

Universidade Federal de Pernambuco

Centro de Informatica

Graduação em Engenharia da Computação

**Investigação e implementação de técnicas em Análise de Sentimentos**

Hallan Cosmo dos Santos

Trabalho de Graduação

Recife, Julho de 2010

Universidade Federal de Pernambuco

Centro de Informática

**Investigação e implementação de técnicas em Análise de Sentimentos**

Hallan Cosmo dos Santos

Trabalho de Graduação

*Monografia apresentada ao Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do Grau em Engenharia da Computação.*

*Orientador: Ricardo Bastos Prudêncio*

Recife, Julho de 2010

“Nothing is as far away as one minute ago”

Jim Bishop

**AGRADECIMENTOS**

 Esta pesquisa reflete a conclusão de uma fase importante e marcante em minha vida, pois representa o resultado de anos de grande esforço, dedicação e perseverança. Marca o término de mais uma fase – graduação – e início de outra em minha caminhada acadêmica e profissional.

 Agradeço aos meus pais, que me presentearam com a educação que possuo, e ainda que eu agradecesse eternamente, jamais seria suficiente pelo apoio que me deram em todos os momentos.

 Agradeço ao professor Ricardo Prudêncio pela disponibilidade, receptividade e orientações que tornaram este trabalho possível e que, ao longo de suas disciplinas cursadas na graduação, despertou em mim o interesse na área que sigo academicamente e profissionalmente.

 Agradeço ao Centro de Informática que possibilitou minha formação em Engenharia da Computação, num dos melhores centros de referência do país.

 Um agradecimento especial a alguns amigos que me incentivaram e participaram tanto da minha vida acadêmica quanto pessoal para que eu conseguisse chegar até aqui. Obrigado, Agenor Facundes, Leandro Henrique Espíndola, Phillip César, Rafael Leão, Rodrigo Porto e todos os outros que direta e indiretamente me apoiaram durante toda esta jornada.

**RESUMO**

 Saber o que as pessoas acham sempre foi um fator importante nos processos de tomar decisão. Mesmo há alguns anos, antes da Internet se tornar tão popular, a maioria de nós já perguntava a nossos amigos em quem eles iriam votar nas eleições ou o que eles achavam de determinada marca de um produto. No entanto, hoje a Internet está tornando possível obter a opinião e a experiência de uma quantidade enorme de pessoas, profissionais ou não, que nem conhecemos em assuntos como estes. O fato é que cada vez mais pessoas estão deixando suas opiniões disponíveis na internet.

 A **análise de sentimentos** classifica uma sentença ou texto como positivo, negativo. Esta tarefa básica é um sub-problema da **mineração de opinião**, que recebeu muita atenção recentemente graças à grande quantidade de documentos online, postados na Internet por usuários comuns ou especialistas, como discussões em fóruns, *blogs* e redes sociais.

 O objetivo deste trabalho é apresentar e implementar técnicas que dão suporte à mineração de opinões e, em mais detalhes, à análise de sentimentos. Em geral, elas lidam com o tratamento de opiniões, sentimentos e subjetividade em documentos.

 Para alcançar este objetivo, serão apresentadas abordagens para a busca de opiniões online e algoritmos para sua posterior classificação. Em seguida, será apresentado o problema de análise de sentimentos propriamente dito, suas técnicas, abordagens e dificuldades envolvidas. E, finalmente, serão realizados experimentos e análises dos resultados obtidos neles.

**Palavras-chave**: Mineração de Opinião, Análise de Sentimentos, Classificação, Opinião na Internet, Processos de Tomar Decisão.

**ABSTRACT**

 Knowing what people think has always been an important factor in decision-making process. Even a few years ago, before the Internet became so popular, most of us have asked our friends about who they would vote in elections or what they thought of a particular product’s brand. However, today the Internet is making possible to obtain the opinion and experience of a huge amount of people, professional or not, who we don't even met, in matters like these. More and more people are posting their opinions on the internet.

 The **sentiment analysis** classifies a sentence or text as positive, negative or neutral. This basic task is a sub-problem of **opinion mining**, which has received much attention recently due the large amount of online documents, posted or not by common users of the Internet, like discussions on forums, blogs and social networks.

 The aim of this paper is to introduce and implement techniques that support opinion mining and, detailed, sentiment analysis. In general, they deal with the computational treatment of opinions, sentiment and subjectivity in documents.

 To achieve this goal, this research will present some approaches for the search of online reviews and algorithms for their subsequent classification. Then, the problem of sentiment analysis will be presented, its techniques, approaches and difficulties involved. And, finally, it will be performed experiments and analysis of its results.

**Keywords**: Opinion Mining, Sentiment Analysis, Classification, Reviews in the Internet, Decision-making Process

**SUMÁRIO**

1. **Introdução** ................................................................................................. 101.1 Mineração de opinião ...................................................................... 10
1.2 Análise de Sentimentos .................................................................. 12
1.3 Aplicações ....................................................................................... 14
2. **Busca por opiniões** ................................................................................... 17
3. **Classificação** ............................................................................................. 203.1 Principais definições para classificação de sentimentos .................. 21

3.2Classificação utilizando *frases com sentimentos* .............................. 233.3 Classificação utilizando os métodos de *classificação de texto* ......... 263.4 Classificação utilizando uma função *score* ....................................... 27

1. **Experimentos** ............................................................................................. 28

4.1 Extração ............................................................................................ 28

4.2 Classificação ..................................................................................... 29

1. **Resultados obtidos** ................................................................................... 30
2. **Conclusão e trabalhos futuros** ................................................................. 32
3. **Referência Bibliográfica** ............................................................................ 34

**LISTA DE FIGURAS**

Figura 1 – Representação de uma árvore

Figura 2 – Epinions - Sistema de agregação de opiniões

Figura 3 – Exemplo de um sumário de opiniões sobre postos de combustível

Figura 4 – Classificação de sentimentos utilizando thumbs up e thumbs down

**LISTA DE TABELAS**

Tabela 1 – Definição de mineração de opinião

Tabela 2 – Definição de um objeto e seus componentes

Tabela 3 – Definição das características de um objeto

Tabela 4 – Padrão Penn Treebank POS Tags

Tabela 5 – Padrões para extração de duas palavras consecutivas de opiniões

Tabela 6 – Equação PMI para cálculo da orientação semântica de um par

Tabela 7 – Orientação semântica de uma frase

Tabela 8 – Função *score*

Tabela 9 – Função final para classificação de um documento

Tabela 10 – Função score para classificação de uma opinião

Tabela 11 – Resultados obtidos com o experimento

1. **Introdução**

 Antes de começarmos a expor nossos experimentos computacionais, julgamos relevante abordar alguns tópicos relacionados à análise de sentimentos tais como: Mineração de opinão e suas aplicações.

* 1. Mineração de opinião

 A Internet contém uma quantidade enorme de documentos não-estruturados. Analisar estes documentos (textos, de uma forma geral) é de grande importância pelo seu papel na prática. Por exemplo, o mercado deseja saber a opinião dos seus consumidores a respeito de um determinado produto ou serviço. E também, usuários querem saber a opinião de alguns outros consumidores antes de usar ou comprar um produto. A tarefa de analisar estes documentos, no entanto, é desafiadora, pois se trata de processamento de linguagem natural.

 A mineração de opinião pode fornecer informação, por exemplo, sobre quais páginas podem ser mais interessantes para se colocar algum anúncio. De forma que se em uma determinada página, as pessoas expressam opiniões boas – positivas – sobre algum produto, lá pode ser um bom lugar para se fazer uma propaganda dele. Ao mesmo tempo, se as pessoas falam mal e criticam o mesmo produto, provavelmente não deve ser uma boa idéia colocar um anúncio lá. No entanto, poderia ser viável a um concorrente fazer um anúncio neste local.

 O fato é que a Internet está mudando a maneira que as pessoas expressam suas opiniões. Eles postam críticas ou sugestões de produtos em sites dos comerciantes e dão seu ponto de vista em vários fóruns, grupos de discussão e *blogs*. Ou seja, usuários comuns estão criando conteúdo na Web. E, esta é uma fonte de informações com muitas aplicações práticas. Técnicas estão sendo desenvolvidas para explorar isso e para ajudar as pessoas a obter tais informações de uma forma fácil e eficiente. Talvez a mais completa definição para mineração de opinião atualmente seja esta:

Dado um conjunto de documentos de texto D que contém opiniões (ou sentimentos) a respeito de um objeto, a mineração de opinião busca extrair características (atributos) e componentes do objeto que foram comentados em cada documento d ∈ D e determinar se os comentários são positivos, negativos ou neutros.[1]

Tabela 1 – Definição de mineração de opinião

 Basicamente, podemos listar três tarefas da mineração de opinião: a mineração em sentenças comparativas, a mineração de opinião baseada em atributos e sumarização, e por último, a classificação de sentimentos.

1. **Mineração em sentenças comparativas**

 Dentre os vários tipos de avaliação de objetos, a comparação é uma delas. De uma forma direta, ela compara um objeto a outro ou a vários outros objetos similares. Por exemplo, “O tamanho do telefone X é muito menor que o tamanho do telefone Y”. O objetivo aqui é identificar as sentenças do tipo comparativa e extrair a relação de comparação expressas nelas.

 A comparação pode ser do tipo superlativa, que é quando um objeto tem um atributo melhor ou pior que todos os outros, ou do tipo comparativa, quando um objeto é melhor do que outro objeto em alguma coisa. Normalmente, primeiro precisamos determinar o tipo de comparação envolvida. A estrutura de uma palavra comparativa normalmente vem de algum adjetivo ou advérbio, acrescido de modificadores como “o mais“, “o menos”, “mais...que”, “menos...que”, “tão...quanto”, etc. Portanto, a classificação do tipo de comparação pode se dar, determinando se há uma relação de igualdade ou desigualdade entre os objetos, diferentemente de uma relação superlativa.

1. **Mineração de opinião baseada em atributos e sumarização**

 Este tipo de mineração vai ao nível das sentenças dos documentos para descobrir sobre o que as pessoas estão expressão suas opiniões. Pode ser um produto, um serviço, uma pessoa ou empresa. Numa opinião sobre um produto, por exemplo, a tarefa é identificar quais atributos dos produtos estão sendo comentados e determinar se eles suas opiniões são positivas, negativas ou neutras. Assim, a sentença: “O tempo de vida da bateria deste celular é muito curto.” seria um comentário negativo a respeito da bateria do telefone.

 A sumarização trata de como será apresentado um relatório com os resultados produzidos na mineração.

1. **Classificação de sentimentos**

 Este tipo reduz a mineração de opinião em um problema de classificação de textos. Ou seja, dado um documento de texto, ele vai classificá-lo como positivo ou negativo. Assim, o sistema é capaz de determinar se a opinião a respeito de um determinado produto é positiva ou negativa. No entanto, a classificação fica a nível do documento como um todo, sem determinar o que exatamente as pessoas gostam ou não gostam sobre o objeto das sentenças.

 Este é o tipo de mineração o qual este trabalho se propõe investigar em mais detalhes. Portanto, quando não estiver indicado o tipo de mineração utilizado em alguma técnica apresentada, fica implicito que ela se refere à classificação de sentimentos, exceto quando expresso o contrário.

* 1. Análise de Sentimentos

 A análise de sentimentos é o sub-problema da mineração de opinião que mais tem sido estudado. Ela classifica textos e documentos como positivo, negativo e, eventualmente, neutro. O ponto desafiador é o fato de que os sentimentos podem ser expressos a respeito de qualquer coisa. Por algumas vezes já foi utilizado o termo *objeto* neste trabalho. Ele é a parte da sentença que está sendo comentada. Um objeto possui um conjunto de *componentes*, que são as partes que formam o objeto, e um conjunto de *atributos*, suas características.

 Formalmente, podemos fazer a seguinte definição:

Um *objeto* O é uma entidade que pode ser um produto, pessoa, evento, empresa ou um tópico. Está associado a um par, O: (T,A), onde T é uma hierarquia de *componentes* (ou partes), *sub-componentes*, e assim por diante, e A é o conjunto de atributos de O. E cada componente tem seu próprio conjunto de sub-componentes e atributos.[2]

Tabela 2 – Definição de um objeto e seus componentes

 De acordo com a definição acima podemos representar um objeto como uma árvore (ver figura 1). A raiz, é representada pelo objeto propriamente dito, os outros nós são os componentes do seu nó pai. O fato é que um sentimento pode ser expresso a respeito de qualquer nó ou de qualquer atributo de um nó.



Figura 1 – Representação de uma árvore

 Por exemplo, a sentença: “Eu não gosto deste aparelho celular” está se expressando a respeito do nó raiz. Mas, ele poderia estar citando um atributo do telefone, como na sentença: “A qualidade da câmera deste telefone não é boa”. Ou sobre atributos dos componentes como em: “A nitidez da câmera deste telefone é ruim”.

 Em muitas situações confudem-se os conceitos de componentes e os atributos. Por isso, na prática costuma-se chamar *características* do objeto, ambos os componentes e atributos.

Característica de um objeto representa ambos componentes e atributos. Pode ser de dois tipos: a) **explicitas** – Quando aparece na sentença. b) **implícita** – Quando não aparece, mas o texto implica nela. Como por exemplo: “Este telefone é muito pequeno”. O tamanho é uma característica implícita do objeto telefone. [1]

Tabela 3 – Definição das características de um objeto

 Uma vez definidos os elementos que fazem parte das sentenças que devem ser classificadas. Podemos investigar algumas das aplicações nesta área.

* 1. Aplicações

 Recentemente tem crescido bastante as pesquisas na área de análise de sentimentos. Podemos listar inúmeras das utilidade práticas. A seguir serão apresentadas algumas de suas aplicações.

1. **Aplicações em sites relacionados à opiniões**

 Uma ferramenta de busca pode servir muito bem para a criação e manutenção automática de páginas que agregam opiniões. É uma alternativa para sites como o *Epinions* [13] que solicitam *feedbacks* e armazenam opiniões.

 Nestes sistemas pode ser solicitado ao usuário que avalie produtos, como também pode ser armazenada a opinião sobre candidatos à eleição, por exemplo. Com um sistema que avalia a opinião do usuário, eventuais erros, da parte do usuário, no momento de atribuir um *rating* (determinar a qualidade do produto numa escala de uma a cinco estrelas) poderiam ser corrigidos, nos casos em que usuários acidentalmente selecionaram um *rating* baixo, mas expressam uma opinião positiva.

****

Figura 2 – Epinions - Sistema de agregação de opiniões

1. **Aplicações como sub-componentes de algumas tecnologias**

A análise de sentimentos pode ter um papel importante na usabilidade de algumas tecnologias. Uma possibilidade é uma melhora nos sistemas de recomendação, fazendo com que sistemas não recomendem itens com *feedbacks* negativos.

 Na deteccção de *flames* em e-mails é um outro possível uso da classificação de subjetivdade e sentimentos nos meios de comunicação.

 Em sistemas *online* que fazem propagandas é estremamente interessante detectar quando as páginas contém algum conteúdo inadequado para este tipo de propaganda. Em geral, apresentar propagandas apenas quando opiniões positivas forem emitidas das páginas ou opiniões negativas dos seus concorrentes.

1. **Aplicações em Business ou Inteligência governamental**

As técnicas em análise de sentimentos são bastante adequadas para muitas aplicações que necessitam de inteligência. Na verdade, o *business intelligence* parece ser um dos principais fatores por trás do interesse nesta área.

 Considere, por exemplo, que um grande fabricante de automóveis deseja saber o motivo de seu novo modelo não estar sendo vendido como esperado. Existem uma série de critérios objetivos que devem ser avaliados para detectar-se a razão das baixas vendas, como o preço dos concorrentes, os ruídos do motor, etc. Além disso, há critérios subjetivos que precisam ser considerados da mesma forma, como *design*.

 A análise de sentimentos, neste caso, pode ser aplicada em serviços da Internet como o *Epinions* [13] (a busca também pode se dar em *blogs*, sites de notícias,...) para encontrar opiniões. Em seguida, é produzido um relatório ao analista para que ele não tenha seu tempo tomado com a leitura de dezenas de críticas.

 No caso da gerência e relações públicas, é possível avaliar o ponto de vista do público alvo da empresa, permitindo assim, a predição de vendas ou de outros dados relevantes.

1. **Aplicações no domínio político**

 Como se sabe, opinião é um assunto de grande interesse na política. Em Goldberg [4] um trabalho de predição da afiliação política foi feito baseado em listas de discussão de caracter político. Em geral, busca-se a posição (apoio ou rejeição) das figuras públicas a respeito de propostas políticas.

 Análise de sentimentos é o princípio chave por trás do *eRulemaking*, que é o uso de tecnologia para se produzir leis. Esse processo se dá avaliando as opiniões que as pessoas submetem a respeito das propostas que o governo propõe.

1. **Busca por opiniões**

 O uso de ferramentas de busca de propósito geral já está bastante disseminado na sociedade. Hoje, não precisamos mais memorizar endereços eletrônicos, pois com os serviços de busca, podemos localizar facilmente a maioria dos endereços que desejamos. Assim como a busca de propósito geral, as pessoas podem buscar por conteúdos postados por usuários comuns. A Internet pode fornecer também um serviço de busca por opiniões sobre qualquer objeto.

 Vamos dividir estas buscas em dois tipos:

1. A busca por opiniões de um objeto específico ou de uma característica deste objeto. Como, por exemplo, a pesquisa pela qualidade da imagem de uma câmera digital ou pela opinão pública em algum tópico político. Neste caso um objeto pode ser um produto, organização ou um tópico.
2. A busca pela opinião de alguém. Alguém pode desejar saber a opinão do Ministro da Saúde em relação ao aborto, por exemplo. Este tipo de busca é relevante quando o autor das opinões é explicitamente divulgado, já que sabemos que a Internet é um espaço onde as pessoas costumam se expressar de forma anônima.

 Para o primeiro tipo de busca, o usuário vai fornecer o nome do objeto ou a característica que ele deseja buscar. E para o segundo tipo de busca, o usuário precisa informar o nome do dono da opinião que ele busca.

 O fato é que nos sistemas de busca de opinião não podemos simplesmente aplicar a busca por palavras-chave como nas ferramentas de busca de propósito geral. Para entender o que acontece, basta observar que ao buscar simplesmente por palavras-chave iremos encontrar muitos documentos que não estão expressando opinião. Isso porque muitas discussões e postagens de *blogs* não expressam opinião, às vezes apenas perguntas e respostas sobre algum tópico. Portanto, sentenças e documentos que realmente emitem uma opinião precisam ser identificados antes da busca ser realizada, e utilizar palavras-chave seria a forma mais simples e inocente de fazer isto.

 Os buscadores tradicionais, usam um sistema de *ranking* [5] baseados na relevância das páginas. Esta premissa básica diz que páginas com ranking alto, que idealmente são as primeiras páginas obtidas como resultado, fornecem informações suficientes para as necessidades do usuário. Em geral, isto satisfaz o tipo de buscas b), já que, normalmente, a opinião de um determinado autor sobre um objeto encontra-se em um único local. No entanto, para o tipo a), documentos listados por *rank* apenas apresentam representam a opinião de poucas pessoas. É preciso que eles reflitam uma distribuição normal de sentimentos positivos e negativos da população como um todo.

 Em muitos casos, os documentos que expressam opinião são grandes. Isso faz com que o usuário gaste muito tempo lendo para obter um cenário completo do que ele deseja pesquisar. Portanto é desejável que haja alguma forma de sumário que pode conter, por exemplo, uma média das avaliações positivas e negativas. Ou seja, o usuário terá em suas páginas de resultado, opiniões positivas e negativas distribuidas proporcionalmente ao conjunto inteiro de opiniões. Uma alternativa mais simples, é forncer dois *rankings*, um com as opiniões positivas e outro com opiniões negativas.

 Fornecer um sumário do objeto para cada pesquisa do usuário é uma solução ideal. Vale observar o cenário do Censo realizado no Brasil. Ao visitar um domicílio, o pesquisador realiza uma série de perguntas e preenche um grande formulário relativos aquele morador. Ao final do recenseamento, são produzidos relatórios em forma de gráficos que apresentam cada um dos pontos relevantes da pesquisa realizada. Seria raro que algum analista resolvesse avaliar cada uma das milhares de respostas para se chegar a alguma conclusão ou tomar uma decisão.

 Podemos enxergar a pesquisa por opinião de forma parecida com o Censo. Se analisarmos o formulário que o pesquisador preenche como um documento de opinião de um determinado objeto. Temos na Internet um conjunto enorme de “formulários” que precisam ser sumarizados, apresentando resultados ao usuário. Sumarizar é, portanto, um problema bastante desafiador e importante.

 As opiniões possuem, em geral, um aspecto temporal. Ou seja, com o passar do tempo é natural que as pessoas mudem sua forma de pensar em relação aos objetos. Logo, é bastante útil fornecer um sumário que mostre a mudança de sentimento das pessoas ao longo do eixo de tempo.

 Da mesma forma, um sistema de busca por opiniões podem fornecer uma busca comparativa de objetos. Por exemplo, quando alguém deseja abastecer o carro numa avenida em que há uma grande quantidade de postos de combustível. Provavelmente, é desejável saber qual o melhor local para se abastecer ou apresentar um comparativo de características entre eles como apresentado na figura 3.



Figura 3 – Exemplo de um sumário de opiniões sobre postos de combustível

1. **Classificação**

 De acordo com a tabela 1, a análise de sentimentos é um sub-problema da mineração de opinião que recebe como entrada um conjunto de documentos D, classifica cada d ∈ D em uma das classes, positivo ou negativo. Uma classificação positiva significa que d expressa uma opinião positiva. Uma classificação negativa significa que d expressa uma opinião negativa. Por exemplo, fornecendo-se algumas opiniões de um filme, o sistema classifica elas em opiniões positivas ou negativas.

 Sem dúvida, a principal aplicabilidade da classificação na análise de sentimentos é determinar a opinião que prevalece a respeito de algum objeto. A tarefa é similar àquela de classificar textos em alguns tópicos, que classifica documentos em algumas classes predefinidas, como *política*, *ciência* ou *esporte*. No entanto, na classificação de sentimentos, substantivos e verbos são pouco relevantes. O que indica uma opinião positiva ou negativa são as palavras de sentimento, como *ótimo*, *ruim*, *melhor*, *pior*, *excelente*,etc.

 Nesta seção discutiremos algumas abordagens existentes na classificação de documentos como positivos ou negativos.

* 1. Principais definições para classificação de sentimentos

 Como foi apresentado na seção 1.3, há inúmeras aplicações diferentes no mundo real para análise de sentimentos. Vamos fazer um estudo em conceitos importantes envolvidos nestes problemas.

1. **Polaridade dos Sentimentos e Graus de Positividade**

 Alguns problemas em análise de sentimentos compartilham a seguinte propriedade: dado um documento com algumas opiniões, onde supõe-se que elas giram em torno de um único objeto, pode haver duas alternativas de classificação: classificar esta opinião em uma das duas polaridades opostas de sentimento, vale dizer, positiva ou negativa; Ou localizar sua posição no *continuum* entre estas duas polaridades. O trabalho de Eguchi [6] em análise de sentimento está relacionado nesta última categoria. De uma forma geral, esta abordagem afirma que a polaridade ou positividade podem ser utilizadas para sumarização dos documentos com opinião, seja ela positiva ou negativa.

 A opção pela técnica *binária*, que apenas classifica o documento como positivo ou negativo, é aquela apresentada como a classificação de sentimento propriamente dita em 1.1, também costuma ser chamada de *classificação de polaridade* [7,8].

 Na classificação binária de sentimentos há algumas aplicações interessantes, como os conhecidos *"thumbs up"* ou *"thumbs down"* para opiniões sobre filmes. A figura 4 ilustra um exemplo. Embora, neste caso, "positivo" e "negativo" estão diretamente ligados com “gostar” e “não gostar”, respectivamente, nem sempre é assim. Podemos observar um discurso político e solicitar ao sistema de classificação que determine se o discurso é de um partido da direita ou de esquerda. Uma vez que este tipo de classificação trata os problemas com duas classas opostas e bem definidas – ao menos em tese, deveriam ser – podemos utilizar técnicas tradicionalmente utilizadas em apredizagem de máquina.



Figura 4 – Classificação de sentimentos utilizando *thumbs up* e *thumbs down*

1. **Detecção de subjetividade e Identificação de Opinião**

 Na análise de sentimentos, muitas vezes fazemos a suposição de que os documentos contêm apenas opiniões. O fato é que muitas vezes é preciso determinar se o documento contém informações subjetivas ou não. E ainda, identificar que partes são subjetivas. Mihalcea afirma em seu trabalho [9] que distinguir uma instância objetiva de um subjetiva é frequentemente mais difícil que a classificação de polaridade. E que, portanto, avanços nas técnicas de detecção de subjetividade podem implicar em avanços na classificação de sentimentos.

 Hatzivassiloglou e Wiebe [10] estudaram os efeitos dos adjetivos na subjetividade das sentenças. O objetivo era afirmar se uma sentença é ou não subjetiva, julgando os adjetivos contidos nela.

 Outro ponto importante é determinar a intensidade da opinião. Enquanto que no problema de classificação de polaridade, é preciso encontrar o ponto de sentimento entre o positivo e o negativo neste problema, uma sentença pode expressar uma opinião muito positiva ou pouco positiva.

* 1. Classificação utilizando *frases com sentimentos*

 Este método realiza a classificação de cada sentença do documento baseado no fato de cada palavra ser positiva ou negativa. O algoritmo descrito nesta seção está no artigo publicado por Turney [11]. Lá, ele é utilizado para classificação da opinião de consumidores.

 Este algoritmo utiliza o procedimento de *análise morfológica*, que é responsável por estudar as palavras de acordo com a classe gramatical a qual elas pertencem. Quando nos referimos às classes gramaticais, logo sabemos que se refere àquelas dez, que são: substantivos, artigos, pronomes, verbos, adjetivos, conjunções, interjeições, preposições, advérbios e numerais. No entanto estas podem assumir uma quantidade imensa de formas. Por exemplo, um verbo estar presente em muitos tempos, como o pretérito, futuro, etc. Nesta seção, utilizaremos o padrão **Penn Treebank POS Tags**, apresentado na Tabela 4, que atribui um identificador a alguns dos elementos morfológicos em inglês, e que está traduzido aqui para português, cobrindo uma grande quantidade das variações destas classes gramaticais.

|  |  |  |  |
| --- | --- | --- | --- |
| **Identificador** | **Classe Gramatical** | **Identificador** | **Classe Gramatical** |
| CC | Conjunção coordenativa | POS | Pronome Possessivo |
| CD | Numeral Cardinal | PRP | Pronome Pessoal |
| DT | Artigo definido | RB | Advérbio |
| FW | Palavra estrangeira | RBR | Advérbio comparativo |
| IN | Preposição ou Conjunção subordinada | RBS | Advérbio superlativo |
| JJ | Adjetivo | UH | Interjeição |
| JJR | Adjetivo comparativo | VB | Verbos, presente |
| NN | Substantivo, singular | VBD | Verbos, passado |
| NNS | Substantivo, plural | VBG | Verbos, gerúndio ou particípio |
| NNP | Substantivo próprio singular | VBP | Verbos, singular |
| NNPS | Substativo próprio plural | VBZ | Verbos, plural |

Tabela 4 – Padrão Penn Treebank POS Tags

 O algoritmo sugerido em [11] consiste de três passos:

 **Passo 1:** São extraídas as frases que contêm adjetivos ou advérbios. Isto é feito pois as pesquisas que existem até hoje, mostram que adjetivos e advérbios são bons indicadores de subjetividade e opiniões. No entanto, eles ainda não são determinantes para classificação de polaridade. Por exemplo, o adjetivo *imprevisível* pode ser negativo na sentença “O comportamento dos freios deste carro é imprevisível”, ou positivo em “O final do filme é imprevisível”. O fato é que são extraídas da frase um par de palavras que obedeça as restrições apresentadas na Tabela 5. Por exemplo, o padrão na terceira linha indica que duas palavras consecutivas serão extraídas se a primeira for um adjetivo, a segunda for um adjetivo, mas se a terceira (que não é extraída) não for um substantivo. Os substantivos são evitados para que o nome dos objetos não influenciem na classificação.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Primeira palavra | Segunda Palavra | Terceira Palavra (Não é extraída) |
| 1. | JJ | NN ou NNS | Qualquer uma |
| 2. | RB, RBR ou RBS | JJ | Exceto NN e NNS |
| 3. | JJ | JJ | Exceto NN e NNS |
| 4. | NN ou NNS | JJ | Exceto NN e NNS |
| 5. | RB, RBR, RBS | VB, VBD, VBN ou VBG | Qualquer uma |

 Tabela 5 – Padrões para extração de duas palavras consecutivas de opiniões

 **Passo 2:** Estimar a orientação semântica de uma sentença é obtida dos pares extraídos utilizando a equação **PMI** (Pointwse Mutual Information):

PMI ( *term1* , *term2* ) = log 2 [ Pr (*term1 ^ term2) / (* Pr *(term1) \** Pr *(term2) ) ]*

 Tabela 6 – Equação PMI para cálculo da orientação semântica de um par

 Na equação PMI, Pr (*term1 ^ term2)* é a probabilidade de uma ocorrência simultânea dos termos *term1* e *term2*. Enquanto quePr *(term1) \** Pr *(term2)* é a medida do grau de dependência estatística entre eles. O *log* é da razão entre estas duas medidas é a quantidade de informação que nós adquirimos da presença de uma das palavras quando nós observamos a outra.

 Logo, a orientação semântica de uma sentença é computada baseada sua ligação com uma palavra positiva (“excelente”) e com sua ligação com uma palavra negativa (“péssimo”).

SO (*frase*) = PMI (frase,”excelente”) - PMI (frase,”péssimo”)

Tabela 7 – Orientação semântica de uma frase

 Para o cálculo das probabilidades citadas pode ser feito através de uma ferramenta de busca de propósito geral, coletando o número de *hits.* Para cada busca, uma ferramenta de busca, retorna o número de documentos relevantes daquela busca. Este é considerado o numero de *hits*.

 **Passo 3:** Dada uma opinião, o algoritmo computa a orientação semântica média de todas as frases dela, e classifica a opinião como recomendada se o SO médio é positivo, e como não recomendado caso contrário.

 Ao aplicar este algoritmo em vários domínios, foi obtida uma precisão de 84% para opiniões sobre automóveis e 66% em opiniões sobre filmes.

* 1. Classificação utilizando os métodos de *classificação de texto*

 O caminho mais simples para classificação de sentimentos é tratar o problema como um problema de classificação de textos baseados em tópicos. Há alguns algoritmos que podem ser empregados nesta abordagem do problema, como o algoritmo ingênuo de Bayes, SVM, *k*NN, etc.

 Em [12] esta abordagem foi empregada utilizando opiniões sobre filmes de duas classes, positiva e negativa. Foi utilizado o algoritmo ingênuo de Bayes e SVM. A partir de 700 opiniões positivas e 700 opiniões negativas, foi obtido 81% e 82.9% de precisão com validação cruzada.

* 1. Classificação utilizando uma função *score*

 Uma função score para classificação de sentimentos em uma opinião foi dada por Dave et al. [3]. O algoritmo consiste de dois passos:

 **Passo 1:** Este primeiro passo atribui a cada termo no conjunto de treinamento o valor da função score daquele termo.

*score* ( *ti* ) = [ Pr (*ti* | C ) - Pr (*ti* | C’ ) ] / [ Pr (*ti* | C ) + Pr (*ti* | C’ ) ]

Tabela 8 – Função score

 Nesta função, *ti* é um termo e *C* é uma classe e *C*’ é seu complemento, ou seja, a negação de *C*. *Pr(ti|C)* é a probabilidade condicional do termo *ti* na classe C. Isto é calculado tomando-se o número de vezes que o termo *ti* ocorre na classe *C* e dividindo o resultado pelo número total de termos nas opiniões da classe *C*. O *score* de um termo é portanto uma medida que varia de -1 a 1.

 **Passo 2:** Para classificar um novo documento *di = t1 ... tn ,* o algoritmo soma o *score* de todos os termos e usa o sinal do resultado para determinar a classe. Ou seja, este passo utiliza a equação indicada na Tabela 9 para classificação.

$ $ 

Tabela 9 – Função final para classificação de um documento

 Experimentos utilizando este algoritmo utilizaram milhares de opiniões de sete tipos de produtos. Os resultados mostraram que utilizar *bigrams* (duas palavras consecutivas) ou *trigrams* (três palavras consecutivas) produziu a melhor precisão, cerca de 84.6% a 88.3%, em dois conjuntos diferentes de opiniões.

1. **Experimentos**

 Reservamos esta seção para aplicar extração e classificação em uma base de opiniões disponibilizada por *Mark Dredze* da universidade Johns Hopkins [14]. A seguir será apresentada cada etapa do experimento realizado.

* 1. Extração

Na literatura de mineração de opinião é comum representar o conteúdo dos textos como um vetor de características de um determinado objeto, onde cada elemento deste vetor é um termo individual. A idéia é registrar a frequência de cada termo. Embora esta idéia seja bastante difundida, em [12] foi obtido um resultado melhor utilizando um vetor de presença no lugar de um vetor de frequência. Ou seja, em algumas situações pode ser mais conveniente utilizar um vetor binário em que cada elemento apenas indica a ocorrência (valor 1) ou não (valor 0) de um determinado termo.

 A base de opiniões em [14] encontra-se pré-processada, de forma que podemos obter opiniões apenas sobre livros ou filmes, por exemplo. Nesta etapa de extração vamos obter uma lista de adjetivos (que como vimos são bons indicadores de subjetividade) referentes ao objeto.

 O primeiro passo é realizar uma varredura em todo o documento da base de opiniões, identificando e selecionando os termos que estão exercendo função de adjetivo. Para esta etapa utilizamos um **Treebank POS Tags**, como citado na seção 3.2, e formamos uma tupla de *n* termos.

 A segunda etapa é varrer todas as opiniões da base e gerar um vetor de *n* posições para cada uma, em que a i-ésima posição vale 1, se o adjetivo indicado da i-ésima posição da tupla obtida acima aparece na opinião atual, ou vale 0 caso contrário.

* 1. Classificação

 Uma vez obtido o vetor binário que representa cada opinião da base, é preciso determinar suas polaridades através de um classificador. Estudamos neste trabalho que uma das maneiras mais simples é tratar este problema como um problema comum de classificação de textos. O fato é que possuímos duas classes, a positiva e a negativa, e temos um conjunto de vetores binários que devem ser classificados em uma destas classes.

 Para tal tarefa, utilizaremos o *SentiWordNet* [15] que é uma base de dados de termos em que identificamos um índice de positividade e negatividade de cada palavra que vai de 0 a 1. Então podemos aplicar uma função *score* para cada opinião que é o somatório da diferença entre o grau positividade e negatividade.

Tabela 10 – Função score para classificação de uma opinião

*score* ( *opinião* ) = *PosScorei – NegScorei  [para todos os i-ésimos termos]*

 Após esta etapa obtemos um conjunto resultado com nossas classificações. Este processo indica não só a polaridade, como também o grau de sentimento existente na opinião.

1. **Resultados obtidos**

 O experimento descrito na seção anterior foi implementado em linguagem C e para analisar seus resultados utilizamos um programa para ler e armazenar memória toda base de opinião. Foram utilizadas bases de opinião de 4 tipos, todas obtidas em [14], vale dizer que se tratam de livros, filmes, produtos eletrônicos e receitas.

 A precisão para base de opinião pode ser observada na tabela 11.

Livros

p = 746, n = 226
p = 211, n = 147

Eletrônicos

p = 817, n = 172
p = 553, n = 424

Receitas

p = 157, n = 28
p = 560, n = 406

Filmes

p = 141, n = 38
p = 67, n = 54

Tabela 11 – Resultados obtidos com o experimento

 Cada base apresentada recebeu como entrada um conjunto de opiniões positivas e outro de opiniões negativas. Cada linha observada corresponde, respectivamente, a estes conjuntos. Para a base de produtos eletrônicos podemos observar que do conjunto de 989 opiniões positivas, 817 foram identificadas pelo experimento como positivas de fato, ou seja, uma precisão de *82%*. No entanto, para opiniões negativas a precisão caiu para *41%*.

 O fato é que a ordem destas taxas se repete nas demais bases. Em média é obtida uma precisão de *83%* para uma base de opiniões positivas, e *44%* para opiniões negativas. Isto é fruto de ter utilizado a função *score*, indicada na tabela 10, com um valor de fronteira no zero. Isto significa que se uma opinião tem *score* negativo (menor que zero), ela é negativa, caso o *score* seja positivo (maior que zero) será classificada como positiva.

 Experimentamos neste trabalho avançar esta fronteira para melhorar os resultados obtidos até aqui. Aumentar o valor fronteira para algum valor maior que zero, na prática vai tornar mais difícil que uma opinião seja considerada positiva. Para a base de opiniões sobre livros, analisamos na tabela 12 como se comporta a precisão para uma faixa de valores de fronteira entre 0 e 5. Embora os resultados sejam bastante discretos, obtemos uma diferença menos abrupta entre cada classe.

fronteira := 0

p = 746, n = 226 - 75%
p = 211, n = 147 - 40%

fronteira := 0.1

p = 746, n = 246 - 75%
p = 211, n = 154 - 42%

fronteira := 0.5

p = 720, n = 269 – 72.5%
p = 203, n = 162 – 44.3%

fronteira := 1

p = 671, n = 315 – 67.6%
p = 188, n = 176 – 48.2%

fronteira := 2

p = 561, n = 421 – 56.5%
p = 145, n = 215 – 58.9%

Tabela 12 – Base de opiniões sobre livros. Variação do valor fronteira entre 0 e 2

1. **Conclusão e trabalhos futuros**

 Assim como toda a área de mineração de opiniões, a análise de sentimentos tem recebido muita atenção nos dias atuais graças ao grande volume de documentos em toda Internet contendo informações geradas por eleitores, consumidores, etc.

 Nosso objetivo neste trabalho foi cobrir técnicas e abordagens que estão diretamente relacionadas com a análise de sentimentos. O problema é muito desafiador e possui muitas aplicações como apresentamos neste trabalho.

 Foram realizados experimentos que conduziram a resultados positivos, utilizando bases de dados fornecidas gratuitamente na Internet. E, a partir destas bases, analisamos variações da técnica implementada para uma melhoria nos resultados.

 Sem dúvida, esta é uma área que irá crescer bastante nos anos que seguem. Espera-se um crescimento de pesquisas que proponham novas abordagens e uma comparação (até mesmo união) das técnicas já existentes para progressos em relação ao que temos hoje.

1. **Referência Bibliográfica**

[1] Liu, Bing, **Web Data Mining – Exploring Hyperlinks, Contents and Usage Data.** Springer, December, 2008

[2] **Sentiment Analysis and Subjectivity - Handbook of Natural Language Processing**, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010

[3] K. Dave, S. Lawrence e D. Pennock. Mining the Peanut Gallery: **Opinion Extraction and Semantic Classification of Product Reviews**. In *Proc.* *Of the 12th Intl. World Wide Web Conference (WWW’03)*

[4] A. B. Goldberg, X. Zhu, and S. Wright, “**Dissimilarity in graph-based semisupervised classification**,” in *Artificial Intelligence and Statistics (AISTATS)*, 2007.

[5] Brin, Sergey, Page, Lawrence. **The Anatomy of a Large-Scale Hypertextual Web Search Engine**. Computer Science Department, Stanford University, Stanford, CA 94305

[6] K. Eguchi and V. Lavrenko, **Sentiment retrieval using generative models**, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

[7] Salvetti, Lewis, Reichenbach. **Automatic Opinion Polarity Classification of Movie Reviews.** *Department of Computer Science, University of Colorado at Boulder*

## [8] **Zhang, Changli.** Polarity Classification of Public Health Opinions in Chinese. Springer Berlin / Heidelberg

[9] R. Mihalcea, C. Banea, and J. Wiebe, **Learning multilingual subjective language**

**via cross-lingual projections**, in *Proceedings of the Association for Computational Linguistics*

[10] V. Hatzivassiloglou and J. Wiebe, **Effects of adjective orientation and gradability on sentence subjectivity**, in *Proceedings of the International Conference*

*on Computational Linguistics*

[11] P. Turney. Thumbs Up or Thumbs Down? **Semantic Orientation Applied to Unsupervised Classification of Reviews**. In *Proc. Of the Meeting of the Association for Computational Linguistics (ACL’ 02)*

[12] B. Pang, L. Lee e S. Vaithyanathan. Thumbs up? **Sentiment Classification Using Machine Learning Techniques**. In *Proc. Of the EMNLP’02*, 2002.

[13] [www.epinions.com](http://www.epinions.com), último acesso em 04/07/2010

[14] [http://www.cs.jhu.edu/~mdredze/datasets/sentiment](http://www.cs.jhu.edu/~mdredze/datasets/sentiment/), último acesso em 04/07/2010

[15] <http://wordnet.princeton.edu>, último acesso em 06/07/2010

[16] <http://www.cs.waikato.ac.nz/ml/weka>, último acesso em 06/07/2010

**Assinaturas**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ricardo Bastos Prudêncio

**Orientador**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Hallan Cosmo dos Santos

**Aluno**