Classificação de documentos do Exército Brasileiro utilizando o algoritmo de *Naive Bayes* e técnicas de Seleção de Sentenças

Sander P. Pivetta[1], **Sergio L. S. Mergen**[1]

[1]Campus Alegrete - Universidade Federal do Pampa (UNIPAMPA)
Caixa Postal 810 – 97.546-550 – Alegrete – RS – Brazil

sanderpivetta@gmail.com, sergiomergen@unipampa.edu.br

Abstract. *One of the needs of the Brazilian Army is the automated classification of documents called Boletins Internos (BIs), which must be grouped in order to produce summarized reports about the military. In this paper, we propose a solution based on the Naive Bayes classifier. To archive this goal, there is a need to select the text sentences that are related to each military, so that only those sentences are used during the training. In this sense, we propose two sentence selection heuristics that choose text blocks that appear close to the military name. The experiments show the benefits of using the Bayes classifier along with the proposed sentence selection techniques.*

Resumo. *Uma das necessidades do Exército Brasileiro é a classificação automatizada de documentos chamados Boletins Internos (BI), que devem ser agrupados a fim de gerar relatórios sumarizados a respeito de militares. Neste trabalho, propõe-se uma solução baseada no classificador Bayesiano. Além disso, é necessário identificar as sentenças que são relativas a cada militar, de modo que apenas elas sejam usadas durante o treinamento do classificador. Nesse sentido, o trabalho propõe duas heurísticas de seleção de sentenças que escolhem trechos de texto que apareçam próximas ao nome de cada militar. Os experimentos mostram os benefícios do uso do classificador bayesiano aliado às técnicas propostas de seleção de sentenças.*

1. Introdução

A popularização do uso dos computadores teve como consequência a existência de uma maior quantidade de documentos digitais. Documentos que antes eram publicados em papel, agora passam a ser representados como sequências de bits em formatos compreendidos por computadores. Com essa mudança, tarefas que costumavam ser feitas manualmente podem ser auxiliadas por meio de abordagens computacionais automatizadas. Uma dessas tarefas envolve a classificação da informação. A classificação visa separar documentos de acordo com algum critério, o que facilita a tomada de decisões sobre os dados agrupados.

Várias organizações possuem a necessidade de classificar documentos. O Exército Brasileiro é um exemplo destas organizações, onde documentos chamados de Boletins Internos (BI) devem ser agrupados a fim de gerar relatórios sumarizados a respeito de militares. Os BIs são documentos confeccionados periodicamente que contém informações relacionadas às atividades realizadas pela instituição e pelos seus integrantes [Exército 2002]. A partir dos BIs são gerados documentos chamados de Folha de

Alterações, existindo um exemplar para cada militar, relatando o histórico referente as atividades por ele desempenhada e sobre a sua vida pessoal [Exército 2001].

Conforme as normas vigentes, encontradas em [Exército 2001], nem todos BIs possuem informações consideradas relevantes para a elaboração das Folhas de Alterações. Dessa forma, dado um militar, é preciso realizar pesquisas sobre todos os BIs produzidos durante o período de um semestre, buscando informações relativas ao militar, para analisar se estas devem ser usadas na produção das respectivas Folhas de Alterações.

Visando agilizar esta atividade, necessita-se encontrar uma forma de realizar a separação automática dos BI possuidores de informações relevantes para cada militar. Neste artigo, propõe-se que esta separação seja realizada com o emprego do aprendizado de máquina. Uma vez que dispõe-se de informações já classificadas em semestres anteriores, torna-se oportuno o emprego do Aprendizado Supervisionado [Mitchell 1997]. Mais especificamente, a tarefa será realizada através do classificador *Naive Bayes* [Mitchell 1997].

Além disso, outro problema pesquisado envolve escolher, para cada documento, quais trechos serão utilizados para realizar a tarefa de classificação. Como os BI são compostos por um conjunto de pequenas informações, referentes a assuntos e pessoas distintas, torna-se necessário identificar quais sentenças são relativas a cada militar. Conforme será demonstrado, a escolha equivocada das informações pode distorcer o treinamento do classificador, levando-o a fazer uma separação menos precisa.

O artigo está organizado da seguinte forma: na seção 2 são mencionados trabalhos que realizam a classificação textual usando o método de *Bayes* e algumas técnicas de seleção de sentenças. Na seção 3 é apresentado o método de classificação proposto. As técnicas utilizadas para realizar a seleção de sentenças são vistas na seção 4. Os experimentos realizados são descritos na seção 5. Já na seção 6 são tecidas as considerações finais.

2. Trabalhos Relacionados

Diversos algoritmos de aprendizado supervisionado podem ser utilizados na classificação de documentos textuais. Dentre eles, um que possui bom desempenho é o classificador *Naive Bayes*. Ele é uma algoritmo baseado no *Teorema de Bayes* que propõem uma maneira de calcular a probabilidade da ocorrência de um evento baseando-se nas probabilidades obtidas com a análise dos eventos anteriores. Um motivo de seu sucesso está na forma em que ele trata cada informação, pois estas são consideradas independentes entre si, o que diminui o espaço de busca usado para encontrar uma solução [Mitchell 1997].

O desempenho dessa técnica na classificação de texto é analisada em [Koga 2011], que compara o classificador com outros métodos de classificação existentes. O objetivo do experimento descrito envolveu a classificação automática do sujeito das frases. Para o treinamento foi utilizado um conjunto de atributos morfológicos e estruturais extraídos de frases pré-processadas. Os algoritmos foram implementados dentro do *software* "WEKA"[1]. Os resultados obtidos demostram um melhor desempenho do classificador *Naive Bayes*, o que foi justificado pelo fato de que a maioria das informações analisadas não possuírem dependência entre si.

[1] www.cs.waikato.ac.nz/ml/weka

Um exemplo clássico de classificação de textos em que se usa o classificador *Bayesiano* é a análise de mensagens do tipo *spam* [Silva and Vieira 2007]. Normalmente existem mensagens eletrônicas previamente categorizadas como *spans*, o que permite com que abordagens supervisionadas de classificação sejam utilizadas. No trabalho descrito em [Rabelo et al. 2011], foi verificado que em 85% dos *emails* analisados a resposta retornada foi correta, porém, na medida que o conteúdo destas mensagens ficava mais denso, a precisão da aplicação diminuía.

A seleção das sentenças a serem treinadas pelo classificador também merece destaque, uma vez que apenas algumas informações contidas em um documento podem estar relacionados ao assunto alvo da classificação. Nessa linha, o trabalho de [Goldstein et al. 1999] demonstra que a remoção de *stop words* pode melhorar a classificação, assim como a seleção de sentenças com poucas palavras. Em [McDonald and Chen 2002] desenvolve a ferramenta TXTRACTOR para realizar a sumarização de textos, realizando a separação de informações através da segmentação das informações. Para realizar uma melhor sintetização, ele separa as informações contidas nos documentos em grupos conforme as semelhanças apresentadas, facilitando o seu processamento.

É importante destacar que a seleção de sentenças tem diversas aplicações além de servir a uma etapa de pré-processamento em uma tarefa de classificação. Por exemplo, [Wang et al. 2012] procura encontrar sentenças distintas que caraterizam um tópico específico, através da análise de um *corpus* de texto. Também vale a pena mencionar que até mesmo a seleção das sentenças pode utilizar os algoritmos de aprendizado de máquina, como apresentado no trabalho de [Metzler and Kanungo 2008], que objetiva selecionar sentenças para realizar a sumarização dos documentos extraídos da *Web*.

3. Método de Classificação Proposto

O objetivo principal deste trabalho envolve selecionar os documentos portadores de informações importantes para compor as Folhas de Alterações de cada militar. Dentre os algoritmos de aprendizado supervisionado existentes, escolheu-se o *Naive Bayes*, devido ao bom desempenho quando aplicado na classificação de documentos textuais, conforme destacado na seção anterior.

O classificador *bayesiano* divide o processamento em duas fases: o **Treinamento** e a **Classificação**. As etapas que devem ser executadas nessas duas fases, assim como as informações necessárias, são ilustradas na Figura 1. A descrição de cada um dos componentes da Figura encontra-se a seguir.

Base de Treinamento: Contém trechos extraídos dos BIs, sendo que os trechos já foram classificados como “relevantes”e “não relevantes”. São utilizados pelo classificador para descobrir a probabilidade das evidências estarem associadas às classes de interesse. Não são realizadas distinções com relação aos indivíduos e aos trechos que se referem.

Documentos a Classificar: Contém BIs que devem ser classificados como “relevantes”ou “não relevantes”. Os BIs já são previamente separados em bases menores, relacionados ao semestre em que foram confeccionados.

Conversão do PDF: Como os BIs encontram-se no formato PDF, é necessário realizar uma conversão para um formato textual que possa ser compreendido pelas etapas

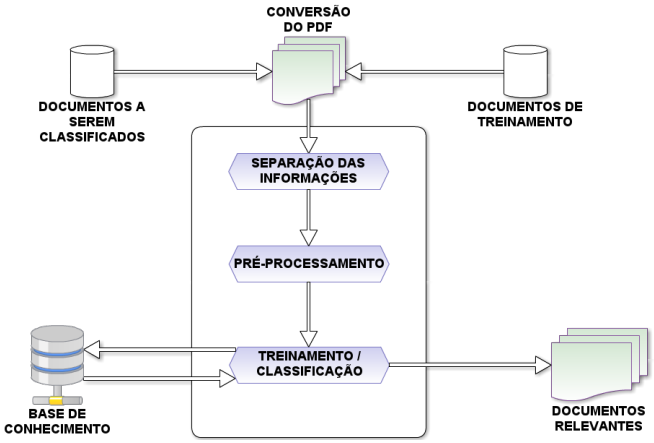


Figura 1. Treinamento e Classificação do algoritmo *Naive Bayes*

posteriores. Para isto foi utilizada uma biblioteca *open source* para java chamada *PDFBox*[2], que possibilita a manipulação e criação de arquivos PDF.

Separação das Informações: Após a leitura dos arquivos, as informações passam por uma pré-separação. Primeiramente são selecionados os arquivos que possuem alguma referência à pessoa analisada, onde são selecionados aqueles que possuem o 'nome completo' do militar (ex. Sander Pes Pivetta) ou a 'graduação mais o nome de guerra' (ex. 3º Sgt Sander). Estas informações são selecionadas devido os militares serem referenciados desta forma. Além disso, é necessário dividir o documento em trechos menores, cuja informação esteja relacionada ao militar em questão. A próxima seção descreve como essa divisão é feita.

Pré-processamento: Devido a língua portuguesa possuir uma grande variação morfológica, com o acréscimo de prefixos, sufixos, variação de tempos verbais, singular e plural, existe uma grande diversidade de palavras com sentidos semelhantes, o que pode interferir no processo de classificação. Na busca por diminuir a gama de palavras que irá compor a base de conhecimento, são utilizadas a técnica de *stemming* [Rezende 2005] para realizar uma normalização linguística das palavras, com a remoção das variações morfológicas descritas acima, e a técnica de remoção de *stop words* [Rigo et al. 2007], para a exclusão das palavras consideradas inúteis (preposições, artigos, numerais, pronomes e algumas palavras de contexto especificamente militar).

Treinamento: Para todos os trechos escolhidos, esta etapa calcula a probabilidade de ocorrência dos eventos dentro das classes analisadas (relevante e não relevante). No caso em questão, cada palavra do vocabulário presente nos trechos escolhidos corresponde a um evento.

Base de Conhecimento: É o resultado obtido com o treinamento, onde são armazenadas as probabilidade de cada palavra do vocabulário estar associada a classe dos "relevantes"e dos "não relevantes".

Classificação: Nesta etapa, é calculada a probabilidade de um trecho pertencer a cada uma das classes. Para encontrar este valor, é usado o *Teorema de Bayes*, em que a

[2]http://pdfbox.apache.org

probabilidade é calculada com base em evidências coletadas anteriormente (armazenadas na base de conhecimento). Se em pelo menos um dos trechos escolhidos do documento, a probabilidade *bayesiana* de ele ser relevante for maior, o documento é considerado relevante para o militar em questão.

Documentos Relevantes: São os BI retornados como relevantes pela fase de classificação, para cada militar.

4. Seleção de Sentenças

De todas as etapas realizadas pelo algoritmo *Naive Bayes*, a forma como os trechos são selecionados, tanto na fase de classificação como na de treinamento, ganha destaque devido a sua influência na obtenção dos resultados. Uma inapropriada seleção das informações acarreta na utilização de dados errados para o cálculo da probabilidade e uma categorização errada dos Boletins Internos.

Neste trabalho são propostas duas técnicas para seleção dos trechos, chamadas de "Janela Fixa"e "Janela Deslizante". Em ambas, o ponto de partida são os pontos no texto onde o nome do militar (pivô) aparece. Dado um pivô, cada técnica utiliza regras diferentes para selecionar o texto que está relacionado ao militar, como será descrito a seguir.

Janela Fixa: Esta técnica considera como informações importantes aquelas que encontram-se mais próximas ao nome pesquisado. Para isso, a partir do pivô, é realizada a seleção dos κ caracteres anteriores e posteriores ao nome. Caso a seleção selecione apenas parte de uma palavra, toda a palavra é considerada como parte integrante da sentença.

Janela Deslizante: Esta técnica leva em consideração que a informação importante possa estar afastada do nome pesquisado, principalmente quando ele estiver contido em uma lista de nomes.

Em primeiro lugar, deve-se verificar se o pivô encontra-se dentro de uma linha válida. Uma linha é assim considerada se o número de palavras válidas da linha for igual ou superior a λ. Para ser válida, a palavra deve possuir mais do que μ caracteres.

A obtenção da linha analisa o texto anterior ao pivô até encontrar um símbolo de término de parágrafo e o posterior até encontrar outro símbolo de término de parágrafo. Caso esta seleção possua $\lambda > 6$ ela é considerada válida e selecionada, caso contrário, ela é descartada e realizam-se verificações sobre os parágrafos anteriores a ele, até que a condição de validação do texto seja satisfeita, sendo o texto selecionado. Nos casos onde a sentença analisada não ser válida, são analisados os parágrafos anteriores pelo motivo das sentenças referentes ao militar estarem no mesmo paragrafo ou em parágrafos anteriores.

Para exemplificar as técnicas de seleção, considere as Figuras 2 e 3, que selecionam o texto tendo "Sander Pes Pivetta"como pivô. Na Figura 2 foi utilizada a "Janela Fixa"com κ = 150. Ou seja, foram selecionados os 300 caracteres mais próximos ao nome. Já na Figura 3 foi utilizada a "Janela Deslizante", com $\lambda = 6$ e μ = 3. A escolha destes valores fundamentou-se em testes realizados, onde os resultados obtidos com o uso destes valores foram os mais eficientes.

Já as Figuras 4 e 5 mostram um outro exemplo em que o nome do militar está em uma tabela, juntamente com o nome de outras pessoas. Nesse caso, a informação

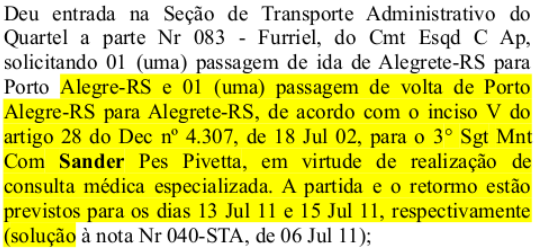

Deu entrada na Seção de Transporte Administrativo do Quartel a parte Nr 083 - Furriel, do Cmt Esqd C Ap, solicitando 01 (uma) passagem de ida de Alegrete-RS para Porto Alegre-RS e 01 (uma) passagem de volta de Porto Alegre-RS para Alegrete-RS, de acordo com o inciso V do artigo 28 do Dec nº 4.307, de 18 Jul 02, para o 3° Sgt Mnt Com **Sander** Pes Pivetta, em virtude de realização de consulta médica especializada. A partida e o retorno estão previstos para os dias 13 Jul 11 e 15 Jul 11, respectivamente (solução à nota Nr 040-STA, de 06 Jul 11);

Figura 2. Janela Fixa

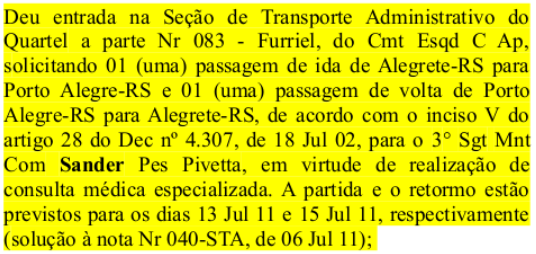
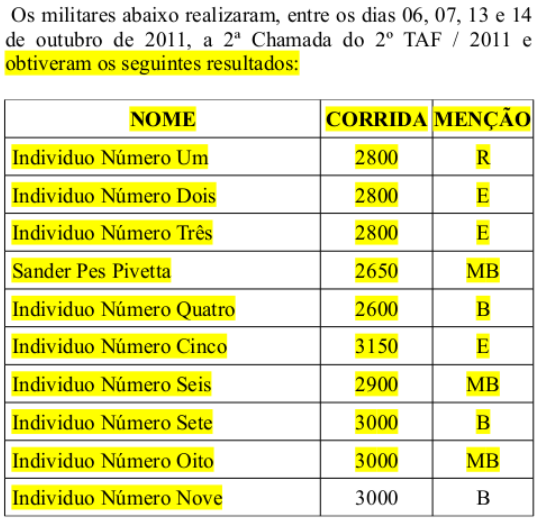

Deu entrada na Seção de Transporte Administrativo do Quartel a parte Nr 083 - Furriel, do Cmt Esqd C Ap, solicitando 01 (uma) passagem de ida de Alegrete-RS para Porto Alegre-RS e 01 (uma) passagem de volta de Porto Alegre-RS para Alegrete-RS, de acordo com o inciso V do artigo 28 do Dec nº 4.307, de 18 Jul 02, para o 3° Sgt Mnt Com **Sander** Pes Pivetta, em virtude de realização de consulta médica especializada. A partida e o retorno estão previstos para os dias 13 Jul 11 e 15 Jul 11, respectivamente (solução à nota Nr 040-STA, de 06 Jul 11);

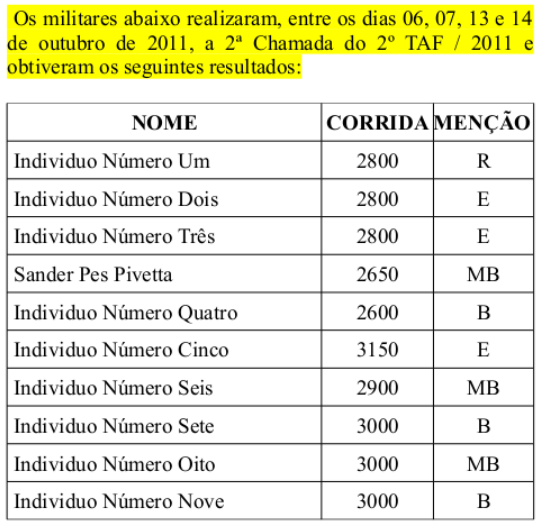
Figura 3. Janela Deslizante

referentes a esses militares encontra-se no parágrafo que aparece antes da tabela. Na Figura 4 foi utilizada a "Janela Fixa"com $\kappa = 150$. Como pode-se ver, a técnica seleciona praticamente todas as informações contidas na tabela e ignora a sentença anterior a ela, a qual possui a informação pertinente. Já na Figura 5 foi utilizada a "Janela Deslizante", com $\lambda = 6$ e $\mu = 3$. Observa-se que, nesse caso, como a linha onde o nome do militar ocorre não é considerada válida, a janela de texto deslizou para cima até o encontro de uma linha válida.


Os militares abaixo realizaram, entre os dias 06, 07, 13 e 14 de outubro de 2011, a 2ª Chamada do 2º TAF / 2011 e obtiveram os seguintes resultados:

NOME	CORRIDA	MENÇÃO
Individuo Número Um	2800	R
Individuo Número Dois	2800	E
Individuo Número Três	2800	E
Sander Pes Pivetta	2650	MB
Individuo Número Quatro	2600	B
Individuo Número Cinco	3150	E
Individuo Número Seis	2900	MB
Individuo Número Sete	3000	B
Individuo Número Oito	3000	MB
Individuo Número Nove	3000	B

Figura 4. Janela Fixa


Os militares abaixo realizaram, entre os dias 06, 07, 13 e 14 de outubro de 2011, a 2ª Chamada do 2º TAF / 2011 e obtiveram os seguintes resultados:

NOME	CORRIDA	MENÇÃO
Individuo Número Um	2800	R
Individuo Número Dois	2800	E
Individuo Número Três	2800	E
Sander Pes Pivetta	2650	MB
Individuo Número Quatro	2600	B
Individuo Número Cinco	3150	E
Individuo Número Seis	2900	MB
Individuo Número Sete	3000	B
Individuo Número Oito	3000	MB
Individuo Número Nove	3000	B

Figura 5. Janela Deslizante

5. Resultados Obtidos

Nesta seção é avaliada a qualidade dos métodos propostos para a classificação de BIs disponibilizados pelo Exército Brasileiro, usando as técnicas de seleção de sentenças "Janela Fixa"e "Janela Deslizante". Para efeitos de comparação, também é verificada a qualidade de outros dois métodos básicos, chamados de "Pesquisa Nominal"(que tem a finalidade de comparar os resultados obtidos com o emprego da seleção de sentenças) e "Documento Inteiro"(que tem a finalidade de apresentar a necessidade de realizar uma correta seleção das informações). A descrição de cada método empregado é apresentada a seguir:

Pesquisa Nominal: Esse método realiza a categorização dos BIs sem a aplicação

do aprendizado de máquina. Dado um militar, são considerados relevantes todos os documentos onde o seu nome ocorre.

Documento Inteiro: Esse método realiza o treinamento dos BIs sem utilizar a seleção de sentenças. Ou seja, caso um documento contenha informações relevantes a respeito de um militar, todas as palavras do documento também são consideradas como eventos e são associados a classe de documentos relevantes.

Janela Fixa: Esse método realiza o treinamento dos boletins internos utilizando a técnica de seleção de sentenças "Janela Fixa", com κ = 150. Ou seja, caso um documento contenha informações relevantes a respeito de um militar, apenas as palavras que pertencem a um trecho, selecionado por essa técnica, são consideradas eventos associados a classe de documentos relevantes.

Janela deslizante: Esse método realiza o treinamento dos BIs utilizando a técnica de seleção de sentenças "Janela Deslizante", com $\lambda = 6$ e μ = 3. Ou seja, caso um documento contenha informações relevantes a respeito de um militar, apenas as palavras que pertencem ao trecho selecionado por essa técnica são consideradas eventos, sendo associados a classe de documentos relevantes.

Na etapa de treinamento foram usados 214 BIs confeccionados durante um período de dois semestres, para 64 militares selecionados aleatoriamente. A marcação dos documentos em "relevantes"e "não relevantes"foi realizada manualmente, com o auxílio das Folhas de Alterações destes militares. Cada Folha de Alteração possui a identificação dos BIs usados para a sua confecção, o que facilitou o processo de marcação.

A base de treinamento gerada depende do método de classificação usado. No método "Documento Inteiro", a lista de relevantes e não relevantes é composta por BIs inteiros. Por exemplo, se um BI tiver sido usado para confeccionar uma folha de alterações, todo o BI é incorporado à base de documentos relevantes. Já nos métodos "Janela Fixa"e "Janela Deslizante", a lista de relevantes e não relevantes é composta por trechos dos BIs. Por exemplo, se um BI tiver sido usado para confeccionar uma folha de alterações, um método de seleção de sentenças é usados para extrair os trechos do documento onde o nome do militar ocorre. Os trechos são incorporados a base de documentos relevantes.

O indicador de desempenho utilizado nos experimentos é o *F-Measure*, que fornece uma medida balanceada dos escores de precisão e cobertura. Valores próximos a zero indicam que tanto a precisão quanto a cobertura foram pobres, enquanto valores próximos a 100 indicam que tanto a precisão quanto a cobertura obtiveram resultados satisfatórios. A precisão é calculada em função do número de BIs classificados como "relevantes"que realmente são. Já a cobertura é calculada em função do número de BIs relevantes que assim foram classificados.

A etapa de testes utilizou BIs e folhas de alterações de um semestre específico que não foi utilizado durante o treinamento. Os métodos de classificação foram encarregados de classificar 113 BIs para um conjunto de 27 militares selecionados aleatoriamente.

Ressalta-se que dentre os BIs empregados no treinamento e na classificação, aproximadamente 12% possuem pelo menos uma informação relevante. Os outros não possuíam nenhuma sentença relevante.

A Tabela 1 apresenta os valores de precisão, cobertura e *F-Measure* encontrados

para cada um dos métodos de classificação empregados. A Figura 6 compara os valores de *F-Measure* alcançados pela técnica de seleção.

Método	**Precisão**	**Cobertura**	**Medida** F
Pesquisa Nominal	33%	100%	49,6%
Documento Inteiro	11%	21%	13,7%
Janela Fixa	57%	65%	60,7%
Janela Deslizante	76%	64%	69,5%

Tabela 1. Comparativo entre os resultados

Observe que a "Pesquisa Nominal"atingiu uma *F-Measure* próxima a 50%. Apesar de encontrar todos os documentos relevantes, a precisão é baixa, ou seja, muitos dos documentos retornados não possuem relação para a confecção das folhas de alterações de militares específicos. Já os métodos de classificação *bayeasiana* baseados em técnicas de seleção de sentença obtiveram um desempenho superior. Dentre os dois, o método da "Janela Deslizante"se saiu melhor, atingindo uma *F-Measure* igual a 69,5%. Esse resultado reflete o fato de que, em muitos casos, o nome do militar aparece dentro de uma lista, sendo que nessas situações o método de "Janela Deslizante"consegue selecionar melhor o trecho significativo para o militar.

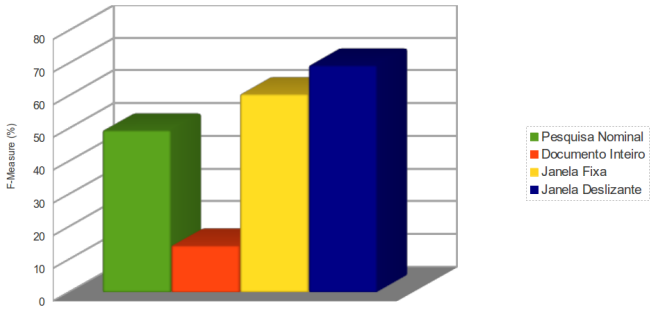


Figura 6. Comparativo entre os resultados

Dos quatro métodos, o "Documento Inteiro"obteve o pior desempenho, perdendo inclusive para o método "Pesquisa Nominal"que não utiliza algoritmos de aprendizado de máquina. Isso ocorre porque o treinamento leva em consideração muitos eventos (ocorrência de palavras) que não possuem relação nenhuma com os militares usados durante o treinamento.

Para realizar uma boa classificação, é necessário que o treinamento utilize uma base de conhecimento de tamanho adequado. Bases com pouca informação podem não dispor de evidências suficientes para realizar a classificação da forma correta, levando a uma situação conhecida como *underfitting*. Já bases com muita informação podem se especializar nos dados de treinamento e falhar quando novos dados precisarem ser classificados, levando a uma situação conhecida como *overfitting*.

Nesse sentido, o próximo experimento tem o intuito de verificar como o tamanho da base de treinamento afeta a classificação. A Figura 5 apresenta os resultados obtidos. Os gráficos medem a *F-Measure* obtida quando se usa bases de treinamento de tamanho variável.

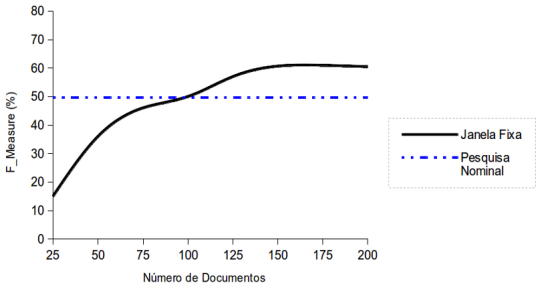


Figura 7. Variação dos resultados utilizando seleção Janela Fixa

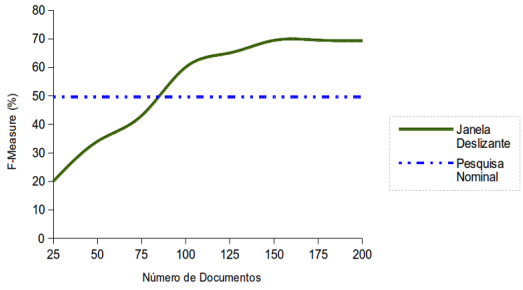


Figura 8. Variação dos resultados utilizando seleção Janela Deslizante

O gráfico mostra que o desempenho de ambos os classificadores é pior ao que usa "Pesquisa nominal"quando a base de treinamento dispõe de menos do que 75 BIs, o que caracteriza o *underfitting*. A partir de 150 BIs, o desempenho cai, mas se mantém superior a "Pesquisa Nominal". Isso sugere que ambas técnicas de seleção de sentenças são capazes de reduzir o ruído na fase de treinamento, o que reflete em uma taxa de precisão e cobertura razoavelmente altas mesmo quando se usa uma quantidade elevada de documentos para treinar o classificador.

A leve queda percebida no desempenho indica que ainda existe um certo ruído na base de treinamento, mas que não chega a gerar o problema de *overfitting*. A explicação para isso pode derivar do fato de que os BIs são produzidos por pessoas diferentes, que escrevem de modo particular, usando um vocabulário próprio. Assim, é possível que os BIs de teste tenham sido produzidos por pessoas diferentes das que produziram os BIs de treinamento, e algumas evidências relevantes não tenham sido devidamente identificadas pelo classificador.

6. Conclusão

Este artigo apresentou uma aplicação de classificação de documentos que emprega o algoritmo de aprendizado de máquina supervisionado *Naive Bayes*. O objetivo da aplicação é selecionar os Boletins Internos que devem compor as Folhas de Alterações de militares do Exército Brasileiro. Para auxiliar no treinamento, também foram propostas duas técnicas de seleção de sentença, chamadas de "Janela Fixa"e "Janela Deslizante". Essas técnicas tem a função de delimitar as palavras dos BIs que são usadas tanto na etapa de treinamento quanto na etapa de classificação.

Para validar a proposta, os métodos de classificação apresentados foram empregados na classificação de um conjunto de Boletins Internos. Analisando os resultados, verifica-se que o algoritmo de *Naive Bayes*, combinado com uma técnica de seleção de sentenças, consegue realizar uma classificação satisfatória dos documentos, comprovando que a atividade mais influente no resultado é a forma como as informações são selecionadas.

Como trabalhos futuros, pretende-se analisar o desempenho das técnicas de "Janela Fixa"e "Janela Deslizante"quando são utilizados valores diferentes para κ, λ e μ. Além disso, pretende-se analisar a possibilidade de descobrir os melhores parâmetros

automaticamente através de algoritmos de aprendizado de máquina baseado em redes neurais.

Outra possibilidade de trabalho futuro envolve estender o classificador *bayesiano* utilizado para que as evidências coletadas englobem a frequência de conjuntos de palavras subjacentes em vez da frequência de palavras individuais. Essa estratégia pode mostrar-se útil caso existam sequencias de palavras que costumem aparecer mais frequentemente em documentos de uma certa classe (relevante ou não relevante). Além de implementar essa versão estendida, pretende-se descobrir se existe um tamanho adequado para a sequências de palavras que maximize a *F-Measure*.

Referências

Exército (2001). *Boletim do Exército 02*. Secretaria Geral do Exército, Brasilia.

Exército (2002). *Separata ao Boletim do Exército Número 08: Instruções Gerais para a Correspondência, as Publicações e os Atos Administrativos no Âmbito do Exército (IG 10-42)*. Gabinete do Comandante do Exército, Brasilia.

Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York. ACM.

Koga, M. L. (2011). Classificadores Bayesianos: Aplicados a análise sintática da língua portuguesa. In *Escola Politécnica da Universidade de São Paulo*, São Paulo.

McDonald, D. and Chen, H. (2002). Using sentence-selection heuristics to rank text segments in txtractor. In *Joint Conference on Digital Libraries - JCDL*, New York.

Metzler, D. and Kanungo, T. (2008). Machine learned sentence selection strategies for query-biased summarization. In *SIGIR Learning to Rank Workshop*.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Rabelo, J. P., Filho, M. A., and Oliveira, T. (2011). Mineração de Textos Através do Algoritmo de Classificação. In *Instituto de Matemática. Universidade Federal da Bahia (UFBA)*, Salvador.

Rezende, S. O. (2005). *Sistemas Inteligentes. Fundamentos e Aplicação*. Editora Manole Ltda, Barueri.

Rigo, S. J., Oliveira, J. P. M., and Barbieri, C. (2007). Classificação de Textos Baseada em Ontologias de Domínio. In *Anais do XXXVII Congresso da Sociedade Brasileira de Computação - V Workshop em Tecnologia da Informação e da Linguagem Humana*, Rio de Janeiro.

Silva, C. F. and Vieira, R. (2007). Categorização de Textos da Língua Portuguesa com Árvores de Decisão, SVM e Informações Linguísticas. In *Anais do XXVII Congresso da Sociedade Brasileira de Computação. V Workshop de Tecnologia da Informação e da Linguagem Humana*, Rio de Janeiro.

Wang, D., Zhu, S., Li, T., and Gong, Y. (2012). Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data*.