláteis, o que torna inviável trabalha-los manualmente. Nessa situação, a melhor alternativa é utilizar a clusterização, posto que ela automaticamente agrupa as notícias de acordo com suas similaridades de conteúdo.

Será apresentado nessa seção uma solução de aprendizado não supervisionado para análise de sentimentos que utiliza *Part-of-Speech Tagging* e o algoritmo *Pointwise Mutual Information* (PMI)

Support Vector Machines - SVM

Consiste em uma técnica muito útil de classificação de dados, baseada em aprendizado supervisionado. É um considerado um dos melhores algoritmos existentes no ramo [32].

A versão mais atual do algoritmo foi proposta por Cortes e Vapnik [6]. De maneira informal, o algoritmo funciona da seguinte forma: representar cada documento analisado por um ponto ou vetor em um espaço n-dimensional e traçar um hiperplano que separe da melhor forma possível as duas classes de documentos em questão, isto é, deve-se maximar a distância do hiperplano e dos elementos (documentos) de borda das classes, que são os chamados *support-vectors* [9]. Emanuel [9] diz sobre o funcionamento do algoritmo:

> "[...] O algoritmo tenta estimar uma função $f : R_n \rightarrow \{-1, +1\}$ usando um conjunto de treinamento, onde cada elemento desse conjunto é um vetor n-dimensional $(x_i, y_i) \in R_n X \{-1, +1\}$, de forma que essa função será capaz de classificar corretamente uma nova instância (x, y), ou seja, $f(x) = y$."

Liu[17] trabalha de uma maneira mais formal os conceitos relacionados ao SVM. Ele explica que este consiste em um sistema de aprendizado linear que constrói duas classes de classificadores. Para melhor compreensão deste fato, considere o exemplo de conjunto de treino:

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \tag{3.1}$$

onde $x_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$ é um vetor de entrada (é o mesmo que instância de dados explicado anteriormente) n-dimensional em um espaço com valores reais $X \in R_n$, y_i é seu rótulo de classe (é o mesmo que a constante C explicada na seção de aprendizado supervisionado) e $y_i \in \{-1, +1\}$. -1 denota a classe negativa e 1 denota a classe positiva.

Para construir um classificador, o algoritmo SVM visa encontrar uma função linear que possui o seguinte formato:

$$f(x) =< w \cdot x > +b \tag{3.2}$$

Dessa forma um vetor de entrada é direcionado para uma classe positiva se $f(x_i) >= 0$ e para a classe negativa em no outro caso possível. A fórmula abaixo resume este conceito:

$$Y_i = \begin{cases} 1, \text{se } < w \cdot x_i >>= 0 \\ \\ -1, \text{se } < w \cdot x_i >< 0 \end{cases} \tag{3.3}$$

Nestas fórmulas, temos que f(x) é uma função com valores reais $f : R_n \rightarrow R$, $w = (w_1, w_2, ..., w_r) \in Rn$ e é chamado de vetor de peso. $b \in R$ e é chamado de fator de viés. $< w \cdot .x >$ é o produto vetorial de w e x, que pode ser representado da seguinte forma:

$$f(x_1, x_2, ..., x_n) = w_1x_1 + w_2x_2 + ... + w_nx_n + b \tag{3.4}$$

Onde x_i representa a i-ésima coordenada do vetor x.
A essência do SVM é encontrar o hiperplano, ou superfície de decisão:

$$< w \cdot x > +b = 0 \tag{3.5}$$

que separa os exemplos de treinamento em planos positivo e negativo.
A figura 3.9 mostra um exemplo de uso de SVM em um plano bidimensional. As entradas positivas são representadas pelos retângulos e as entradas negativas são representadas pelos círculos. A linha ao meio do gráfico é a superfície de decisão, que separa as instâncias negativas das positivas.

Classificação *Naive Bayes*

O aprendizado supervisionado pode ser estudado com naturalidade a partir do ponto de vista probabilístico. Os classificadores bayesianos são um exemplo disso. Eles têm ganhado bastante popularidade, pois mostraram ter um desempenho supreendentemente bom [18].

Andrew [18] explica que estas abordagens probabilísticas fazem fortes suposições sobre como os dados são produzidos e postulam um modelo probabilístico que engloba estas suposições. Então elas usam a coleção de dados rotulados para estimar os parâmetros do modelo gerador.

O autor também afirma que o classificador de *Naive-Bayes* é o mais simples destes modelos bayesianos. Esse classificador tem a premissa de que todos os atributos dos exemplos de treino são independentes dos outros, dado um contexto de classe. Por mais que essa premissa seja falsa para a maioria das atividades do mundo real, *Naive-Bayes* tem um desempenho muito bom frequentemente. Essa contradição é explicada pelo fato de que a estimativa de classificação é apenas uma função para dar o sinal da função estimativa. Em outras palavras, a função de aproximação pode ser pobre enquanto a eficácia da classificação continua alta.

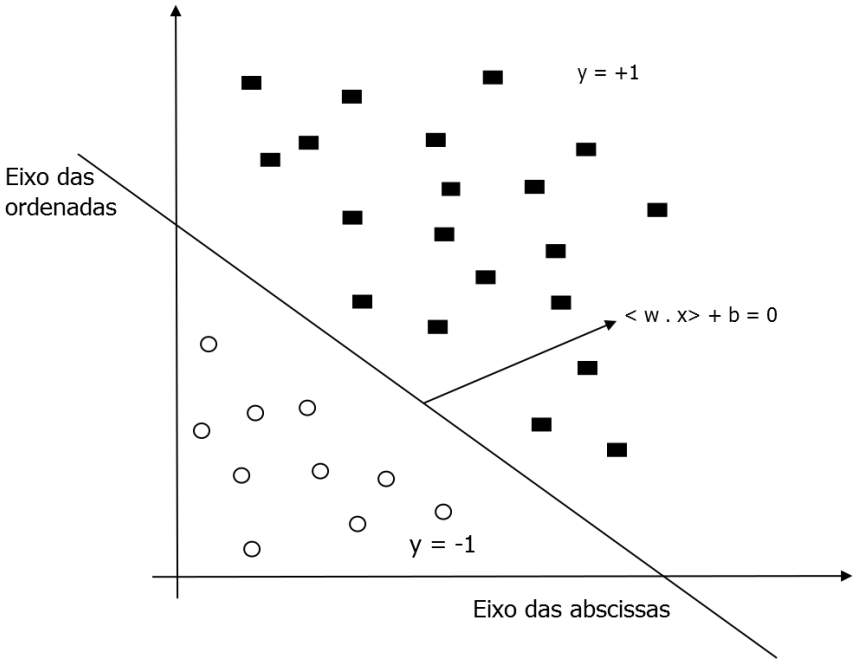

Figura 3.9: Exemplo de *Support Vector Machine*. Adpatado de Liu [17]

Para demonstrar os conceitos que regem a classificação de *Naive-Bayes*, StatSoft [7] apresenta um exemplo ilustrativo.

Considere a figura 3.10. É possível classificar os objetos como sendo X ou O. O desafio aqui é classificar novas ocorrências de um elemento na medida em que chegam, definindo a qual classe ele pertence baseado nos objetos que existem atualmente.

Percebe-se que existe o dobro de elementos O em relação aos elementos X. Dessa forma, é possível imaginar que um novo caso (que ainda não foi observado) tem duas vezes mais chances de ser membro do grupo O ao invés de X. Na análise bayesiana, essa premissa é entendida como uma probabilidade prévia. Probabilidades prévias são baseadas em experiências anteriores, no caso do exemplo a porcentagem de elementos O e X, e utilizadas para predizer novas ocorrências antes mesmo de elas acontecerem. Logo, podemos representar essa afirmação da seguinte forma:

$$\text{Probabilidade prévia de O} = \frac{n^{\circ}\text{de objetos O}}{n^{\circ}\text{total de objetos}} \tag{3.6}$$

$$\text{Probabilidade prévia de X} = \frac{n^{\circ}\text{de objetos X}}{n^{\circ}\text{total de objetos}} \tag{3.7}$$

Como temos um total de 60 elementos, sendo 40 O e 20 X, nossas probabilidades prévias para ser membro de uma das classes são:

$$\text{Probabilidade prévia de O} = \frac{40}{60} \tag{3.8}$$

[7] http://www.statsoft.com/textbook/naive-bayes-classifier/

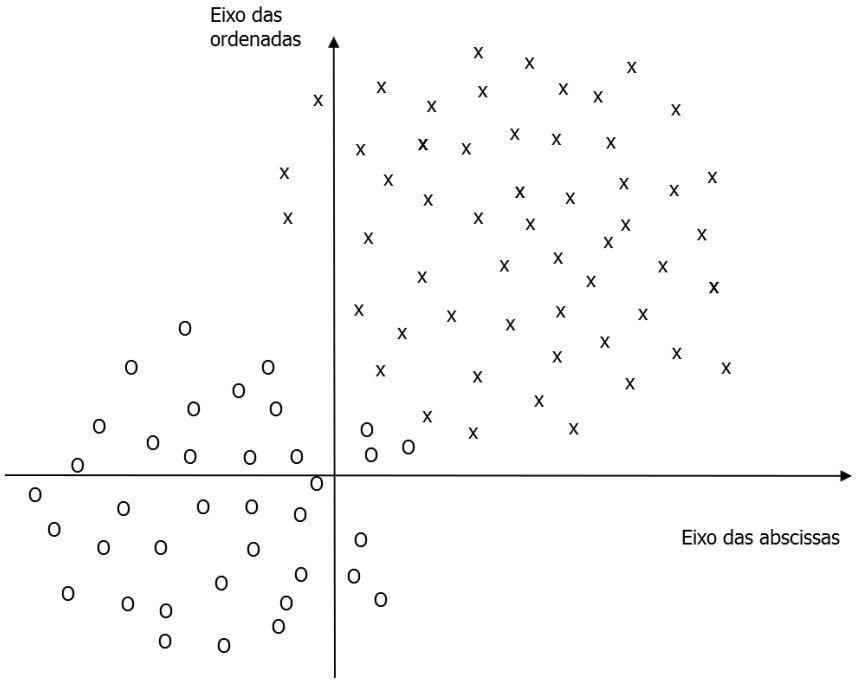

Figura 3.10: Distribuição de elementos em um plano. Adpatado de Statsoft

$$\text{Probabilidade prévia de X} = \frac{20}{60} \tag{3.9}$$

Posta as relações de probabilidade prévias, é possível agora classificar um novo objeto Y (Z mostrado na figura 3.11). Uma vez que os objetos estão bem agrupados em suas classes, é possível supor que quanto mais objetos O (ou X) existem na vizinhança de Y, maior a probabilidade de Y pertencer a essa classe. A mensuração dessa probabilidade se dá traçando um círculo em torno de Y e que engloba um número de pontos, que é escolhido a priori, independente dos rótulos de classe. Depois é calculado o número de pontos no círculo pertencentes a cada etiqueta de classe. A partir dessa informação, é calculado:

$$\text{Probabilidade Y dado O} = \frac{n°\text{de O na vizinhança de Y}}{n°\text{total de O}} \tag{3.10}$$

$$\text{Probabilidade Y dado X} = \frac{n°\text{de X na vizinhança de Y}}{n°\text{total de X}} \tag{3.11}$$

É possível afirmar que a probabilidade de Y dado O é menor que a probabilidade de Y dado X a partir da visualização da figura 3.11. Logo:

$$\text{Probabilidade Y dado O} = \frac{1}{40} \tag{3.12}$$

$$\text{Probabilidade Y dado X} = \frac{3}{20} \tag{3.13}$$

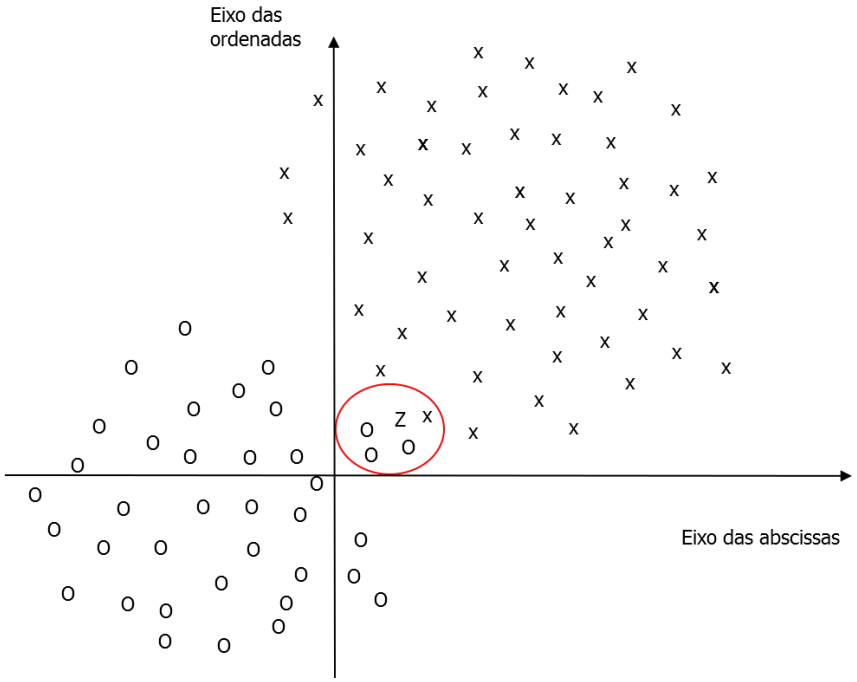

Figura 3.11: Distribuição de elementos em um plano. Adpatado de Statsoft

Apesar de as probabilidades prévias informarem que Y tende a ser da classe de O, a probabilidade calculada acima mostra o contrário. Na análise bayesiana, a classificação final é produzida a partir da combinação das duas fontes de informação obtidas nas equações (3.8), (3.9), (3.12) e (3.13) para formar a probabilidade posterior. Veja as equações (3.14) e (3.15)

$$\text{Prob. Post. de Y ser O} = \text{Prob. Prévia de O x Probabilidade de Y dado O} \tag{3.14}$$

$$\text{Prob. Post. de Y ser X} = \text{Prob. Prévia de X x Probabilidade de Y dado X} \tag{3.15}$$

Substituindo os valores, temos:

$$\text{Prob. Post. de Y ser O} = \frac{4}{6} \times \frac{1}{40} = \frac{1}{60} \tag{3.16}$$

$$\text{Prob. Post. de Y ser X} = \frac{2}{6} \times \frac{3}{20} = \frac{1}{20} \tag{3.17}$$

Finalmente, Y é calculado como pertencente a classe X.
StatSoft e Liu[17] também definem matematicamente este classificador.

Definição: Seja $D = \{A_1, A_2, ..., A_{|A|}\}$ o conjunto de atributos com valores discretos de um conjunto de dados D. Seja C o atributo de classe com |C| valores possíveis,

$C = \{c_1, c_2, ..., c_n\}$. Dado um exemplo de teste d com valores de atributos conhecidos a_1 até $a_{|A|}$, onde ai é um possível valor para A_i, ou seja,

$$D =< A_1 = a_1, ..., A_{|A|} = a_{|A|} > \tag{3.18}$$

A probabilidade posterior $P(C = c_j | A_1 = a_1, ..., A_{|A|} = a_{|A|})$ pode ser avaliada como:

$$\begin{aligned} P(C = c_j | A_1 = a_1, ..., A_{|A|} = a_{|A|}) \propto \\ P(A_1 = a_1, ..., A_{|A|} = a_{|A|} | C = c_j) P(C = c_j) \end{aligned} \tag{3.19}$$

Como *Naive-Bayes* assume que todos os atributos são condicionalmente independentes, dada uma classe $C = c_j$ é possível decompor a probabilidade de A dado $C = c_j$ no produto dos termos:

$$P(A | C = c_j) \propto \prod_{i=1}^{|A|} P(A_i = a_i | C = c_j) \tag{3.20}$$

E reescrever a equação (3.19) como:

$$P(C = c_j | A_1 = a_1, ..., A_{|A|} = a_{|A|}) \propto P(C = c_j) \prod_{i=1}^{|A|} P(A_i = a_i | C = c_j) \tag{3.21}$$

A modelagem do algoritmo *Naive-Bayes* pode ser feita de várias formas. Para visualizá-las, recomenda-se que o leitor veja StatSoft.

Entropia Máxima

Consiste em uma técnica genérica de aprendizado de máquina que fornece uma estimativa o menos tendenciosa possível baseado nas informações dadas. Em outras palavras, ela evita ao máximo fazer premissas em cima de informações em falta. Também vale a ressalva de que não há suposição de independência condicional de atributos, como o classificador de *Naive Bayes* faz.

O princípio primordial que rege a Entropia Máxima (EM) é o de que, quando nada é sabido, a distribuição de dados deve ser a mais uniforme possível. Dados rotulados de treinamento são utilizados para derivar um conjunto de restrições para o modelo. Estes limites são representados como os valores esperados dos atributos, que consistem em uma função de valores reais. O melhor algoritmo de escala interativa encontra a distribuição de entropia máxima que está em harmonia com as restrições impostas inicialmente [25].

Pang e Lee [29] afirmam em várias aplicações de processamento de linguagem natural essa técnica é muito efetiva. Dizem também que algumas vezes esse algoritmo apresenta desempenho melhor que o classificador *Naive Bayes*. Neste mesmo trabalho, estes pesquisadores fazem uma comparação do desempenho dos três algoritmos citados até agora nessa seção no ambiente de análise de sentimentos.

Nigam[25] aborda em seu trabalho a matemática por trás desse tipo de classificação. Os próximos parágrafos servirão para descrever tal ponto.

Seja $f_i(d, c)$ uma função de valores reais. Considere-a como um atributo do documento e de sua classe. A técnica da máxima entropia permite restringir o modelo de distribuição a fim de ter o mesmo valor esperado para esse atributo que o visto nos dados de treinamento, chamados aqui de D. É possível estipular que a distribuição condicional aprendida deve ter a propriedade:

$$P(c|d) = \frac{1}{|D|} \sum_{d \in D} f_i(d, c(d)) = \sum_{d} P(d) \sum_{c} P(c|d) f_i(d, c) \tag{3.22}$$

Na prática, a distribuição do documento P(d) não é conhecida. Então, os dados de treino são utilizados, sem rótulos de classe, como uma aproximação da distribuição do documento e produz-se a restrição:

$$P(c|d) = \frac{1}{|D|} \sum_{d \in D} f_i(d, c(d)) = \sum_{d \in D} \sum_{c} P(c|d) f_i(d, c) \tag{3.23}$$

Portanto, o primeiro passo quando se está utilizando esse classificador é identificar o conjunto de funções de atributo que são importantes para a classificação. Após isso, é preciso medir o valor esperado de cada atributo de acordo com os dados de treinamento, capturar o resultado e utilizar isso como uma restrição para modelo distributivo.

Após as restrições serem estimadas no passo anterior, é garantido que existe apenas uma única distribuição que tem entropia máxima. O pesquisador também afirma que a distribuição sempre tem um formato exponencial:

$$P(c|d) = \frac{1}{Z(d)} exp(\sum_{i} \lambda_i f_i(d, c)) \tag{3.24}$$

Onde cada $f_i(d, c)$ é um atributo, λ_i é um parâmetro estimado e Z(d) é um fator de normalização que pode ser definido como:

$$P(c|d) = \sum_{c} exp(\sum_{i} \lambda_i f_i(d, c)) \tag{3.25}$$

Para finalizar, o pesquisador sugere uma abordagem possível para encontrar a solução de máxima entropia:

1. Suponha inicialmente qualquer distribuição exponencial no formato correto como um ponto de largada
2. Execute o algoritmo *hillclimbing* no espaço de probabilidade
3. Como não existe nenhum valor local máximo, os passos anteriores convergirão para a solução de probabilidade máxima em modelos exponenciais, que também será a solução de máxima entropia global.

Part-of-speech **(POS) e** ***Pointwise Mutual Information*** **(PMI)**

Nesta subseção será explicado como essa técnica é utilizada em conjunto com a *Pointwise Mutual Information* (PMI) para determinar a orientação semântica de textos. Estes per-

meiam a etapa de definição dos atributos até os modelos matemáticos aplicados para determinação da orientação semântica.

Liu[17] explica tal ponto em três passos:

1. Passo 1: o algoritmo extrai frases contendo adjetivos ou advérbios, posto que esses são elementos sao bons indicadores de opiniões. Porém, mesmo um adjetivo ou advérbio podendo indicar uma opinião, há casos em que eles podem ter contextos insuficientes para determinar a orientação semântica da opinião. Por exemplo, a palavra "imprevisível"pode ter uma orientação negativa no contexto de análise automotiva, como na expressão "direção imprevisível", mas também pode ter uma orientação positiva, caso se esteja no contexto de de análise de filmes, como na expressão "enredo imprevisível". Portanto, o algoritmo extrai duas palavras consecutivas, onde um membro no par é um adjetivo ou advérbio, e o outro é uma palavra de contexto.

   Duas palavras consecutivas são extraidas se suas *tags* POS obedecerem a qualquer um dos padrões da tabela 3.2. Por exemplo, o padrão da linha 2 significa que duas palavras consecutivas são extraídas se a primeira palavra for um advérbio e a segunda palavra é um adjetivo, mas a terceira palavra (que não é extraída) não pode ser um nome. NNP e NNPS são evitadas evitadas, em outras palavras, o nome das entidades não pode influenciar a classificação.

2. Passo 2: o algoritmo extrai a orientação semântica das frases extraídas usando a medida chamada de *Pointwise mutual information* (PMI):

$$PMI(termo_1, termo_2) = \log_2(\frac{Pr(termo_1 \wedge termo_2)}{Pr(termo_1)Pr(termo_2)}) \tag{3.26}$$

   aqui, $Pr(termo_1 \wedge termo_2)$ é a probabilidade de co-ocorrência de $termo_1$ e $termo_2$, e $Pr(termo_1)Pr(termo_2)$ nos informa a probabilidade de os dois termos acontecerem ao mesmo tempo presumindo que são estatisticamente independentes. A razão entre essas expressões dadas é então uma medida de dependênia estatística entre o numerados e o denomidador da razão. O logaritmo dessa razão é a quantidade de informação que é adquirida sobre a presença de uma das palavras quando se observa a outra.

   A orientação semântica (OS) da frase é computada baseada na sua associação com a palavra de referência positiva "excelente"e sua assosciação com a palavra de referência negativa "pobre":

$$OS(\text{frase}) = \text{PMI(frase, "excelente")} - \text{PMI(frase, "pobre")} \tag{3.27}$$

   as probabilidades são calculadas através da execução de *queries* em uma base de dados determinada e coletando o número de resultados. Na descrição de Liu[17], este afirma que neste experimento foi utilizado o mecanismo de busca do Alta Vista, pois ele possuía o operador NEAR, que limitava a busca para documentos que contivessem palavras em um grupo de palavras e que estivessem a uma distância de até dez palavras uma da outra. Para cada consulta de pesquisa, a máquina de

Tabela 3.2: Padrões de extração de frases

Primeira Palavra	Segunda Palavra	Terceira Palavra
JJ	NN ou NNS	qualquer coisa
RB, RBR ou RBS	JJ	não-NN e não-NNS
JJ	JJ	não-NN e não-NNS
NN ou NNS	JJ	não-NN e não-NNS
RB, RBR ou RBS	VB, VBD, VBNou VBG	qualquer coisa

busca geralmente fornece o número de documentos relevantes para a consulta, que é o número de resultados. Então, através da pesquisa de dois termos em conjunto e separadamente, nós podemos estimar as probabilidades na equação (3.26).

3. Passo 3: dada uma análise de filme, o algoritmo computa a média de OS de todas as frases da análise e classifica a análise como uma recomendação caso a média de OS é positiva, e como não recomendada caso contrário.

Vale a ressalva de que a técnica descrita aqui não pode ser mais utilizada pois atualmente o mecanismo de busca do Alta Vista não possui mais o operador NEAR.

3.3 Anotação

O aprendizado supervisionado conta com dados de treino rotulados para realizar suas inferências de classificação. Para alguns domínios estão disponíveis documentos rotulados com sentimento, como os produzidos por Pang e Lee[29] sobre críticas de filmes, e Dave e colaboradores [8]. Por outro lado, para domínios que não possuem tais documentos rotulados, faz-se necessária a presença de anotadores para prover julgamentos sobre os sentimentos presentes em documentos, frases ou sentenças [26]. Estes textos, anotadores com base em sentimentos, podem ser utilizados para treinar um algoritmo classificador.

Neil e colaboradores [26] afirma que anotar sentimentos pode ser uma atividade difícil, posto que a interpretação do sentimento está sujeita a diversos fatores humanos, como a experiência dentro de domínio, estados privados do anotador e as inferências que este faz sobre o texto que será anotado.

Wiebe e colaboradores[37] investiga em seu trabalho o uso de opiniões e emoções na linguagem através do estudo de um *corpus* de anotação. Puderam identificar dentro do *corpus* as principais palavras que identificam frases subjetivas e as principais que identificam frases objetivas. Encontraram também que há um percentual das sentenças que carregam uma mistura de subjetividade e objetividade. E por último, descobriram que, dentro do *corpus* utilizado as sentenças negativas tendiam a ter uma intensidade de sentimento maior e que as sentenças com maior intensidade, tendem a ser mais claras na classificação quanto ao sentimento.

Wilson e colaboradores [41] testam a hipótese de que anotadores podem ser treinados para que consigam anotar de forma confiável expressões de estados privados e seus atributos. Para comprovar a hipótese, a pesquisadora treinou os anotadores e posteriormente

comparou os resultados de suas anotações a fim de compreender o grau de concordância entre estas. Estas duas atividades serão descritas nos próximos parágrafos.

O treinamento dos anotadores começa com a leitura do manual de codificação [36]. Posteriormente, o treinamento segue em dois estágios. No primeiro estágio, o anotador pratica aplicando o esquema de anotação (apresentado por Wiebe [36]) a seis documentos de treino, utilizando papel e caneta para marcar os atributos dos estados privados. O segundo estágio consiste no aprendizado, por parte do anotador, de como aplicar o esquema de anotação dentro da ferramenta GATE.

Para analisar o grau de concordância, a pesquisadora avaliou três aspectos principais das anotações:

1. Identificação de âncoras de texto - elementos dentro do texto que revelam características sobre o mesmo - em um conjunto de anotação de textos subjetivos e objetivos.
2. Distinção entre anotações de textos subjetivos e objetivos.
3. Julgamento de qual a intensidade de atributos de estados privados e expressões que os representam.

Como resultados, Wilson e colaboradores [41] encontraram que para o tópico 1 descrito acima, houve uma concordância média de 81% entre os anotadores. Para o tópico 2, houve 89% de concordância entre os anotadores e o tópico 3 houve 53% de concordância.

Capítulo 4

O projeto

A partir deste capítulo o leitor encontrará informações referentes às atividades cumpridas para compreensão do estudo científico que permeia este projeto.

Como citado no capítulo 1, o objetivo geral deste trabalho é comprovar que a análise de sentimentos possui resultados mais precisos quando aplicada a determinado contexto através da utilização de um algoritmo de aprendizado supervisionado.

Este capítulo foi dividido em duas seções, a saber: concepção e desenvolvimento.

4.1 Concepção

A principal ideia motriz deste projeto é a construção de uma plataforma, o Analisador de Sentimentos, capaz de fazer inferências a respeito da polaridade de frases. A polaridade está distribuída em três categorias: positivo, negativo e neutro. Ela também apresenta um módulo de anotação e retreino do classificador.

A plataforma possui dois processos dominantes: o de classificação inicial de dados e o de anotação e retreino. O primeiro seguiu uma adaptação do modelo descrito na seção 3.1. A etapa de filtro de subjetividade foi retirada do processo utilizado nesse projeto, pois neste projeto é considerado que as frases objetivas pertencem a categoria "neutro" de polaridade. A figura 4.1 mostra o seu funcionamento. O segundo também possui uma estrutura linear e trabalha sobre a mesma base de dados armazenada no primeiro processo. A figura 4.2 mostra o seu funcionamento.

As subseções 4.1.1 e 4.1.2 explicam o funcionamento de cada um destes elementos do Analisador de sentimentos.

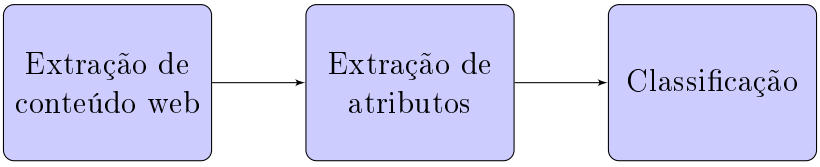

Figura 4.1: Modelo esquemático do primeiro processo da plataforma Analisador de Sentimentos

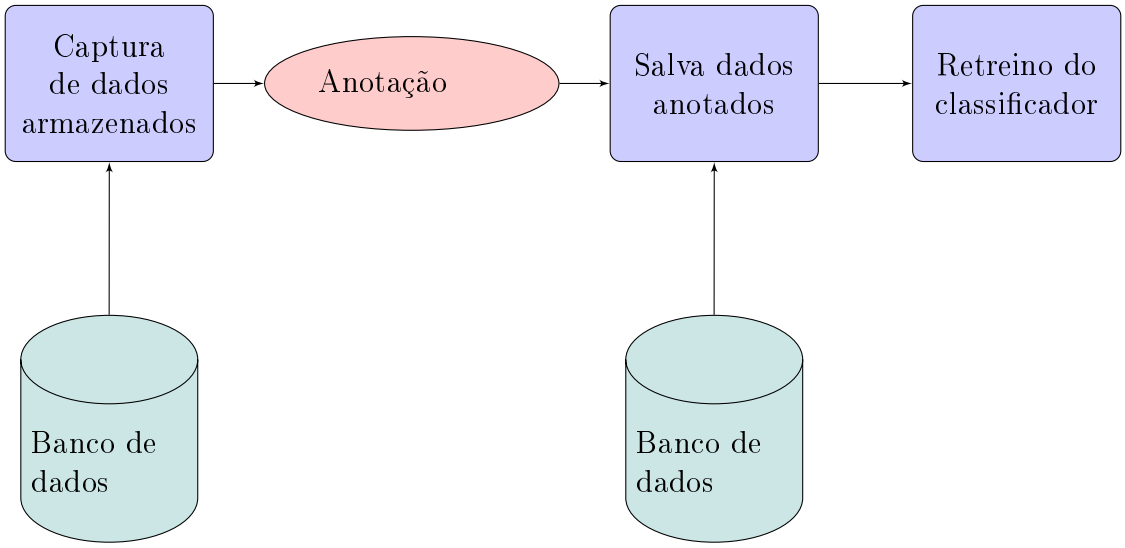

Figura 4.2: Modelo esquemático do segundo processo da plataforma Analisador de Sentimentos

4.1.1 Processo de classificação

A primeira fase, de extração de conteúdos da web é composta por dois elementos principais: um *web crawler* e um *web scraper*. Cada um destes será explorado nos parágrafos a seguir.

O *web crawler* desenvolvido trabalha de forma diferente do mostrado na figura 3.6 na medida em que não possui uma característica recursiva, ou seja, ele recebe como entrada a URL raiz e a quantidade de páginas que se deseja que o *crawler* percorra. A figura 4.3 apresenta seu modelo de funcionamento.

O *web scraper* extrai o conteúdo dos textos que existem nas URL's armazenadas pelo *crawler*. É importante frisar que este componente executa filtros de limpeza de texto, nos quais são removidas imagens, vídeos, códigos javascript, propaganda inseridas dentro do texto. A figura 4.4 apresenta o modelo de funcionamento do *web scraper*.

Vale a ressalva também de que estes componentes foram desenvolvidos para atuar de forma especifica, levando em consideração o *layout* do blog utilizado no experimento.

Para compreensão da segunda fase, é interessante que seja feita uma breve análise de algumas frases que compõem a base de dados montada, cuja temática principal é *startups*, a fim de identificar quais os elementos destas impactam na polaridade das frases. Utilizaremos os exemplos abaixo para esta atividade.

1. *"Skit is an iOS app that allows you to import images [...] on the Internet and string them together into fun little animated cut-out movies [...]"*
2. *"We're thrilled to have him on the team."*
3. *"But the data behind an experiment can be so messy."*
4. *"In a statement, Tom Impallomeni, CEO of Swapit, adds: "With SuperAwesome we've created the biggest kids and teens discovery platform in the UK which is safe, compliant and effective".*

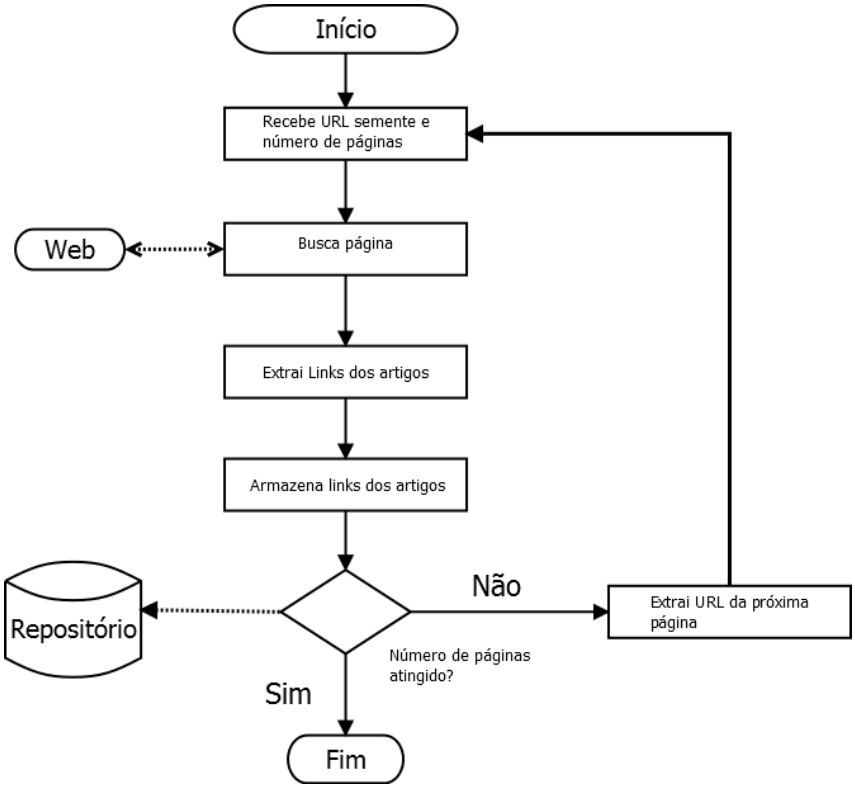

Figura 4.3: Modelo esquemático do *crawler* do Analisador de Sentimentos

5. *"Until we can zap Bitcoins to the cashier at Arby's, we're not really living in the future"*

6. *"And yes, the writing on English-language blogs can be pretty rough, too"*

A tabela 4.1 apresenta os elementos das frases sob o ponto de vista de alguns dos conceitos descritos na seção 2.3.

A partir dos dados da tabela 4.1 , é interessante notar que as palavras-chave que são capazes de identificar a polaridade das frases são adjetivos, advérbios e algumas preposições. Outro fator significativo é o fato de que estas palavras possuem tamanho maior ou igual a três. As entidades são representadas por um substantivo ou um pronome.

Outro fator a ser analisado é o fato de que artigos como *an, a, the* e preposições como *of, to, in, on* não contribuem para a determinação da polaridade, trabalhando como conectores dentro de textos.

Esta breve análise apresentada acima faz parte de um estudo feito com cinco textos da base de dados a fim de identificar a melhor forma de representar os atributos. Estes textos contém ao todo 108 frases e estas possuem ao todo 2662 palavras. Destas palavras, 460 possuem tamanho menor que três. Este último grupo é formado por siglas, artigos, preposições, conjunções e alguns símbolos textuais, como parênteses. Estes elementos, assim como descrito acima não contribuem para a determinação da polaridade das frases. No conjunto de 2202 palavras restantes há palavras que impactam e que não impactam na polaridade das frases.

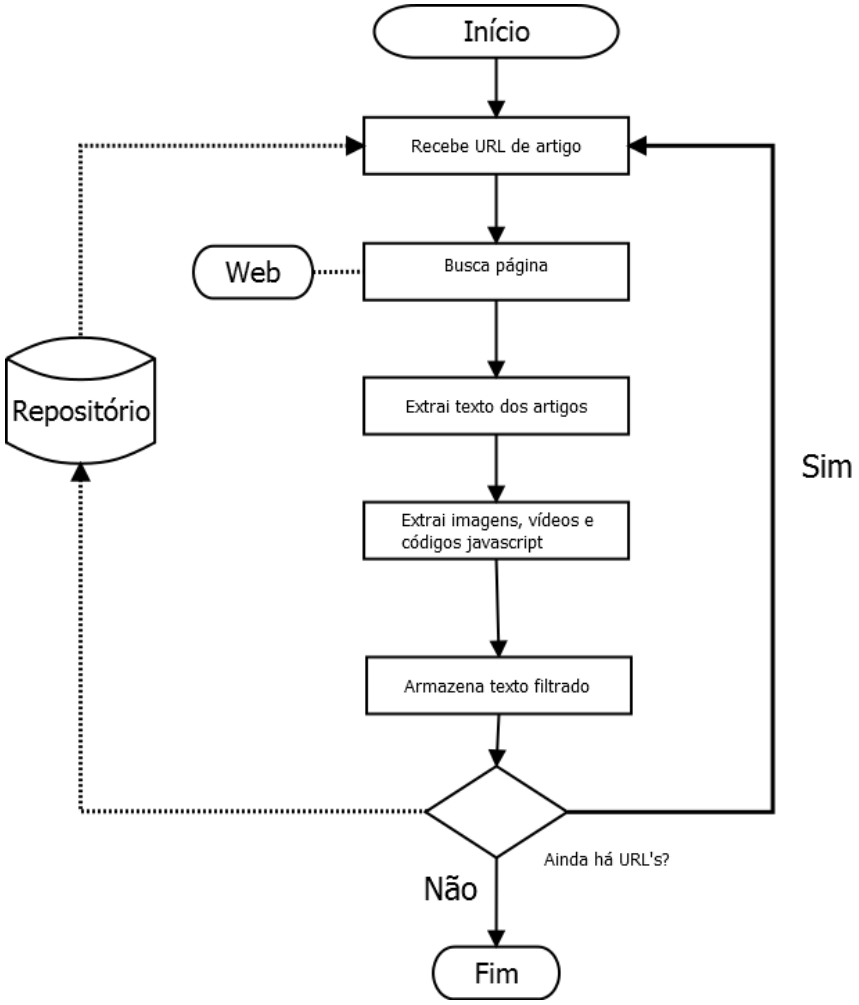

Figura 4.4: Modelo esquemático do *scraper* do Analisador de Sentimentos

Tabela 4.1: Estudo das frases do blog *Techcrunch*

Entidade	**Dono da opinião**	**Polaridade**	**Palavras-chave**
`Skit`	autor do texto	positiva	*fun, little, animated, allows*
`Pronome "him"`	pronome "we"	positiva	*thrilled*
`data`	autor do texto	negativa	*but, messy*
`SuperAwesome`	Tom Impallomeni	positiva	*biggest, safe, compliant, effective*
`Bitcoins`	autor do texto	negativa	*until, not, really*
`writing`	autor do texto	negativa	*pretty, rough*

Dado tal fato, é plausível dentro do contexto deste trabalho definir como atributos n-gramas compostos por palavras únicas (chamados unigramas) e que possuem tamanho (quantidade de letras) maior ou igual a três. Esta abordagem é explicada por Laurent Luce [1] em seu site. Dessa forma, a estrutura geral do conjunto de atributos é:

$$F = (n - grama_1, n - grama_2, n - grama_3, ...) \tag{4.1}$$

Para a frase "*We're thrilled to have him on the team.*" O conjunto F de atributos é:

$$F = ([We're], [thrilled], [have], [him], [the], [team]) \tag{4.2}$$

A limitação desta abordagem se encontra no fato de que ela não realça os elementos que efetivamente determinam a polaridade das frases, ela apenas elimina alguns dos elementos que não contribuem. Tal fato pode impactar nos resultados do classificador.

A última fase, de classificação, utiliza o algoritmo de Naive-Bayes para definir a polaridade das frases. Este algoritmo, aplicado a análise de sentimentos pode ser implementado conforme descrito na documentação da biblioteca de processamento de linguagem natural da linguagem de programação Python – NLTK [2]:

Seja C o conjunto as categorias de classificação (positivo, negativo e neutro) e $C_n, 0 < n <= 3$ cada um dos elementos do conjunto.

1. Utilize a regra de Bayes para expressar $P(C_n|atributos)$ em termos de $P(C_n)$ e $P(atributos|C_n)$.

$$P(C_n|atributos) = \frac{[P(C_n) \times P(atributos|C_n)]}{P(atributos)} \tag{4.3}$$

2. Assuma que todos os atributos são independentes, dada uma categoria C_n.

$$P(C_n|atributos) = \frac{P(C_n) \times P(atributo_1|C_n) \times ... \times P(atributo_n|C_n)}{P(atributos)} \tag{4.4}$$

3. Substitua P(atributos) pelo cálculo do denominador para cada categoria e então some estes valores

$$P(C_n|atributos) = \frac{P(C_n) \times P(atributo_1|C_n) \times ... \times P(atributo_n|C_n}{\sum_C (P(C) \times P(atributo_1|C) \times ... \times P(atributo_n|C)} \tag{4.5}$$

Conforme citado anteriormente, a polaridade das frases possui três categorias. Em especial a polaridade neutra é definida para identificar três tipos de frases:

1. Frases subjetivas com polaridade neutra
2. Frases objetivas

[1] http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/
[2] http://nltk.org/

3. Resíduos textuais que possam não ter sido eliminados no filtro realizado na fase de extração de conteúdo web.

Percebe-se então que essa polaridade é capaz de trabalhar sobre as limitações dos filtros das fases anteriores.

É importante frisar que o classificador neste processo se encontra previamente treinado. Para realizar essa atividade, o classificador precisa receber como entrada uma base de dados de frases polarizadas. Em especial, a base de dados selecionada foi a da Universidade de Cornell, chamada *"sentence polarity dataset"*, que contém frases com polaridades positiva e negativa. Esta é a base padrão deste e não está contextualizada com a temática de startups, portanto, é genérica. Dados esses fatos, é correto afirmar que o classificador gerado para esta fase é genérico. Quando aplicado às frases armazenadas previamente na etapa de coleta de dados, produz uma resposta que consiste na categoria a qual esta pertence, à saber: positiva ou negativa.

4.1.2 Processo de anotação e retreino

O objetivo deste processo é criar uma base de dados de treino nova para gerar classificadores, que também utilizam o algoritmo de *Naive-Bayes* e que são adequados ao contexto da base de dados, cuja temática principal é *startups*. Sua primeira etapa consiste apenas em recuperar os textos armazenados na base de dados. Tais textos são apresentados a usuários anotadores, que podem inserir novas classificações para cada uma das frases dos textos. Esses dados modificados são novamente armazenados na base de dados.

A próxima etapa de processamento consiste na análise de congruência entre anotações para geração dos arquivos de treino. De forma distinta da proposta apresentada por Wilson e colaboradores [41], que buscaram treinar anotadores a fim de garantir a padronização e assertividade entre os mesmos (conforme descrito na seção 3.3), nesta etapa o objetivo é atingir essa assertividade entre anotadores não treinados previamente. Para concretizar essa meta, a proposta é que um mesmo texto seja anotado por um númreo mínimo de pessoas. A congruência entre as anotações destas pessoas garante a padronização e a assertividade desejados. As frases que não possuem congruência nas avaliações dos anotadores são consideradas ambíguas em relação a polaridade. O número mínimo de anotadores é determinado a partir da quantidade de anotadores utilizada no experimento, como será explicado na seção 5.1. É interessante ressaltar também que a abordagem proposta evita que os arquivos de retreino montados estejam condicionados a opinião de apenas uma pessoa. Tal fato pode gerar um classificador enviesado ao final deste processo. Seguindo esta regra, os novos arquivos de treino são criados. Este novo conjunto de arquivos contém frases com polaridade positiva, negativa e neutra e estas frases estão contextualizadas à temática de *startups*.

Por fim, o último passo consiste em utilizar o conjunto de arquivos de treino gerado na etapa anterior para criar um novo classificador. Vale a ressalva de que para cada conjunto de arquivos criado, um novo classificador é gerado.

4.2 Desenvolvimento

O Analisador de Sentimentos, que é base para a execução de todo o experimento deste trabalho, foi construída utilizando a linguagem de programação Python [3], criada para agilizar o desenvolvimento de aplicações e de rápida curva de aprendizado. Por se tratar de um sistema web, utilizou-se o framework Django [4] desta mesma linguagem para aumentar a praticidade da construção da plataforma. O armazenamento de dados é feito com o auxílio do Sistema de Gerencimento de Banco de Dados (SGBD) MySQL [5], por seu excelente desempenho e compatibilidade com a linguagem Python.

Houve a preocupação também de se desenvolver uma plataforma de fácil manuseio por parte do usuário. Esse requisito foi atingido através do emprego da biblioteca Twitter Bootstrap [6], que contém estruturas de layout de páginas web, folhas de estilo (CSS) e códigos javascript pré-definidos e que tornam as aplicações web mais amigáveis aos usuários. O layout das páginas da plataforma foram derivados dos exemplos presentes na documentação desta biblioteca.

No que tange aos algoritmos utilizados para a realização da análise de sentimentos dentro da plataforma, empregou-se a biblioteca de processamento de linguagem natural do Python, NLTK (*Natural Language ToolKit*), que contém funções úteis para "tokenização"de frases, classificação de textos, rotulação (*tagging*), entre outras funcionalidades. Esta biblioteca é código aberto e não paga.

O módulo de *crawler* e *web scraper* foi construído com o auxílio da biblioteca *Beautiful Soup* [7] do Python, que consiste em um *parser* de HTML e XML. Ela provê métodos simples de navegar, pesquisar e modificar a árvore de análise dos sites, o que otimizou o tempo de implementação do módulo.

Ademais, a tabela 4.2 apresenta as principais funcionalidades do Analisador de Sentimentos:

As figuras 4.5, 4.6, 4.6, 4.7 , 4.8, 4.9 e 4.10 apresentam as telas da plataforma Analisador de Sentimentos referentes às funcionalidades descritas acima.

[3] http://www.python.org/
[4] https://www.djangoproject.com/
[5] http://www.mysql.com/
[6] http://getbootstrap.com/
[7] http://www.crummy.com/software/BeautifulSoup/

Tabela 4.2: Funcionalidades da plataforma Analisador de Sentimentos

Funcionalidade	**Descrição**
Crawler	1. Capturar as páginas web do blog *Techcrunch* referentes a seção de startups 2. Armazenar na base de dados informações relevantes sobre tais páginas no banco de dados
Scraper	1. Capturar, dentro das páginas web da seção de startups, apenas o texto referente ao artigo de uma página. 2. Dividir o texto automaticamente em frases. 3. Armazenar as frases dos textos capturados no banco de dados
Classificar	Classifica as frases armazenadas no banco de dados em positivas, negativas e neutras.
Anotar	Permite ao usuário classificar as frases dos textos armazenados no banco de dados como positivas, negativas e neutras
Criar arquivos de treino	Cria arquivo de treino baseados nos dados fornecidos pelas anotações dos usuários
Criar classificador	A partir de dados de treino, criar um classificador que utiliza o algoritmo Naive-Bayes
Comparação entre classificadores e anotadores	Interface com o usuário capaz de mostrar as semelhanças e diferenças entre a classificação das frases entre anotadores e os classificadores criados na plataforma
Associação	Permite ligar os anotadores cadastrados no sistema com os textos que irão anotar
Tutorial	Apresenta uma sequência de passos que ensina ao usuário os principais conceitos relacionados a temática de análise de sentimentos e como realizar o processo de anotação dentro da plataforma

Figura 4.5: Tela que permite ao usuário executar o *crawler* e o *web scraper*

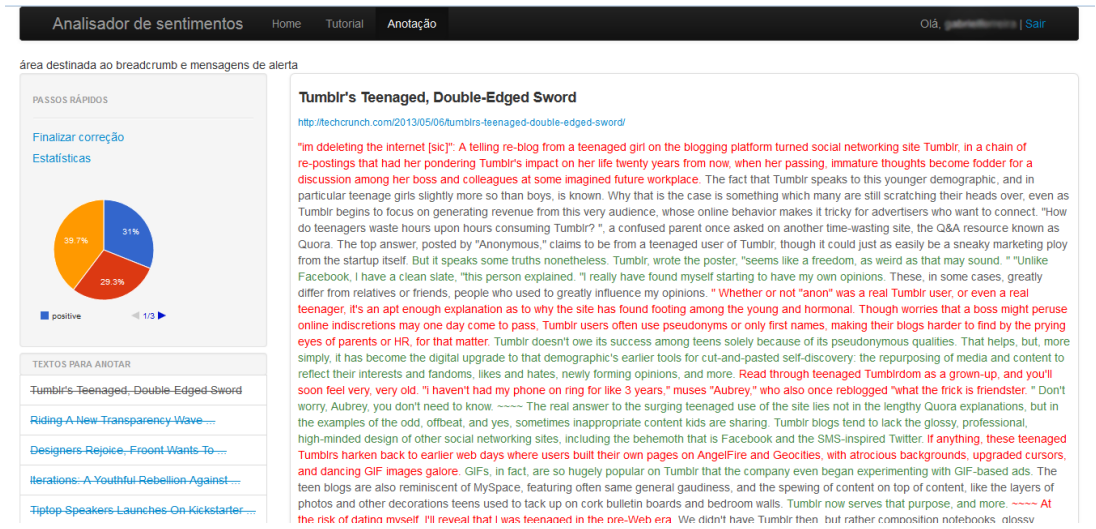

Figura 4.6: Tela de anotação de um texto após a finalização da atividade do anotador

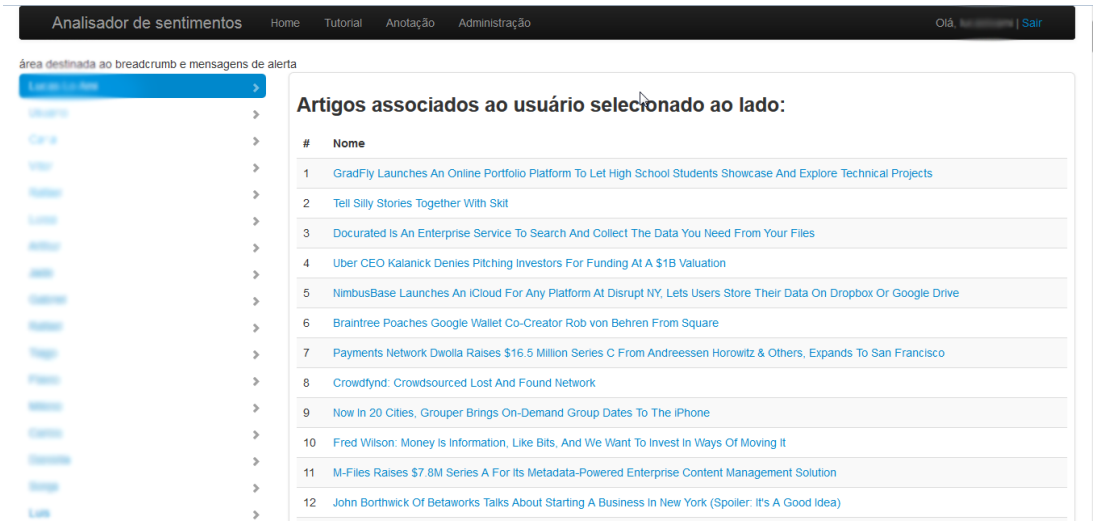

Figura 4.7: Tela de associação entre usuários anotadores e os textos que irão anotar

Figura 4.8: Tela de tutorial, passo 1: explicação de conceitos

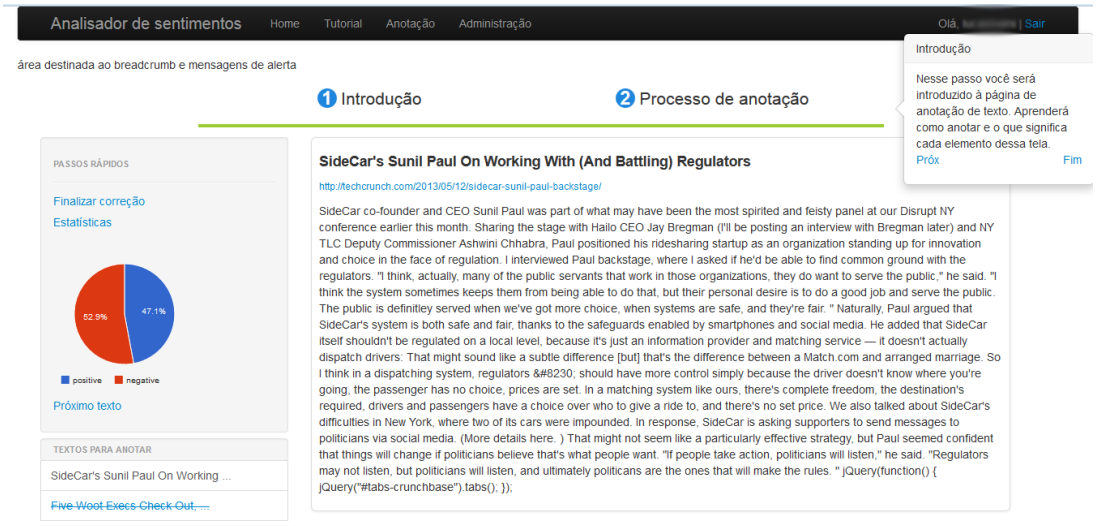

Figura 4.9: Tela de tutorial, passo 2: explicação de conceitos

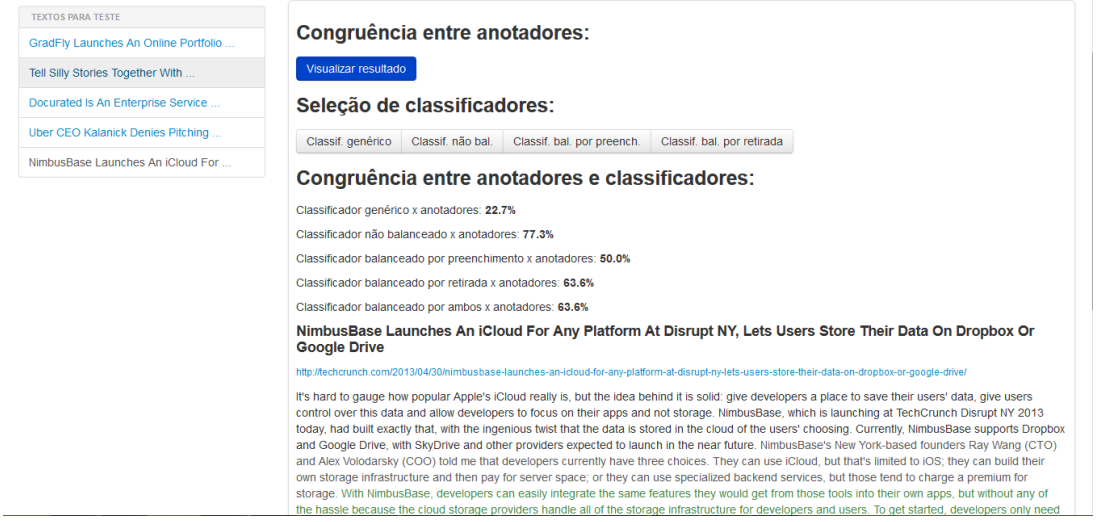

Figura 4.10: Tela de comparação entre classificadores e anotadores

Capítulo 5

Estudo de caso

Após a construção, junto ao leitor, do conhecimento relacionado ao projeto aqui desenvolvido, este é o momento para testar a precisão dos resultados obtidos a partir da aplicação dos processos descritos no capítulo 4 dentro de um experimento. Neste capítulo serão apresentados o experimento que foi realizado, os dados 5.1 coletados e uma reflexão sobre as informações obtidas 5.2.

5.1 O experimento

A coleta inicial de dados para alimentação da plataforma Analisador de Sentimentos foi realizada através da execução do *crawler* e do *web scraper* desenvolvidos. Ao todo foram coletadas 180 páginas do blog, e que fazem parte da temática de *startups*. Os textos destas páginas produziram 4563 frases.

Cada uma destas frases extraídas dos textos foi classificada pelo classificador genérico, como denominaremos o classificador criado a partir dos arquivos de treino disponibilizados por Pang e Lee - dados da Universidade de Cornell - e que será utilizado como base de comparação para os resultados dos novos classificadores. Lembre-se que o classificador genérico é capaz de classificar frases apenas duas categorias: positiva e negativa. A figura 5.1 apresenta o percentual de inferências do classificador genérico para cada uma das categorias em relação às 4563 frases. A figura 5.2 apresenta um exemplo de como o sistema apresenta estas frases classificadas previamente para o usuário.

Posteriormente foram selecionados anotadores para avaliar as frases dos textos capturados na coleta inicial de dados. Como a temática dos textos diz respeito a empreendedorismo e administração de negócios, buscou-se selecionar anotadores que possuíssem conhecimentos nessas áreas. De forma geral, o perfil buscado atende aos seguintes requisitos:

1. ser brasileiro
2. dominar a leitura na língua inglesa
3. ter experiência de pelo menos dois anos com empreendedorismo ou administração de negócios.

Ao todo foram convidados 12 anotadores que completavam o perfil desejado. Estes foram divididos em dois grupos, chamados "*Startup* 1" e "*Startup* 2", cada um com 6 inte-

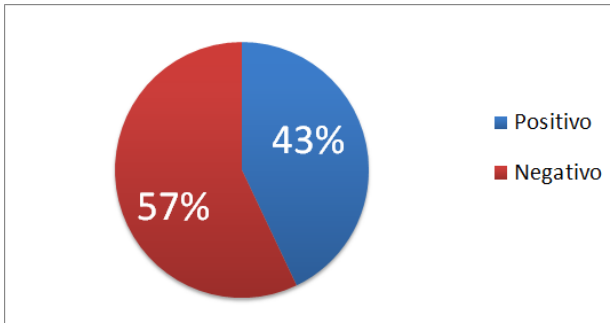

Figura 5.1: Gráfico com percentual entre categorias positiva e negativa após a classificação das frases com o classificaador genérico

Figura 5.2: Apresentação do sistema de frases classificadas como positivas e negativas. Frases em verde são positivas e frases em vermelho negativas

grantes. Dos 12 anotadores convidados, apenas 9 participaram efetivamente do processo de anotação. Portanto, a nível de experimento, o grupo "*Startup* 1" continha 6 membros e o grupo "*Startup* 2" continha 3 membros.

Vale a ressalva de que os anotadores não receberam treinamento prévio sobre como realizar suas tarefas dentro da plataforma Analisador de Sentimentos. Para suprir este fator, foi recomendado a todos os anotadores que lessem o tutorial presente na plataforma.

Para o grupo "*Startup* 1" foram selecionados 10 textos para anotação. A tabela 5.1 apresenta os dados de congruência entre os resultados das anotações, ou seja, as frases que foram classificadas com a mesma polaridade por mais de um anotador. Cada linha da tabela apresenta a quantidade de frases que foram avaliadas com a mesma polaridade por um certo número de anotadores.

Já para o grupo "*Startup* 2" foram selecionados 5 textos para anotação. Estes são diferentes dos textos selecionados para o outro grupo. A tabela 5.2 apresenta os dados de congruência entre os resultados das anotações do grupo "*Startup* 2". Cada linha da tabela apresenta a quantidade de frases que foram avaliadas com a mesma polaridade por um certo número de anotadores.

É interessante citar também que os dados fornecidos pelos anotadores são considerados *ground truth* neste experimento. Ou seja, serão tomados como verdadeiros os dados fornecidos por eles.

Os 15 textos selecionados para os dois grupos de anotadores possuíam 439 frases nos mesmos. O grupo "*Startup* 1"anotou 309 frases e o grupo "*Startup* 2"anotou 130 frases. Estes dados, acrescidos da quantidade de membros em cada grupo de anotação, são significativos para trabalharmos a concordância entre anotadores dentro do espaço

Tabela 5.1: Congruência de anotação no grupo *Startup* 1

Nº de anotadores	**Número de frases para as quais os anotadores concordaram na avaliação**
6	86
5	72
4	90
3	60
2	1

Tabela 5.2: Congruência de anotação no grupo *Startup* 2

Nº de anotadores	**Número de frases para as quais os anotadores concordaram na avaliação**
3	13
2	83

amostral aqui trabalhado. Comparado ao trabalho de Wilson e colaboradores [41], que utilizaram em seu experimento três anotadores diferentes e que anotaram 13 documentos com um total de 210 frases, a quantidade de dados produzida neste trabalho foi maior, o que indica que a amostra utilizada é suficiente para avaliar as etapas seguintes deste trabalho.

Posto que as anotações foram realizadas, o passo seguinte consistiu em construir os arquivos de treino para o gerar os novos classificadores. Para esta atividade, utilizou-se as frases produzidas pelas anotações do grupo "*Startup* 1". Conforme citado na seção 4.1.2, a construção desses arquivos deve seguir uma regra de congruência mínima entre anotadores. O fator de congruência mínimo utilizado foi 3 anotadores. Dado isso, foi produzido o primeiro conjunto de treino, composto por:

1. Um arquivo de treino de frases positivas que continha 98 frases
2. Um arquivo de treino de frases negativas que continha 37 frases
3. Um arquivo de treino de frases neutras que continha 228 frases

Os arquivos citados acima foram utilizados para treinar o **classificador não balanceado**.

É perceptível que existe um desnivelamento entre a quantidade de frases presente em cada um dos arquivos de treino. A fim de trabalhar sobre esse fato, três novos conjuntos de arquivo de treino foram criados: o conjunto de treino com balanceamento por preenchimento, o conjunto de treino com balanceamento por retirada, conjunto de treino

Tabela 5.3: Congruência entre classificadores e anotadores para frases não ambíguas

Texto	**Classif. genérico**	**Classif. não bal.**	**Classif. bal. por preench.**	**Classif. bal. por retirada**	**Classif. bal. por ambos**
texto 1	28,6%	67,9%	32,1%	53,6%	32,1%
texto 2	50%	80%	60%	40%	40%
texto 3	19%	71,4%	19%	66,7%	38,1%
texto 4	6%	73,3%	26,7%	53,3%	26,7%
texto 5	22,7%	77,3%	50%	63,6%	68,9%
Média	**25,3%**	**74%**	**37,4%**	**55,4%**	**41,2%**

com balanceamento por ambos. O primeiro consiste em adicionar ao conjunto aleatório de frases positivas e negativas do primeiro conjunto de treino frases das respectivas categorias e que vieram da base de dados genérica de Cornell até que o número de frases de cada uma das categorias se iguale ao número de frases neutras do primeiro conjunto de treino. O segundo consiste em retirar do conjunto de frases positivas e neutras do primeiro conjunto de treino o excesso de frases para que a quantidade de frases de cada categoria se iguale a quantidade de frases negativas. O terceiro consiste em mesclar as duas outras abordagens de forma que os arquivos de treino gerados não contenham número de frases advindas da base genérica maior do que o número de frases advindas dos anotadores. A partir destes novos conjuntos foram gerados o **classificador balanceado por preenchimento**, o **classificador balanceado por retirada** e o **classificador balanceado por ambos**.

Gerados os classificadores, buscou-se testar a precisão dos mesmos. Para implementar esta atividade, foram utilizados os textos do grupo "*Startup* 2". Sobre esses textos foi aplicada a regra de congruência mínima entre anotadores também. O fator de congruência mínimo utilizado foi 2 anotadores.

A robustez dos classificadores foi definida como o grau de congruência entre as classificações feitas pelos anotadores e as classificações feitas pelos classificadores em relação às frases não ambíguas. A tabela 5.3 mostra o grau de congruência para cada um dos textos de teste.

A figura 5.3 mostra a tela do Analisador de Sentimentos através da qual é possível averiguar a robustez dos classificadores.

Usando o classificador não balanceado foi determinada a polaridade dos textos de teste como um todo e não apenas das frases. Esta polaridade é definida a partir da captura da maior frequência de polaridade presente no texto analisado. Também se analisou a tendência destes textos. Entende-se por tendência como a polaridade de segunda maior frequência dentro de um texto. A figura 5.4 apresenta o resultado da atividade descrita neste parágrafo.

Usando o classificador balanceado por retirada, fez-se o mesmo exercício descrito no parágrafo anterior. A figura 5.5 apresenta os resultados obtidos.

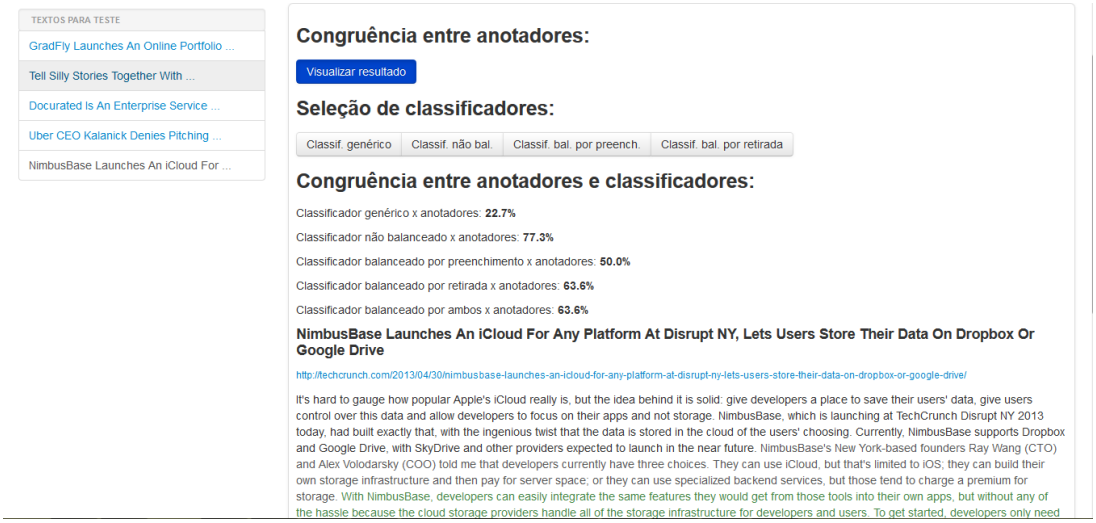

Figura 5.3: Tela de validação de classificadores do Analisador de Sentimentos

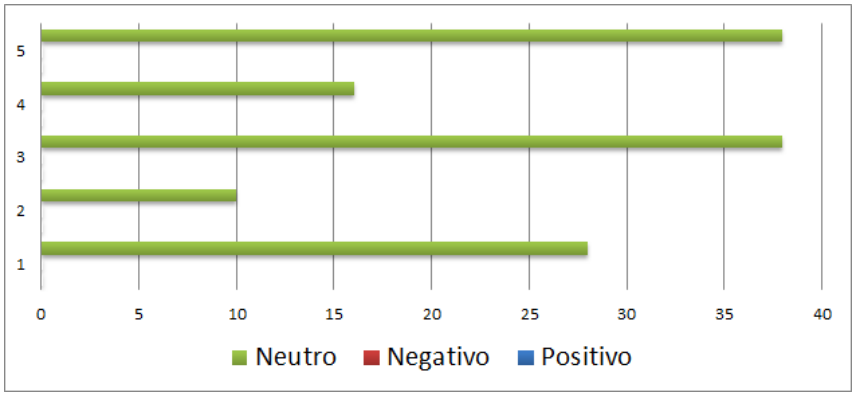

Figura 5.4: Gráfico com polaridade e tendência dos textos classificados com o classificador não balanceado

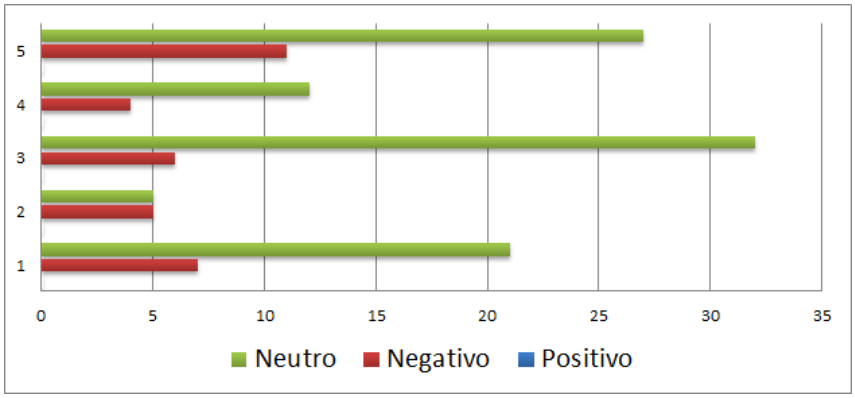

Figura 5.5: Gráfico com polaridade e tendência dos textos classificados com o classificador balanceado por retirada

5.2 Análise dos resultados

Ao observar os dados obtidos durante o experimento, percebe-se que o classificador genérico possui uma precisão muito baixa diante dos textos de teste. Tal fato confirma o problema exposto no início deste trabalho. Por outro lado, os três novos classificadores gerados se mostraram mais vigorosos diante dos mesmos textos, vide tabela 5.3.

O classificador não balanceado apresentou os melhores resultados. É importante ressaltar que esse classificador tem tendência a classificar as frases como neutras, tendo em vista o número exacerbadamente maior de frases neutras utilizadas para o treino em relação ao número de frases positivas e neutras. A consequência disso é o fato de que foram produzidos mais atributos que indicam neutralidade nas frases do que os demais. Isso afeta os cálculos de probabilidade do algoritmo de classificação Naive-Bayes.

Por outro lado, o classificador balanceado por preenchimento obteve os piores resultados. Apesar de ter sido criado a partir de arquivos de treino com a mesma quantidade de frases, a maioria dos atributos produzidos a partir das frases positivas e negativas advém dos arquivos de treino da base da universidade de Cornell, ou seja, não estão contextualizados a temática de *startups*. Dado este fator, é possível afirmar que quando o número de frases não contextualizadas é maior que o de frases contextualizadas, o classificador produzido recebe influência negativa.

É interessante perceber que o classificador balanceado por ambos obteve um resultado intermediário em relação aos demais balanceados. Este resultado indica que, quanto menor a interferência de atributos não contextualizados dentro do conjunto de treinamento, mais preciso se torna o classificador dentro do contexto de aplicação

Quanto a polaridade dos textos, a partir da análise dos dois melhores resultados, percebe-se que, de forma geral, ambos os classificadores definiram que os textos possuem a mesma polaridade: neutra. O classificador não balanceado identificou que não havia tendência dentro dos textos de teste. Isso implica que os autores dos textos são totalmente imparciais. Por outro lado, o classificador balanceado por retirada identificou uma tendência negativa naqueles. Tal fato implica que os mesmos autores possuem um leve pessimismo em relação às temáticas sobre as quais escrevem.

Como o novo classificador genérico é o melhor, entende-se que sua inferência em relação à polaridade dos textos é a verdadeira.

Capítulo 6

Conclusão

Os resultados experimentais obtidos neste trabalho validam a hipótese de que classificadores de análise de sentimentos podem se tornar mais precisos a partir do treinamento destes com arquivos de treino contextualizados. Todos os novos classificadores gerados apresentaram resultados melhores que o classificador genérico utilizado como base de comparação. É interessante frisar também que este trabalho expôs que existe uma relação direta entre a robustez dos classificadores e a quantidade de frases ão contextualizadas utilizadas para treiná-los.

Ademais, a plataforma Analisador de Sentimentos é uma importante contribuição científica na medida em que pode ser utilizada por outros pesquisadores para aprofundamento de estudos na área de Análise de sentimentos. Ela consiste em uma plataforma código aberto desenvolvida com ferramentas de atuais e disponibilizada sob licença GPL (GNU *General Public License*) [1]. Ela esta disponível no repositório Bitbucket [2]. Ademais, cada etapa dos processos da plataforma pode ser utilizada em trabalhos científicos que envolvam áreas correlatas.

6.1 Principais contribuições

As principais contribuições deste trabalho foram:

1. O desenvolvimento de uma plataforma código aberto capaz de analisar sentimentos de frases e criar classificadores mais robustos.
2. Constatação da influência de atributos não contextualizados na robustez de classificadores.
3. Comprovação de que a análise de sentimentos contextualizada é mais precisa.

6.2 Trabalhos futuros

Durante a execução desse trabalho, foram identificados possíveis assuntos que podem ser aprofundados por meio de novos trabalhos:

[1] http://www.gnu.org/licenses/gpl.html
[2] https://lucasloami@bitbucket.org/lucasloami/analisador-sentimentos-publico.git

1. Como citado em capítulos anteriores, a abordagem para construção dos atributos da plataforma foi simples, posto que não levava em consideração a identificação das palavras mais importantes para a definição do sentimento em torno de uma frase. Um novo trabalho poderia investigar a melhoria dos resultados a partir da utilização de atributos mais complexo, como por exemplo, combinar n-gramas e POS *tags*.
2. É possivel repetir o experimento utilizando um número maior de anotadores e grupos com a mesma quantidade de integrantes a fim de explorar como este fator pode impactar na melhoria dos resultados.
3. Pode-se explorar a coleta de uma quantidade maior de frases anotadas para treino de novos classificadores a fim de averiguar o quanto a quantidade de frases utilizadas para treino pode impactar os resultados.

Referências

[1] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 1–4, 2007. Short paper. 18

[2] Carlos Castillo. Effective web crawling. *SIGIR Forum*, 39(1):55–56, June 2005. 15

[3] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Comput. Netw.*, 31(11-16):1623–1640, May 1999. 16

[4] Junghoo Cho, Hector Garcia-molina, and Lawrence Page. Efficient crawling through url ordering. In *Computer networks and ISDN systems*, pages 161–172, 1998. 15, 16

[5] The Nielsen Company. Global faces and networked places, a nielsen report on social networking's new global footprint. Technical report, Nielsen Company, March 2009. 1

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995. 25

[7] Sanjiv Das and Mike Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *In Asia Pacific Finance Association Annual Conf. (APFA)*, pages 1375–1388, 2001. 20

[8] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM. 17, 33

[9] Emanuel de Barros Albuquerque Ferreira. Análise de sentimento em redes sociais utilizando influência de palavras. Trabalho de Graduação - Universidade Federal de Pernambuco - UFPE. Departamento de Ciência da Computação, Dezembro 2010. 22, 23, 25

[10] Wenjing Duan, Bin Gu, and Andrew B. Whinston. Do online reviews matter? - an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016, November 2008. 1

[11] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 9–14, 2007. 22

[12] Vasileios Hatzivassiloglou and Kathleen McKeown. Predicting the semantic orientation of adjectives. In Philip R. Cohen and Wolfgang Wahlster, editors, *ACL*, pages 174–181. Morgan Kaufmann Publishers / ACL, 1997. 18

[13] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *COLING*, pages 299–305. Morgan Kaufmann, 2000. 18

[14] Soo-Min Kim and Eduard Hovy. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of IJCNLP-05, the Second International Joint Conference on Natural Language Processing*, pages 61–66, Jeju Island, KR, 2005. 17, 21

[15] Soo-Min Kim and Eduard Hovy. Crystal: Analyzing predictive opinions on the web. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1056–1064, 2007. 20

[16] Vipin Kumar. *Introdução ao data mining - mineração de Dados*. Ciência Moderna, segunda edition, 2005. vi, 11

[17] Bin Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 2008. vi, 4, 6, 7, 15, 16, 22, 23, 24, 25, 27, 29, 32

[18] Andrew McCallum and Kamal Nigam. A comparison of event models for naive Bayes text classification. In *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, volume 752, pages 41–48, 1998. 26

[19] Yelena Mejova. Sentiment analysis: An overview - comprehensive exam paper. Technical report, Department of Computer Science, University of Iowa, 2009. 20

[20] F Menczer, G Pant, and P Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology*, 4(4):378–419, 2004. 16

[21] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 169–176, New York, NY, USA, 2011. ACM. 4

[22] Subhabrata Mukherjee. Sentiment analysis - a literature survey, June 2012. Indian Institute of Technology, Bombay. Roll No: 10305061. 1, 4, 7, 17

[23] Avik Sarkar Mukund Deshpande. Bi and sentiment analysis. *Business Intelligence Journal*, 15, 2010. 5

[24] Jin-Cheon Na, Haiyang Sui, Christopher Khoo, Syin Chan, and Yunyun Zhou. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In *Conference of the International Society for Knowledge Organization (ISKO)*, pages 49–54, 2004. 20

[25] Kamal Nigam. Using maximum entropy for text classification. In *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999. 30, 31

[26] Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F. Smeaton. Topic-dependent sentiment analysis of financial blogs. *In Proc. of CIKM workshop on Topic-sentiment analysis for mass opinion (TSA '09)*, pages 09–16, November 2009. 33

[27] Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, and Ian Soboroff. Overview of the trec-2006 blog track. In *Text Retrieval Conference*, pages 1–4, 2006. 12, 21

[28] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008. 5, 14, 16, 20, 21

[29] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 12, 17, 18, 21, 23, 30, 33

[30] Eric Ries. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. Crown Business, first edition edition, 2011. 1

[31] Beatrice Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990. 18

[32] Leandro Matioli Santos. Protótipo para mineração de opiniões em redes sociais: estudo de casos selecionados usando o twitter. Trabalho de Graduação - Universidade Federal de Lavras - MG, 2010. 25

[33] Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, pages 300–307, 2007. 18

[34] Ian Soboroff and Donna Harman. Overview of the trec 2003 novelty track. In *Proceedings of TREC-2003 (2003)*, pages 38–53, 2003. 21

[35] Ana Sufian and Ranjith Anantharaman. Social media data mining and inference system based on sentiment analysis. Master's thesis, Chalmers University Of Technology - Department of Applied Information Technology, 2011. 16

[36] J. Wiebe. Instructions for annotating opinions in newspaper articles. Department of computer science technical report tr-02-101, University of Pittsburgh, 2002. 34

[37] Janyce Wiebe and Claire Cardie. Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation (formerly Computers and the Humanities)*, pages 165–210, 2005. 33

[38] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, September 2004. 17, 21

[39] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. 21

[40] Theresa Wilson. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354, 2005. 12, 20

[41] Theresa Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of private states.* PhD thesis, Intelligent Systems Program, University of Pittsburgh, 2007. 33, 34, 40, 48