**BARBARA MARTINAZZO**

# UM MÉTODO DE IDENTIFICAÇÃO DE EMOÇÕES EM TEXTOS CURTOS PARA O PORTUGUÊS DO BRASIL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

**CURITIBA**

**2010**

**BARBARA MARTINAZZO**

# UM MÉTODO DE IDENTIFICAÇÃO DE EMOÇÕES EM TEXTOS CURTOS PARA O PORTUGUÊS DO BRASIL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito para obtenção do título de Mestre em Informática.

Área de Concentração: *Descoberta de Conhecimento e Aprendizagem de Máquina*

Orientador: Prof. Dr. Emerson Cabrera Paraiso

**CURITIBA**

**2010**

Martinazzo, Barbara
Um Método De Identificação De Emoções Em Textos Curtos Para O Português Do Brasil. Curitiba, 2011. 68p.

Um Método De Identificação De Emoções Em Textos Curtos Para O Português Do Brasil – Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática Aplicada.

1. processamento de linguagem natural 2. análise de emoções 3. identificação de emoções 4. computação afetiva. I.Pontifícia Universidade Católica do Paraná. Centro de Ciências Exatas e de Tecnologia. Programa de Pós-Graduação em Informática Aplicada II-t

À minha família, pelo apoio e presença em todos os momentos de minha vida.

Aos meus pais, pelas oportunidades e pela educação que me foram dadas.

À minha irmã, pelo incondicional companheirismo e amizade durante todos estes anos.

Aos meus amigos, pelos muitos momentos em que estiveram ao meu lado.

# Agradecimentos

Embora não seja uma tarefa fácil, é de extrema importância que alguns agradecimentos sejam feitos. Durante o desenvolvimento deste trabalho, muitas pessoas passaram pela minha vida e contribuiram, de uma forma ou de outra, para que o mesmo tivesse bons resultados e se tornasse algo digno de ser mencionado e citado para outras pesquisas. Tentarei, então, agradecer a todos que me auxiliaram, procurando não me esquecer de ninguém.

Meu maior agradecimento é dirigido à minha família, especialmente meus pais, por terem sido o contínuo apoio em todos estes anos, dando-me as ferramentas e as bases necessárias para que eu me tornasse a pessoa que sou hoje. Nicole, minha irmã, que sempre com muita paciência e disponibilidade, esteve sempre presente para me dar uma opinião ou um auxílio quando precisei.

Não poderia deixar de mencionar alguns amigos, que me acompanharam durante toda minha infância e adolescência e que, apesar de alguns desentendimentos, não deixaram de acreditar em mim e na nossa amizade. Catherine, a amiga de jardim de infância, e que está sempre ao meu lado; Diana, Danelise, Lorena, Mayara, Tatiana, Gisele, Monica, Fábio, Tatiane, Thiago, Cíntia e Lucas. Também gostaria de citar algumas pessoas que conheci em minha vida profissional, e que de uma forma ou de outra participaram desse processo: Rocha, Bortolotto, Nico, Ale e Dulce, Leoni e Ilson, que me acompanharam desde minha primeira experiência profissional e hoje, apesar do pouco contato, são amigos muito queridos. Ao pessoal da Hi Technologies, especialmente os sócios Alfredo, Marcus e Sérgio, que me incentivaram a iniciar o mestrado e investir em minha formação.

Por fim, no âmbito acadêmico, agradeço principalmente ao meu orientador, professor Emerson Paraiso, pela oportunidade e auxílio no desenvolvimento do projeto, à secretaria do PPGIa (Programa de Pós Graduação em Informática) pela disponibilidade, e à colega Lilian, pela amizade e pelos conselhos.

Àqueles que eu não mencionei, peço desculpas, mas infelizmente meu espaço é limitado. Sintam-se, então, agradecidos por terem feito parte de minha história e por terem feito a diferença em minha vida.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas

| | |
|---|---|
| BC | Base de conhecimento |
| IA | Inteligência Artificial |
| FSA | *Finite State Automata* |
| LSA | *Latent Semantic Analysis* |
| PLN | Processamento de Linguagem Natural |
| IR | *Information Retrieval* |
| SVD | *Singular Value Decomposition* |

# Resumo

Os avanços recentes na análise automática de textos conduziram ao surgimento de uma área responsável por reconhecimento de aspectos subjetivos, tais como opiniões, sentimentos e emoções do autor do texto analisado. Pesquisas nessa área remetem ao desenvolvimento de métodos que possibilitam que sistemas computacionais sejam capazes de reconhecer e detectar fatores afetivos no texto. Entretanto, por ser uma área relativamente nova, estes métodos ainda estão em fase de desenvolvimento e são, em sua grande maioria, para a língua inglesa. Desta forma, observa-se a necessidade de adaptação dos mesmos para outros idiomas, como o português. O presente documento tem por objetivo apresentar um método baseado em Latent Semantic Analysis de identificação de emoções em bases textuais em língua portuguesa. No caso dessa pesquisa, textos curtos serão manchetes de notícias diversas, extraídas de sites da internet, seguidas de uma breve descrição. Em algumas experimentações, o método obteve uma taxa média de identificação de emoções na ordem de 70%.

**Palavras-Chaves**: processamento de linguagem natural, análise de emoções, identificação de emoções, computação afetiva.

# Abstract

Recent advances in texts analysis lead to the emergence of a new area responsible for the recognition of subjective aspects, such as opinions, feelings and emotions in texts. Research in this area refer to the development of methods to allow computational systems to be able to recognize and detect affective factors in texts. However, as it is a relatively new area, these methods are still in the development phase and are, in its vast majority, only for the English language. Thus, we notice the need for adaptation of these methods to other languages, such as Portuguese. The main goal of this research is to propose a method, based on the Latent Semantic Analysis, for emotion detection within textual basis in Portuguese. In the context of this research, the texts that will be used are news headlines, extracted from the internet, followed by its short description. Experimentations have shown that this method can find the correct emotions in a text in 70% of the cases.

**Keywords:** natural language processing, emotion analysis, emotion annotation, affective computing.

# Capítulo 1

# Introdução

As emoções são objeto de pesquisas em diferentes áreas, tais como a psicologia e outras ciências responsáveis pelo estudo do comportamento. Isso se deve ao fato de que uma emoção é um elemento extremamente importante da natureza e da conduta humana em qualquer sociedade e cultura. Recentemente, esse tipo de estudo tem atraído também a atenção de pesquisadores da Ciência da Computação, especialmente os interessados no processamento de textos, recuperação de informação e na interação humano-computador.

Nos últimos anos, bases de dados em formato textual têm crescido e se disseminado muito rapidamente, devido ao desenvolvimento dos meios digitais como CD-ROMs, publicações eletrônicas, e-mails e a própria Internet. Esse crescimento motivou o estudo de novos métodos para suprir esta necessidade de extrair informações úteis de textos [HAN e KAMBER, 2001].

Uma das principais aplicações da mineração de textos consiste em classificar ou comparar textos de acordo com um ou mais critérios determinados pelo sistema de mineração de textos. Ainda segundo Han e Kamber [HAN e KAMBER, 2001], somente uma pequena porção dos documentos existentes será realmente relevante para um determinado fim. Entretanto, sem se saber o que está contido em cada texto, é difícil extrair deles qualquer informação útil. Devido a este problema foram criadas ferramentas para analisar diversos documentos e classificá-los de acordo com categorias pré-estabelecidas ou encontrar padrões que os conectem a outros textos. Uma das utilidades que vem sendo bastante explorada é a identificação de emoções em textos.

O reconhecimento automático de emoções em informação textual [STRAPPARAVA e MIHALCEA, 2008], conhecida como *Sentiment Analysis*, é uma das áreas que tem atraído a atenção dos pesquisadores, pois sabe-se que, além de informação, textos podem conter também opiniões e teores emocionais [ALM et al., 2005]. Esse assunto é crítico para o desenvolvimento de interfaces inteligentes e de várias aplicações de multimídia como, por exemplo, a síntese da fala a partir de textos. Entretanto, por ser uma área relativamente nova, estes métodos ainda estão em fase de desenvolvimento e, em sua grande maioria, estão sendo desenvolvidos para a língua inglesa. Desta forma, observa-se a necessidade de desenvolver tais estudos para outros idiomas, como o português.

Uma das aplicações possíveis para essa finalidade de pesquisa está em sistemas de classificação de produtos, por exemplo, que não trazem fatos absolutos, mas sim opiniões pessoais. Dessa forma, a análise sentimental de tais revisões pode auxiliar o sistema a recomendar ou não determinado produto a uma pessoa que busque informações sobre o mesmo, baseando-se nas avaliações fornecidas por outros usuários. Um exemplo disso é a classificação de revisões literárias ou cinematográficas como positivas ou negativas, atribuindo-se a um filme uma nota baseada não somente no sistema tradicional de ranqueamento, mas também com base em avaliações textuais fornecidas por pessoas que já o assistiram [PANG e LEE, 2004]. Isto seria bastante útil para estabelecer um comparativo entre as resenhas avaliativas fornecidas pelos usuários e as notas atribuídas pelos mesmos no ato da avaliação. Através da Figura 1, é possível observar que nem sempre a nota atribuída pelo consumidor é compatível com a sua real opinião.

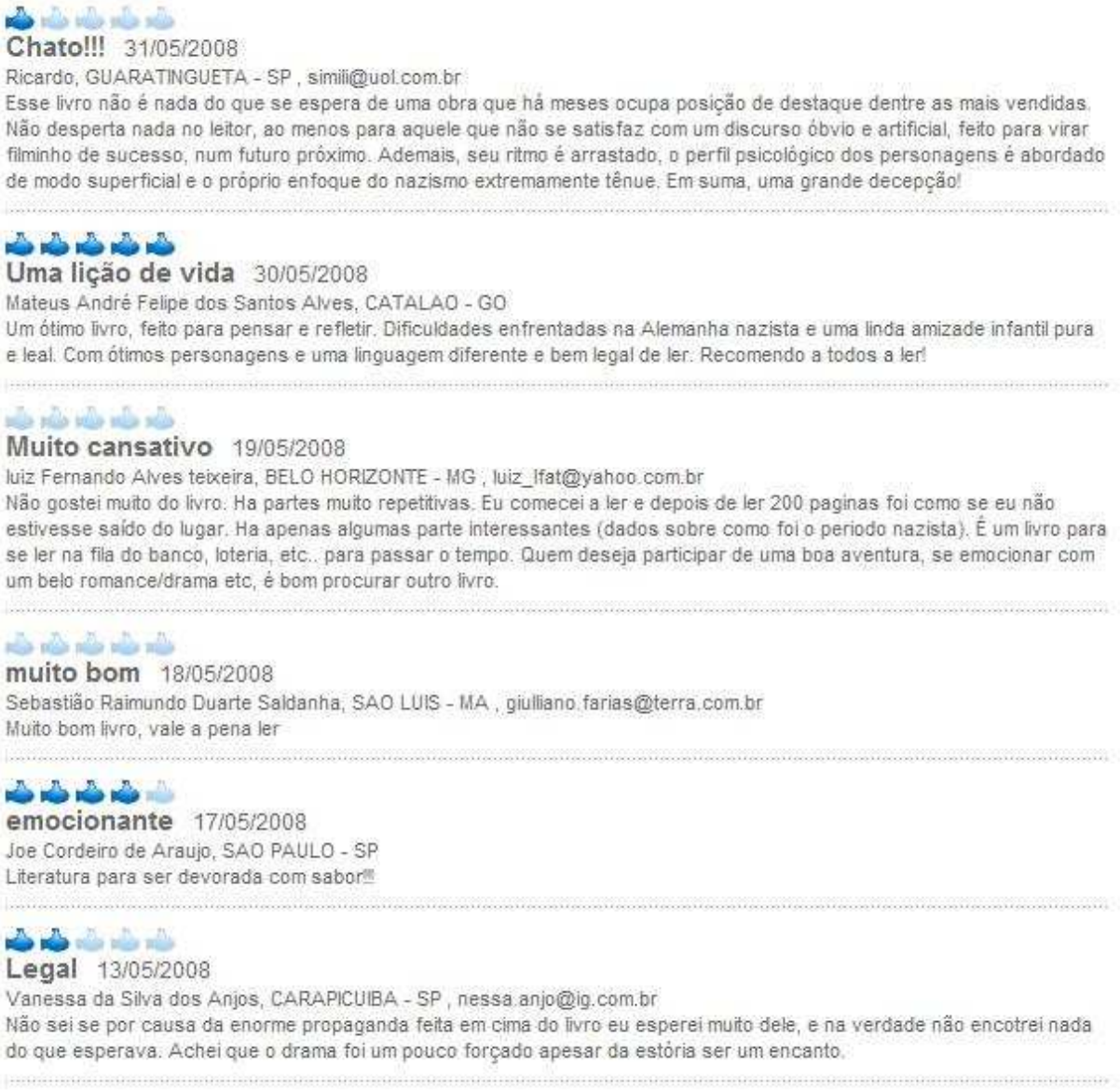


Figura 1: Resenhas avaliativas de um produto em um *site* de compras *online*[1].

Como um exemplo, pode ser visto na Figura 1 que uma avaliação intitulada “muito bom” (e avaliada positivamente através da resenha textual) foi ranqueada com nota 0, igualmente a outra que realmente demonstrava desinteresse pelo produto analisado (“muito cansativo”). Este cenário é comum em qualquer sistema avaliativo, o que torna útil a integração entre a classificação por nota e a análise textual no ato de atribuição automática de notas e conceitos a determinado produto.

[1] http://www.submarino.com.br/produto/1/1900640/menina+que+roubava+livros,+a#readRatings

## 1.1. Objetivos

O trabalho descrito no presente documento propõe o desenvolvimento de um sistema de identificação de emoções em bases textuais escritas em português do Brasil. O objetivo é identificar uma das seis emoções básicas descritas por Paul Ekman e Wallace Friesen [EKMAN e FRIESEN, 1978] (alegria, raiva, tristeza, desgosto, medo e surpresa) em notícias curtas. Uma vez identificadas, estas emoções serão posteriormente utilizadas para a animação facial de um avatar (agente conversacional animado) que lê tais textos (nesta dissertação, notícias). O avatar que será responsável pela leitura do texto modificará seu comportamento, basicamente suas expressões faciais, de acordo com as emoções encontradas pelo sistema de identificação de emoções no decorrer do texto.

Para a realização do objetivo acima proposto, foram executadas algumas tarefas de mineração de textos que serão posteriormente explicadas. Entre elas, citam-se a definição de um método para identificação de emoções a partir de notícias curtas e a validação do método através de experimentos.

## 1.2. Hipóteses de Trabalho

Demonstrar que, através da implementação de um método baseado em LSA, é possível realizar, de forma automática, a identificação de emoções em textos curtos escritos em português do Brasil.

## 1.3. Contribuições Científicas e Tecnológicas

A principal contribuição científica deste trabalho é um método para realizar a identificação de emoções em textos no idioma português, uma vez que os métodos disponíveis na literatura trabalham com o idioma inglês.

Além do método de identificação de emoções, este trabalho resultou na disponibilização de um algoritmo computacional, escrito em Java, capaz de realizar a identificação de emoções contidas em textos curtos, e seis conjuntos de palavras, em português, que descrevem as seis emoções utilizadas no trabalho.

## 1.4. Organização do Documento

A organização do documento será da seguinte forma: no Capítulo 2 será feita uma breve introdução teórica de conceitos fundamentais para o entendimento do trabalho aqui proposto. Os conceitos abordados referem-se, primeiramente, ao conceito básico de emoções; em seguida, serão abordados tópicos tais como a recuperação de informação em textos, mineração de textos e seu princípio base, além de algoritmos e ferramentas já utilizados e conhecidos para a finalidade; também serão apresentados alguns trabalhos já realizados nessa área de identificação de emoções em textos. O Capítulo 3 tratará especificamente da explanação do método utilizado para o desenvolvimento do trabalho; primeiramente, serão apresentados os conceitos principais por trás do método e, em seguida, o mesmo será apresentado em detalhes, de forma a deixar claro como é feita a identificação de emoções em notícias escritas em português do Brasil. No Capítulo 4 serão apresentadas as ferramentas utilizadas para a implementação do método descrito no Capítulo 3. O Capítulo 5 tem por objetivo relatar experimentos e resultados obtidos com base nos estudos realizados e no método implementado. Por fim, serão apresentadas, no Capítulo 6, as conclusões e propostas para trabalhos futuros.

# Capítulo 2

# O Processo de Identificação de Emoções em Textos

O presente capítulo tem por objetivo estabelecer uma introdução sobre os conceitos que foram necessários para o desenvolvimento do trabalho. Inicialmente será feita uma breve apresentação do conceito de Emoções utilizado nesta pesquisa. No decorrer deste tópico, os assuntos serão tratados da seguinte forma: em um primeiro instante, é abordado o conceito básico de emoção, visando informar ao leitor o motivo de escolha das emoções que serão avaliadas no decorrer do processo. Em seguida, é iniciado o processo de introdução de conceitos mais específicos ao trabalho, como Recuperação de Informação (também chamada de *Inforation Retrieval* - IR) e mineração de textos, bem como seu princípio base, alguns dos algoritmos utilizados e algumas das aplicações. Em seguida, este conteúdo é estendido para a identificação de emoções em textos. Por fim, são discutidos alguns trabalhos desenvolvidos na área.

## 2.1. Emoções

Fehr e Russell [FEHR e RUSSEL, 1984] afirmam que todas as pessoas sabem o que é uma emoção, até que seja solicitada uma definição. Eles ainda questionam se as emoções são eventos psicológicos, mentais ou comportamentais e, além disso, se existem emoções mais "básicas" que outras. Embora ainda não exista um consenso sobre sua definição, pode-se dizer que emoções são estados mentais e psicológicos associados com uma grande variedade de sentimentos, pensamentos e comportamentos. Gazzaniga e Heatherton [GAZZANIGA e HEATHERTON, 2005] afirmam que as emoções são objeto de estudo de diversas áreas do conhecimento humano já há bastante tempo. Segundo Strongman [STRONGMAN, 2003], os filósofos gregos, como Platão e Aristóteles, foram os primeiros a questionar o tema.

Aristóteles acreditava que a emoção é o lado mais interessante da existência humana. Já Darwin, em seu livro “A Expressão da Emoção em Homens e Animais”, enfatizou o papel fundamental das emoções no processo evolutivo dos seres vivos.

O estudo das emoções se divide em várias áreas distintas. A utilizada neste trabalho é chamada de Emoções Básicas (ou Puras). Esse conceito diz respeito às emoções inatas compartilhadas por todas as culturas, e foi proposto na década de 1970 por Paul Ekman e Wallace Friesen [EKMAN e FRIESEN, 1978]. Uma vez que não existe um acordo sobre quantas e quais são as emoções básicas, o modelo proposto por Paul e Wallace foi composto por seis: tristeza, raiva, alegria, medo, desgosto e surpresa. Em título experimental, uma pesquisa foi realizada em vários países, onde os autores pediram às pessoas que identificassem respostas emocionais apresentadas em fotografias de expressões faciais. Foi descoberto, a partir desse estudo, que as seis emoções propostas no modelo foram facilmente interpretadas em todos os países onde o teste foi aplicado. Sendo assim, é natural que se possa estender as mesmas emoções para modificar o comportamento de um avatar, que tem por objetivo fundamental fazê-lo através de expressões faciais.

As emoções têm sido pesquisadas em diferentes ramos, tais como a psicologia e outras ciências responsáveis pelo estudo do comportamento. Isso se deve ao fato de elas serem um elemento extremamente importante da natureza e da conduta humana. Recentemente, esse tipo de estudo tem atraído também a atenção de pesquisadores do ramo da Ciência da Computação, especialmente no que tange a interação entre homens e máquinas [STRAPPARAVA e MIHALCEA, 2008]. Entre os assuntos e as pesquisas realizadas encontra-se o reconhecimento automático de emoções em informação textual, conhecida como *Sentiment Analysis.*

Na sequência deste capítulo são apresentados os elementos relacionados a identificação de emoções em textos.

## 2.2. Recuperação de Informação

A recuperação de informação (IR) possui um significado muito amplo e é aplicada em uma grande variedade de tarefas, desde sistemas de bancos de dados até páginas de busca na web [VARELAS et al., 2005]. Segundo [MANNING et al., 2008], o simples fato de se consultar o número de um cartão de crédito pode ser considerado como uma forma de recuperação de informações. Entretanto, num âmbito acadêmico, ela deve ser considerada

como a busca por informações ou dados desestruturados dentro de documentos ou conjuntos de documentos para satisfazer determinada necessidade de informação. Pode-se dizer que ela é responsável pela representação, armazenamento, organização de informação, além do acesso a mesma. [YATES e RIBEIRO NETO, 1999].

A busca por informações é realizada há muitos anos e a idéia principal consiste em localizar documentos através de termos especificados pelo usuário [VARELAS et al., 2005]. O homem organiza informações para busca e utilização posteriores. Um exemplo típico pode ser visto nos trabalhos realizados por bibliotecários, assistentes e pesquisadores; nesse aspecto, também se pode citar o índice de um livro. Com o crescimento das bases de informações, houve a necessidade de implementação da primeira estrutura de acesso às informações armazenadas.

Por séculos, esses sistemas foram criados e administrados manualmente, através de hierarquias de categorização tais como o índice alfabético. Entretanto, com os avanços tecnológicos e as mudanças nas necessidades individuais, fez-se necessária a evolução dessa área para que ela fosse capaz de atender, de forma mais eficiente, às demandas de cada um. Com isso, na década de 1990, surgiram os sistemas de busca, que permitiram maior facilidade no acesso à informação por parte de qualquer indivíduo com acesso à internet [MANNING et al., 2008], [YATES e RIBEIRO NETO, 1999].

Foi com o surgimento e a popularização da internet que a recuperação de informação ficou realmente conhecida. Nessa época surgiram os primeiros sistemas de busca on-line, como o Google. Nesses sistemas é implementado o método mais conhecido de extração de informações, denominado "*ad-hoc querying*", onde um termo (ou palavra-chave) é utilizado para pesquisar um grande conjunto de documentos armazenados em uma base de dados. Na maioria dos casos, a precisão nos resultados não é satisfatória, cabendo ao usuário filtrar a informação e separar aquilo que é relevante ao que ele busca [SHAH et al., 2002].

Yates e Ribeiro Neto [YATES e RIBEIRO NETO, 1999] afirmam que a recuperação de informações relevantes está diretamente relacionada com a tarefa do usuário e com a forma de representação dos documentos utilizada no sistema de recuperação de informações. O usuário de um sistema de recuperação de informações deve saber dizer ao sistema a informação exata que ele precisa. Em outras palavras, ele deve ser capaz de traduzir a sua necessidade em uma linguagem específica do sistema. Geralmente, essa linguagem consiste em um determinado conjunto de palavras que, semanticamente, exprimem a necessidade do

usuário. Já a representação lógica de documentos, devido a razões históricas, ocorre, muito frequentemente, através de conjuntos de palavras-chave (que podem ser extraídas diretamente dos documentos analisados ou especificadas pelo usuário) e índices. Outro tema de pesquisa recorrente no processo de identificação de emoções em texto pode ser genericamente chamado de Mineração de Textos.

## 2.3. Mineração de Textos

A mineração de textos é, por alguns autores, considerada uma área da mineração de dados. De fato, segundo Feldman e Sanger [FELDMAN e SANGER, 2007], analogamente à mineração de dados, a mineração de textos busca extrair informações úteis de textos através da identificação e exploração de padrões por meios computacionais. Entretanto, como afirma Hearst [HEARST, 2003], a mineração de textos é diferente daquilo que se conhece por busca na web através de sistemas como o Google. Nessas buscas, o usuário geralmente busca algo que já foi pesquisado e escrito por outras pessoas e, portanto, localiza várias informações através de palavras-chave. O problema dessas buscas é filtrar o conteúdo irrelevante. Ao contrário, a mineração de textos busca informações até então implícitas, porém desconhecidas e que, portanto, ainda não foram formalmente documentadas.

A diferença básica entre mineração de dados e de textos consiste em que estes padrões não são encontrados em bases de dados formalizadas, mas em dados textuais não estruturados presentes nos documentos da base. Estes dados textuais podem ser completamente desestruturados ou parcialmente estruturados; um exemplo seria um artigo, que possui alguns dados estruturados (título, nomes dos autores, data de publicação, categoria, etc.), mas também uma grande quantidade de texto não estruturado (resumo, introdução, conclusão, etc.) [HAN e KAMBER, 2001].

A seguir, descreve-se o princípio base da mineração de textos, algumas ferramentas disponíveis e algumas aplicações correntes.

### 2.3.1. Premissa

Nos processos de mineração de dados, a etapa de pré-processamento é dedicada, basicamente, à normalização dos dados já estruturados. Entretanto, em um processo de mineração de textos (ilustrado na Figura 2), a etapa de pré-processamento objetiva identificar e extrair informações representativas dos textos não estruturados. Esta etapa é, então,

responsável por transformar o texto desestruturado de uma base de documentos em um formato intermediário, mais formalizado, o que não é comum aos sistemas de mineração de dados. Em outras palavras, de uma maneira bem simplificada, um sistema de mineração de textos adquire como entrada os documentos em forma de textos e gera diversos tipos de saída.

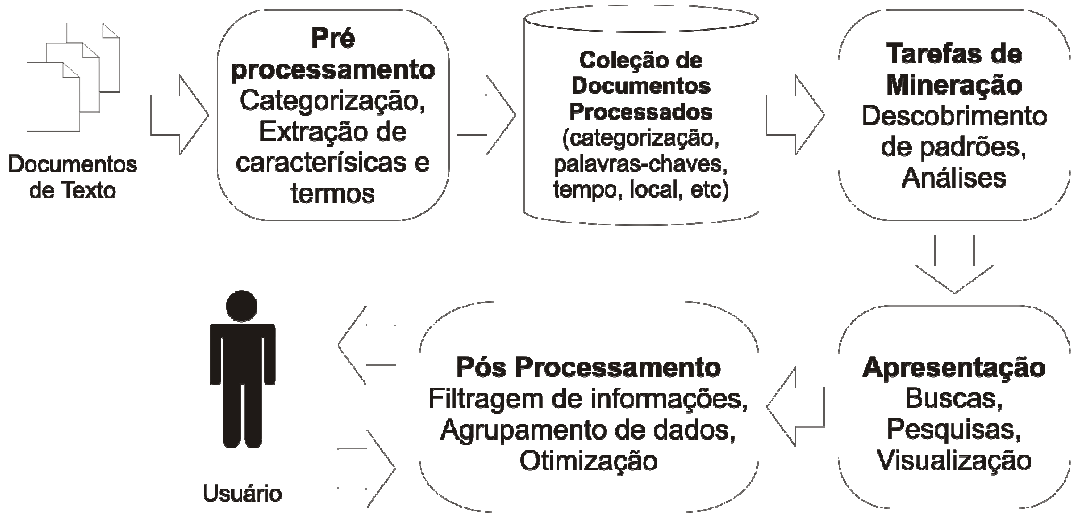


Figura 2: Processo de mineração de textos (adaptada de [FELDMAN e SANGER, 2007]).

Segundo Feldman, em um nível funcional, os sistemas de mineração de texto seguem um modelo geral de algumas aplicações de mineração de dados e são divididos em quatro estágios principais, como pode ser visto na Figura 2. Destes, os dois primeiros são os mais críticos em qualquer sistema de mineração de textos. São eles:

1. **Pré-processamento**

   São todas as rotinas, processos e métodos necessários para preparar os dados para o próximo estágio e geralmente são focadas em atividades de categorização. Seu objetivo é formatar a informação original de forma a torná-la acessível aos métodos de mineração. Em alguns casos é possível executar tarefas capazes de extrair ou aplicar regras para facilitar a formatação.

2. **Operações de mineração (ou *Core Mining Operations)***

   Trata-se do núcleo de qualquer sistema de mineração de textos. Entre as tarefas executadas nesse estágio, pode-se citar o descobrimento de padrões, análises de tendência e algoritmos de descoberta de conhecimento. Os padrões mais utilizados para descoberta de conhecimento em textos baseiam-se em conceitos de associação, frequência e distribuição dos dados ao longo da base;

eles podem ser aplicados individualmente ou comparados entre si visando a obtenção de melhores resultados.

3. **Camada de apresentação**

   Os componentes desse estágio são as funcionalidades de busca e pesquisa e, também, as ferramentas de visualização. Está englobado na camada de apresentação tudo que diz respeito à visualização, como o leiaute do usuário, e as ferramentas otimizadoras.

4. **Pós Processamento**

   Essa etapa é composta de técnicas de refinamento que incluem métodos para a filtragem de informações redundantes e agrupamento de dados associados. Ela pode crescer em um sistema de mineração de dados a ponto de representar um conjunto de aproximações para ordenação, poda, generalização, entre outras tarefas, visando a otimização dos sistemas que implementam essa camada.

Em adição a estas etapas, é comum que exista uma base de conhecimento adquirida em experimentos e processos anteriores chamada de Base de Conhecimento. Estas bases podem ser criadas de várias formas, sendo a mais comum delas através da execução de rotinas de *parsing* em bases externas (como, por exemplo, ontologias) para a identificação de termos nos documentos da coleção de documentos do sistema de mineração. Elas são armazenadas e se tornam disponíveis por muitos elementos do sistema. O objetivo dessas é fornecer informações que facilitem o processo de mineração de textos da base desejada. Posteriormente, elas podem ser utilizadas pelo usuário para criar buscas com restrições ou refinar buscas já criadas, adicionando ou removendo restrições já impostas.

### 2.3.2. Ferramentas Utilizadas

Nesta seção são apresentadas algumas ferramentas testadas, principalmente, na camada 2 descrita na seção anterior (Operações de Mineração). Visto que esta é a camada principal dos sistemas de mineração de textos, como apresentado anteriormente, os algoritmos de extração de características e análises estão, em sua grande maioria, nesta camada.

A seguir são comentadas as principais ferramentas ou bibliotecas empregadas nesse tipo de pesquisa. A maioria dos itens descritos a seguir consiste em grupos de ferramentas disponíveis para o processamento e a mineração de textos. Cada um destes grupos possui suas

próprias ferramentas para cada etapa de processamento, visando a unificação e a facilitação do trabalho do pesquisador.

#### *2.3.2.1. GATE*

Gate, ou *General Architecture for Text Engineering* [CUNNINGHAM et al., 2002], é um ambiente de desenvolvimento que visa auxiliar na criação, avaliação e distribuição de sistemas de mineração de textos. Por ser gratuito, de código aberto e conter uma interface gráfica de desenvolvimento, além de oferecer suporte a diversos idiomas, ele tornou-se popular tanto entre empresas como no meio científico e acadêmico. O GATE fornece aos usuários não somente suporte a aplicações padrão de mineração de textos como extração de informações, mas também tarefas tais como construção e marcação de comentários ao longo do texto e a avaliação de aplicações.

Entre a variedade de ferramentas inclusas para processamento de textos, pode-se citar a tokenização, fragmentação de sentenças, classificação de palavras (verbos, adjetivos, nomes, etc.), análise de frases e recuperação de informações. Também é possível, por meio do GATE, obter acesso a vários recursos linguísticos como ontologias e dicionários. Juntamente com o GATE é distribuído um sistema de extração de informações, ANNIE, que é capaz de detectar nomes de pessoas e organizações, localizações geográficas, datas, horários e informações monetárias tais como valores. Ele possui um dicionário geográfico com nomes de cidades e países e *tokens* como dias da semana e meses. Os padrões podem ser especificados através de strings particulares ou notas previamente criadas por módulos como *tokenizer*, dicionário geográfico ou análise de formato de documento. Também estão inclusos módulos capazes de reconhecer relacionamentos entre entidades e detectar co-referência.

A ferramenta, que recebeu sua versão 6.0 beta 1 em agosto de 2010, é utilizada em vários projetos e possui clientes como a Brittish Telecom, Imperial College, Hewlet Packard, AT&T, entre outros.

#### *2.3.2.2. Natural Language Toolkit*

O NLTK [LOPER e BIRD, 2002] é uma suíte de aplicativos e módulos de código aberto, tutoriais e exemplos que provê o aprendizado da linguística computacional. Essa suíte foi concebida segundo os seguintes critérios: facilidade de uso, consistência, extensibilidade, documentação, simplicidade e é, por causa disso, a mais utilizada por professores da área de PLN. A NLTK é implementada como uma coleção de módulos independentes, sendo que

cada um define uma estrutura de dados específica ou tarefa, podendo ser citados como principais os módulos a seguir: parser, que define uma interface de alto nível que representa as estruturas dos textos através de árvores; rotulador, que é responsável por expandir as características e informações de um determinado token com dados adicionais; fsa, responsável pela codificação e criação de autômatos finitos para expressões regulares; classificador, uma interface voltada para a classificação de textos em categorias, que pode ocorrer através de Naive Bayes, ou através de um modelo de entropia máxima.

O NTLK pode ser executado nas plataformas que suportam Pyton, incluindo Windows, OS X, Linux e Unix.

#### *2.3.2.3. Text-Garden*

Trata-se de uma coleção de ferramentas para resolver problemas com dados estruturados, não estruturados e semi-estruturados, com ênfase em mineração de textos. Text-Garden possibilita ao usuário fácil manipulação de documentos com o objetivo de análise de dados [GROBELNIK e MLADENIC]. Através dessas análises é possível a geração de modelos; classificação, agrupamento, indexação e visualização de documentos; entre outras coisas. Os códigos foram escritos em C++ para ambiente Windows; é possível a utilização das mesmas em ambientes Unix através de emuladores como o Wine [GROBELNIK e MLADENIC].

São utilizados, basicamente, três formatos próprios para a análise e manipulação dos textos e, para tanto, são disponibilizadas as ferramentas de pré-processamento responsáveis por converter os arquivos aos formatos reconhecidos; entre elas estão os conversores de HTML para XML e de HTML para texto em formato TXT. As demais ferramentas possuem como finalidades o agrupamento de documentos, o aprendizado de modelos para classificação de textos, a classificação de documentos, a visualização baseada em agrupamento e em espaço semântico, mineração simples da web, e um sistema de busca; todas elas com diversas opções de configuração e utilização.

#### *2.3.2.4. TextMine*

O kit de ferramentas TextMine [KONCHADY, 2006] consiste em uma coleção de módulos e *scripts* escritos em Perl para executar tarefas de mineração de textos. É possível fazer uso das ferramentas através do navegador de internet ou de linhas de comando em qualquer máquina que possua configurados o Perl, Apache e MySQL.

O conjunto de ferramentas permite que várias tarefas diferenciadas sejam executadas. Entre elas, pode-se citar a busca da *web* para a criação de uma coleção de dados em qualquer tópico desejado; a extração de informações como nomes de pessoas e lugares de textos; busca e indexação de arquivos na máquina local; coleção e organização de artigos e outros textos oriundos da internet; busca por respostas em sistemas de esclarecimento de dúvidas mais comuns de clientes (FAQ) através da extração de palavras-chave de páginas que podem conter respostas às questões; acompanhamento e categorização de *e-mails*; resumos de artigos, páginas web e outros textos e um dicionário baseado no WordNet para definição de palavras e sinônimos.

### 2.3.3. Algumas Aplicações

Com o desenvolvimento dos segmentos mercadológicos, a mineração de textos começou a ser aplicada neste ramo, especialmente no que tange a análise de relacionamento do cliente com a empresa. Uma boa aplicação foi idealizada por Coussement e Van den Poel (http://www.textmining.UGent.be), e consiste em prever possíveis atritos com os clientes, visando evitá-los.

Existem algumas iniciativas comerciais envolvendo a classificação de textos que expressem emoção. Em exemplo é o produto OpSys, desenvolvido por Thomas Jefferson Pereira Lopes. Trata-se de um sistema de mineração de opiniões em conteúdo web: ele monitora redes sociais, blogs e portais, extrai conteúdo relevante e classifica o sentimento desse conteúdo, mostrando para o usuário como está o desempenho de sua marca ou empresa na web (http://www.opsys.com.br/).

Outro projeto em desenvolvimento é o Eleitorando, que está sendo desenvolvido pelo estudante de mestrado Marcel Pinheiro Caraciolo na Universidade Federal de Pernambuco. A ideia do sistema é entender sentimentos em redes sociais sobre determinados assuntos. Nos testes feitos até agora, o programa apresentou nível de acerto de 80% nas classificações de mensagens como positivas, negativas ou neutras (site do projeto: http://www.eleitorando.com.br).

Na próxima seção, serão apresentados alguns trabalhos relacionados ao processo de identificação de emoções em texto.

## 2.4. A Identificação de Emoções em Textos

Para exemplificar diferentes processos de identificação de emoções em textos, apresenta-se nesta seção alguns trabalhos desenvolvidos com este objetivo. Vale ressaltar que todos trabalham com a língua inglesa e que, até o momento de escrita desta dissertação, apenas dois trabalhos realizados para a língua portuguesa foram encontrados.

### 2.4.1. Discernimento de Emoções em Textos

O estudo apresentado por [RUBIN et al., 2004] combina teoria e métodos de pesquisa da psicologia social e da personalidade humana com técnicas analíticas de PLN como uma possível aproximação para isolar, quantificar e descrever de forma qualificativa as emoções encontradas por leitores em textos escritos. A psicologia descreve um modelo empiricamente verificado de emoções discerníveis chamada de "*Watson and Tellegen's Circumplex Theory of Affect*", ilustrado na Figura 3.

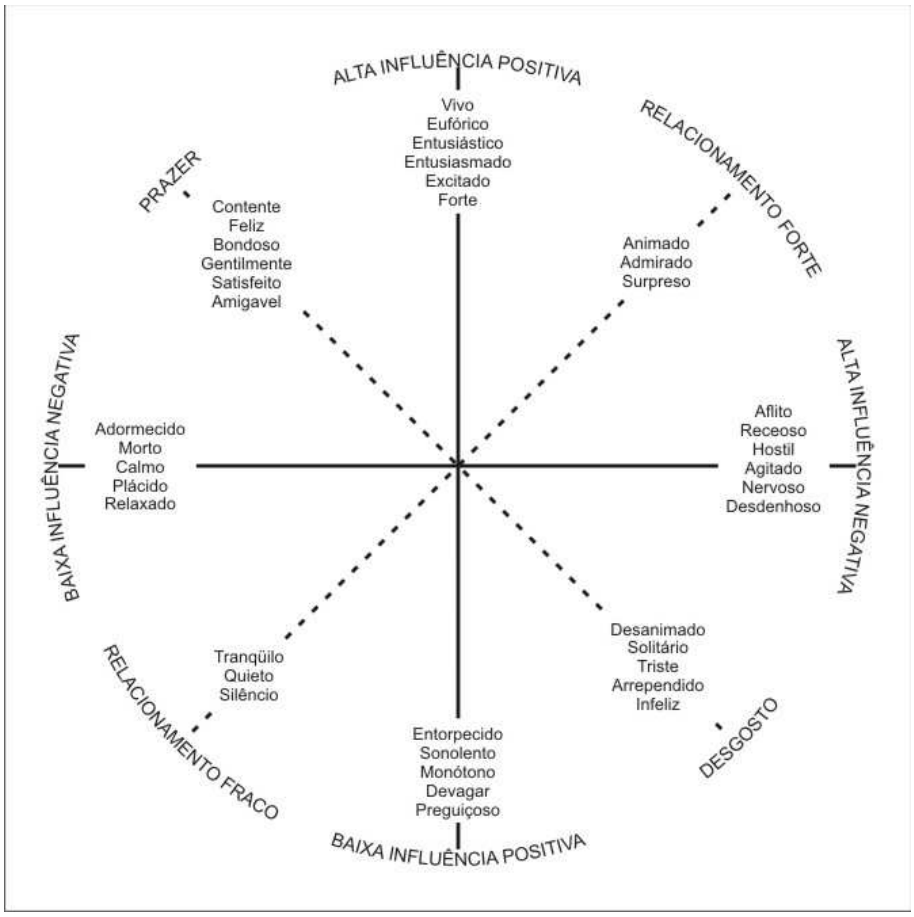


Figura 3: Circunferência afetiva proposta por Watson e Telegen, traduzido e adaptada de [RUBIN et al., 2004].

Antes que se possa automaticamente efetuar uma tentativa de discernimento de emoções em textos escritos é importante certificar-se de que as pessoas concordem, em sua maioria, no tipo de emoções identificáveis presentes nos documentos em análise. O estudo realizado por Watson e Tellegen acerca da estrutura de influências afirma que existem duas dimensões bipolares principais: influência positiva e influência negativa. As influências positivas refletem uma combinação de alta energia e avaliação positiva caracterizadas em emoções como alegria. Influências negativas compreendem sentimentos como aflição, transtorno e angústia. Ambas as influências ocorrem de forma bipolar e contínua, como pode ser observado na Figura 3, atingindo níveis altos e baixos.

O estudo foi realizado em duas etapas: a primeira, que consistia em um pré-teste que garantiu a usabilidade e a acessibilidade da base de dados; a segunda, que envolvia a base completa de textos e uma grande quantidade de participantes, aos quais foi solicitado que identificassem tipos de emoções, avaliassem a influência da presença dessas emoções e selecionassem indícios associativos nos textos.

A base de dados era composta por 100 textos, sendo que o número mínimo de palavras era 12, o máximo era 348 e a média, 86. Ao todo, foram avaliadas 679 sentenças. Todos os textos foram extraídos de diversos recursos públicos como, por exemplo, web-blogs e revisões de consumidores, por serem tipicamente escritos em primeira pessoa e conterem uma grande variedade de traços emocionais. Aspectos gramaticais e estilos pessoais de escrita foram preservados ao máximo.

Os exemplos continham de 2 a 21 sentenças cada. Isso foi delimitado pelo autor do texto ou, quando necessário, pelos autores do experimento em algo próximo a 20 frases. Esse limite foi estabelecido porque, para análises mais precisas, sentenças avulsas são melhores enquanto que textos contendo mais do que duas frases proporcionam um maior ruído no resultado final, aumentando assim as chances de erros.

Na segunda fase do experimento foi pedido aos participantes que associassem uma emoção da circunferência mostrada na Figura 3 ao texto lido. Em seguida, os resultados foram analisados e somente foram considerados aqueles textos que receberam pelo menos quatro opiniões iguais. Aqueles que alcançaram a maior porcentagem de concordância entre os participantes receberam maior prioridade na análise, e foram analisadas tanto a consistência como o comprimento do argumento anotado pelo usuário.

Ao final da segunda etapa do experimento, os idealizadores observaram que houve uma taxa de aproximadamente 70,7% dos textos cujas anotações feitas pelos participantes atenderam às exigências impostas para que o texto pudesse ser considerado para análises posteriores.

A conclusão retirada pelos autores foi de que a circunferência proposta por Watson e Tellegen, exibida na Figura 3 é bastante útil como um guia para o desenvolvimento de um algoritmo de PLN para identificações automáticas de emoções categorizadas em oito classes, ou seja, de um "*emotion-miner*".

### 2.4.2. Identificando Emoções em Textos em Inglês

O estudo realizado por [STRAPPARAVA e MIHALCEA, 2008] descreve um experimento baseado em uma grande base de dados (composta por manchetes de notícias) e seis emoções: raiva, desgosto, medo, alegria, tristeza e surpresa. O objetivo era avaliar métodos para a identificação automática de tais emoções em materiais escritos em inglês. Os autores afirmam que a detecção automática de emoções em textos está se tornando cada vez mais importante num ponto de vista prático. Segundo eles, ainda, pode-se tomar como exemplos tarefas como mineração de opinião, análise de mercado, ambientes de *e-learning* e jogos educativos.

A atividade consiste em classificar manchetes de notícias extraídas de páginas da internet e jornais. Estas manchetes, em sua maioria, além de curtas, são escritas com o objetivo de provocar emoções e, conseqüentemente, atrair a atenção de leitores. Por essas razões, elas foram escolhidas para a realização do experimento. A base de dados foi dividida em duas partes, sendo que a primeira continha 250 manchetes e foi utilizada para o desenvolvimento; e a segunda, com 1000 manchetes, foi empregada como conjunto de testes.

A experiência foi realizada de maneira não-supervisionada. Segundo os autores, o objetivo era o estudo de semânticas léxico-emocionais, evitando uma simples categorização dos textos por parte dos participantes. Entretanto, a restrição do treinamento supervisionado não foi um fator crítico ao desenvolvimento do trabalho. Nestes casos, os participantes foram autorizados a escolher e montar suas próprias regras de treinamento. Também, aos participantes foi dada a liberdade de escolha dos recursos a serem utilizados. Foi fornecido um conjunto de palavras relevantes às emoções escolhidas para a pesquisa, cujo uso pelos participantes era também totalmente optativo.

Para os testes, foram implementados cinco sistemas diferentes para análise de emoções, a saber: *WN-Affect-Presence*, que consiste em um sistema de referência para definir as emoções baseando-se apenas na presença ou não das palavras do dicionário léxico; *LSA* [LANDAUER et al., 2005], [DEERWESTER et al., 1998] *Single Word*, que calcula a similaridade LSA entre o texto analisado e cada emoção; *LSA Emotion Synset*, onde, além da palavra que denomina uma emoção, seus sinônimos armazenados em um dicionário também são usados; *LSA All Emotion Words*, que incrementa o conjunto a cada interação adicionando a ele todos os sinônimos encontrados num dicionário para uma determinada emoção; e, por último, *NB Trained on Blogs*, que nada mais é do que o classificador Naive Bayes treinado para uma base de registros extraídos de blogs.

Os resultados obtidos mostram que a eficiência de cada método varia. Por exemplo, o *LSA Single Word* mostrou ser o mais preciso, com baixo custo de processamento. Em contrapartida, os demais sistemas demandam maiores esforços computacionais, enquanto que a precisão se mostra inferior. Para análise de blogs, verificou-se que o sistema é bastante eficiente para sentimentos como Alegria e Raiva, uma vez que esses são mais facilmente encontrados nos posts, o que implica em um maior número de amostras dessas duas categorias. A Tabela 1, a seguir mostra a média de resultados obtidos com cada um dos métodos testados.

Tabela 1: Resultados obtidos no experimento (adaptado de [STRAPPARAVA e MIHALCEA, 2008])

| ***Método*** | ***Fine-Grained*** | | ***Coarsed-Grained*** | |
|---|---|---|---|---|
| | ***r*** | ***Precisão*** | ***Recall*** | ***F1*** |
| ***WN-Affect presence*** | 9.54 | 38.28 | *1.54* | 4.00 |
| ***LSA Single Word*** | 12.36 | 9.88 | *66.72* | 16.37 |
| ***LSA Emotion Synset*** | 12.50 | 9.20 | *77.71* | 13.38 |
| ***LSA All Emotion Words*** | 9.06 | 9.77 | *90.22* | 17.57 |
| ***NB Trained on Blogs*** | 10.81 | 12.04 | *18.01* | 13.22 |

Como pode ser concluído através dos dados apresentados na Tabela 1, os melhores resultados em *coarse-grained* foram obtidos com o método *LSA All Emotion Words*, que ofereceu os melhores resultados para *recall* (cobertura) e F-measure. Já na precisão, o melhor resultado foi obtido com o método *WN-Affect Presence*.

### 2.4.3. Aprendizagem de Máquina para Predição de Emoções em Textos

Segundo [ALM, 2005], os textos não são somente fonte de conteúdo, mas também de informação "atitudinal" e emocional. Isso pode ser fortemente observado na literatura, principalmente em contos de fadas, onde emoções como felicidade, raiva, amor e ódio são partes importantes das histórias. Quem lê uma história procura interpretar as emoções de forma a torná-la o mais interessante possível para quem ouve. De fato, quando da expressão oral, procura-se manifestar as emoções através de várias formas, como por exemplo, a entonação da voz.

Visando, através disso, transformar a leitura automática de texto em processos que soem mais naturais, é importante identificar e demonstrar o significado emocional adequado da passagem em questão. Para isso, uma aplicação que se destina a resolver esse problema deverá ser capaz de cumprir duas tarefas: reconhecer as emoções que mais se adéquam à passagem lida e, em seguida, saber como utilizar mecanismos de entonação de voz e de articulação das palavras para transmitir a emoção correta.

O objetivo principal desta pesquisa foi realizado em duas etapas: primeiro, classificar passagens e sentenças entre NEUTRA (sem emoções) e EMOCIONAL (que transmite algum tipo de emoção) e, por fim, em dividir o conjunto EMOCIONAL em duas partes, de tal forma que as sentenças que transmitem qualquer tipo de emoção seriam classificadas positiva ou negativamente. Dessa forma, no primeiro caso ter-se-ia um conjunto de casos de emoções E={N, E} e, no segundo caso, o conjunto E={N, EP, EN}, sendo N=neutro, E=emoção, EP=emoção positiva e EN=emoção negativa. Para esse trabalho, somente a existência e a característica de ser boa ou má foram consideradas porque o conjunto de dados disponível não era grande; este conjunto consistia em aproximadamente 185 contos infantis de autores como H.C. Andersen, Irmãos Grimm e B. Potter.

Como o treinamento foi supervisionado, foi necessária a pré-classificação das histórias do conjunto de treinamento. Os classificadores trabalharam em duplas sobre os textos, mas de forma independente, visando a não interferência da opinião de um sobre a avaliação do outro. Eles foram orientados a qualificar as sentenças de acordo com o ponto de vista encontrado no texto, ou seja, pela personagem que está submetida à emoção na passagem analisada.

Os testes foram feitos com rede neural Perceptron e Naive Bayes, sendo utilizada a abordagem *cross validation* fator 10% para testes. Os resultados obtidos com o Perceptron

não foram comentados no artigo porque, segundo os autores, os resultados foram piores. Com Naive Bayes, os resultados obtidos para o primeiro teste, onde E={N, E}, foram satisfatórios apesar da base analisada ser bastante reduzida, e podem ser observados na Tabela 2.

Tabela 2: Primeiros resultados (adaptado de [ALM, 2005])

| ***Medida*** | ***N*** | ***E*** |
|---|---|---|
| ***Precisão média*** | 66% | 56% |
| ***Recuperação média*** | 75% | 42% |
| ***F médio*** | 70% | 47% |

Foi observado que as emoções não podem ser facilmente classificadas e, por esse motivo, os resultados da segunda etapa da pesquisa, E={N, EP, EN}, foram inferiores. Observou-se que menos de 10% das sentenças foram classificadas como emocionalmente positivas, o que implicou em um pior resultado para a classe PE. Observou-se, além disso, que as passagens rotuladas de forma semelhante pelos revisores foram corretamente classificadas nas duas etapas do experimento. Tais resultados podem ver vistos na Tabela 3.

Tabela 3: Segundos resultados (adaptado de [ALM, 2005])

| ***Medida*** | ***N*** | ***NE*** | ***PE*** |
|---|---|---|---|
| ***Precisão média*** | 64% | 45% | 13% |
| ***Recuperação média*** | 75% | 27% | 19% |
| ***F médio*** | 69% | 32% | 13% |

### 2.4.4. Detecção de Emoções em Informação Textual Através de Rotulação Semântica e Técnicas de Web Mining

O experimento descrito por [LU et al., 2006] trata de um sistema que utiliza um classificador semântico para detectar emoções contidas em textos. Para isso, e visando a validação do ensaio, os autores utilizaram apenas a língua inglesa durante os testes. As ferramentas escolhidas estão disponíveis para acesso público e se tratam de um classificador semântico (desenvolvido na Universidade de Illinois) e um sistema de busca como o Google, que permite a busca por palavras-chave específicas. A função utilizada do Google foi a "*define*", que fornece definições atualizadas da internet acerca das palavras pesquisadas. As definições mais significativas, geralmente adjetivos, são armazenadas em uma tabela e

indexadas de acordo com suas classes gramaticais. Se o sistema detecta mais de uma palavra com a mesma definição, ele imediatamente usará o mesmo adjetivo para referenciá-las.

Este experimento, segundo seus autores, é diferente da maioria. Isso porque ele leva em consideração características semânticas para o auxílio na análise das sentenças. Nos demais, os sistemas somente são capazes de classificar a sentença de acordo com uma categoria de emoção através de palavras-chave. Um exemplo do funcionamento deste sistema é citado com a expressão: “Uma garota encontrou um tigre”. Através de análise semântica é possível descobrir a que classes gramaticais pertencem as palavras da sentença. Em seguida, sabe-se que o sujeito e o objeto (“uma garota” e “tigre”, respectivamente) estão ligados a adjetivos definidos através do sistema do Google. Como a palavra “garota” está relacionada a “jovem” e o objeto “tigre” a “predatório”, através da combinação desses dois adjetivos com o verbo “encontrar”, o resultado da sentença é “medo”.

Outra aplicação importante utilizada no decorrer do ensaio é o *Concept Net*, que serviu para extrair informações como localizações associadas aos objetos. Nesse caso, estas informações podem ser usadas para ajustar os planos de fundo, tanto no plano auditivo quanto no visual, de uma aplicação multimídia. Os testes da pesquisa foram realizados através de um programa para bate papo, onde, para cada mensagem trocada, o programa é capaz de alterar a imagem da janela de acordo com as emoções identificadas. No artigo aqui referenciado, não se encontrou resultados que indicassem a eficácia do método.

## 2.5. Discussão

Esse capítulo apresentou conceitos principais e fundamentais para o desenvolvimento do trabalho proposto, tais como a mineração de textos e como é possível a extração de informações úteis de bases textuais.

O capítulo seguinte apresenta o método para a identificação de emoções em textos curtos para o português desenvolvido para o presente trabalho.

# Capítulo 3

# Um Método Para Identificar Emoções em Notícias Curtas Para o Português do Brasil

O presente capítulo apresenta detalhadamente o método desenvolvido para realização da tarefa proposta. O método é baseado no conceito *Latent Semantic Analysis* (LSA), que será apresentado nos parágrafos seguintes. Entretanto, antes da explanação do método, faz-se necessária a introdução de dois conceitos que serão abordados ao longo do desevolvimento do trabalho.

## 3.1. Remoção de Termos Irrelevantes

*Stop word* é o termo utilizado para designar palavras consideradas irrelevantas num processo de mineração de textos. Considera-se que são palavras que ocorrem com muita frequência e, portanto, possuem significados menos importantes que palavras chaves; presume-se, então, que essas palavras podem ser filtradas durante alguma das etapas durante a mineração (em geral, no pré processamento), uma vez que sua permanência nos textos não é importante e só aumentaria tempo e esforço de processamento. Nesse ponto do processo, uma lista de palavras é criada manualmente pelo desenvolvedor do sistema, e a mesma será utilizada para excluir palavras que afetariam o desempenho do algoritmo [CARROLTON, 2002 ].

Embora essa relação de palavras seja algo variável de sistema para sistema, existem disponíveis algumas listas pré-definidas, com palavras que são consideradas irrelevantes na maior parte dos casos. Essas palavras são, em sua grande maioria, artigos, preposições, numerais, nomes de meses ou dias da semana. Exemplos de *stop words* são:

| |
|---|
| *à, ainda, ano, ao, as, às, bem, bom, brasileiro, centro, com, como, da, das, de, do, dos, depois, dia, e, é, ela, ele, em, então, entre, essa, esse, esta, está, estão, eu, há, já, mês, muito, na, não, nem, neste, no, num, número, para, pela, pelo, por, primeira, qual, se, sem, também, um, uma, você* |

No caso do presente trabalho, como já mencionado anteriormente, a lista de *stop words* foi retirada do site da Linguateca[2], pois a mesma contém as palavras mais comuns na língua portuguesa, independente da classe gramatical, e cobre facilmente as maiores ocorrências de termos encontrados em notícias jornalísticas.

Para remoção de *stop words* de um texto, foi utilizada uma ferramenta auxiliar, já desenvolvida e agregada no Weka. Essa ferramenta faz parte de uma maior, chamada *Rainbow*, que conta com um eficiente removedor de palavras irrelevantes. Com a ferramenta escolhida, é possível definir um arquivo personalizado de palavras consideradas irrelevantes e, através deste, editá-lo, visualizá-lo ou, no caso do presente trabalho, aplicá-lo para remoção das palavras indesejadas dos textos analisados.

## 3.2. Lematização

O algoritmo de *stemmer*, responsável pela lematização dos termos, efetua um processo que consiste em reduzir as palavras para os seus radicais. Pode-se dizer, então que a lematização nada mais é do que a redução de variações de uma mesma palavra a uma representação única, com o objetivo de aumentar o nível de recuperação de documentos. Pode-se citar, como exemplo, a família de palavras:

| |
|---|
| *terra, terrinha, terriola, térreo, terráqueo, terreno, terreiro, terroso* |

Para todas elas existe um elemento comum, chamado radical: ***terr-***.

Essa representação tem a intenção de isolar o semantema (ou radical: elemento portador de significado, comum a um grupo de palavras da mesma família) das palavras dos

[2] www.linguateca.pt

seus morfemas (elemento lingüístico que, isolado, não possui nenhum valor; serve somente para relacionar semantemas, definir a categoria gramatical etc) [VIEIRA e VIRGIL, 2007].

Atualmente, segundo [VIEIRA e VIRGIL, 2007], poucos algoritmos de lematização estão disponíveis para o idioma trabalhado nesse projeto (Português). Um deles, é o algoritmo de Porter, criado na década de 1980 e adaptado em 2005 para o português, que é baseado em regras e critérios. Sua execução está baseada nos seguintes passos:

1. Remoção dos sufixos;
2. Remoção dos sufixos verbais, se o primeiro passo não realizou nenhuma alteração;
3. Remoção do sufixo i, se precedido de c;
4. Remoção dos sufixos residuais os, a, i, o, á, í, ó;
5. Remoção dos sufixos e, é, ê e tratamento da cedilha.

A ferramenta escolhida para realizar esse procedimento no projeto é a *Snowball* [SNOWBALL], contida no Weka, e que possui uma versão do algoritmo de Porter adaptado para diversos idiomas, inclusive o Português.

## 3.3. Latent Semantic Analysis

*Latent Semantic Analysis* (LSA também chamado de *Latent Semantic Indexing* - LSI) é um método matemático/estatístico para identificação de relações entre palavras em textos [LANDAUER et al., 2005]. A partir dessas relações, visa-se estabelecer associações entre os termos encontrados [DEERWESTER et al., 1998]. Tradicionalmente, não é considerado como um método de PLN, pois não utiliza nenhum tipo de dicionário ou base confeccionada por humanos, redes semânticas, analisadores gramaticais, etc. Além disso, como única entrada são usados textos ou pequenas passagens [LANDAUER et al., 2005]. Entretanto, o LSA tem mostrado resultados satisfatórios em várias áreas de aplicação, entre elas a categorização de documentos com base em similaridades conceituais [BRADFORD, 2003].

Para o desenvolvimento do modelo, parte-se do pressuposto que as palavras sempre possuirão significados semelhantes e estarão inseridas dentro de contextos parecidos [DEERWESTER et al., 1998], [BRADFORD, 2003].

O primeiro passo para a execução do LSA é representar o texto através de uma matriz chamada de termo-documento (originalmente *term-document matrix*) onde cada linha representa uma única palavra e cada coluna representa um dos documentos, seja ele uma frase, parágrafo, etc. Essa matriz identifica a ocorrência de termos dentro de um conjunto de documentos. Dessa forma, cada célula da matriz contém a frequência com que cada palavra de determinada linha aparece na passagem de uma coluna qualquer.

Em seguida, cada célula será submetida a um cálculo preliminar onde a cada frequência será atribuído um peso. Este peso será, na verdade, um para a importância da palavra em cada documento (colunas) e outro para o conjunto total de documentos analisados. Por fim, é aplicado na matriz o teorema SVD (*Singular Value Decomposition*, ou decomposição em valor singular) para determinar padrões e relacionamentos entre os termos encontrados nos documentos. [LANDAUER et al., 2005], [BRADFORD, 2003].

A seguir, o conceito do teorema SVD será apresentado para, por fim, ser demonstrada a implementação do método utilizado no trabalho.

## 3.4. Singular Value Decomposition

*Singular Value Decomposition* (SVD) é um método utilizado para a fatorização de matrizes retangulares (de ordem *m* x *n*), comumente empregado em sistemas estatísticos ou de processamento de sinais. Essas fatorizações transformam a matriz analisada em uma série de aproximações lineares, que mostram a estrutura básica da matriz e facilitam a sua compreensão e análise. A finalidade do método consiste em obter uma decomposição da matriz A, de forma que ela possa ser descrita da seguinte forma:

$$A = U.S.V^T \tag{1}$$

Sendo:

A é a matriz que se deseja decompor;

U é uma matriz que descreve as linhas da matriz A; no caso desse trabalho, é a matriz que descreve os termos;

S é uma matriz diagonal que contém os “valores singulares” (*singular values*), que definem a relação entre os termos de U e $V^T$. Eles podem ser vistos como “controle de ganho”;

$V^T$ é a matriz V, transposta, que descreve as colunas da matriz A; no caso desse trabalho, á a matriz que descreve os documentos analisados.

O cálculo do SVD consiste em encontrar os autovalores e os autovetores de $AA^T$ e $A^TA$, onde os autovetores de $A^TA$ constroem as colunas de V, e os autovetores de $AA^T$ definem as colunas de U. Os valores de S são obtidos através das raízes quadradas dos autovalores de $AA^T$ e $A^TA$.

Para exemplificar como o método funciona, será feita uma adaptação da explicação existente em [MIT, 2002]

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Para se encontrar os autovalores da matriz, faz-se os cálculos de $AA^T$ e $A^TA$. Como já foi mencionado, dos autovetores de $AA^T$ extrai-se a matriz U. Então, pode-se fazer a seguinte análise:

$$AA^T = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 20 & 14 & 0 & 0 \\ 14 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Agora que se tem uma matriz quadrada, pode-se obter os autovalores de $AA^T$ e, como resultado, obtém-se a matriz U, como:

$$U = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

O processo é repetido para $A^TA$, de forma a obter-se a matriz V:

$$A^TA = \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.40 & -0.91 \\ 0.91 & 0.40 \end{bmatrix}$$

E, finalmente, S é obtido através das raízes dos autovalores de $AA^T$ e $A^TA$, como já foi mencionado:

$$S = \begin{bmatrix} 5.47 & 0 \\ 0 & 0.37 \end{bmatrix}$$

## 3.5. Método Estudado e Desenvolvido para a Identificação de Emoções

O objetivo desta pesquisa é identificar automaticamente emoções de um conjunto de notícias curtas em português. As informações identificadas serão posteriormente utilizadas para realizar a animação de um avatar que pode, por exemplo, fazer o papel de um apresentador de um noticiário televisivo.

Para fins da presente pesquisa, define-se como "notícia curta" um texto que, em geral, antecede a reportagem em si em qualquer noticiário. Este texto tem um comprimento limitado (em número de palavras) e pode ser entendido como a manchete e sua respectiva linha fina, que aparecem destacadas no texto e antecedem a explanação da notícia em si. A manchete possui como objetivo atrair a atenção do leitor para o texto e, por essa razão, é bastante destacada com relação ao restante do conteúdo. Como título de apoio, tem-se as "linhas finas", que estão posicionadas imediatamente após a manchete, ainda destacadas, mas menos com relação à manchete em si, e possuem a função de fornecer melhor explicação acerca do título principal (manchete) [CAMPOS].

Como exemplo de manchete e linha fina, tem-se o texto a seguir: "Dois morrem e um fica ferido após carro cair uma altura de 10 metros" (manchete) "Acidente aconteceu neste sábado em Caxias do Sul. Veículo caiu com as rodas para cima em uma represa vazia" (linha fina).

Para a identificação das emoções de notícias curtas, foi implementado um procedimento baseado no algoritmo LSA [YU et al., 2002], utilizando a linguagem de programação Java, e o ambiente Eclipse[3]. Com o objetivo de ilustrar melhor todo o processo, será feito uso da Figura 2, que explica o procedimento básico padrão utilizado em sistemas de mineração de textos. A mesma será, aqui, decomposta em seus diversos blocos, a fim de se realizar uma melhor apresentação do processo. Para melhor compreensão, ao longo desta seção, será desenvolvido um pequeno experimento explicativo, com um pequeno subconjunto de notícias. Para tal finalidade, tomar-se-ão as seguintes notícias, apresentadas na Tabela 4.

[3] Disponível em http://www.eclipse.org/

Tabela 4: Notícias utilizadas para apresentar o processo.

| *ID* | *Notícia Curta* |
|---|---|
| *1* | Presidente do TCE no RS deixa hospital após assalto: João Vargas foi esfaqueado na barriga e passa bem em São Sepé. Dois suspeitos foram presos pela polícia na cidade de Santa Maria. |
| *2* | Dois morrem e um fica ferido após carro cair uma altura de 10 metros: Acidente aconteceu neste sábado em Caxias do Sul. Veículo caiu com as rodas para cima em uma represa vazia. |
| *3* | PRF flagra menina de 12 anos dirigindo picape em rodovia gaúcha: Segundo os policiais, jovem estava acompanhada da mãe em São Borja. Carro foi retido e a mãe autuada por deixar um menor de 18 anos dirigir. |
| *4* | Sucuri de 8 metros é flagrada após comer uma capivara: Cobra virou atração para moradores de São José do Rio Claro (MT). Ela foi encontrada em um riacho perto de uma fazenda na cidade. |
| *5* | Perito particular questiona imagens de acidente com ex-deputado no PR: Família de vítima contratou profissional para fazer simulação da colisão. Segundo ele, alguns segundos do filme do acidente foram removidos. |
| *6* | Depois da guerra, Faixa de Gaza vira 'ilha à deriva': Quatro meses após ataque de Israel, território palestino segue pressionado. Mas região deve voltar a ser o foco do processo de paz no Oriente Médio. |
| *7* | Prédio desaba e fere ao menos 3 na Bélgica: Acidente aconteceu durante festa local de Ducasse de Doudou. Para bombeiros, pode haver feridos ou mortos nos escombros. |

Esse subconjunto de notícias corresponde à primeira parte da Figura 2, ou seja, o corpus de documentos, como mostra a Figura 4, a seguir.


Figura 4: Etapas de processamento em andamento ou atingidas: documentos de texto.

Tendo-se os documentos a serem utilizados, a primeira etapa a ser realizada, como mostra a Figura 5, consiste em um pré-processamento das notícias curtas.

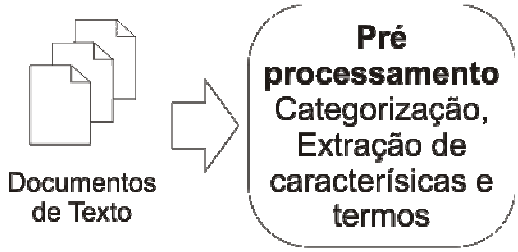


Figura 5: Etapa em andamento no projeto: pré-processamento.

Essa etapa engloba as seguintes tarefas: ler o arquivo original de notícias, converter todos os caracteres em minúsculos, remover caracteres especiais (como pontuação, hífens e números), remover *stop words*[4] e aplicar um algoritmo para realizar a lematização das palavras restantes. O processo de lematização serve para remover os sufixos das palavras, reduzindo, assim, os termos aos seus radicais. Com sua utilização, os termos derivados de um mesmo radical serão contabilizados como um único termo. Por exemplo: ao encontrar-se as palavras guerra e guerrear, as mesmas serão reduzidas a "guerr"; assim sendo, sempre que uma dessas duas for encontrada em um documento, o contador de guerr será incrementado para tal notícia. Essas duas últimas tarefas foram aplicadas com o auxílio da ferramenta Weka [WEKA], que conta com um algoritmo para remoção de *stop words* e com uma extensão do *Snowball Stemmer* [SNOWBALL]. Estes recursos foram escolhidos por serem facilmente integráveis ao projeto em desenvolvimento e pela disponibilidade de configuração para o português.

Terminados os procedimentos descritos acima, ter-se-á uma nova coleção de notícias, já preparadas para o processamento em si, como mostra a Figura 6, abaixo.

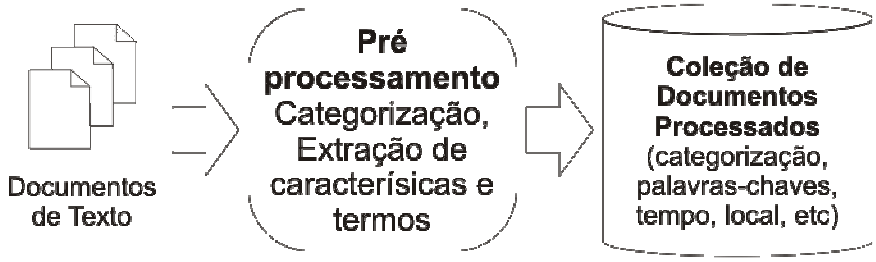


Figura 6: Concluído o pré-processamento, tem-se uma coleção de documentos processados.

A

Tabela 5, a seguir, apresenta o resultado do pré-processamento para o exemplo apresentado na Tabela 4.

[4] Lista de *stop* words extraída do site Linguateca: http://www.linguateca.pt/

Tabela 5: Notícias da Tabela 4 após o pré processamento.

| *ID* | ***Notícia Curta Processada*** |
|---|---|
| *1* | tce rs deix hospital assalt joã varg esfaqu barrig pass sep suspeit pres sant mar |
| *2* | morr fic fer carr cair altur metr acident acontec cax veícul caiu rod cim repres vaz |
| *3* | prf flagr menin dirig picap rodov gaúch polic jov acompanh mã borj carr ret mã autu deix menor dirig |
| *4* | sucur metr flagr com capiv cobr vir atraçã morador clar mt encontr riach pert fazend |
| *5* | perit particul question imagens acident ex deput pr famíl vítim contrat profissional simul colisã segund acident remov |
| *6* | guerr faix gaz vir ilha deriv ataqu israel territóri palestin seg pression volt foc paz orient médi |
| *7* | prédi desab fer bélgic acident acontec fest duc doud bombeir hav fer mort escombr |

Após essa primeira etapa, segue-se para as tarefas de mineração em si, como descreve o próximo passo da mineração de textos, ilustrado na Figura 7. Essa é a parte mais longa e complexa de todo o processo.

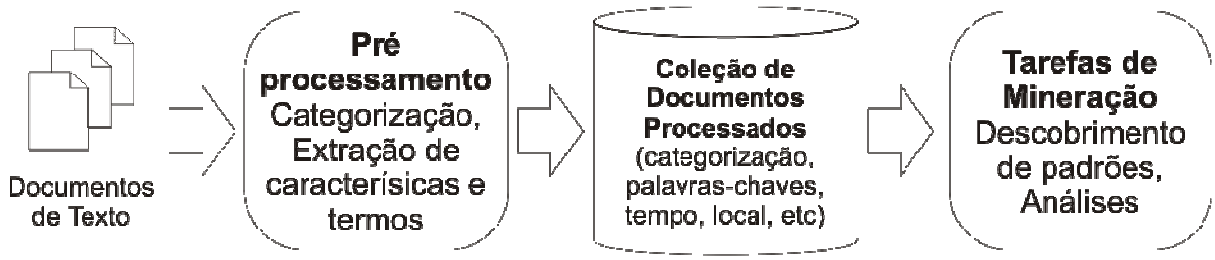


Figura 7: Etapa em andamento (tarefas de mineração)

Para isso, inicialmente dois vetores são gerados. O primeiro conterá todas as notícias curtas (documentos) já processadas. A partir desse, gera-se um segundo, que contém todas as palavras (termos) encontradas em todo conjunto de notícias curtas, sem repetições de termos. Com esses dois vetores, uma matriz (*term-document matrix*) é gerada e, inicialmente, instanciada com zeros. Cada linha dessa matriz corresponde a uma palavra e cada coluna corresponde a um documento. Portanto, cada célula corresponde ao número de ocorrências de um termo dentro de um determinado documento. Nessa etapa, são eliminados os termos que

aparecem somente uma vez em todo conjunto de documentos. A Tabela 6, a seguir, mostra a *term-document matrix* gerada para o exemplo, onde D1, D2, D3,... são as notícias curtas citadas na Tabela 4. Vale lembrar que os termos observados a seguir passaram pelo processo de lematização citado anteriormente, e encontram-se reduzidas ao radical de cada palavra. Esse processo é extremamente necessário para que termos que possuem significados semelhantes e mesmo radical não sejam repetidos, prejudicando os resultados.

Tabela 6: *Term-document matrix* gerada para o exemplo.

| | ***D1*** | ***D2*** | ***D3*** | ***D4*** | ***D5*** | ***D6*** | ***D7*** |
|---|---|---|---|---|---|---|---|
| ***acident*** | 0 | 1 | 0 | 0 | 2 | 0 | 1 |
| ***acontec*** | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| ***carr*** | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| ***deix*** | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| ***dirig*** | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| ***fer*** | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| ***flagr*** | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| ***metr*** | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| ***mã*** | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| ***vir*** | 0 | 0 | 0 | 1 | 0 | 1 | 0 |

A partir da matriz termo-documento, um segundo procedimento é iniciado, também com o auxílio da ferramenta Weka. Esse consiste na aplicação de uma técnica chamada *Singlular Value Decomposition* (SVD). O motivo que leva a utilizar o SVD consiste em encontrar uma representação reduzida (menor número de dimensões) da matriz termo-documento, que enfatize padrões e as ligações mais fortes entre termos e/ou documentos, e descarte as mais fracas, ou ruídos. A execução do SVD decompõe a matriz principal (matriz termo-documento) em três outras. A primeira, chamada U, nos remete a coordenadas de cada termo dentro de um espaço.

A Tabela 7, a seguir, mostra um extrato da matriz U, com as cinco primeiras dimensões e as sete primeiras palavras do conjunto de termos.

Tabela 7: Parte da matriz U obtida

| | | | | | |
|---|---|---|---|---|---|
| ***acident*** | *0.3306* | *0.5410* | *0.2103* | *-0.7412* | *0.0633* |
| ***acontec*** | 0.2182 | 0.3307 | -0.0311 | 0.3146 | -0.1315 |
| ***carr*** | 0.3534 | -0.0282 | -0.0377 | 0.0669 | -0.5667 |
| ***deix*** | 0.2581 | -0.1928 | 0.1013 | 0.0046 | 0.1637 |

| | | | | | |
|---|---|---|---|---|---|
| ***dirig*** | 0.4723 | -0.3481 | 0.1445 | 0.0059 | 0.0755 |
| ***fer*** | 0.3191 | 0.5156 | 0.0476 | 0.5652 | 0.3414 |
| ***flagr*** | 0.2769 | -0.1780 | -0.3898 | -0.0869 | 0.1502 |
| ***metr*** | 0.1580 | 0.1419 | -0.5720 | -0.0259 | -0.4920 |
| ***mã*** | 0.4723 | -0.3481 | 0.1445 | 0.0059 | 0.0755 |
| ***vir*** | 0.0446 | -0.0043 | -0.6481 | -0.1395 | 0.4880 |

Se a referida matriz U fosse representada em um espaço bidimensional, ela poderia ser facilmente observada como mostra a Figura 8, onde cada quadrado representa um termo do conjunto.

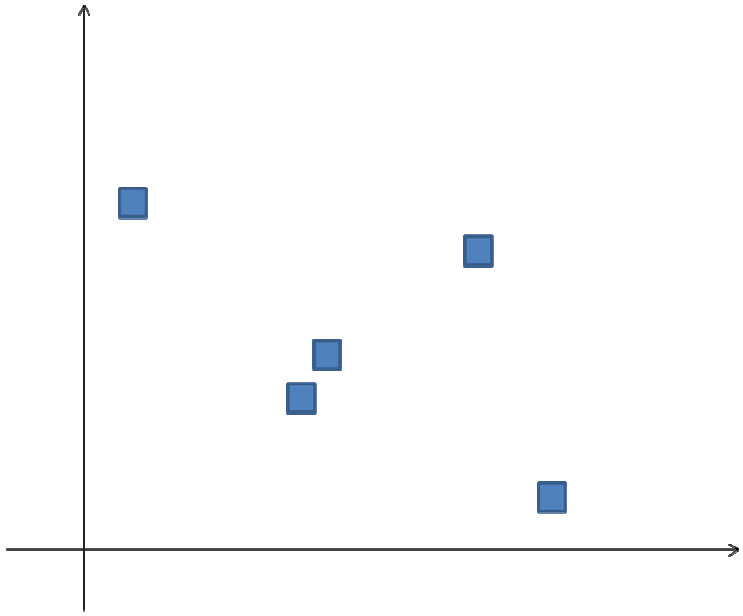

Figura 8: Exemplo de representação gráfica da matriz U.

A segunda matriz, aqui chamada de $V^T$, fornece as coordenadas dos documentos nesse mesmo espaço. A Tabela 8, a seguir, mostra um pedaço da matriz $V^T$, com as cinco primeiras dimensões e os sete primeiros documentos do conjunto de notícias (D1, D2, D3, ...).

Tabela 8: Parte da matriz $V^T$ obtida.

| ***D1*** | ***D2*** | ***D3*** | ***D4*** | ***D5*** | ***D6*** | ***D7*** |
|---|---|---|---|---|---|---|
| 0.0752 | 0.4022 | 0.8099 | 0.1398 | 0.1928 | 0.0130 | 0.3461 |
| -0.0601 | 0.4679 | -0.5584 | -0.0126 | 0.3372 | -0.0014 | 0.5931 |
| 0.0543 | -0.2051 | 0.1348 | -0.8625 | 0.2253 | -0.3472 | 0.1470 |
| 0.0027 | 0.1072 | 0.0050 | -0.1506 | -0.8846 | -0.0833 | 0.4200 |
| 0.1436 | -0.6891 | 0.0430 | 0.1282 | 0.1110 | 0.4280 | 0.5392 |

Analogamente ao processo adotado para a matriz U, se a matriz $V^T$ em questão fosse representada através de um espaço bidimensional, a mesma poderia ser observada como mostra a FIG, onde cada círculo representa um documento do conjunto.

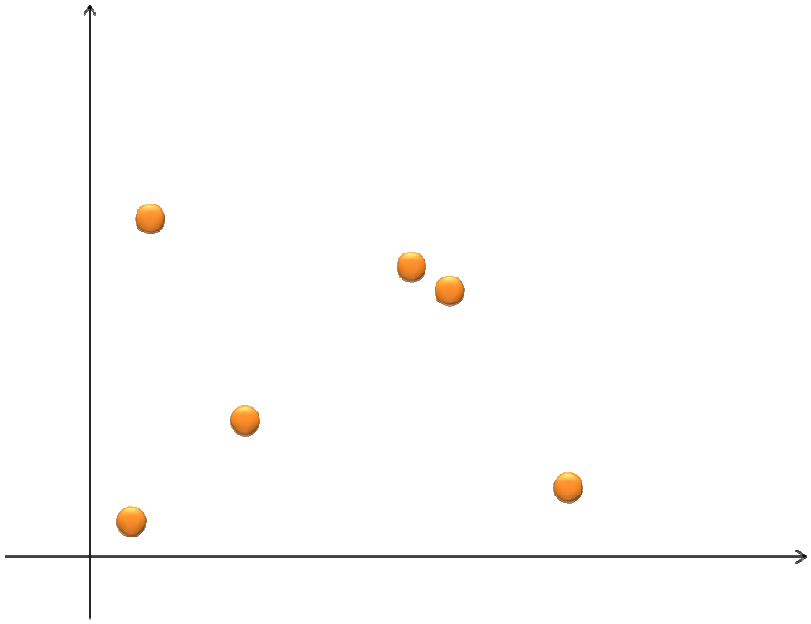

Figura 9: Exemplo de representação gráfica da matriz $V^T$.

Finalmente, a terceira matriz, chamada S, permite estimar quantas dimensões deverão ser utilizadas para a obtenção dos melhores resultados. Essa dimensão (chamada K), segundo Grossman e Frieder [GROSSMAN e FRIEDER, 2004], pode ser estimada arbitrariamente, através de experimentos e comparações. No exemplo que se apresenta no presente trabalho, adota-se apenas cinco dimensões, pois o objetivo é somente ilustrar o funcionamento do algoritmo. Entretanto, para os testes realizados com uma base maior de notícias, adotou-se 50 dimensões. Tal número foi escolhido porque, segundo a literatura, os melhores resultados são obtidos empiricamente, com valores de K não maiores que 100; isso varia de acordo com cada base utilizada e, segundo Grossman e Frieder [GROSSMAN e FRIEDER, 2004], resultados com K superior a 100 não demonstram melhorias significativas nos experimentos realizados para textos escritos em inglês.

A etapa seguinte consiste em definir a localização de cada grupo (emoção) no mesmo espaço criado anteriormente com o SVD. Para isso, fez-se necessária a utilização de seis listas de palavras, sendo cada uma delas relacionada com uma emoção básica descrita anteriormente. Estas listas foram inicialmente disponibilizadas por Strappavara e Mihalcea [STRAPPARAVA e MIHALCEA, 2008], em seis arquivos diferentes (um para cada

emoção), originalmente em inglês. Foi feita, então, a tradução das palavras contidas em cada um dos arquivos, de forma a termos palavras diretamente ligadas a emoções em português. Tal conjunto de palavras pode ser considerado equivalente ao utilizado na terceira metodologia de avaliação utilizada por [STRAPPARAVA e MIHALCEA, 2008], ou seja, LSA Emotion Synset, pois consiste apenas de uma lista fixa de palavras sinonimamente relacionadas a cada emoção. A Tabela 9 contém alguns exemplos para cada emoção e, em sua última linha, apresenta a quantidade total de palavras constantes de cada emoção utilizada.

Tabela 9: Exemplos de palavras contidas nas listas de emoções.

| **Emoção** | ***Alegria*** | ***Desgosto*** | ***Medo*** | ***Raiva*** | ***Surpresa*** | ***Tristeza*** |
|---|---|---|---|---|---|---|
| **Exemplos** | amor<br>amizade<br>brincadeira<br>esperança<br>engraçado | enjoo<br>feio<br>náusea<br>nojo<br>sujo | assombrado<br>cruel<br>medroso<br>pânico<br>terror | assassinar<br>cólera<br>destruir<br>diabólico<br>irritar | deslumbrar<br>embasbacar<br>fantástico<br>pasmo<br>susto | arrepender<br>chorar<br>derrota<br>desamparo<br>luto |
| **Quantidade** | 278 | 72 | 104 | 168 | 40 | 184 |

Como pode-se observar, há uma discrepância entre a quantidade de termos presentes em cada lista de emoções. Entretanto, como será avaliado mais tarde, estes números não afetaram os resultados do experimento.

Na sequência, para cada emoção, busca-se todas as palavras da lista de emoção analisada na lista de termos do nosso conjunto de notícias. Sabendo-se todos os termos que, segundo nossas listas, representam emoções e, com o auxílio da matriz U, calculamos um ponto médio no espaço obtido anteriormente. Esse ponto médio define precisamente a localização do grupo naquele espaço. Como o conjunto de documentos utilizado para exemplificar o algoritmo foi bastante reduzido, não foi possível construir a matriz de centróides. Dessa forma, a mesma será ilustrada na Tabela 10, que mostra parte da matriz gerada com os experimentos reais. Nela, estão exibidas as localizações dos grupos, contendo as seis primeiras coordenadas de cada um deles.

Tabela 10: Parte da matriz de coordenadas de localização dos grupos (centróides).

| ***Alegria*** | ***Desgosto*** | ***Medo*** | ***Raiva*** | ***Surpresa*** | ***Tristeza*** |
|---|---|---|---|---|---|
| 0.0198 | 0.0054 | 0.0022 | 0.0011 | 0.0000 | 0.0035 |
| 0.0136 | -0.0103 | -0.0024 | -0.0009 | 0.0000 | -0.0052 |
| 0.0030 | 0.0088 | -0.0014 | -0.0005 | 0.0000 | 0.0009 |
| 0.0068 | -0.0027 | 0.0000 | -0.0003 | 0.0000 | -0.0023 |

| | | | | | |
|---|---|---|---|---|---|
| -0.0018 | 0.0011 | 0.0025 | 0.0040 | 0.0000 | 0.0114 |
| -0.0031 | 0.0010 | 0.0004 | -0.0028 | 0.0000 | -0.0139 |

Uma representação gráfica pode ser observada na FIG, onde os termos referentes às emoções são representados pelos quadrados, e as emoções são representadas através das circunferências que cercam cada grupo de emoções.

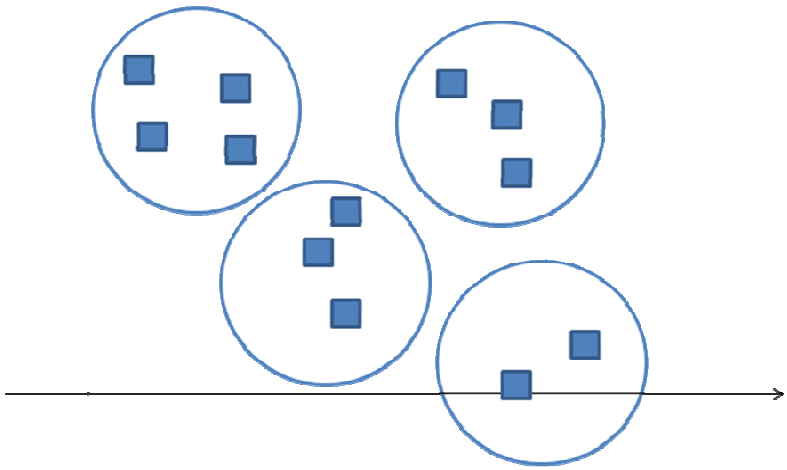

Figura 10: Exemplo de representação gráfica dos grupos de emoções, segundo as listas de emoções.

Por fim, para a última etapa, é utilizada a matriz $V^T$ e o conjunto de coordenadas dos grupos de emoções. Através da similaridade Cossenoidal [GARCIA, 2006], determina-se a distância de cada notícia curta aos grupos definidos, como mostra a equação (2), a seguir:

$$Sim(Dn, Gm) = \frac{\sum_i W_{D_n i} * W_{G_m i}}{\sqrt{\sum_j W_{D_n j}{}^2} * \sqrt{\sum_j W_{G_m j}}} \quad (2)$$

Sendo:

$D_n$ corresponde ao documento n em análise;

$G_m$ corresponde ao grupo sendo considerado;

$W_{Dni}$ e $W_{Dnj}$ são as coordenada i e j do documento n, respectivamente;

$W_{Gmi}$ e $W_{Gmj}$ são as coordenadas i e do grupo m, respectivamente.

Os resultados obtidos através da equação (2) complementam a última fase obrigatória do processo, como ilustrado na Figura 11.

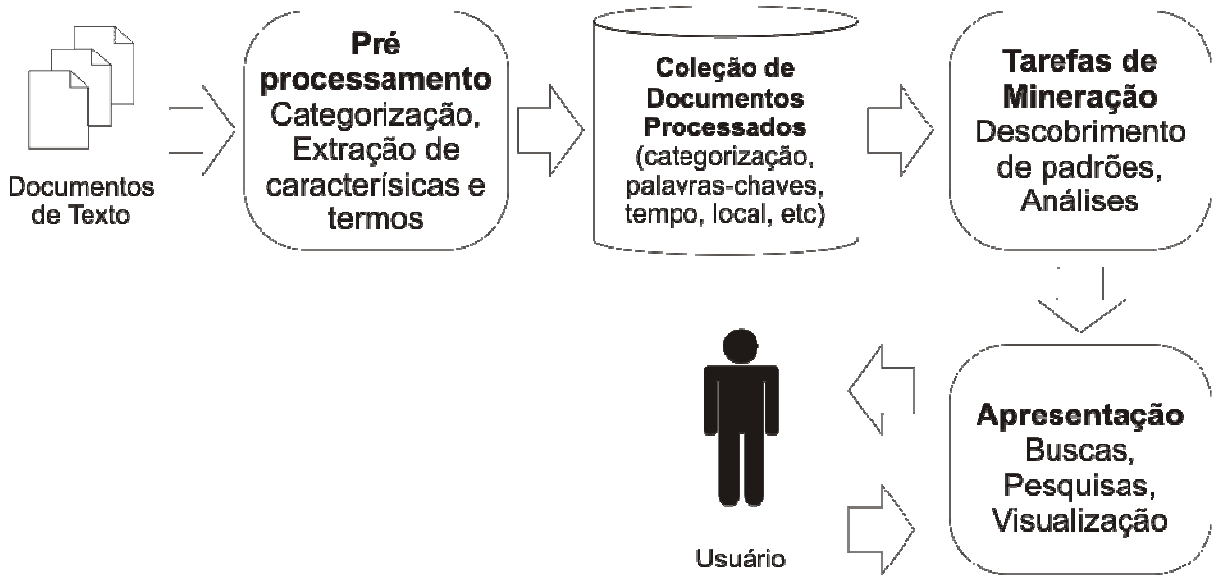


Figura 11: Conclusão das etapas obrigatórias do método.

Por se tratar de uma aproximação cossenoidal, os valores resultantes do cálculo estão compreendidos entre -1 e 1, em que um resultado 1 ou -1 (um, positivo ou negativo) representa que os dois pontos são totalmente idênticos com relação à sua localização no espaço e, 0 (zero) por sua vez, representa que os dois pontos são totalmente distantes entre si. Para facilitar a análise dos resultados, os valores obtidos através da similaridade cossenoidal, como resultados, foram multiplicados por 100. Dessa forma, os resultados exibidos nesse experimento variam, em módulo, entre 0 e 100 (ou seja, de -100 a 100).

De posse desses resultados gerados pelo algoritmo, observa-se, primeiramente, o maior número em módulo. Toma-se, por exemplo, uma notícia cujo maior resultado, em módulo está na coluna da Tristeza. Ou seja, tem-se que o algoritmo identificou, nessa notícia, a emoção Tristeza. Em seguida, pode-se optar por decidir uma emoção secundária nesse processo. Para tanto, procura-se pelo segundo maior (ou menor, caso o primeiro seja negativo) número, dessa vez respeitando-se o sinal da primeira emoção encontrada. Nesse caso, a emoção secundária seria o segundo maior número na linha da notícia em questão.

Essa medida de se respeitar o sinal para avaliar as emoções posteriores à primeira se deve ao fato de que os resultados são obtidos através de aproximações trigonométricas, e os

resultados são exibidos em valores entre 1 e -1. Sendo assim, sabe-se que, no caso do cosseno, -1 e 1 encontram-se em quadrantes distintos, embora possuam a mesma distância. Observou-se que, em casos onde, em módulo, as emoções encontradas como primária e secondária fossem Alegria e Tristeza, antagônicas entre si, cada uma possuía sinais distintos, ou seja, se Alegria fosse positivo, Tristeza seria negativo. Visando evitar que uma notícia fosse classificada ao mesmo tempo com duas emoções contrárias, então tomou-se esssa medida de respeitar os sinais a partir da segunda emoção avaliada.

Os resultados obtidos com a execução do método serão demonstrados e analisados posteriormente, num capítulo destinado especificamente à análise dos resultados obtidos (Capítulo 4).

## 3.6. Avaliação de Uma Nova Notícia

O método descrito anteriormente diz respeito à elaboração de um "espaço" para a avaliação e agrupamento de notícias. Esse procedimento avalia um número pré determinado de textos (corpus de, nesse caso, 700 notícias curtas como apresentado no capítulo seguinte) e, com base nisso, gera resultados para esses textos. Entretanto, esse procedimento pode ser custoso e demorado. Então, para que a aplicação tenha uma finalidade mais justificável e rápida, ela precisa ser capaz de, dada uma nova notícia, identificar a qual grupo ela pertence, com referência naquilo que foi definido anteriormente, sem que para isso todo o processo descrito no item 3.5 precise ser refeito. Para isso, tal procedimento será apresentado na sequência, considerando-se três situações distintas: uma notícia corretamente classificada; uma com um resultado parcialmente correto; e uma terceira com o resultado incorreto. As notícias estão listadas na Tabela 11:

Tabela 11: Novas notícias avaliadas pelo método

| ***ID*** | ***Notícia Curta*** |
|---|---|
| *1* | Novo terremoto volta a sacudir região central da Itália: Tremor ocorreu na região de Abruzzos. Área atingida fica próxima à cidade de L'Aquila. |
| *2* | Cientistas usam palavras em milhões de blogs para monitorar felicidade: Dupla da Universidade de Vermont deu peso para cada expressão. Dia da eleição de Barack Obama foi o mais feliz em 4 anos. |
| *3* | Lula quer reunião com laboratórios sobre vacina para gripe suína: Líderes do Mercosul se disseram preocupados com distribuição de remédios. |

Primeiramente, a nova notícia a ser avaliada passará por todo o pré processamento, que envolve a remoção de algarismos, caracteres especiais, remoção de *stop* words e a

lematização. Concluído esse processo, pode-se ver os resultados para as notícias citadas na Tabela 11 logo abaixo, na Tabela 12:

Tabela 12: Pré processamento das novas notícias.

| ***ID*** | ***Notícia Curta Processada*** |
|---|---|
| *1* | terremot volt sacud itál tremor ocorr abruzz ating fic próxim l aquil |
| *2* | cientist usam palavr blogs monitor felic dupl univers vermont deu pes expressã eleiçã barack obam feliz |
| *3* | lul reuniã laboratóri vacin grip suín líd mercosul diss preocup distribuiçã remédi |

Com o pré processamento concluído, é então montado um vetor *q*, que imaginariamente será anexado à *term-document matrix* mostrada na Tabela 6, visando o cálculo e a localização da notícia dentro do espaço definido anteriormente. Esse vetor *q* pode ser explicado como uma nova coluna da matriz termo x documento, onde cada item desse vetor, assim como as demais, corresponderá à ocorrência da palavra em questão no vetor de termos também formado anteriormente.

Em seguida, a nova notícia precisa ser inserida no espaço criado através da técnica SVD. Para tanto, o procedimento adotado é realizado através da seguinte formulação:

$$V = q^T . U_K . S_K^{-1} \quad (3)$$

Sendo:

V corresponde ao vetor de coordenadas da notícia no espaço criado;

$q^T$ é o vetor (transposto) correspondente à nova coluna da matriz termo x documento, contendo a freqüência de ocorrência dos termos na nova notícia;

$U_K$ é a matriz U, reduzida às K dimensões adotadas na primeira etapa do desenvolvimento. Nesse caso, 50;

$S_K$ é a matriz S, que até o presente momento não havia sido utilizada, reduzida também às K dimensões adotadas na primeira etapa do desenvolvimento.

Concluído esse cálculo, obter-se-á um novo vetor que contém as coordenadas da notícia nova no espaço criado anteriormente. E, com este, basta realizar o procedimento de similaridade cossenoidal já descrito para obter-se a proximidade da notícia analisada com cada um dos seis grupos de emoção.

No caso das notícias aqui exemplificadas, obteve-se os seguintes resultados exibidos na Tabela 13, onde os resultados estão destacados em negrito:

Tabela 13: Emoções identificadas nas novas notícias avaliadas

| *ID* | *Alegria* | *Desgosto* | *Medo* | *Raiva* | *Surpresa* | *Tristeza* |
|---|---|---|---|---|---|---|
| **1** | -2.0569 | **20.7262** | -2.4197 | -3.2550 | 0.7031 | 17.0053 |
| **2** | 0.9345 | 2.2244 | 3.3491 | 1.4608 | 0.6695 | **-3.7151** |
| **3** | 3.5566 | **-16.2259** | 6.6025 | -0.9727 | 1.2290 | 2.3518 |

Avaliando-se esses resultados da mesma forma apresentada no item 3.5, observa-se que o algoritmo identificou as emoções Desgosto, Tristeza e, novamente, Desgosto, respectivamente. Observa-se que, no primeiro caso, o algoritmo identificou corretamente a emoção transmitida pela notícia. Um terremoto significa uma fatalidade e, muito provavelmente, tristeza para os atingidos e seus familiares. No segundo caso, tendo em vista a notícia 2 da Tabela 11, vê-se que, a princípio, não se trata de uma notícia que de fato transmita alguma emoção, visto que é algo relacionado a pesquisas tecnológicas. Nesse caso, o algoritmo detectou índices baixíssimos para todas as emoções, o que a colocaria longe de todos os grupos, mas pendendo para a tristeza. Isso não poderia ser considerado correto, pois a notícia trata de um avanço tecnológico. Em um possível trabalho futuro, para solucionar-se esse problema, uma avaliação dos valores mínimos necessários em cada emoção pode ser estudada, e uma nova categoria pode ser incluída: a de notícias consideradas neutra, ou isentas de emoção. No terceiro caso, a emoção encontrada na notícia é errada, pois uma preocupação acerca da gripe suína não denota necessariamente desgosto. Tal erro deu-se devido a uma limitação do algoritmo, que considera que as palavras terão sempre o mesmo significado e se relacionarão entre si sempre num mesmo contexto. Como na época de coleta das notícias, muitas das encontradas diziam respeito à crise da gripe H1N1, o algoritmo criou uma forte relação entre os termos “gripe” e “suína”, juntamente com a emoção Desgosto, o que gerou o problema aqui encontrado.

## 3.7. Discussão

Depois do primeiro levantamento teórico, realizado no capítulo anterior, esse capítulo teve como finalidade apresentar em detalhes o método desenvolvido para realizar o objetivo proposto. Primeiramente, foi introduzida a idéia e o conceito principal que embasa a técnica,

ou seja, o LSA. Em seguida, uma explicação detalhada do procedimento foi feita, visando uma melhor compreensão do processo como um todo. Por último, foi apresentada a forma como são analisadas novas notícias que não tenham sido utilizadas na construção do método.

O capítulo seguinte apresenta de forma mais detalhada quais foram as ferramentas utilizadas para a implementação do método.

# Capítulo 4

# Implementação do Método

O presente capítulo tem por objetivo descrever as ferramentas utilizadas para a implementação do método descrito anteriormente. Serão comentadas características de plataformas, ferramentas e extensões utilizadas no decorrer das etapas de execução do projeto.

## 4.1. Coletar Notícias

As notícias utilizadas no desenvolvimento do projeto foram extraídas do site www.globo.com. Para facilitar o processo de obtenção desses textos, fez-se necessária a utilização de uma ferramenta. Atualmente, todo site que possui atualizações frequentes, tais como blogs e sites de notícias, fornecem em seus serviços uma maneira de se inscrever chamada *Feed*. Tais feeds, escolhidos pelo usuário, podem ser agrupados em ferramentas chamadas de Agregadores, que tem por objetivo reunir as informações escolhidas e exibir suas atualizações numa única tela ou página da *web*. Na maioria dos casos, os agregadores exibem o título do post e um pequeno extrato do texto (configurado pelo autor do blog); ou a manchete da notícia, seguida por sua linha fina.

Para a aquisição das notícias utilizadas nesse trabalho, instalou-se a ferramenta FeedReader[5], que é um agregador simples de ser utilizado, e que possui diversas funcionalidades que tornam essa parte do trabalho bastante simples. Através dele, é possível conectar-se diretamente às bases de dados da fonte de notícias escolhida e, de forma bastante simplificada, adquirir as notícias curtas. A Figura 12 a seguir mostra um exemplo da tela do FeedReader.

[5] http://www.feedreader.com/

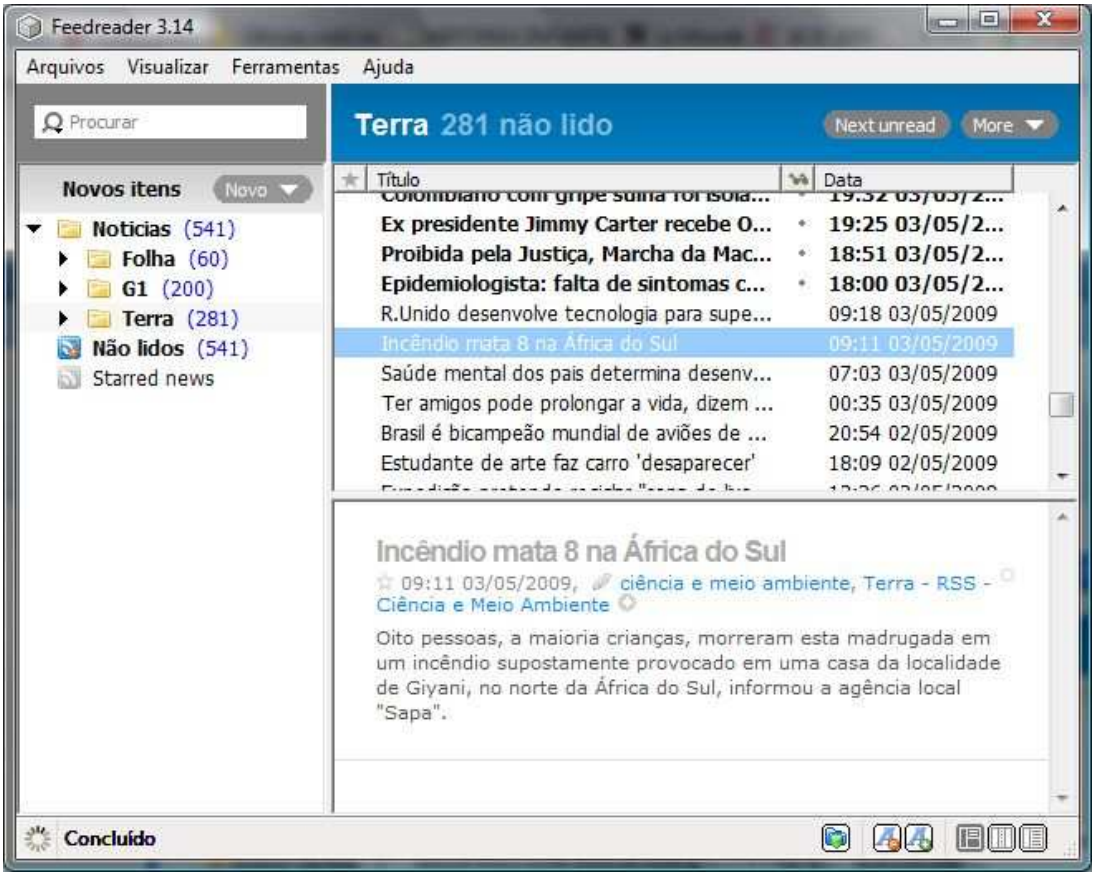


Figura 12: Tela do software que foi utilizado para capturar as notícias.

## 4.2. Implementação do Método

O método apresentado no capítulo anterior foi implementado na linguagem Java, utilizando a plataforma de desenvolvimento Eclipse. Juntamente com essa ferramenta, contou-se com o auxílio de algumas outras, também desenvolvidas em Java, e, consequentemente, facilmente integráveis ao projeto aqui desenvolvido. A seguir serão descritas estas ferramentas.

### 4.2.1. Weka

Weka (*Waikato Environment for Knowledge Analysis*) é uma ferramenta que tem por objetivo principal agregar diversos algoritmos dedicados ao estudo de aprendizagem de máquina e mineração de dados. Essa ferramenta conta com diversos algoritmos para todas as etapas de mineração de dados (pré processamento, processamento, classificação, regressão, agrupamento, associação e visualização), que podem ser usados tanto na própria plataforma Weka, como também podem ser facilmente integrados a qualquer aplicação Java que se

deseje (como foi o caso nesse projeto). [HALL et al., 2009] Tal integração permite uma grande flexibilidade na manipulação dos dados analisados, bem como maior facilidade de customização de todas as etapas presentes em um sistema de mineração de dados, especialmente as posteriores, como a visualização. Isso se deve ao fato de que, uma vez que todo procedimento de execução é definido pelo software Java escrito e desenvolvido pelo usuário, o desenvolvedor do sistema tem total autonomia para exibir seus dados da forma como bem entender, depois de concluídas suas etapas de processamento.

Nesse trabalho, especificamente, foram escolhidas e integradas três ferramentas do Weka, que serão comentadas na sequência. A integração foi escolhida como a alternativa mais viável para a execução do projeto, uma vez que, das ferramentas auxiliares encontradas disponíveis, as mais completas apresentavam uma versão Weka.

### 4.2.2. Manipulação de Matrizes

O método de manipulação de matrizes escolhido para realização dos cálculos necessários é chamado JAMA. Esse pacote é tido como o principal (e mais conhecido) para tal finalidade na linguagem Java, e é utilizado para a maioria dos cálculos essenciais em matrizes reais. O JAMA conta com seis classes para realização de variadas tarefas, entre elas, estão as operações básicas com matrizes e a classe para decomposição SVD. [FOX, 1998] O JAMA também possui uma extensão para Weka. Por apresentar todas as ferramentas de manipulação de matrizes que seriam necessárias no projeto e, pela facilidade de integração ao mesmo através do Weka, ele foi escolhido como utilitário para tal finalidade durante o desenvolvimento do projeto.

# Capítulo 5

# Experimentos e Resultados Obtidos

A seção que segue tem por objetivo apresentar e avaliar os testes que foram realizados no decorrer do projeto. Foram realizadas duas experimentações. A primeira foi realizada exclusivamente pelos pesquisadores, envolvendo a analise de 1000 notícias. No segundo experimento, convidou-se um grupo de voluntários para avaliar os resultados parciais obtidos pelo algoritmo.

## 5.1. Experimentação Envolvendo Notícias Curtas

Os testes realizados com o algoritmo descrito no Capítulo 3 consistiam em avaliar emoções em um corpus contendo 1000 notícias curtas extraídas em um mesmo período no ano de 2008 do site www.globo.com, divididas em dois grupos, sendo: o primeiro, chamado de treinamento, com 700 textos, utilizado para a avaliação do método principal (item 3.5); e o segundo, chamado de teste, com 300 notícias, reservado para a avaliação de novas notícias (item 3.6).

Além do corpus de notícias, utilizou-se seis listas com palavras relacionadas às emoções, além da lista de *stop words*. As listas de emoções foram baseadas na teoria apresentada por [STRAPPARAVA e MIHALCEA, 2008]. Como já comentado, em alguns dos métodos de testes descritos em seus experimentos, fez-se o uso de listas de palavras que remetessem às emoções analisadas para determinação dos grupos e identificação das emoções. Para serem úteis ao trabalho, as listas utilizadas em [STRAPPARAVA e MIHALCEA, 2008] foram traduzidas para o português e adapatadas às necessidades desse projeto.

A seguir, segue-se explanação e análise dos resultados obtidos, primeiro para o conjunto utilizado para construção do espaço com o algoritmo principal (LSA) e, em seguida, para o conjunto utilizado para validar o espaço criado com análise de novas notícias.

### 5.1.1. Base de Treinamento

A coleção de notícias curtas utilizada para construir o espaço e as matrizes do processo descrito anteriormente, no Capítulo 3, consiste de 700 textos. Ao final da execução do algoritmo descrito, obteve-se as matrizes e dados necessários para a realização da segunda etapa do projeto, que consiste na avaliação de uma nova notícia curta. Porém, antes de dar prosseguimento ao experimento avaliando-se uma nova notícia, fez-se necessário avaliar o desempenho do método implementado. Para tanto, adotou-se a seguinte medida: ao final da execução do algoritmo, uma tabela é gerada, contendo a ID da notícia, a notícia e mais seis colunas, cada uma referente a uma das emoções básicas já descritas. Para facilitar a visualização, essa grande tabela foi dividida em duas, e pode ser vista nas tabelas a seguir (Tabela 14 e Tabela 15).

Tabela 14: Exemplos de notícias curtas

| ***ID*** | ***Notícia Curta*** |
|---|---|
| **1** | Dois morrem e um fica ferido após carro cair uma altura de 10 metros: Acidente aconteceu neste sábado em Caxias do Sul. Veículo caiu com as rodas para cima em uma represa vazia. |
| **2** | Presidente do TCE no RS deixa hospital após assalto: João Vargas foi esfaqueado na barriga e passa bem em São Sepé. Dois suspeitos foram presos pela polícia na cidade de Santa Maria. |
| **3** | PRF flagra menina de 12 anos dirigindo picape em rodovia gaúcha: Segundo os policiais, jovem estava acompanhada da mãe em São Borja. Carro foi retido e a mãe autuada por deixar um menor de 18 anos dirigir. |
| **4** | Perito particular questiona imagens de acidente com ex-deputado no PR: Família de vítima contratou profissional para fazer simulação da colisão. Segundo ele, alguns segundos do filme do acidente foram removidos. |
| **5** | Ambulante tem mal súbito e morre no Mineirão: Médico legista crê em enfarto fulminante. Homem vendia camisas do lado de fora do estádio. |

Tabela 15: Emoções encontradas nas notícias curtas, segundo o método

| ***ID*** | ***Alegria*** | ***Desgosto*** | ***Medo*** | ***Raiva*** | ***Surpresa*** | ***Tristeza*** |
|---|---|---|---|---|---|---|
| **1** | 1.8643 | 2.1659 | 22.4061 | 1.6306 | -7.6297 | **31.5120** |
| **2** | 9.3553 | 15.2393 | -14.0036 | -4.0511 | 9.5565 | **16.5808** |

| | | | | | | |
|---|---|---|---|---|---|---|
| **3** | -11.9080 | 0.6329 | **35.5294** | 10.1315 | 10.1518 | -15.7766 |
| **4** | 4.4593 | -12.5701 | 11.1476 | **-21.8743** | -16.3681 | 9.2233 |
| **5** | -18.1971 | -16.2165 | 14.2813 | 8.4324 | 1.8106 | **20.7537** |

De posse desses resultados, cada notícia foi lida individualmente e confrontada com a possível impressão causada a um leitor do texto. Das 700 notícias avaliadas nessa etapa, 507 tiveram suas emoções identificadas corretamente pelo método. Isso equivale a uma taxa de acerto de aproximadamente 72%. Tal resultado pode ser considerado bom, visto que é equivalente às taxas de acerto obtidas em implementações semelhantes voltadas para outros idiomas (principalmente inglês) – pode-se citar como exemplo o trabalho de [STRAPPARAVA e MIHALCEA, 2008]. Outro fator determinante para a consideração de tais resultados é a escassez de métodos e ferramentas já existentes ou adaptadas para o português, pois isso dificulta as pesquisas e comparações de resultados.

### 5.1.2. Base de Teste

A forma de exibição dos resultados utilizada aqui, para avaliação, foi a mesma da seção anterior, ilustrada com a Tabela 14 e a Tabela 15. Das 300 notícias avaliadas nesse estágio, constatou-se, após leitura individual de cada uma, que 204 foram identificadas corretamente pelo método, o que corresponde a uma taxa de acerto de aproximadamente 68%. Observa-se que a taxa de acerto nesse grupo foi ligeiramente inferior ao anterior. O número de notícias avaliadas influencia diretamente na alteração da taxa de acerto, levando a essa diminuição. Entretanto, percebe-se que ainda sendo inferiores, os resultados não sofreram grandes mudanças, o que ainda assim comprova a eficácia do método.

### 5.1.3. Avaliação dos Resultados por Emoção

Terminada a avaliação dos resultados apresentada anteriormente, uma segunda análise foi efetuada: dessa vez, cada emoção foi avaliada individualmente. Sendo assim, todas as notícias foram separadas por emoção, de acordo com a identificação feita pelo sistema. Em seguida, uma análise das identificações foi feita. Ou seja, foi contado o número de notícias que foram identificadas como determinada emoção e o número de notícias que, segundo avaliação humana, realmente podem ser descritas como uma notícia que inspira a emoção em questão. Os resultados obtidos nessa etapa das avaliações podem ser observados na Tabela 16.

Tabela 16: Resultados obtidos, separados por emoção.

| ***Emoções*** | ***Treinamento*** | | | ***Teste*** | | |
|---|---|---|---|---|---|---|
| | ***Total*** | ***Corretas*** | ***%*** | ***Total*** | ***Corretas*** | ***%*** |
| ***Alegria*** | 239 | 159 | 67 | 116 | 69 | 59 |
| ***Desgosto*** | 189 | 146 | 77 | 78 | 60 | 77 |
| ***Medo*** | 30 | 22 | 73 | 20 | 16 | 80 |
| ***Raiva*** | 39 | 27 | 69 | 18 | 9 | 50 |
| ***Surpresa*** | 5 | 3 | 60 | 7 | 6 | 86 |
| ***Tristeza*** | 198 | 150 | 76 | 63 | 45 | 71 |

Uma curiosidade que se observa na análise em questão é a distribuição das notícias coletadas entre os grupos de emoções. Percebe-se que as emoções mais frequentes são: *alegria*, *tristeza* e *desgosto*, com boas taxas de acerto. Ao contrário, as emoções que possuíam menor número de notícias (*medo*, *raiva* e *surpresa*), embora com boas taxas de acerto, obtiveram pior desempenho na avaliação. Durante a leitura das notícias selecionadas para a realização do experimento, constatou-se a facilidade em se encontrar notícias que pudessem inspirar as três emoções mais frequentes, contra uma grande dificuldade em se encontrar notícias que trouxessem ao leitor a sensação inspirada pelas três mais incomumente encontradas.

Constatou-se que a taxa de acerto obtida para cada emoção analisada não está direta e unicamente relacionada ao número de palavras por emoção, descrito na Tabela 9. Como exemplo disso, pode-se citar *Alegria*, que possui uma lista com 278 termos e taxas de acerto para os conjuntos de treinamento e teste de 67% e 59% respectivamente. Em contrapartida, tem-se a emoção *Desgosto*, com 72 termos relacionados, e taxas de acerto de 77% para ambos conjuntos avaliados (treinamento e teste). Acredita-se que, tendo em conta que o método implementado trabalha com a relação entre as palavras nas sentenças [MARTINAZZO e PARAISO, 2010], essas diferenças se dêem mais pelos textos em si do que pelo próprio

número de termos existentes em determinada lista de emoções ou até mesmo pelo número de notícias avaliadas em cada categoria.

## 5.2. Experimentação Com Usuários Voluntários

Concluídos os testes descritos no item 5.1, realizou-se uma nova experimentação, dessa vez buscando aplicar os resultados obtidos a um público de usuários que potencialmente utilizariam o sistema (direta ou indiretamente) em algum momento. Visando a realização deste experimento, contou-se com a participação de 13 voluntários, e um subconjunto do grupo de notícias utilizadas para teste. Tal subconjunto consistia de 140 das 300 notícias utilizadas na base de teste do experimento anterior, divididas em pequenos grupos de 20, cada. Os 13 participantes foram divididos entre os conjuntos, de tal maneira que cada um dos grupos obteve pelo menos uma avaliação e, no máximo, três.

O grupo de voluntários era formado por professores, alunos de mestrado, doutorado e de iniciação científica. Um total de 20 pessoas foi convidado para participar da pesquisa, das quais 13 responderam ao questionário.

Para a elaboração dos conjuntos de notícias enviados aos voluntários, fez-se uso da ferramenta de pesquisa e questionários disponibilizada pelo Google através do Google Docs[6]. Montou-se, então, sete conjuntos com as 140 notícias selecionadas, contendo 20 textos cada, como mostra a Figura 13, a seguir:

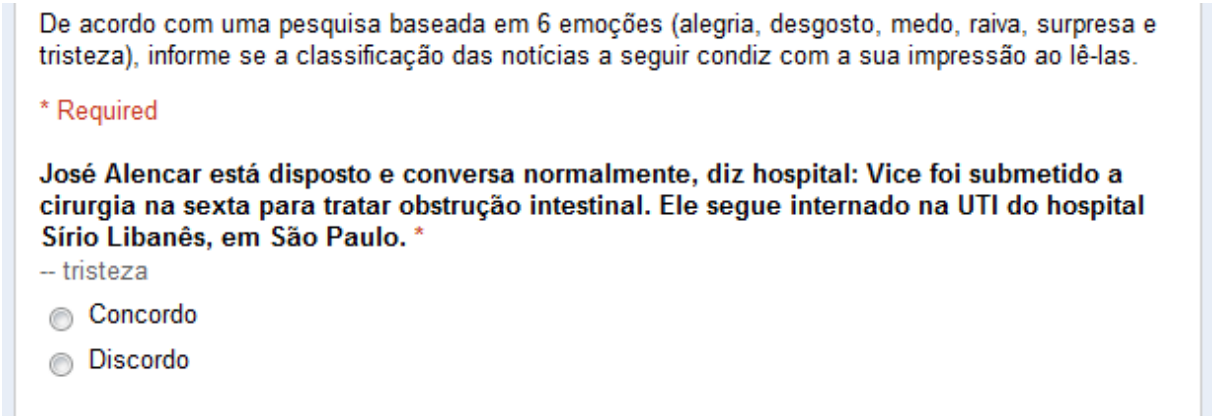


Figura 13: Exemplo de exibição de notícia no formulário criado.

[6] Google Docs está disponível através do endereço https://docs.google.com

Tal medida foi adotada para facilitar a leitura e o entendimento do que era solicitado, uma vez que os voluntários envolvidos neste experimento não estavam cientes de todo processo envolvido nessa pesquisa (processo automático de identificação de emoções). Como pode ser visto na Figura 13, as notícias foram exibidas aos voluntários em forma de questionário, sendo que no início de cada um destes foi escrita uma breve introdução e descrição do que é esperado da participação do voluntário. Após a exibição do texto da notícia, foi citada a emoção encontrada pelo sistema descrito nesse documento e, em seguida, foi dada a opção ao participante de expressar sua opinião a respeito da classificação, indicando se concordava ou discordava da classificação efetuada pelo sistema. Ao final do questionário, havia um campo aberto , onde o participante podia, se desejasse, registrar sugestões e/ou críticas ao projeto.

Para avaliação dos resultados, considerou-se o número de participações em cada grupo de notícias, da seguinte forma:

1. **Um participante respondente:**

   Foi simplesmente considerada a opinião do voluntário participante. Os casos dos grupos onde contou-se somente com a participação de uma pessoa foram os que obtiveram piores índices de concordância com a classificação efetuada pelo sistema, o que indica piores taxas de acerto por grupo.

2. **Dois participantes respondentes:**

   Para a verificação dos resultados em casos onde contou-se com a resposta de dois participantes, considerou-se como respostas em acordo com a classificação do sistema somente aquelas marcadas como “concordo” por ambos os voluntários. Em caso contrário, foi arbitrado que a classificação do sistema estava em desacordo com a identificação humana.

3. **Três participantes respondentes:**

   No caso de haver três respostas para o conjunto de notícias, foram consideradas corretas as identificações do sistema aquelas notícias cujas emoções identificadas automaticamente recebessem como resposta a opção “concordo” de dois ou mais voluntários. Nesse caso, obteve-se as melhores taxas de acerto dentro dos grupos.

Ao final dessa etapa de avaliação individual dos grupos de notícias, contabilizou-se então o número de notícias cujas emoções foram identificadas corretamente segundo os voluntários e, através disso, calculou-se a taxa de acerto para este pequeno grupo experimental. A taxa obtida foi de 41,43%. Percebe-se que houve uma queda significativa entre a taxa obtida com a classificação descrita no experimento anterior. Entretanto, acredita-se que essa queda se deve ao fato de o conjunto amostrado ser bastante reduzido com relação ao conjunto total e, também, ao número de participantes envolvido no experimento, que foi, além de tudo, distribuído de maneira bastante disforme.

# Capítulo 6

# Conclusão e Trabalhos Futuros

O trabalho descrito no presente documento propõe o desenvolvimento de um sistema de identificação de emoções em bases textuais escritas em português do Brasil. O objetivo é identificar uma das seis emoções básicas descritas por Paul Ekman e Wallace Friesen [EKMAN e FRIESEN, 1978] (alegria, raiva, tristeza, desgosto, medo e surpresa) em notícias curtas.

Para realização do trabalho, fez-se pesquisas acerca de emoções e mineração de textos, visando a aquisição de bom embasamento teórico, essencial para o desenvolvimento do método e dos experimentos realizados. Em um segundo momento, fez-se a escolha das ferramentas que foram utilizadas no processo. Optou-se por realizar a implementação do método em Java, com o auxílio de algumas ferramentas externas. As ferramentas foram selecionadas objetivando-se, sempre que possível, maior facilidade de integração umas com as outras, como também com a linguagem de programação escolhida para a implementação.

Concluída a implementação do método, iniciou-se a etapa de testes e experimentos para validação do trabalho elaborado. Para ser possível avaliar o desempenho do método, foi necessária a construção de uma base de notícias. Optou-se por coletar notícias do site www.globo.com, por ser o site com conteúdo mais variado. Esse corpus continha 1000 notícias, e foi dividido em dois grupos. O primeiro, com 700 notícias, foi utilizado para a construção do espaço e definição dos grupos que seriam classificados como cada uma das emoções básicas em avaliação. Após a interação do sistema, cada notícia foi lida juntamente com sua respectiva identificação automática e, com isso, percebeu-se que houve uma taxa de acerto de 72%. Em seguida, fez-se a execução da segunda parte do algoritmo, responsável por realizar a identificação de notícias novas, não classificadas com o algoritmo responsável pela

criação do espaço de cada grupo. Através dessa etapa do projeto, comprovou-se a possibilidade de se classificar uma nova notícia sem a necessidade de efetuar todos os cálculos novamente, o que toma muito tempo. Analogamente ao primeiro grupo, depois de classificadas todas as 300 notícias do segundo, realizou-se a leitura das mesmas juntamente com seus respectivos resultados obtidos pelo sistema e constatou-se uma taxa de acerto de 68%.

Através desses experimentos realizados, comprovou-se que é possível a identificação de emoções em textos de forma automatizada, através de um algoritmo de mineração de textos. Embora a eficácia comprovada do sistema ainda seja inferior a 80%, percebe-se que a média atingida através de outros estudos realizados na área é semelhante a aqui obtida, o que comprova a eficácia do método escolhido. Atualmente, além da finalidade descrita nesse projeto, estuda-se a possibilidade de empregar a técnica em outras áreas, como a identificação de assédio moral em mensagens eletrônicas [NUNES et al., 2009].

Como trabalhos futuros, espera-se realizar um estudo complementar sobre as emoções utilizadas no trabalho, de forma a adaptá-las melhor ao contexto onde estão atualmente inseridas e criar melhor relação entre notícias e emoções mais frequentes nesses textos. Também deseja-se realizar a integração desse projeto a um avatar, visando a animação automática das expressões faciais do mesmo através das notícias identificadas pelo algoritmo. Em um estudo ainda mais avançado, pode-se optar por pesquisar também a forma de conciliar a animação de expressões faciais à entonação de uma voz sintetizada para esse avatar. Várias hipóteses de aplicação para esse tipo de funcionalidade em ambientes *text-to-speech* são possíveis, tais como telejornais ou ainda sistemas de leitura de histórias e contos infantis. Por fim, objetiva-se também a criação de uma ontologia léxica para expressar as emoções a serem utilizadas nas pesquisas futuras.

# Referências


[ALM et al., 2005] ALM, C. O; ROTH, D; SPROAT, R. *Emotions from text: machine learning for text-base emotion prediction*. Human Language Technology Conference, p. 579-586, 2005.

[BRADFORD, 2003] BRADFORD, R. *Why LSI? Latent Semantic Indexing and Information Retrieval*. Content Analysis. Agilex Technologies, Inc. Chantilly, Virginia, 2003.

[CAMPOS] CAMPOS, C. P. *A Imagem no Jornalismo*. Disponível em http://webmail.faac.unesp.br/~pcampos/A%20Imagem%20no%20Jornalismo.htm Acesso em 18 de maio de 2010

[CARROLTON, 2002] CARROLTON, G. *What is Stop Word*. Disponível em http://searchsoa.techtarget.com/sDefinition/0,,sid26_gci798881,00.html. Acesso em 15 de novembro de 2010.

[CUNNINGHAM et al., 2002] CUNNINGHAM, H.; MAYNARD, D.; BONTCHEVA, K.; TABLAN, V. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[DEERWESTER et al., 1998] DEERWESTER, S.; DUMAIS, S.; LANDAUER, T.; FURNAS, G.; BECK, L. *Improving Information Retrieval with Latent Semantic Indexing*. Proceedings of the 51st Annual Meeting of the American Society for Information Science, v.25, pp. 36-40, 1998.

[EKMAN e FRIESEN, 1978] EKMAN, P.; FRIESEN, W. V. *Facial Action Coding System*. Palo Alto: Consulting Psychologists Press, 1978

[FEHR e RUSSEL, 1984] FEHR, B.; RUSSEL, J. A. *Concept of Emotion viewed from a prototype perspective*. Journal of Experimental Psychology, Washington, p. 464-486, 1984.

[FELDMAN e SANGER, 2007] FELDMAN, R; SANGER, J. *The Text Mining Handbook*. New York: Cambridge University Press, 2007.

[FOX, 1998] FOX, G. *JAMA: A Java Matrix Package*. Disponível em http://www.javagrande.org/leapforward/JGFMolerSC98/index.htm. Acesso em 18 de novembro de 2010.

[GARCIA, 2006] GARCIA, E. *Mi Islita: Cosine Similarity and Term Weight Tutorial*. http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html, 2006.

[GAZZANIGA e HEATHERTON, 2005] GAZZANIGA, M. S.; HEATHERTON, T. F. *Ciência Psicológica: Mente, Cérebro e Comportamento*. Porto Alegre: Artmed, 2005.

[GROBELNIK e MLADENIC] GROBELNIK, M.; MLADENIC, D. *Text-Garden -- Text-Mining Software Tools*. Department of Knowledge Technologies, Jozef Stefan Institute, Slovenia. Disponível em http://kt.ijs.si/Dunja/textgarden/. Acesso em 21 de maio de 2008.

[GROSSMAN e FRIEDER, 2004] GROSSMAN, D. A.; FRIEDER, O. *Information retrieval: Algorithms and Heuristics*. Second Edition. Springer, The Netherlands, 2004.

[HALL et al., 2009] HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. *The WEKA Data Mining Software: An Update; SIGKDD Explorations*. Volume 11, Issue 1. 2009.

[HAN e KAMBER, 2001] HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers, 2001.

[HEARST, 2003] HEARST, M. *What is Text Mining?*. UC Berkeley School of Information. Disponível em http://people.ischool.berkeley.edu/~hearst/text-mining.html. Acesso em 21 de maio de 2008.

[KONCHADY, 2006] KONCHADY, M. *Text Mining Application Programming*. Charles River Media, 1 ed. 2006.

[LANDAUER et al., 2005] LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. *Introduction to Latent Semantic Analysis*. Discourse Processes 25, p. 259-284, 2005.

[LOPER e BIRD, 2002] LOPER, E.; BIRD, S. *NLTK: The Natural Language Toolkit*. CoRR, 2002.

[LU et al., 2006] LU, C. Y; HONG, J. S.; LARA, S. C. *Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques*. 2006.

[MANNING et al., 2008] MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[MARTINAZZO e PARAISO, 2010] MARTINAZZO, B.; PARAISO, E. C. *Identificação de Emoções em Notícias Curtas*. CLEI - Conferência Latino-ameriaca de Informática, vol. 1, p. 1-10, 2010.

[MIT, 2002] MIT. *Singular Value Decomposition (SVD) Tutorial*. Massachusetts Institute of Technology, 2002.

[NUNES et al., 2009] NUNES, A.; FREITAS, C. O.; PARAISO, E. C. *Detecção de Assédio Moral em e-mails*. I Student Workshop on Information and Human Language

Technology - 7th Brazilian Symposium in Information and Human Language Technology, 2009, São Carlos. vol. 1, p. 1-6, 2009.

[PANG e LEE, 2004] PANG, B; LEE, L. *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*. Proceedings of the 42nd ACL, p. 271-278, 2004.

[RUBIN et al., 2004] RUBIN, V. L.; STANTON, J. M.; LIDDY, E. D. *Discerning emotions in texts*. AAAI Spring Symposium. Stanford, CA, 2004.

[SHAH et al., 2002] SHAH, U.; FININ, T.; JOSHI, A.; COST, R. S.; MATFIELD, J. *Information retrieval on the semantic web*. CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management, p. 461-468, 2002.

[SNOWBALL] *Snowball Stemmer*. http://snowball.tartarus.org/

[STRAPPARAVA e MIHALCEA, 2008] STRAPPARAVA, C.; MIHALCEA, R. *Learning to Identify Emotions in Text*. ACM Symposium on Applied Computing, p. 1556-1560, 2008.

[STRONGMAN, 2003] STRONGMAN, K. T. *The Psychology of Emotion*. 5. ed. Chichester: John Wiley & Sons Ltd., 2003

[VARELAS et al., 2005] VARELAS, G.; VOUTSAKIS, E.; RAFTOPOULOU, P.; PETRAKIS, E. G. M.; MILIOS, E. E. *Semantic similarity methods in wordNet and their application to information retrieval on the web*. WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management, p. 10-16, 2005.

[VIEIRA e VIRGIL, 2007] VIERA, A. F. G.; VIRGIL, J. *Uma revisão dos algoritmos de radicalização em língua portuguesa*. Information Research, volume 12 (3), paper 315,

Abril 2007. Disponível em http://InformationR.net/ir/12-3/paper315.html. Acesso em 01 de novembro de 2010.

[WEKA] WEKA: *Data Mining Software in Java*. The University of Waikato. Dispoível em http://www.cs.waikato.ac.nz/ml/weka/

[YATES e RIBEIRO NETO, 1999] YATES, R. B.; RIBEIRO NETO, B. *Modern Information Retrieval*. New York: ACM Press, 1999.

[YU et al., 2002] YU, C.; CUADRADO, J.; CEGLOWSKI, M.; PAYNE, J. S. *Patterns in Unstructured Data: Discovery, Aggregation, and Visualization*. National Institute for Technology and Liberal Education, 2002.

# Anexo I

## Lista de Palavras Referentes à Emoção “Alegria”

abundante
acalmar
aceitável
aclamar
aconchego
adesão
admirar
adorar
afável
afeição
afeto
afortunado
agradar
ajeitar
alívio
amabilidade
amado
amar
amável
amenizar
ameno
amigável
amistoso
amizade
amor
animação
ânimo
anseio
ânsia
ansioso
apaixonado
apaziguar
aplausos
apoiar
aprazer
apreciar
aprovação
aproveitar
ardor
armirar
arrumar
atração
atraente
atrair
avidamente
avidez
ávido
belo
bem-estar
beneficência
beneficiador
benefício
benéfico
benevocência
benignamente
benígno
bom
bondade
bondoso
bonito
brilhante
brincadeira
calma
calor
caridade
caridoso
carinho
cativar
charme
cheery
clamar
cofortar
coleguismo
comédia
cômico
comover
compaixão
companheirismo
compatibilidade
compatível
complacência
completar
compreensão
conclusão
concretização
condescendência
confiança
confortante

congratulação
conquistar
consentir
consideração
consolação
contentamento
coragem
cordial
considerar
consolo
contente
cuidadoso
cumplicidade
dedicação
deleitado
delicadamente
delicadeza
delicado
desejar
despreocupação
devoção
devoto
diversão
divertido
encantar
elogiado
emoção
emocionante
emotivo
empatia
empático
empolgação
enamorar
encantado
encorajado
enfeitar
engraçado
entendimento
entusiasmadamente
entusiástico
esperança
esplendor
estima
estimar
estimulante
euforia
eufórico
euforizante
exaltar
excelente
excitar
expansivo
extasiar
exuberante
exultar
fã
facilitar
familiaridade
fascinação
fascínio
favor
favorecer
favorito
felicidade
feliz
festa
festejar
festivo
fidelidade
fiel
filantropia
filantrópico
fraterno
ganhar
generosidade
generoso
gentil
glória
glorificar
gostar
gostoso
gozar
gratificante
grato
hilariante
honra
humor
impressionar
incentivar
incentivo
inclinação
incrível
inspirar
interessar
interesse
irmandade
jovial
jubilante
júbilo
lealdade
legítimo
leveza
louvar
louvável
louvavelmente
lucrativo
lucro
maravilhoso
melhor
obter
obteve
ode
orgulho
paixão
parabenizar
paz
piedoso
positivo
prazenteiro
prazer
predileção
preencher
preferência
preferido
promissor
prosperidade
proteção

proteger
protetor
proveito
provilégio
querer
radiante
realizar
recomendável
reconhecer
recompensa
recrear
recreativo
recreação
regozijar
respeitar
ressuscitar
revigorar
risada
risonho
romântico
romantismo
saciar
saciável
satisfação
satisfatoriamente
satisfatório
satisfazer
satisfeito
sedução
seduzir
sereno
simpaticamente
simpático
sobrevivência
sobreviver
sorte
sortudo
sucesso
surpreender
tenro
ternura
torcer
tranqüilo
tranquilo
triunfo
triunfal
triunfante
vantagem
vantajoso
vencedor
veneração
ventura
vida
vigor
virtude
virtuoso
vitória
vitorioso
viver
vivo
zelo
zeloso

# Anexo II

## Lista de Palavras Referentes à Emoção “Desgosto”

abominável
adoentado
amargamente
antipatia
antipático
asco
asqueroso
aversão
chateação
chatear
desagradável
desagrado
desprezível
detestável
doença
doente
enfermidade
enjoativo
enjoo
enjôo
feio
fétido
golfar
grave
gravidade
grosseiro
grosso
horrível
ignóbil
ilegal
incômdo
incomodar
indecente
indisposição
indisposto
inescrupuloso
maldade
maldoso
malvado
mau
náusea
nauseabundo
nauseante
nausear
nauseoso
nojento
nojo
obsceno
obstrução
obstruir
ofensivo
patético
perigoso
repelente
repelir
repugnante
repulsa
repulsão
repulsivo
rude
sujeira
sujo
terrível
terrivelmente
torpe
travesso
travessura
ultrajante
vil
vomitar
vômito

# Anexo III

## Lista de Palavras Referentes à Emoção “Medo”

abominável
afugentar
alarmar
alerta
ameaça
amedrontar
angustia
angústia
angustiadamente
ansiedade
ansioso
apavorar
apreender
apreensão
apreensivo
arrepio
assombrado
assombro
assustado
assustadoramente
atemorizar
aterrorizante
brutal
calafrio
chocado
chocante
consternado
covarde
cruel
crueldade
cruelmente
cuidado
cuidadosamente
cuidadoso
defender
defensor
defesa
derrotar
desconfiado
desconfiança
desencorajar
desespero
deter
envergonhado
escandalizado
escuridão
espantoso
estremecedor
estremecer
expulsar
feio
friamente
fugir
hesitar
horrendo
horripilante
horrível
horrivelmente
horror
horrorizar
impaciência
impaciente
impiedade
impiedoso
indecisão
inquieto
insegurança
inseguro
intimidar
medonho
medroso
monstruosamente
mortalha
nervoso
pânico
pavor
premonição
preocupar
presságio
pressentimento
recear
receativamente
receio
receoso
ruim
suspeita
suspense
susto

| | | | |
|---|---|---|---|
| temer | tenso | terror | tremor |
| temeroso | terrificar | timidamente | vigiar |
| temor | terrível | timidez | vigilante |
| tensão | terrivelmente | tímido | |

# Anexo IV

## Lista de Palavras Referentes à Emoção “Raiva”

abominação
aborrecer
adredido
agredir
agressão
agressivo
amaldiçoado
amargor
amargura
amolar
angústia
animosidade
antipatia
antipático
asco
assassinar
assassinato
assediar
assédio
atormentar
avarento
avareza
aversão
beligerante
bravejar
chateação
chato
cobiçoso
cólera
colérico
complicar
contraiedade
contrariar
corrupção
corrupto
cruxificar
demoníaco
demônio
descaso
descontente
descontrole
desenganar
desgostar
desgraça
desprazer
desprezar
destruição
destruir
detestar
diabo
diabólico
doido
encolerizar
energicamente
enfurecido
enfuriante
enlouquecer
enraivecer
escandalizar
escândalo
escoriar
exasperar
execração
ferir
frustração
frustrar
fúria
furioso
furor
ganância
ganancioso
guerra
guerreador
guerrilha
hostil
humilhar
implicância
implicar
importunar
incomodar
incômodo
indignar
infernizar
inimigo
inimizade
injúria
injuriado
injustiça

insulto
inveja
ira
irado
irascibilidade
irascível
irritar
louco
loucura
magoar
mal
maldade
maldição
maldito
maldizer
maldoso
maleficência
maléfico
malevolência
malévolo
malícia
malicioso
malignidade
malígno
maltratar
maluco
malvadeza
malvado
matar
mesquinho
misantropia
misantrópico
molestar
moléstia
mortal
morte
mortífero
mortificar
nervoso
odiar
odiável
ódio
odioso
ofendido
ofensa
opressão
opressivo
oprimir
perseguição
perseguir
perturbar
perverso
provocar
rabugento
raivoso
rancor
reclamar
repressão
reprimir
repulsa
repulsivo
resmungar
ressentido
revolta
ridículo
tempestuoso
tirano
tormento
torturar
ultrage
ultrajar
vexatório
vigoroso
vingança
vingar
vingativo
violência
violento
zangar

# Anexo V

## Lista de Palavras Referentes à Emoção “Surpresa”

admirar
afeição
apavorante
assombro
chocado
chocante
desconcertar
deslumbrar
embasbacar
emudecer
encantamento
enorme
espanto
estupefante
estupefato
estupefazer
expectativa
fantasticamente
fantástico
horripilante
imaginário
imenso
impressionado
incrível
maravilha
milagre
mistério
misterioso
ótimo
pasmo
perplexo
prodígio
sensacional
surpreendente
surpreender
suspense
susto
temor
tremendo

# Anexo VI

## Lista de Palavras Referentes à Emoção “Tristeza”

abandonar
abatido
abominável
aborrecer
abortar
aflição
afligir
aflito
agoniar
amargo
amargor
amargura
ansiedade
arrepender
arrependidamente
atrito
azar
cabisbaixo
chorão
choro
choroso
coitado
compassivo
compunção
contrição
contristador
contrito
culpa
defeituoso
degradante
deplorável
deposição
depravado
depressão
depressivo
deprimente
deprimir
derrota
derrubar
desalentar
desamparo
desanimar
desânimo
desapontar
desconsolo
descontente
desculpas
desencorajar
desespero
desgaste
desgosto
desgraça
desistência
desistir
deslocado
desmoralizar
desolar
desonra
despojado
desprazer
desprezo
desumano
discriminar
disforia
disfórico
dissuadir
dó
doloroso
dor
enfadado
enlutar
entediado
entristecedor
entristecer
envergonhar
errante
erro
errôneo
escurecer
escuridão
escuro
esquecido
estragado
execrável
extirpar
falso
falsidade
falta

fraco
fraqueza
fricção
frieza
frio
fúnebre
funesto
grave
horror
humilhar
inconsolável
indefeso
infelicidade
infeliz
infortúnio
isolar
lacrimejante
lacrimoso
lágrima
lamentar
lástima
lastimoso
lúgubre
luto
lutoso
mágoa
magoar
martírio
martirizar
mau
melancolia
melancólico
menosprezar
miseravelmente
miséria
mistério
misterioso
morre
morte
mortificante
negligentemente
nocivo
obscuro
opressão
opressivo
oprimir
pena
penalizar
penitente
penoso
penumbra
perder
perturbado
perverso
pervertar
pesaroso
pessimamente
piedade
pobre
porcamente
prejudicado
prejudicial
prejuízo
pressão
pressionar
quebrar
queda
queixoso
rechaçar
remorso
repressão
repressivo
reprimir
ruim
secreto
servil
só
sobrecarga
sobrecarregado
sofrer
sofrimento
solidão
sombrio
soturno
sujo
suplicar
suplício
tédio
timidez
tímido
torturar
trevas
triste
tristemente
vazio