Identificação de Emoções em Notícias Curtas

Barbara Martinazzo, Emerson Cabrera Paraiso

Pontifícia Universidade Católica do Paraná - PUCPR
Programa de Pós-Graduação em Informática
CEP: 80.215-901, Curitiba, Paraná, Brasil
{b.martinazzo, paraiso}@ppgia.pucpr.br

Abstract. The automatic detection of emotions in texts has presented significant results in several and different situations. In this paper, we present an approach to identify automatically emotions in short texts (basically, news headlines and a brief description of it) written in Portuguese. In the work presented here, we process each text before the avatar reads it, in order to find the emotions presented on it and to prepare the avatar to behave according to the emotions found. Each text is processed using an algorithm based on the Latent Semantic Analysis theory.

Keywords: Emotion detection, Avatar, LSA, Text Mining.

Resumo. A detecção automática de emoções em textos tem mostrado resultados significativos nas mais variadas situações. Nesse artigo, apresentamos um método para a identificação automática de emoções em textos curtos (manchetes e suas respectivas linhas finas) escritos em português. O objetivo principal é identificar emoções em notícias e utilizá-las para controlar o comportamento de um avatar, que lê a notícia analisada. No trabalho apresentado, processamos cada texto integralmente utilizando a técnica *Latent Semantic Analysis*.

Palavras-chave: Identificação de emoções, Avatar, LSA, Mineração de Textos.

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. Recentemente, esse tipo de estudo tem atraído também a atenção de pesquisadores da Ciência da Computação, especialmente os interessados no processamento de textos e na interação humano computador. Entre as pesquisas realizadas, encontra-se o reconhecimento automático de emoções em informação textual [14], conhecida como *Sentiment Analysis*. Sabe-se que, além de informação,

textos podem conter também opiniões e teores emocionais [1]. 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.

Neste artigo, apresentamos um processo de identificação de emoções em notícias curtas escritas em português. O objetivo é, a partir da identificação das emoções presentes no texto, dar a uma figura artificial animada (avatar) a capacidade de emitir expressões faciais de acordo com o conteúdo do texto. O avatar neste caso lê o texto, adaptando seu comportamento as emoções encontradas. Neste artigo, nos concentraremos somente com a apresentação do processo de identificação de emoções, não detalhando o processo de controle do comportamento do avatar.

Na sequência deste artigo, a seção 2 apresenta os principais conceitos necessários para o entendimento do trabalho, como, por exemplo, o conceito de emoções e a identificação de emoções a partir de textos. Na seção 3 apresentamos uma descrição aprofundada do método e da forma como ele foi construído. Na seção 4, apresentamos os resultados obtidos em um experimento realizado utilizando um corpus com 700 notícias curtas. Ao final, oferecemos uma conclusão e as perspectivas de trabalhos futuros.

2 Identificação de Emoções em Textos

Essa seção tem por objetivo apresentar os principais conceitos presentes no processo de identificação de emoções em texto. Iniciaremos pelo conceito de Emoções.

2.1 Emoções

Fehr e Russell [7] afirmam que todas as pessoas sabem o que é uma emoção, até que lhes seja solicitada por 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 [9] afirmam que as emoções são objeto de estudo de diversas áreas do conhecimento humano já há bastante tempo. Segundo Strongman [15], 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.

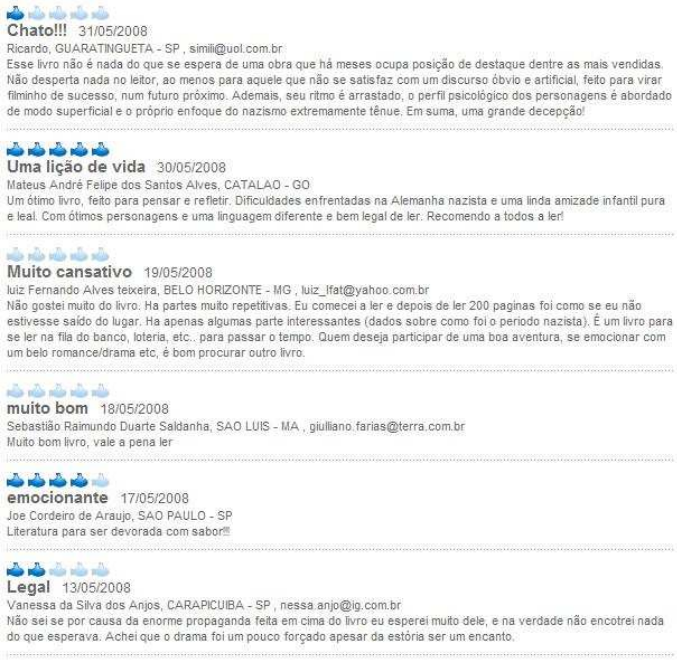
Atualmente, o estudo das emoções se divide em várias áreas distintas, mas a utilizada neste trabalho é a mais simples delas, chamada de Emoções Básicas (ou Puras). Esse conceito diz respeito às emoções inatas compartilhadas por todas as

culturas, e foi proposto na década de 1970 por Paul Ekman e Wallace Friesen [6]. 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. A 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 possamos estender as mesmas emoções para modificar o comportamento de um avatar, que o fará fundamentalmente 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 [14]. Entre os assuntos e as pesquisas realizadas encontra-se o reconhecimento automático de emoções em informação textual, conhecida como *Sentiment Analysis*.

2.2 Identificação de Emoções em Textos

Sabe-se que, além de informação, textos podem conter também, a expressão da opinião e do estado emocional de seu autor. Os sistemas de classificação de produtos, por exemplo, muitas vezes não trazem apenas fatos absolutos, mas sim opiniões pessoais. Desta 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 disto 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 [12]. 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. Na Fig. 1, é possível observar que nem sempre a nota atribuída pelo consumidor é condizente com a sua real opinião.



Chato!!! 31/05/2008
Ricardo, GUARATINGUETA - SP , simili@uol.com.br
Esse livro não é nada do que se espera de uma obra que há meses ocupa posição de destaque dentre as mais vendidas. Não desperta nada no leitor, ao menos para aquele que não se satisfaz com um discurso óbvio e artificial, feito para virar filminho de sucesso, num futuro próximo. Ademais, seu ritmo é arrastado, o perfil psicológico dos personagens é abordado de modo superficial e o próprio enfoque do nazismo extremamente tênue. Em suma, uma grande decepção!

Uma lição de vida 30/05/2008
Mateus André Felipe dos Santos Alves, CATALAO - GO
Um ótimo livro, feito para pensar e refletir. Dificuldades enfrentadas na Alemanha nazista e uma linda amizade infantil pura e leal. Com ótimos personagens e uma linguagem diferente e bem legal de ler. Recomendo a todos a ler!

Muito cansativo 19/05/2008
luiz Fernando Alves teixeira, BELO HORIZONTE - MG , luiz_lfat@yahoo.com.br
Não gostei muito do livro. Ha partes muito repetitivas. Eu comecei a ler e depois de ler 200 paginas foi como se eu não estivesse saído do lugar. Ha apenas algumas parte interessantes (dados sobre como foi o periodo nazista). É um livro para se ler na fila do banco, loteria, etc.. para passar o tempo. Quem deseja participar de uma boa aventura, se emocionar com um belo romance/drama etc, é bom procurar outro livro.

muito bom 18/05/2008
Sebastião Raimundo Duarte Saldanha, SAO LUIS - MA , giulliano.farias@terra.com.br
Muito bom livro, vale a pena ler

emocionante 17/05/2008
Joe Cordeiro de Araujo, SAO PAULO - SP
Literatura para ser devorada com sabor!!

Legal 13/05/2008
Vanessa da Silva dos Anjos, CARAPICUIBA - SP , nessa.anjo@ig.com.br
Não sei se por causa da enorme propaganda feita em cima do livro eu esperei muito dele, e na verdade não encotrei nada do que esperava. Achei que o drama foi um pouco forçado apesar da estória ser um encanto.

Fig. 1. Resenhas avaliativas de um produto em um *site* de compras *online*.

Como um exemplo, na avaliação “muito bom” (e avaliada positivamente através da resenha textual) foi ranqueada com nota 0, mesma nota atribuída a avaliação que demonstrou 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.

Embora pareça algo relativamente simples, a tarefa é complicada. A simples busca de palavras relacionadas a emoções para a classificação de um documento, de acordo com o número de ocorrências destas palavras (como, por exemplo, “bom”, “ruim”, etc.), não é suficiente. Tomemos a seguinte frase como exemplo: “O protagonista tenta proteger seu bom nome”, que contém a palavra “bom”. Como não há nenhuma informação relevante, esta frase pode ser considerada uma afirmação objetiva. Pang [12] e CUCS [4] introduzem uma nova abordagem à atividade, visando, desta forma, obter melhores resultados. Tal abordagem, segundo Pang 12], consiste em: (1) efetuar a distinção entre sentenças subjetivas e objetivas, devendo estas ser descartadas; (2) aplicar às sentenças subjetivas um classificador. CUCS [4] afirma que os problemas enfrentados consistem, basicamente, em definir quando uma sentença é subjetiva ou objetiva, em determinar se o sentimento é positivo ou negativo e, finalmente na valência do sentimento.

A riqueza da linguagem humana possibilita que uma única informação seja expressa de várias formas, o que dificulta a tarefa de encontrar os indicadores

emocionais corretos. Como exemplo, pode-se citar as três sentenças a seguir (extraídas de CUCS [4]):

- *This laptop is a great deal.*
- *A great deal of media attention surrounded the release of the new laptop model.*
- *If you think this laptop is a great deal, I've got a nice bridge for you to buy.*

Nos três casos a expressão "*great deal*" é encontrada, mas as opiniões expressas são, respectivamente, positiva, neutra e negativa. Nos dois primeiros casos, a expressão é utilizada com significados diferentes; já no último há sarcasmo, que é algo muito comum em textos subjetivos como *posts* em blogs e listas de discussão, entre outros.

Uma das técnicas utilizadas para a identificação de relações entre palavras em um texto é conhecida como *Latent Semantic Analysis* e é apresentada a seguir.

2.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) é um método matemático/estatístico para identificação de relações entre palavras em textos [11]. A partir dessas relações, visa-se estabelecer associações entre os termos encontrados [5]. Tradicionalmente, não é considerado como um método de processamento de linguagem natural, 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 [11]. Entretanto, o LSA tem mostrado resultados consistentes em várias áreas de aplicação, entre elas a categorização de documentos com base em similaridades conceituais [2].

Para o desenvolvimento do modelo, parte-se do pressuposto que as palavras sempre possuirão significados semelhantes e estarão inseridas dentro de contextos parecidos ([5] e [2]). 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 (*Single Value Decomposition* ou decomposição em valor único) para determinar padrões e relacionamentos entre os termos encontrados nos documentos ([11] e [2]). Com base nesses padrões encontrados, serão feitas as aproximações necessárias para agrupamentos e/ou classificações.

O LSA foi utilizado em nosso método, como apresentado na seção seguinte.

3 Identificação de Emoções em Notícias Curtas

O objetivo desta pesquisa é identificar automaticamente emoções em um conjunto de notícias curtas em português. As informações extraídas serão posteriormente utilizadas para realizar a animação de um avatar que faz papel de um apresentador de um programa de notícias.

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, com menor destaque, e possuem a função de fornecer melhor explicação acerca do título principal (manchete) [3].

Como exemplo de manchete e linha fina, temos 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 em notícias curtas, implementamos um procedimento baseado no algoritmo *Latent Semantic Analysis* (LSA) [17], utilizando a linguagem Java. Com o objetivo de ilustrar melhor todo o processo, tomaremos como exemplo as seguintes notícias apresentadas na Tabela 1:

Tabela 1. Notícias utilizadas para exemplificar 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.

A primeira etapa consiste em um pré-processamento das notícias curtas, e engloba as seguintes tarefas: ler o arquivo original de notícias, converter todas as palavras para minúsculas, remover caracteres especiais (como pontuação e números), remover *stop words*[1]e aplicar um *Stemmer* (utilizado para reduzir os termos aos seus radicais). Com o *Stemmer*, os termos derivados de um mesmo radical serão contabilizados como um único termo, como no exemplo: guerra, guerrear = guerr. Essas duas últimas etapas (remoção de *Stop words* e *Stemmer*) foram desenvolvidas com o auxílio da ferramenta Weka [16], que conta com um algoritmo para remoção de *stop words* e com uma extensão do *Snowball Stemmer* [13]. Estes recursos foram escolhidos por serem facilmente integráveis ao projeto desenvolvido e serem configuráveis para o português. A Tabela 2, apresenta o resultado do pré-processamento para os exemplos apresentados na Tabela 1.

Tabela 2. Notícias da Tabela 1 após o préprocessamento.

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, 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, preenchida 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 3, a seguir, mostra a

[1] Lista para o português obtida no *site* Linguateca: http://www.linguateca.pt/

matriz gerada para o exemplo, onde D1, D2, D3,..., são as notícias curtas da Tabela 1. Vale lembrar que os termos observados a seguir passaram pelo processo de *stemming* citado anteriormente e, portanto, podem parecer palavras estranhas. Esse processo é extremamente necessário para que termos que possuem significados semelhantes e mesmo radical não sejam repetidos, pois isso alteraria os resultados.

Tabela 3. Matriz 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 desta matriz, um segundo procedimento é iniciado, também com o auxílio da ferramenta Weka. Esse consiste na aplicação de uma técnica chamada *Single Value Decomposition* (SVD). O motivo que nos leva a utilizar o SVD consiste em encontrar uma representação dimensionalmente reduzida da matriz, 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 (*term-document matrix*) em três outras. A primeira, chamada U, nos remete a coordenadas de cada termo dentro de um espaço.

A Tabela 4, a seguir, mostra um pedaço da matriz U, com as cinco primeiras dimensões e as sete primeiras palavras do conjunto de termos.

Tabela 4. 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

A segunda matriz, aqui chamada de V^T, fornece as coordenadas dos documentos nesse mesmo espaço. A Tabela 5, 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 5. 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

Finalmente, a terceira matriz, chamada S, nos permite estimar quantas dimensões deverão ser utilizadas para a obtenção dos melhores resultados. Essa dimensão, segundo Grossman e Frieder [10], pode ser estimada arbitrariamente, através de experimentos e comparações. No exemplo que ilustra o trabalho, adotamos apenas 5 dimensões, pois o objetivo é somente ilustrar o funcionamento do algoritmo. Entretanto, para o experimento apresentado a seguir, adotamos 50 dimensões, pois é o número que nos forneceu os melhores resultados.

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 no item 2.1. Estas listas foram inicialmente disponibilizadas por Strappavara e Mihalcea [14], em seis arquivos diferentes (um para cada emoção), originalmente em inglês. Fizemos, então, a tradução das palavras contidas em cada um dos arquivos, de forma a termos palavras diretamente ligadas a emoções em português. A Tabela 6 contém alguns exemplos e, ao final, demonstra a quantidade total de palavras contidas em cada lista de emoções.

Tabela 6. Exemplos de palavras contidas nas listas de emoções.

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

Na sequência, para cada emoção, busca-se todas as palavras da lista da 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 7, 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 7: Parte da matriz de coordenadas de localização dos grupos.

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

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 [8], determinamos a distância de cada notícia curta aos grupos definidos, como mostra a equação (1), a seguir:

$$Sim(Dn, Gm) = \frac{\sum_i W_{D_n i} * W_{G_m i}}{\sqrt{\sum_j W_{D_n j}{}^2} * \sqrt{\sum_j W_{G_m j}}} \quad (1)$$

Onde:
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 (1), por se tratar de uma aproximação cossenoidal, resultam em valores, em módulo, entre 0 e 1 (ou seja, de -1 a 1), onde 1 (um) 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, foram multiplicados por 100. Dessa forma, nossos resultados variam, em módulo, entre 0 e 100 (ou seja, de -100 a 100).

Na seção seguinte apresentamos alguns exemplos de resultados obtidos a partir das experimentações realizadas com o algoritmo que acabamos de descrever.

4 Experimentos e Resultados

Para a elaboração dos testes, foi selecionado um corpus contendo 700 notícias curtas escritas em português do Brasil, do ano de 2009, extraídas do site Globo.com. Este corpus de notícias foi manualmente anotado, ou seja, as emoções de cada notícia foram identificadas, de forma que possamos calcular a taxa de acerto na identificação das emoções, por parte de nosso método. Após a aplicação do algoritmo descrito na seção 3, obtivemos uma taxa de acerto de aproximadamente 67%. Vale ressaltar que,

no momento que a pesquisa foi iniciada, não foram encontradas referências para trabalhos semelhantes no idioma aqui descrito (a saber, português).

Para exemplificar, escolhemos cinco textos. Após o processo de pré-processamento e cálculo dos pesos, um arquivo é construído contendo um registro para cada notícia avaliada. Cada registro tem a seguinte estrutura:

ID: alegria, desgosto, medo, raiva, surpresa, tristeza

onde ID corresponde ao índice da manchete no arquivo e cada uma das emoções corresponde ao seu respectivo peso na manchete em questão. Para exemplificar, a Tabela 8 apresenta algumas notícias curtas analisadas pelo algoritmo.

Tabela 8. 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.

Na Tabela 9 temos os registros obtidos para cada notícia.

Tabela 9: Emoções encontradas nas notícias curtas.

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

Para a análise dos resultados, utilizamos a Tabela 8 e a Tabela 9. Ignora-se o sinal dos pesos, ou seja, consideramos apenas o módulo dos mesmos, sempre do maior resultado para o menor. Geralmente, os dois maiores resultados (em módulo) são suficientes para definir as emoções presentes em uma notícia. Dessa forma, comparando-se os resultados obtidos com as manchetes, percebe-se que houve uma aproximação nos mesmos, de acordo com o que o texto realmente deseja transmitir. Por exemplo, a notícia curta 1 possui como maior resultado o número 31.5120 (tristeza) e, como segundo maior, 22.4061 (medo). Concluímos, assim, que a mesma é uma notícia que transmite tristeza e medo a quem a lê.

5 Conclusões e Trabalhos Futuros

Neste artigo apresentamos um processo de identificação de emoções em notícias curtas escritas em português, com o objetivo de utilizá-las para melhorar a naturalidade de figuras virtuais animadas. Aqui, selecionamos um grupo de emoções básicas para representar comportamentos e expressões distintos da figura animada: alegria, desgosto, medo, raiva, surpresa e tristeza. Para cada emoção encontrada no texto, um comportamento compatível foi definido. Um método de extração de emoções a partir de textos curtos foi apresentado.

Na sequência deste projeto estamos trabalhando para melhorar os resultados obtidos pelo método, principalmente, tentando melhor estimar o número de dimensões da matriz de documentos. Também estamos trabalhando no desenvolvimento de uma técnica que informe a ordem de aparecimento das emoções no texto, caso o mesmo apresente mais de uma emoção ao longo do seu conteúdo.

Referências

1. 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).
2. Bradford, R. Why LSI? Latent Semantic Indexing and Information Retrieval. Content Analysis. Agilex Technologies, Inc. Chantilly, Virginia, (2003).
3. Campos, C. P. A Imagem no Jornalismo. http://webmail.faac.unesp.br/~pcampos/A%20Imagem%20no%20Jornalismo.htm Consultado em 18 de maio de 2010
4. Cornell University (CU) Computer Science (CS). Sentiment Analysis. CS 40th anniversary Symposium, p. 26-27, (2005).
5. 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).
6. Ekman, P., Friesen, W. V. Facial Action Coding System. Palo Alto: Consulting Psychologists Press, (1978).
7. Fehr, B., Russel, J. A. Concept of Emotion viewed from a prototype perspective. Journal of Experimental Psychology, Washington, p. 464-486, (1984).
8. Garcia, E. Mi Islita: Cosine Similarity and Term Weight Tutorial. http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html, (2006).
9. Gazzaniga, M. S., Heatherton, Todd F. Ciência Psicológica: Mente, Cérebro e Comportamento. Porto Alegre: Artmed, (2005).
10. Grossman, D. A., Frieder, O. Information retrieval: Algorithms and Heuristics. Second Edition. Springer, The Netherlands, (2004).
11. Landauer, T. K.; Foltz, P. W.; Laham, D. Introduction to Latent Semantic Analysis. Discourse Processes 25, p. 259-284, (2005).
12. 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).
13. Snowball Stemmer. http://snowball.tartarus.org/
14. Strapparava, C., Mihalcea, R. Learning to Identify Emotions in Text. ACM Symposium on Applied Computing, p. 1556-1560, (2008).

15.Strongman, K. T. The Psychology of Emotion. 5. ed. Chichester: John Wiley & Sons Ltd., (2003).
16.WEKA: Data Mining Software in Java. The University of Waikato
17.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)