Identificando Emoções em Textos em Português do Brasil usando Máquina de Vetores de Suporte em Solução Multiclasse

Mariza Miola Dosciatti, Lohann Paterno Coutinho Ferreira, Emerson Cabrera Paraiso

Programa de Pós-Graduação em Informática – PPGIa, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba – PR – Brasil
{mariza.dosciatti, paraiso}@ppgia.pucpr.br; lohann.ferreira@pucpr.br

Resumo. *A identificação automática de emoções em textos tem apresentado resultados significativos em diversas aplicações. Neste artigo, é apresentada uma abordagem utilizando Máquinas de Vetores de Suporte para identificar emoções em textos escritos em Português do Brasil. O corpus utilizado no experimento é composto de notícias extraídas de um jornal online. Os textos são previamente rotulados e submetidos a um classificador SVM em configuração multiclasse, obtendo uma taxa de acerto de 61%.*

Abstract. *The automatic identification of emotions in texts has shown significant results in several applications. In this article, we present an approach using Support Vector Machines to identify emotions in texts written in Brazilian Portuguese. The corpus used in the experiment consists of news extracted from an online newspaper. The texts were labeled and subjected to a SVM classifier in a multiclass configuration, obtaining an accuracy rate of 61%.*

1. Introdução

Identificar emoções em textos é um dos objetivos da área de Análise de Sentimento. Segundo [Liu 2010], a Análise de Sentimento é o estudo de opiniões, sentimentos e emoções expressas em textos. Nesta área, que ganhou impulso com a difusão da web, muitas pesquisas vêm sendo desenvolvidas e grande parte delas visam criar métodos computacionais que sejam capazes de identificar fatores afetivos em textos. A grande maioria das pesquisas disponíveis atualmente são para a língua Inglesa, focando principalmente na identificação da polaridade nos textos, ou seja, em identificar se os textos são "positivos" ou "negativos"[Pang *et al.* 2002]. Desta forma, métodos que sejam capazes de identificar emoções em textos escritos em Português do Brasil e que classifiquem emoções em categorias (alegria, tristeza, raiva, etc.) são uma contribuição relevante para a área.

Atualmente as Máquinas de Vetores de Suporte vêm sendo utilizadas com sucesso na classificação de textos [Joachims 2002]. Neste trabalho são apresentados os resultados obtidos com a aplicação do algoritmo Máquina de Vetores de Suporte (ou *Support Vector Machine* (SVM)) na identificação das seis emoções básicas de Ekman e Friesen [Ekman e Friesen 1978] (alegria, tristeza, raiva, medo, desgosto e surpresa) em textos escritos em Português do Brasil. Neste método, o algoritmo SVM trata o problema de forma multiclasse (as seis emoções básicas além de uma sétima classe chamada

de “neutro”). Para testar o método proposto, um corpus formado por 1.750 textos foi construído. Os experimentos realizados mostram que o método é capaz de identificar a emoção predominante em 61% dos textos. Destaca-se ainda que o método não utiliza nenhum recurso linguístico adicional, como um léxico especialmente preparado para relacionar palavras ligadas à emoções (como o *WordnetAffect* [Strapparava e Valitutti 2004], por exemplo), isso faz com que o método se torne ainda mais independente.

Este trabalho está organizado da seguinte forma: a seção 2 mostra o conceito de emoções e cita alguns trabalhos que realizam análise de emoções em textos. A seção 3 descreve o método que utiliza o SVM para identificar emoções em textos para o Português do Brasil. A seção 4 relata um experimento bem como uma discussão sobre os resultados obtidos. Na seção 5 são apresentadas as conclusões e os trabalhos futuros.

2. Identificando Emoções em Textos

Esta seção apresenta os principais conceitos referentes ao processo de identificação de emoções em textos. Começamos por definir emoções nesse contexto.

2.1.Emoções

A noção precisa do que se conhece por emoção é algo ainda tão incompleto quanto o conhecimento acerca de sua importância [Roman 2007]. Várias são as definições, dependendo da área do conhecimento de onde surgem. Em termos psicológicos e comportamentais, emoções podem ser vistas como “respostas sistêmicas que ocorrem quando ações altamente motivadas são proteladas ou inibidas” [Lang 1995] ainda que estas ações não tenham ocorrido realmente [Roman 2007]. Assim, as emoções dizem respeito à execução de algo importante ao organismo [Lang 1995].

Paul Ekman e Wallace Friesen em [Ekman e Friesen 1978], ao realizar uma série de replicações dos experimentos de Darwin sobre as expressões faciais, chegaram a um padrão morfológico de cada emoção. Os estudos incluem experimentos com diferentes culturas, e seus trabalhos, assim como os de Darwin, tratam de emoções básicas como medo, surpresa, raiva, desgosto, tristeza e alegria [Roman 2007]. Em um desses experimentos, 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.

Plutchik em [Plutchik 2001] criou um modelo circunflexo tridimensional que descreve as relações entre os conceitos de emoção, que são análogas às cores de uma roda de cores (Figura 1). A dimensão vertical do cone representa a intensidade, e o círculo representa o grau de similaridade entre as emoções. Os oito setores são projetados para indicar que há oito dimensões de emoções primárias. As emoções nos espaços em branco são misturas de duas emoções primárias.

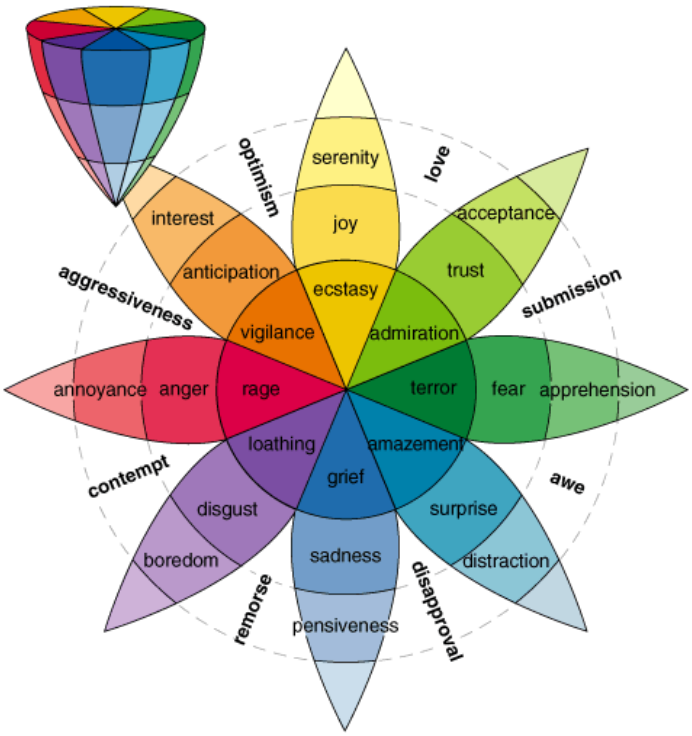

Figura 1: Circulo de emoções de Plutchik. Fonte: [Plutchik 2001]

Recentemente, o estudo das emoções tem atraído também a atenção de pesquisadores da Ciência da Computação, especialmente no que tange a interação entre homens e máquinas [Grossman e Frieder 2004]. Entre as pesquisas realizadas encontra-se a identificação automática de emoções em informação textual, conhecida como Análise de Sentimento.

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. Embora pareça algo relativamente simples, a tarefa não é trivial. 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”, onde apesar da frase conter a palavra “bom”, ela não apresenta nenhum cenário emocional, podendo ser considerada uma afirmação objetiva. [Pang e Lee 2004] e [CUCS 2005] introduzem uma nova abordagem à atividade, visando, desta forma, obter melhores resultados. Tal abordagem, segundo [Pang e Lee 2004], consiste em: (1) efetuar a distinção entre sentenças subjetivas e objetivas, devendo estas ser descartadas e; (2) aplicar às sentenças subjetivas um classificador. [CUCS 2005] afirma que os problemas enfrentados consistem, basicamente, em definir quando uma sentença é subjetiva ou objetiva e em determinar se a emoção é positiva ou negativa.

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. Mesmo assim, métodos baseados nos conceitos da Aprendizagem de Máquina

podem ser utilizados. Gupta e demais autores em [Gupta *et al*. 2010] utilizaram técnicas de Aprendizagem de Máquina supervisionadas para identificar emoções em e-mails melhorando, dessa forma, o atendimento dado a clientes que utilizam esse meio de comunicação para entrar em contato com a central de atendimento de empresas. Finatto e demais autores em [Finatto *et al*. 2011], utilizaram algoritmos de Aprendizagem de Máquina supervisionados para identificar as características que diferenciavam o texto de um jornal popular de um jornal tradicional. No trabalho de Cavalcanti e demais autores apresentado em [Cavalcanti *et al.* 2012] foi utilizada Análise de Sentimento para classificar citações, uma abordagem baseada em léxico foi utilizada e o *SentiWordNet* foi usado para identificar o grau de positividade e negatividade para cada termo extraído da citação.

Procurou-se buscar na literatura trabalhos que apresentassem semelhanças com o método desenvolvido afim de realizar a comparação dos resultados. Os trabalhos procurados na literatura deveriam ter as seguintes características: 1) identificar emoções em textos escritos em língua Portuguesa do Brasil, 2) identificar emoções e suas categorias (alegria, tristeza, raiva, medo, surpresa e desgosto) presentes nos textos e, 3) identificar emoções utilizando o algoritmo SVM ou qualquer outro algoritmo de classificação de dados. Entretanto, nessa busca não foram encontrados trabalhos onde os resultados pudessem ser comparados com os resultados do trabalho desenvolvido neste artigo.

Atualmente, existem poucos trabalhos voltados para o Português do Brasil que realizam a identificação das seis emoções básicas em textos e isso dificulta ainda mais a comparação do trabalho desenvolvido com trabalhos já existentes.

Alguns trabalhos já desenvolvidos para o Português do Brasil foram publicados por [Martinazzo 2010], [Martinazzo e Paraiso 2010] e [Dosciatti *et al.* 2012]. Nestes trabalhos foi utilizado um algoritmo baseado no método *Latent Semantic Analysis* (LSA) para identificar as emoções em textos. O LSA é um método matemático/estatístico usado na Análise de Sentimento para identificar as relações entre as palavras em textos. O resultado alcançado nestes trabalhos ficou em torno de 70% e é mostrado na Tabela 1.

Tabela 1. Resultados do LSA para cada emoção. Fonte: [Dosciatti *et al*. 2012].

Emoção	Nº. de textos	Nº. de acertos	Acurácia
Alegria	116	69	59%
Desgosto	78	60	77%
Medo	20	16	80%
Raiva	18	9	50%
Surpresa	7	6	86%
Tristeza	63	45	71%

A Tabela 1 corresponde ao conjunto de testes composto por 302 textos que foram submetidos ao algoritmo LSA. Para o treinamento foi utilizado um conjunto de 700 textos. Ambos os conjuntos são referentes à notícias jornalísticas extraídas de jornais online.

Existem outros trabalhos de identificação de emoções em textos desenvolvidos para o Português, porém estes focam principalmente a identificação da polaridade das emoções, ou seja, verificam se o texto possui uma emoção positiva ou negativa. Um desses trabalhos é o de [Souza e Vieira 2012] que utiliza um corpus de textos em Portu-

guês, extraídos de *Twitter,* onde utiliza léxicos de sentimento e avalia técnicas de pré-processamento para classificar os *tweets* em positivos e negativos. Outro trabalho desenvolvido para o Português é o de [Freitas e Vieira 2013] onde os autores propõem e avaliam métodos para identificar a polaridade das emoções em comentários de usuários de filmes e hotéis. A identificação é realizada de acordo com as características descritas em ontologias de domínio (os experimentos consideram os domínios cinema e hotel). As opiniões dos usuários foram extraídas de três sites relacionados aos domínios.

2.3. Máquina de Vetores de Suporte

O SVM é uma técnica de Aprendizagem de Máquina desenvolvida por Vapnik em 1995, que é fundamentada na Teoria de Aprendizado Estatístico e utilizada para a classificação de dados [Lorena e Carvalho 2003]. Smola e demais autores em [Smola *et al.* 1999] citam que a boa capacidade de generalização e a robustez em grandes dimensões de dados são algumas das principais características do SVM que tornam o seu uso atrativo.

Originalmente o SVM foi projetado para a classificação binária de dados. Para duas classes, poder haver muitos possíveis separadores lineares. Intuitivamente, a fronteira de decisão traçada no meio do espaço vazio entre os itens de dados das duas classes parece ser melhor do que uma que se aproxima muito de exemplos de uma ou de ambas as classes. Enquanto alguns métodos de aprendizagem como o algoritmo *Perceptron* encontra apenas um separador linear outros, como o *Naive Bayes*, procura o melhor separador linear de acordo com algum critério. Em particular, o SVM define o critério que maximiza a superfície de decisão. Esta distância, a partir da superfície de decisão para o ponto mais próximo de dados, determina a margem do classificador. Este método de construção implica necessariamente que a função de decisão para uma SVM esteja completamente especificada por um subconjunto (geralmente pequeno) dos dados, que define a posição do separador. Estes pontos são referidos como os vetores de suporte [Manning *et al*. 2008].

Diversas técnicas de Aprendizagem de Máquina foram originalmente formuladas para a solução de problemas de classificação contendo apenas duas classes. Entre elas podem se citar as SVMs e as RNAs *Perceptron*. Muitos problemas de classificação, contudo, apresentam mais de duas classes. Uma estratégia para solucionar esse problema é combinar os classificadores gerados em subproblemas binários, essa estratégia é conhecida como decomposição [Facelli *et al.* 2011]. A estratégia decomposicional que foi utilizada neste trabalho é chamada de *Um-Contra-Um* (OAO, do inglês *One-Against-One*).

3. Identificando Emoções em Textos para o Português do Brasil Usando SVM

Este artigo apresenta um método para identificar emoções em textos escritos em Português do Brasil. Um conjunto de notícias curtas (que nada mais são do que a manchete da notícia e uma pequena descrição da mesma), previamente rotuladas, é submetido ao classificador SVM. Para desenvolver e testar este método, as etapas descritas a seguir foram necessárias.

3.1. Construção do Corpus

O corpus é formado por notícias extraídas do site www.globo.com. Os textos se enquadram em variadas categorias, tradicionalmente divididas em: mundial, nacional, política, policial e econômica. Para selecionar e coletar as notícias foi utilizada uma ferramenta chamada *FeedReader*[1], que é um sofware agregrador de *Feeds*.

Para a construção do corpus, respeitou-se uma proporção de tal forma que haja um mesmo número de textos para cada emoção e mais um conjunto de textos classificados como neutro, ou seja, textos que não têm uma das seis emoções básicas como emoção predominante. Assim, o corpus é composto por 1.750 textos sendo 250 textos rotulados para cada uma das seis emoções e mais 250 para a classe "neutro".

O corpus passou por um processo de rotulação. No contexto desta pesquisa, rotular os textos é atribuir a cada texto uma das seis emoções básicas (alegria, tristeza, raiva, medo, desgosto e surpresa), ou seja, classificá-lo a partir da emoção predominante no mesmo. Caso o texto não possa ser classificado em nenhuma delas, há uma classe chamada "neutro" que poderá ser selecionada. O processo de rotulação foi realizado por dois anotadores. Um deles com experiência em linguística e outro com experiência em linguística computacional. Após o processo de rotulação individual, em função de alguns conflitos (um mesmo texto recebendo rotulação diferente), houve um processo de rotulação "consensual" para resolver as divergências encontradas. O corpus está disponível em http://www.ppgia.pucpr.br/~paraiso/mineracaodeemocoes/.

A Tabela 2 mostra alguns exemplos de notícias que compõem o corpus.

Tabela 2. Exemplos de notícias

	Trechos de Notícias	Emoção predominante
1	Bala perdida mata menina de 13 anos em Belo Horizonte: Tiro atingiu adolescente que levava irmã mais nova para escola. Disparos foram feitos durante acerto de contas de traficantes, segundo PM.	Tristeza
2	Manifestantes queimam carro em Honduras: Grupo pede o retorno de Zelaya à Presidência do país. Encontro na Costa Rica nessa quinta pode ser início do fim da crise.	Desgosto
3	Zoológico alemão apresenta tigresas quadrigêmeas ao público: Mãe deu à luz bebês sem ajuda de veterinários, afirma parque. Felinas já têm pouco mais de um mês de vida.	Alegria
4	Cobra jararaca é encontrada enrolada em carrinho de bebê no Paraná: Ainda assustada com o incidente, a dona de casa Solange Fretta disse que acordou, tomou café e ao levar o copo até a lavanderia, se deparou com a serpente.	Surpresa

3.2. Pré-processamento dos Textos

Para que os dados textuais do corpus sejam submetidos a um processamento subsequente, é preciso transformá-los em vetores numéricos. Dessa forma, técnicas de pré-processamento de texto se fazem necessárias.

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

O pré-processamento consistiu em transformar o texto para letras minúsculas, remover os acentos, remover os caracteres especiais (sinais de pontuação, hífens e números), remover as *stopwords* e aplicar *stemmer* para reduzir as palavras ao seu radical.

3.3. Construção do Vetor de Características

Dois processos importantes serão tratados aqui: a redução da dimensionalidade dos dados e a representação dos dados textuais em um vetor de características.

Quando é utilizado um corpus de textos é natural que se tenha dados de alta dimensionalidade. Com o objetivo de diminuir o número de dimensões, antes de transformar os dados pré-processados em um modelo vetorial, foram implementados dois filtros. No primeiro filtro são excluídos todos os termos que possuem um número de ocorrências inferior a um limiar pré-estabelecido, esse limiar varia de acordo com o tamanho da base de documentos e leva em consideração a hipótese de que termos muito raros são irrelevantes para a classificação do documento.

No segundo filtro é implementado o ganho de informação [MITCHELL 1997], que é usado para selecionar os termos mais representativos no conjunto de características. O ganho de informação de um termo é definido a partir das equações 1 e 2 a seguir.

A equação 1 mede a entropia, que calcula o grau de mistura de cada termo em relação às classes. Na equação 1, o conjunto de treinamento *S* pode ter *c* classes distintas.

$$Entropia(S) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i \qquad (1)$$

onde *pi* é a proporção de dados em *S* que pertencem à classe *i*.

A equação 2 define o ganho de informação de um termo *t* em relação ao conjunto de treinamento *S*.

$$GanhoDeInformação(S,t) = Entropia(S) - \sum \frac{|S_v|}{|S|} Entropia(S_v) \qquad (2)$$

onde *v* é o valor do termo.

Após o cálculo do ganho de informação para todos os termos, é realizado um ranqueamento dos termos com base no seu valor em ordem decrescente, onde são excluídos aqueles que possuem um valor abaixo de um limiar pré-estabelecido. Atualmente esse limiar é identificado através da observação dos termos com baixo teor emocional, porém se estuda uma forma de automatizá-lo. No experimento realizado neste artigo utilizou-se os termos com ganho de informação igual ou superior a 68% e também foi configurado o primeiro filtro, que se refere a frequência dos termos, para quatro, ou seja, somente permanecem no conjunto de características os termos que se repetem quatro ou mais vezes no conjunto de documentos.

Após a aplicação dos filtros é gerada a lista de características que será utilizada para submeter os dados a uma representação vetorial de características. A representação dos documentos é obtida a partir do modelo TF-IDF (*Term Frequency – Inverse Document Frequency*) [SALTON e BUCKLEY 1998]. O modelo TF-IDF gera uma eficaz representação de documentos na forma vetorial. Tal modelo expressa a ideia de que: (1)

quanto maior a frequência de um termo em um documento, mais representativo ele é para o conteúdo, e (2) quanto mais documentos contiverem um termo, menos discriminante ele é para o conteúdo. Esse modelo atribui um peso $w_{t,d}$ para um termo t em um documento d conforme as equações 3 e 4 a seguir.

$$w_{t,d} = tf_{t,d} * idf_t \tag{3}$$

$$idf_t = \log \frac{N}{df_t} \tag{4}$$

onde $tf_{t,d}$ é a quantidade de vezes que o termo t aparece no documento d, sendo N a quantidade total de documentos que compõem o conjunto de dados e df_t é a quantidade de documentos que contém o termo t.

3.4. Treinamento e Teste

Tendo definido a lista de características e os documentos devidamente rotulados e representados em um vetor de características, pode-se passar para a etapa de treinamento e teste do classificador SVM.

Para submeter o conjunto de dados ao algoritmo SVM optou-se por utilizar a implementação LibSVM criada por [Chang e Lin 2011] e incorporada ao Weka [Weka] por [EL-Manzalawy e Honavar 2005]. O motivo dessa escolha deu-se pelo fato deste recurso ser facilmente integrável ao projeto, desenvolvido em linguagem Java.

4. Experimento e Discussão dos Resultados

Um experimento foi realizado com um conjunto de 1.750 textos previamente rotulados e balanceados. Para treinar e testar o classificador SVM foi utilizado o método de validação cruzada com 10 partições. O conjunto de características é composto por 1.565 termos. O resultado obtido neste experimento, utilizando SVM com kernel Linear é apresentado na Tabela 3.

As medidas de avaliação utilizadas no experimento foram *Precision* (equação 5), *Recall* (equação 6) e *F-measure* (equação 7) e são mostradas na sequência.

$$\mathrm{Pr}\,ecision = \frac{VP}{(VP + FP)} \tag{5}$$

$$\mathrm{Re}\,call = \frac{VP}{(VP + FN)} \tag{6}$$

$$F - measure = \frac{2 \cdot \mathrm{Pr}\,ecision \cdot \mathrm{Re}\,call}{(\mathrm{Pr}\,ecision + \mathrm{Re}\,call)} \tag{7}$$

onde VP são verdadeiros positivos, FP são falsos positivos e FN são falsos negativos.

Neste experimento foi obtida uma precisão de 60.7% do classificador SVM, na configuração multiclasse, ao identificar a emoção predominante em cada texto.

Tabela 3. Resultados obtidos com LibSVM (kernel Linear)

Emoção	Acurácia	*Precision*	*Recall*	*F-Measure*
Neutro	0,50	0.52	0,50	0.51
Alegria	0,45	0.48	0,45	0.46
Desgosto	0,39	0.42	0,39	0.40
Medo	0,81	0.72	0,81	0.76
Raiva	0,75	0.76	0,75	0.75
Surpresa	0,81	0.75	0,81	0.78
Tristeza	0,54	0.55	0,54	0.54
Média	**0,61**	**0.58**	**0,61**	**0.60**

O corpus utilizado no experimento também foi testado com outros dois algoritmos de classificação, o *Naive Bayes* e o *K-Nearest Neighbors* (KNN). Os resultados podem ser visualizados nas tabelas 4 e 5.

Tabela 4. Resultados obtidos com *Naive Bayes*

Emoção	Acurácia	*Precision*	*Recall*	*F-Measure*
Neutro	0,44	0.49	0.44	0.46
Alegria	0,40	0.40	0.40	0.40
Desgosto	0,34	0.35	0.34	0.34
Medo	0,63	0.55	0.63	0.58
Raiva	0,62	0.60	0.62	0.61
Surpresa	0,54	0.60	0.54	0.56
Tristeza	0,46	0.44	0.46	0.45
Média	**0,49**	**0.49**	**0.49**	**0.49**

Tabela 5. Resultados obtidos com KNN (n=5)

Emoção	Acurácia	*Precision*	*Recall*	*F-Measure*
Neutro	0,44	0.46	0.44	0.45
Alegria	0,24	0.41	0.24	0.31
Desgosto	0,47	0.36	0.47	0.41
Medo	0,72	0.71	0.72	0.72
Raiva	0,74	0.77	0.74	0.75
Surpresa	0,70	0.73	0.70	0.72
Tristeza	0,45	0.38	0.45	0.41
Média	**0,54**	**0.55**	**0.54**	**0.54**

O *Naive Bayes* (Tabela 4) apresentou uma acurácia de 49% ao identificar as emoções nos textos, enquanto o KNN (Tabela 5) obteve uma acurácia de 54%.

Analisando os dados obtidos, podemos verificar que os classificadores tiveram resultado inferior a 50% ao classificar textos referentes às emoções “alegria” e “desgosto”. Isso se deve ao fato do classificador ter confundido as emoções “alegria” com “desgosto” e “desgosto” com “neutro”, “alegria” e “tristeza”. Com relação a confusão das emoções “desgosto” e “tristeza”, esse fato é explicável uma vez que, conforme pode-se observar no círculo de emoções de Plutchik (Figura 1), essas duas emoções são muito próximas. Isso significa que textos que se referem a essas duas emoções podem conter

termos iguais ou muito parecidos entre si. Todavia, o fato do classificador ter confundido as emoções “alegria” com “desgosto” e “desgosto” com “neutro” e “alegria”, leva-se a acreditar que isso ocorra em função de existirem termos presentes nessas duas emoções que também estão presentes em textos neutros.

5. Conclusões e Trabalhos Futuros

Este artigo apresenta uma abordagem multiclasse para identificar emoções presentes em textos jornalísticos. Os resultados obtidos neste trabalho não puderam ser comparados com trabalhos já existentes devido a inexistência de trabalhos publicados na literatura que façam a identificação das seis emoções básicas em textos escritos em Português do Brasil. Assim, foi realizada a comparação dos resultados obtidos com o classificador SVM com os classificadores *Naive Bayes* e KNN, sendo que o SVM apresentou um melhor resultado.

Além do método de classificação de emoções em textos em Português do Brasil, uma das principais contribuições deste trabalho é tornar disponível um corpus de textos que poderá ser utilizado em outras pesquisas de Análise de Sentimento.

Como trabalhos futuros, pretende-se melhorar os resultados obtidos neste artigo ampliando-se o número de textos do corpus e aperfeiçoando-se a forma de extrair as melhores características dos textos utilizando Algoritmos Genéticos. Também pretende-se utilizar outros tipos de textos que possam conter emoções, como posts de blogs, por exemplo.

Outra questão que se pretende abordar na pesquisa é estudar o impacto dos regionalismos no processo de identificação das emoções. Textos escritos por pessoas de diferentes regiões do país podem apresentar diferenças ao expressar as emoções?

6. Referências

Cavalcanti, D. C. Prudêncio, R. B. C. Pradhan, S. S. Shah, J. Y. Pietrobon, R. S. (2012) Análise de Sentimento em Citações Científicas para Definição de Fatores de Impacto Positivo In: IV International Workshop on Web and Text Intelligence, Curitiba, Brazil. October/2012.

Chang, C. C. Lin, C. J. (2011) LIBSVM - A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, vol 2.

Cornell University (CU) Computer Science (CS).(2005) Sentiment Analysis. CS 40th anniversary Symposium, p. 26-27.

Dosciatti, M. M. Martinazzo, B. Paraiso, E. C. (2012) Identifying Emotions in Short Texts for Brazilian Portuguese. In: IV International Workshop on Web and Text Intelligence, Curitiba

Ekman, P. Friesen, W. V. (1978) Facial Action Coding System. Palo Alto: Consulting Psychologists Press.

EL-Manzalawy, Y. Honavar, V. (2005) WLSVM: Integrating LibSVM into Weka Environment. Software disponível em http://www.cs.iastate.edu/~yasser/wlsvm

Facelli, K. Lorena, A. C. Gama, J. Carvalho, A. C. P. L. F. (2011) Inteligência Artificial - Uma abordagem de Aprendizagem de Máquina. LTC. Rio de Janeiro.

Finatto, M. J. B. Scarton, C. E. Rocha, A. Aluísio, S. (2011) Características do Jornalismo Popular: Avaliação da Inteligibilidade e Auxílio à Descrição do Gênero. Simpósio Brasileiro em Tecnologia da Informação e da Linguagem Humana – STIL, Mato Grosso.

Freitas, L. A. Vieira, R. (2013) Ontology-based Feature Level Opinion Mining for Portuguese Reviews. In: 22nd International World Wide Web Conference. Rio de Janeiro.

Grossman, D. A. Frieder, O. (2004) Recuperação de Informação: Algoritmos e Heurísticas. 2 ed. Springer, Holanda.

Gupta, N. Gilbert, M. Di Fabbrizio, G. (2010) Emotion Detection in Email Customer Care. In NAACL-HLT (Conference of the North American Association for Computational Linguistics: Human Language Technologies), Los Angeles, California.

Joachims, T. (2002) Learning to Classify Text using Support Vector Machines, Cornell University - Department of Computer Science, Kluwer Academic Publishers/Springer.

Lang, P. (1995) The emotion probe: Studies of motivation and attention. American Psychologist, 50(5): 372–385.

Liu, B. (2010) Sentiment Analysis and Subjectivity. In. Handbook of Natural Language Processing. Segunda Edição.

Lorena, A. C. de Carvalho, A. C. P. L. F. (2003) Introdução às Máquinas de Vetores Suporte (Support Vector Machines). Relatório Técnico nº 192 do Instituto de Ciências Matemáticas e de Computação da USP.

Manning, C. D.Raghavan, P. Schütze, H. (2008) Introduction to Information Retrieval, Cambridge University Press.

Martinazzo, B. (2010) Um Método de Identificação de Emoções em Textos Curtos para o Português do Brasil. PUCPR, Curitiba.

Martinazzo, B. Paraiso, E. C. (2010) Identificação de Emoções em Notícias Curtas. In: CLEI - Conferência Latino-Americana de Informática, Assunção, Paraguai.

Mitchell, T. (1997) Machine Learning, McGraw-Hill.

Pang, B. Lee, L. (2004) A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of the 42nd ACL, pp. 271-278.

Pang, B. Lee, L. Vaithyanathan, S. (2002) Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). Philadelphia, Pennsylvania, pp. 79–86.

Plutchik, P. (2001) A Nature of Emotions. American Scientis 89:344-350.

Roman, N. T. (2007) Emoção e a Sumarização Automática de Diálogos. Tese de doutorado do Instituto de Computação da Unicamp.

Salton, G. Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. Information Processing and Management – Cornell University, Ithaca.

Smola, A. J. Barlett, P. Schölkopf, B. Schuurmans, D. (1999) Introduction to Large Margin Classifiers. MIT Press.

Souza, M. Vieira, R. (2012) Sentiment Analysis on Twitter Data for Portuguese Language.International Conference on Computational Processing of the Portuguese Language - PROPOR 2012. Coimbra, Portugal, 241-247.

Strapparava, C. Valitutti, A. (2004) WordNet Affect: an affective extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisboa, Portugal.

Weka: Data Mining Software in Java. The University of Waikato.Disponível em http://www.cs.waikato.ac.nz/ml/weka/