Estudo do Impacto de um Corpus Desbalanceado na Identificação de Emoções em Textos

Lohann Paterno Coutinho Ferreira[1], Mariza Miola Dosciatti[2], Emerson Cabrera Paraiso[3]
Pontifícia Universidade Católica do Paraná, Curitiba, PR, Brasil
lohann.ferreira@pucpr.br, {mariza.dosciatti, paraiso}@ppgia.pucpr.br

Resumo - A identificação automática de emoções em textos tem mostrado resultados significativos em várias aplicações. Os classificadores SVM (Support Vector Machine) têm sido utilizado em muitos métodos para identificar emoções em textos por apresentarem boa capacidade de generalização e robustez com dados com alta dimensionalidade. No entanto, a maioria dos corpora textuais submetidos aos métodos são naturalmente desbalanceados em relação ao número de textos de emoção e assim os classificadores SVM geralmente classificam a maioria dos textos para as classes majoritárias. Neste artigo apresentamos uma abordagem baseada em Algoritmo Genético desenvolvida para equilibrar um corpus de texto e analisar o impacto do uso dessa abordagem em um método de identificação de emoções em textos.

Palavras-chave - Desbalanceamento; Algoritmo Genético; SVM (Support Vector Machine); Análise de Sentimento.

I. INTRODUÇÃO

Os recentes avanços na análise de texto levam ao surgimento de uma nova área responsável pelo reconhecimento dos aspectos subjetivos tais como opiniões, sentimentos e emoções em textos. A Análise de Sentimento é o campo de pesquisa dedicado ao estudo de emoções e mineração de opiniões. Trata-se de um problema desafiante de processamento de linguagem natural ou mineração de textos. Devido ao seu valor para aplicações práticas, houve um crescimento tanto em pesquisas na área acadêmica quanto em aplicações comerciais [Liu, 2012]. Conforme Liu em [Liu, 2010], a investigação na área começou com classificação de sentimento e subjetividade, que tratou o problema como classificação de texto: classifica-se um documento opinativo (por exemplo, análise de produtos) ou frase que expressa uma opinião positiva ou negativa [Pang e Lee, 2008].

Emoção, um elemento importante da natureza humana, tem sido amplamente estudada em psicologia e ciências do comportamento. Ela também tem atraído a atenção de pesquisadores da ciência da computação e, em particular, da linguística computacional. Houve progressos na pesquisa sobre polaridade e análise de sentimentos, mas menos trabalhos têm sido desenvolvidos em reconhecimento automático de emoções em textos [Ghazi et al., 2010]. Supomos que as emoções presentes em um texto não são independentes de sua polaridade; portanto, tentamos encontrá-las e queremos aplicar métodos de classificação para isso. Além disso, a maioria das pesquisas desenvolvidas são para o idioma Inglês [Strapparava e Mihalcea, 2008], [Chaffar e Inkpen, 2011], [Ulinski et al., 2012]. Para suprir esta lacuna, estamos desenvolvendo métodos para identificar emoções em textos para o Português do Brasil [Dosciatti et al., 2012], [Dosciatti et al., 2013]. O fato dos métodos já existentes serem desenvolvidos para outros idiomas nos motiva a criar métodos para o Português Brasileiro, e a partir disso, podemos comparar tais métodos com os demais e também analisar o seu comportamento ao ser submetido a outros idiomas.

Atualmente, as pesquisas sobre as emoções são divididas em diferentes áreas, mas o que será discutida e utilizada neste trabalho é a mais simples delas, chamada de Emoções Básicas (ou puras). Este conceito está relacionado com as emoções inatas compartilhadas entre todas as culturas do mundo e foi proposto na década de 1970 por Paul Ekman e Wallace Friesen [Ekman e Friesen, 1978]. Como não há consenso sobre quantas e quais são essas emoções, o modelo proposto por Ekman e Friesen se refere a seis delas: alegria, tristeza, raiva, medo, desgosto e surpresa.

Em muitos domínios, os conjuntos de dados textuais utilizados para a identificação de emoções em textos são desbalanceados. Em [Strapparava e Mihalcea, 2008] os autores utilizaram um corpus de texto extraído de blogs composto de 8.761 textos em que 55% deles foram rotulados pela emoção (ou classe) alegria e apenas 0,8% com desgosto. Em [Ghazi et al., 2010] foi utilizado um corpus de texto extraído de blogs contendo 4.090 sentenças rotuladas sendo 68% delas rotuladas como não-emocionais e 3% delas rotuladas como pertencentes às emoções medo e surpresa. Em [Dosciatti et al., 2012] foi utilizado um corpus de texto contendo 1.002 notícias curtas, onde cerca de 38% delas foram rotuladas com alegria e cerca de 2% com surpresa. Neste artigo, apresentamos uma abordagem que se baseia em Algoritmos Genéticos para balancear um corpus de textos, a fim de compreender o impacto dessa ação do ponto de vista da classificação. O corpus equilibrado por meio desta abordagem foi submetido a um método desenvolvido para identificar emoções em textos e, dessa forma, foi possível avaliar o impacto de tal operação.

O restante deste artigo está organizado da seguinte forma: a seção 2 define dados desbalanceados e apresenta algumas referências nas quais o problema é tratado no contexto da aprendizagem de máquina, a seção 3 introduz o método de identificação de emoções em que a abordagem apresentada neste artigo poderá ser aplicada, a seção 4 apresenta efetivamente a abordagem baseada em Algoritmo Genético para fazer o balanceamento a nível de dados, a Seção 5 apresenta os resultados obtidos a partir dos experimentos que

foram realizados e a seção 6 apresenta as conclusões e os trabalhos futuros.

II. Dados Desbalanceados em Aprendizagem de Máquina

Os dados desbalanceados correspondem a domínios onde algumas classes de um conjunto de dados estão representadas por um grande número de exemplos, enquanto que outras classes são representadas por poucos exemplos [Batista, 2003]. Conjuntos de dados não balanceados podem ser encontrados em muitas áreas, como por exemplo na detecção de fraudes em transações de cartão de crédito [Fawcett e Provost, 1997], na análise de risco para as seguradoras [Stolfo et al., 1997], no diagnóstico médico [Silva et al., 2009], na detecção de falhas [Carvalho et al., 2008], no reconhecimento de assinaturas [Souza et al., 2010], na categorização de texto [Li e Shawe-Taylor, 2003], [Manevitz e Yousef, 2007], entre outras áreas.

Vários pesquisadores investigaram o problema e propuseram algumas abordagens para minimizar os efeitos do mesmo [Japkowicz e Stephen, 2002], [Batista et al., 2004], [Khoshgoftaar et al., 2010]. A partir desses estudos, duas abordagens principais emergiram. A primeira abordagem tenta equilibrar a distribuição das classes no conjunto de dados. Duas técnicas principais são utilizadas neste caso: *undersampling*, que elimina aleatoriamente exemplos da classe majoritária [Kubat e Matwin, 1997] e *oversampling*, que replica de forma aleatória exemplos da classe minoritária [Ling e Li, 1998]. Ambas as técnicas podem ser combinadas em um único conjunto de dados [Estabrooks et al., 2004]. Ambas as técnicas têm desvantagens. A técnica de *undersampling* pode causar a perda de informação útil, enquanto que a *oversampling* pode aumentar o tamanho do conjunto de dados sem qualquer ganho de informação [Provost, 2000]. Com base nestas técnicas, algumas pesquisas propõem o uso de heurísticas para selecionar os exemplos a serem replicados ou removidos a partir de um conjunto de dados, a fim de minimizar a quantidade de dados úteis descartados [Chawla et al., 2002], [Milaré et al., 2012], [Beckmann, 2010].

Na segunda abordagem, também conhecida como balanceamento de dados a nível de algoritmo, os algoritmos de aprendizagem são modificados ou ajustados para possibilitar o aprendizado com conjuntos de dados desbalanceados. Por exemplo, no caso do algoritmo SVM, o hiperplano pode ser "empurrado" para mais perto da fronteira das classes majoritárias por meio da realização de ajustamentos especiais durante o cálculo do limite [Akbani et al., 2004]. Outra estratégia no caso do SVM é aplicar penalidades quando o erro é cometido com os exemplos da classe minoritária [Wang e Japkowicz, 2008].

Em Análise de Sentimento, percebe-se que os corpora são desbalanceados em relação ao número de textos para cada emoção. Em [Rangel e Rosso, 2013] os autores utilizaram um conjunto de dados onde 29% dos textos extraídos do Facebook foram rotulados com surpresa e apenas 0,2% foram rotulados com medo. Nesses experimentos, foi utilizado um algoritmo de classificação SVM. Os resultados relatam que o classificador é muito sensível ao desequilíbrio das classes, como era esperado.

Em muitos domínios, os conjuntos de dados são naturalmente desbalanceados. Identificar as emoções em notícias, blogs ou tweets pode ser o caso. No entanto, nesta pesquisa, estamos tentando entender como isso afeta o processo de classificação. Para tal, inicialmente precisamos de um método de identificação de emoções em textos. Este será apresentado na próxima seção.

III. Um Método de Identificação de Emoções em Textos em Português Brasileiro

O método de identificação de emoções em textos apresentado aqui tem a finalidade de identificar a emoção predominante em cada texto. Esses textos são normalmente extraídos da internet e as emoções identificadas pelo método são as seis emoções básicas de [Ekman e Friesen, 1978]: alegria, tristeza, raiva, medo, desgosto e surpresa, e também uma categoria "neutro" para indicar os textos que não possui nenhuma das emoções básicas. O método também permite identificar a polaridade das emoções. O método foi desenvolvido em linguagem Java e utiliza uma abordagem supervisionada baseada no SVM para classificar os textos.

Ao submeter um conjunto de dados textuais rotulados ao método, ele pré-processará os textos a partir de três etapas principais: a preparação dos dados, a seleção dos atributos e redução da dimensionalidade e a representação vetorial.

Na etapa de preparação dos dados os textos são modificados para conter apenas letras minúsculas, são removidos os acentos, os caracteres especiais e as *stopwords* e é aplicado um *stemmer* por meio de uma extensão do *stemmer Snowball*[1]. Na etapa de seleção dos atributos e redução da dimensionalidade, os dados são submetidos ao processo *Bag-of-Words* [Radovanovic e Ivanovic, 2008] afim de gerar uma lista de termos que será submetida a dois filtros sendo que no primeiro os termos raros são removidos e no segundo é utilizado o Ganho de Informação [Mitchell, 1997] para selecionar os atributos mais representativos do conjunto de termos. A terceira etapa consiste em gerar a lista de atributos que será utilizada para submeter os conjuntos de dados textuais a uma representação vetorial. A representação vetorial foi realizada a partir do modelo TF-IDF (*Term Frequency – Inverse Document Frequency*) [Salton e Buckley, 1988].

Na sequência, com dados pré-processados, o método treina e testa um classificador baseado no SVM e a partir disso pode classificar novos textos não rotulados. O classificador foi configurado com um kernel RBF (*Radial Basis Function*) em uma configuração padrão de gama = 0 e custo = 1 e validação cruzada com 10 *folds*. Esta etapa também poderá ser realizada por meio da integração da abordagem descrita na seção seguinte com o método. Essa abordagem permite que o método treine o classificador com dados balanceados e consiga ter uma maior representatividade para cada classe.

[1] http://snowball.tartarus.org/

IV. UMA ABORDAGEM BASEADA EM ALGORITMO GENÉTICO PARA BALANCEAMENTO DE CORPUS DE TEXTO

A abordagem baseada em Algoritmo Genético (GA) desenvolvida neste artigo visa equilibrar um corpus de texto, de modo que, quando o corpus é usado para treinar e testar um algoritmo de classificação, o classificador gerado consiga aumentar o valor da medida *F-Measure* (F1) para cada uma das classes e balancear o valor de F1 para todas as classes.

O algoritmo funciona de forma iterativa, onde a cada iteração realiza *oversampling* e/ou *undersamping* no conjunto de dados, fazendo o treinamento e o teste do classificador baseado no SVM. O desempenho do classificador gerado é avaliado através de uma função *fitness* que calcula a média geométrica da medida F1 de todas as classes.

Inicialmente, o GA recebe um conjunto de dados rotulados, pré-processados e representados de forma vetorial. Em seguida, ele cria a estrutura do indivíduo, que neste problema é o número de instâncias do conjunto de dados (cada gene de um indivíduo representa uma instância), e de forma aleatória inicializa uma população fixa de indivíduos. A inicialização aleatória do indivíduo é dada pelas seguintes etapas:

1. É inicializado com um número de genes igual a *n* (sendo *n* o número de instâncias do conjunto de treinamento), onde para cada gene é atribuído o valor 1.
2. Para o *undersampling* está associada uma probabilidade baixa (da ordem de 10%) para cada um dos genes do indivíduo e, em seguida, é selecionado aleatoriamente um valor entre 0 e 1. Se o valor selecionado é menor do que a probabilidade predeterminada, é atribuído o valor de 0 para aquele gene.
3. Para o *oversampling* é selecionado aleatoriamente um gene que terá o seu valor incrementado por 1, esse passo é executado um número de vezes igual a um percentual (pré-definido) do número total de genes do indivíduo. Por exemplo, se o indivíduo tem 100 genes e o percentual definido for 20, então serão sorteadas 20 posições que terão seus valores incrementados por 1. É importante ressaltar que uma posição pode ser selecionada e incrementada mais do que uma vez de forma aleatória.

O valor do percentual definido na etapa 3 irá depender do grau de desbalanceamento de cada conjunto de treino. Em experiências com corpus de notícias coletados da internet, observou-se que definir o valor percentual com 20 deixa a aleatoriedade mais homogênea e que esta distribuição, ao final de cada geração, produz em média 10% de *undersampling*, 20% de *oversampling* e 70% dos casos permanecem sem *undersampling* e sem *oversampling*. Assim, o GA converge para melhores resultados de forma mais rápida. Na Figura 1, são mostrados três etapas da iniciação aleatória em indivíduos com 24 genes.

```
Etapa 1: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Etapa 2: [0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1]
Etapa 3: [0, 1, 1, 0, 1, 1, 1, 1, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2]
```

Figura 1. Exemplo de iniciação aleatória de uma população de indivíduos.

Com base no cromossomo de cada indivíduo é gerado um novo conjunto de dados que é usado para treinar e testar o classificador usando validação cruzada com 10 partes. O desempenho do classificador gerado é avaliado pela função de avaliação *(fitness)* apresentada na Equação 1.

$$Fitness = \sqrt[i]{A_1 * ... * A_i} \quad (1)$$

Sendo A_1, ..., A_i, respectivamente, a medida F1 para cada uma das *i* classes.

A função *fitness* calcula efetivamente a média geométrica dos *n* valores da medida F1 (um valor para cada uma das classes) gerados pelo classificador.

Uma estratégia de elitismo é aplicada à população de indivíduos permitindo que os *n* melhores indivíduos de cada geração sejam passados para a próxima geração para assegurar que suas características sejam preservadas. Para selecionar os pais que irão gerar os novos indivíduos da população foi utilizado o método da Roleta Viciada [Linden, 2012].

Nos pais selecionados são aplicados operadores genéticos de modo a formar uma nova população de indivíduos. Os operadores genéticos utilizados neste GA foram o cruzamento aritmético (que pode ocorrer em 80% dos casos) e uma variação da mutação aleatória (com uma probabilidade de ocorrer em 20% dos casos). A mutação aleatória foi adaptada para tratar de forma mais eficaz o problema dos dados desbalanceados.

No cruzamento aritmético é definido randomicamente um parâmetro λ ($0 \leq \lambda \leq 1$) e calculado cada posição do primeiro descendente usando a Equação 2.

$$c_l^{\text{Filho1}} = \lambda c_l^1 + (l - \lambda) c_l^2 \quad (2)$$

sendo *l* o índice da posição do vetor que varia de *1* a *n* onde *n* é o número de instâncias do conjunto de dados.

Na Figura 2 é apresentado um exemplo da aplicação da Equação 2 em pais selecionados para gerar a descendência por cruzamento aritmético.

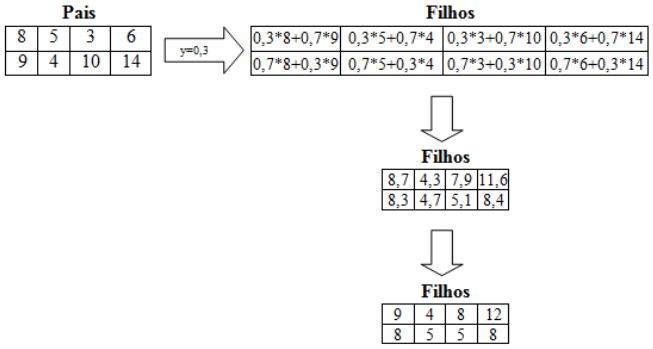

Figura 2. Exemplo da operação do cruzamento aritmético. (Fonte: Adaptado de [Linden, 2012], p. 221)

Para o operador de mutação é definido um índice de variação σ, sendo (σ > 0) para delimitar a faixa mínima e máxima do valor de mutação para cada gene.

O valor do limite inferior (Lim^{Inf}) é definido por:

$$Lim^{Inf} = \begin{cases} 0, & if \quad x-\sigma \leq 0 \\ x-\sigma, & if \quad x-\sigma > 0 \end{cases}$$

O valor do limite superior (Lim^{Sup}) é definido por:

$$Lim^{Sup} = x+\sigma$$

Assim, é associada uma probabilidade baixa (cerca de 0,5%) para cada gene individual e aleatoriamente escolhido um valor entre 0 e 1. Se o valor escolhido ao acaso é menor do que a probabilidade predeterminada, então o operador atua sobre o gene em questão, atribuindo um valor aleatório $y\{y \in N \mid y \in [Lim^{Inf}, Lim^{Sup}]\}$ para o gene. Por exemplo, se σ =3 e o valor do gene é igual a 7, então, $Lim^{Inf} = 4$ e $Lim^{Sup} = 10$. Se o teste de probabilidade resultar em verdadeiro, então o operador de mutação atuará no gene em questão sorteando um novo valor entre 4 e 10 para este gene.

Em seguida são avaliados todos os novos indivíduos e inseridos na nova geração. Se o critério de parada não for alcançado, seleciona-se os *n* melhores indivíduos da população e repete-se todo o processo a partir deste ponto até que o critério de parada seja atingido.

Esta abordagem, baseada em um GA, faz o balanceamento do número de instâncias de cada classe somente nas partes da validação cruzada utilizadas para treinar o classificador. Assim, as instâncias presentes nos *folds* utilizados para testar o classificador não contêm instâncias repetidas. O critério de parada do GA ocorre no momento em que a execução do algoritmo estabilizar, ou seja, quando o algoritmo não avançar para uma melhor solução durante o período de 100 gerações.

V. RESULTADOS EXPERIMENTAIS

Nesta seção são apresentados os resultados de dois experimentos. O objetivo de cada experimento é mostrar o comportamento do método de identificação de emoções em duas situações: primeiro, quando for submetido ao método um conjunto de dados textuais naturalmente desbalanceados e segundo, quando for submetido ao método um conjunto de dados balanceados através da abordagem do GA. Os textos utilizados nos experimentos são procedentes de um corpus de notícias que apresenta um desequilíbrio no número de instâncias em cada classe.

Os textos que compõem o corpus de notícias foram coletados a partir de sites de notícias como *www.globo.com* e pertencem às categorias tradicionalmente divididas em "global", "política", "polícia" e "economia". Cada texto tem uma média 35 palavras e para facilitar o processo de coleta de notícias foi utilizada a ferramenta *FeedReader*[2], que é um agregador de *feeds*.

O corpus de notícias utilizado no primeiro experimento contém 1.531 textos rotulados [Dosciatti et al., 2013] distribuídos em 280 (18%) textos de alegria, 226 (15%) de desgosto, 160 (10%) do medo, 168 (11%) de raiva, 172 (11%) de surpresa, 306 (20%) de tristeza e 219 (14%) textos que não possuem nenhuma dessas seis emoções, sendo rotulados com a classe "neutro". Esse número de textos é o resultado final de um processo baseado em N-Gram onde todos os textos com um grau de similaridade superior a 70% foram removidos. Na etapa de pré-processamento do método foram considerados como atributos somente os termos que apresentaram um ganho de informação superior a 70% e que ocorreram no mínimo cinco vezes no conjunto de documentos.

Os dados naturalmente desbalanceados e os dados balanceados pela abordagem do GA foram submetidos ao método de identificação de emoções e os resultados podem ser visualizados na Tabela 1.

Tabela 1. Experimento 1 - Desempenho do método de identificação de emoções com sete classes.

Classe	*Precisão*		*Revocação*		*F1*		*Desvio Padrão de F1*	
	Sem GA	*Com GA*	*Sem GA*	*Com GA*	*Sem GA*	*Com GA*	*Sem GA*	*Com GA*
Neutro	0,667	0,568	0,110	0,342	0,188	0,427	0,128	0,066
Alegria	0,344	0,433	0,857	0,621	0,491	0,510		
Desgosto	0,571	0,415	0,142	0,301	0,227	0,349		
Medo	0,839	0,670	0,163	0,381	0,272	0,486		
Raiva	0,736	0,717	0,232	0,393	0,353	0,508		
Surpresa	0,745	0,684	0,203	0,378	0,320	0,487		
Tristeza	0,393	0,427	0,784	0,775	0,524	0,551		
Acurácia Sem GA: 41,5%			Acurácia Com GA: 48,7%					

Ao decidir-se utilizar a abordagem do GA para fazer o balanceamento dos dados, consideramos que mais importante do que alcançar um alto valor de acurácia, é tentar obter o maior valor possível da medida F1 para todas as classes. O fato do GA possibilitar esse balanceamento faz com que o classificador seja treinado com um número representativo de instâncias para cada classe reduzindo dessa forma os efeitos negativos da sensibilidade que o SVM tem ao ser treinado com dados desbalanceados. Na Tabela 1 pode-se visualizar o aumento do valor de F1 para todas as classes nos dados

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

balanceados pela abordagem do GA e também um balanceamento do valor de F1 de um modo geral. O desvio padrão dos valores de F1 obtidos pelo método com os dados balanceados foi consideravelmente reduzido.

No segundo experimento foram utilizados apenas os textos com polaridade positiva e negativa do corpus de notícias utilizado no experimento anterior. Os 1.204 textos foram distribuídos em 311 (26%) textos positivos (classe "alegria") e em 893 (74%) textos negativos (demais classes). Os textos neutros não foram considerados neste experimento. Foram mantidas as configurações do classificador SVM do experimento anterior e no pré-processamento foram considerados como atributos apenas os termos que apresentaram um ganho de informação superior a 60% e que ocorreram no mínimo quatro vezes no conjunto de documentos. Os resultados obtidos podem ser visualizados na Tabela 2.

Tabela 2. Experimento 2 - Desempenho do método de identificação de emoções com duas classes.

Classe	*Precisão*		*Revocação*		*F1*		*Desvio Padrão de F1*	
	Sem GA	*Com GA*	*Sem GA*	*Com GA*	*Sem GA*	*Com GA*	*Sem GA*	*Com GA*
Positivo	0,873	0,858	0,331	0,408	0,480	0,553	0,288	0,242
Negativo	0,808	0,826	0,983	0,976	0,887	0,895		
Acurácia Sem GA: **81,5%** Acurácia Com GA: **83,0%**								

Neste experimento com polaridade os dados apresentavam um alto grau de desbalanceamento em relação aos dados do experimento anterior. Mesmo assim os resultados obtidos com abordagem do GA mostram que o valor de F1 das duas classes foram maiores do que quando se utilizou os dados desbalanceados. Esse valor foi maior para a classe minoritária e menor para a classe majoritária, como era esperado. O desvio padrão de F1 também foi reduzido ao se utilizar os dados balanceados.

Considerando que um bom desempenho para o método de identificação de emoções é obter um aumento e um balanceamento do valor de F1 para todas as classes e com base nos resultados obtidos com os experimentos foi possível observar que o desempenho do método ao utilizar a abordagem do GA está diretamente ligado com o grau de desbalanceamento do corpus. Os resultados obtidos com o método utilizando a abordagem do GA no primeiro experimento, em que o corpus original (antes de ser submetido ao GA) apresentava um grau médio de desbalanceamento, foram superiores aos resultados obtidos com o método utilizando a abordagem do GA no segundo experimento, em que o corpus original possuía um alto grau de desbalanceamento.

Apesar do desempenho do método variar conforme o grau de desbalanceamento do corpus, ainda assim é válido o uso dessa abordagem para o balanceamento dos dados. Pois, mesmo num caso como o do segundo experimento, onde o corpus original era altamente desbalanceado, o uso dos dados balanceados pela abordagem permitiu reduzir a diferença do valor de F1 entre as duas classes.

VI. Conclusões e Trabalhos Futuros

Este estudo teve como objetivo verificar o comportamento de um método de identificação de emoções em textos ao utilizar textos naturalmente desbalanceados e textos que foram submetidos a um processo de balanceamento. Dessa forma, foi apresentada neste artigo uma abordagem baseada em GA para balanceamento dos dados.

Através dos experimentos foi possível verificar que o método obteve um aumento dos valores de F1 para todas as classes ao utilizar o corpus balanceado pela abordagem do GA. Também o balanceamento dos valores F1 de um modo geral, contribuiu para diminuir os efeitos negativos que um classificador SVM tem ao ser treinado com dados desbalanceados, isto é, tender a classificação para as classes que tiveram maior representatividade durante o treinamento.

Em trabalhos futuros pretende-se comparar o desempenho da abordagem do GA apresentado neste artigo com outras abordagens utilizadas para reduzir os efeitos do desbalanceamento a nível de dados, por exemplo, soluções que tratam o problema somente através do *undersampling* como é o caso *EasyEnsemble* e *BalanceCascade* propostos por [Liu et al., 2006] e algoritmos e soluções que tratam o desbalanceamento realizando apenas o *oversampling*, como é o caso do algoritmo SMOTE (*Synthetic Minority Over-Sampling Technique*), proposto por [Chawla et al., 2002]. Também pretende-se testar o método de identificação de emoções com abordagens que tratam desbalanceamento dos dados a nível de algoritmo, como é o caso de [Wang e Japkowicz, 2008].

Agradecimentos

Esta pesquisa tem o apoio financeiro da PUCPR e da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

Referências

1. Liu, B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. University of Illinois at Chicago. Morgan & Claypool Publishers, Maio (2012)
2. Liu, B. Sentiment Analysis: A Multifaceted Problem. IEEE Intelligent Systems, vol. 25(3), pp. 76-80 (2010)
3. Pang, B., Lee, L. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, vol. 2(1-2), pp. 1-135 (2008)
4. Ghazi, D., Inkpen, D., Szpakowicz, S. Hierarchical Approach to Emotion Recognition and Classification in Texts. Advances in Artificial Intelligence. 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, pp. 40-50 (2010)
5. Strapparava, C., Mihalcea R. Learning to Identify Emotions in Text. 23rd Annual ACM Symposium on Applied Computing, pp. 1556-1560, Fortaleza (2008)
6. Chaffar, S., Inkpen, D. Using a Heterogeneous Dataset for Emotion Analysis in Text. 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada (2011)

7. Ulinski, M., Soto, V., Hirschberg, J. Finding emotion in image descriptions. Proceeding WISDOM 12, Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, Article No. 8, ACM, New York, USA (2012)
8. Dosciatti, M. M., Martinazzo, B., Paraiso, E. C. Identifying Emotions in Short Texts for Brazilian Portuguese. In: IV International Workshop on Web and Text Intelligence, Curitiba (2012)
9. Dosciatti, M. D., Ferreira, L. P. C., Paraiso, E. C. Identificando Emoções em Textos em Português Brasileiro usando Máquina de Vetores de Suporte em Solução Multiclasse. In: X Encontro Nacional de Inteligência Artificial e Computacional, 2013, Fortaleza, Brazil (2013)
10. Ekman, P., Friesen, W. V. Facial Action Coding System. Palo Alto: Consulting Psychologists Press (1978)
11. Batista, G. E. A. P. A. Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de doutorado. Instituto de Ciências Matemáticas e de Computação, ICMC-USP, São Carlos, São Paulo (2003)
12. Fawcett, T., Provost, F. J. Adaptive Fraud Detection. Data Mining and Knowledge Discovery, v. 1, n. 3, pp. 291-316 (1997)
13. Stolfo, S. J., Fan, D. W., Lee, W., Prodromidis, A. L., Chan, P. K. Credit Card Fraud Detection using Meta_Learning: Issues and Initial Results. In AAAI-97 Workshop on AI Methods in Fraud and Risk Management (1997)
14. Silva, C., Silva, A., Netto, S., Paiva, A., Junior, G., Nunes, R. Lung nodules classification in ct images using simpsons index, geometrical measures and svm. Machine Learning and Data Mining in Pattern Recognition, vol. 5632 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 810-822 (2009)
15. Carvalho, A., Pozo, A. ,Vergilio, S., Lenz, A. Predicting fault proneness of classes trough a multiobjective particle swarm optimization algorithm. Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence, vol. 2, IEEE Computer Society, pp. 387-394 (2008)
16. Souza, M. R. P., Cavalcanti, G. D. C., Tsang, I. R. Off-line signature verification: An approach based on combining distances and one-class classifiers. Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2010, Arras, France, IEEE Computer Society, pp. 7-11 (2010)
17. Li, Y., Shawe-Taylor, J. The svm with uneven margins and chinese document categorization. Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation, pp. 216-227 (2003)
18. Manevitz, L., Yousef, M. One-class document classification via neural networks, Neurocomputing. vol.7, issues 7-9, pp. 1466-1481 (2007)
19. Japkowicz, N., Stephen, S. The class imbalance problem: A systematic study. Journal Intelligent Data Analysis, vol. 6. issues 5, pp. 429-449 (2002)
20. Batista, G. E. A. P. A., Prati, R. C., Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. vol. 6, issue 1, pp. 20-29 (2004)
21. Khoshgoftaar, T. M., Hulse, J. V., Napolitano, A. Supervised neural network modeling: An empirical investigation into learning from imbalanced data with labeling errors. IEEE Trans. on Neural Networks vol. 21, issues 5, pp. 813-830 (2010)
22. Kubat, M., Matwin, S. Addressing the curse of imbalanced data set: One sided sampling. In Proceedings of the Fourteenth International Conference on Machine Learning, Eds. Morgan Kaufmann, pp. 179-186, San Francisco, CA (1997)
23. Ling, C., Li, C. Data Mining for Marketing: Problems and Solutions. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, pp. 73-79 (1998)
24. Estabrooks, A., Jo, T., Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, vol. 20, issue 1, pp. 18-36 (2004)
25. Chawla, N. V., Bowyer, K. W., Hall L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, vol. 16, pp. 321-357 (2002)
26. Milaré, C. R., Batista, G. E. A. P. A., Carvalho, A. C. P. L. F. Descrição de uma Abordagem Híbrida para Aprender com Classes Desbalanceadas Utilizando Algoritmos Genéticos. Instituto de Ciências Matemáticas e de Computação- ICMC - Universidade de São Paulo - USP (2010)
27. Beckmann, M. Algoritmos Genéticos como Estratégia de Pré-Processamento em Conjuntos de Dados Desbalanceados. Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia Civil, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro (2010)
28. Akbani, R., Kwek, S., Japkowicz, N. Applying support vector machines to imbalanced datasets. In Proceedings of the 15th European Conference on Machine Learning, pp. 39-50, Pisa, Italy (2004)
29. Wang, B. X., Japkowicz, N. Boosting support vector machines for imbalanced data sets. ISMIS'08 Proceedings of the 17th international conference on Foundations of intelligent systems, pp. 38-47, Springer-Verlag Berlin, Heidelberg (2008)
30. Rangel, F., Rosso, P. On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style. In proc.: Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI. ESSEM 2013 at XIII Conference of the Italian Association for Artificial Intelligence, pp.4-6, Turin, Italy (2013)
31. Radovanovic, M., Ivanovic, M. Text Mining: Approaches and Applications. Novi Sad J. Math.V. 38, N. 3 (2008).
32. Mitchell, T. Machine Learning. McGraw-Hill, New York (1997).
33. Salton, G. Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management – Cornell University, Ithaca (1988)
34. Linden, R. Algoritmos Genéticos. 3 ed. Editora Ciência Moderna, Rio de Janeiro (2012)
35. Chang, C. C., Lin, C. J. LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, vol. 2 (2011)
36. Liu, X. Y., Wu, J., Zhou, Z. H. Exploratory Under Sampling for Class Imbalance Learning, In: International Conference IEEE on Data Mining (2006)
37. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, vol. 16 (2002)