Uso de Aprendizado de Máquina para a Classificação de Documentos do Exército Brasileiro

Sander P. Pivetta, Sergio L. S. Mergen, Fabio N. Kepler

Universidade Federal do Pampa (UNIPAMPA)
Campus Alegrete – RS – Brasil

{sanderpivetta,sergiomergen,kepler}@unipampa.edu.br

Abstract. *The Brazilian Army produces a summarized report about each military member's activities during a semester. This requires that the references to a specific member are searched through a set of documents produced during the period of six months. This work proposes ways of performing this classification task automatically by using the Naive Bayes learning method. It is also necessary to identify which sentences inside a document refer to the military member in question, such that only these sentences are used for training the classifier. We propose two techniques for sentence selection that choose portions of texts that surrounds the target names. The experiments show that it is possible to achieve an f-measure of* 76.7% *in retrieving relevant documents, and that sentence selection and training data size play important roles in the task.*

Resumo. *A cada semestre o Exército Brasileiro gera relatórios sumarizados a respeito de cada militar e suas atividades. Para isso é necessário encontrar referências relevantes a cada militar dentro de um conjunto de documentos produzidos periodicamente no intervalo de seis meses. Este trabalho propõe formas de realizar essa classificação de maneira automática, utilizando o método* Naive Bayes *de aprendizado probabilístico. Para isso, também é necessário identificar quais sentenças em um documento são relativas a cada militar, de modo que apenas elas sejam usadas durante o treinamento do classificador. Assim, este trabalho propõe duas heurísticas de seleção de sentenças que escolhem trechos de texto que aparecem próximos ao nome de cada militar. Os experimentos mostram que é possível atingir* $76,7\%$ *de medida-f na recuperação de documentos relevantes, e que a seleção de sentenças e o tamanho da base de treinamento desempenham papéis importantes na tarefa.*

1. Introdução

O desenvolvimento tecnológico e a popularização dos computadores provocou um aumento no número de documentos digitais existentes. Assim, documentos passaram a ser confeccionados em formatos digitais, e não mais diretamente em papéis, possibilitando automatizar atividades através de métodos computacionais. Uma tarefa que pode ser realizada é a classificação de informações, ou seja, separar documentos em subconjuntos baseando-se em alguns critérios. Tal atividade facilita a tomada de decisões a respeito de determinados assuntos.

Várias organizações necessitam realizar a classificação de seus documentos. Um exemplo é o Exército Brasileiro, que produz os Boletins Internos (BI) e as Folhas

de Alterações (FA). Os BIs são documentos confeccionados periodicamente contendo informações relacionadas às atividades realizadas pela instituição e por seus integrantes. A partir deles são extraídas informações referentes a militares a fim de gerar um relatório das atividades por eles desempenhadas [Exército 2002].

Esses relatórios, chamados de Folha de Alterações (FA), são confeccionados semestralmente para cada militar, e relatam o histórico de atividades desempenhada pelo militar durante o período analisado [Exército 2001]. Conforme as normas vigentes que normatizam a confecção das FAs encontradas em [Exército 2001], nem todas as informações presentes no BIs são relevantes para a sua elaboração. Desta forma, a partir de um militar, é necessário realizar uma pesquisa sobre todos os BIs produzidos durante um semestre. Para cada informação encontrada a respeito do militar, deve-se analisar se ela é relevante na produção da sua FA.

Como os BIs são disponibilizados em formato digital, é possível agilizar esta tarefa, encontrando os documentos que possuem apenas as informações pertinentes a um militar. Este trabalho propõe o emprego de técnicas de aprendizado de máquina como forma de automatizar esta atividade. Como dispõe-se de informações previamente classificadas em semestres anteriores, é possivel o emprego de abordagens de aprendizado supervisionado [Mitchell 1997]. Dentre estas abordagens, será utilizado o algoritmo de *Naive Bayes* (NB) [Mitchell 1997].

Problemas clássicos de classificação de documentos consideram todo conteúdo de um documento para verificar sua relevância. Entretanto, no caso de uso aqui apresentado, a tarefa é dificultada pelo fato dos BIs não serem documentos exclusivos para cada militar, isto é, eles são compostos por conjuntos de pequenas informações referentes a assuntos e pessoas distintas. Assim, é necessário identificar quais desses conjuntos são relevantes a cada militar. Caso contrário, informações errôneas podem ser consideradas relevantes durante o treinamento. Este é o problema da Seleção de Sentenças, e envolve escolher quais trechos de cada BI serão utilizados para realizar a classificação de forma mais precisa.

O restante deste artigo está organizado da seguinte forma: Na Seção 2 é apresentado o funcionamento do classificador *Naive Bayes*, juntamente com trabalhos que realizam classificação textual usando este algoritmo e trabalhos que abordam seleção de sentenças. Na Seção 3 é apresentado o método de classificação escolhido. Na Seção 3.1 as técnicas propostas para realizar a seleção de sentenças são demonstradas. Na Seção 4 são descritos os experimentos realizados, e na seção 5 são tecidas as considerações finais.

2. Trabalhos Relacionados

O aprendizado de máquina é uma área que pode ser empregada em diversas atividades. Quando têm-se a disposição uma conjunto de informações a serem utilizadas na obtenção do conhecimento, torna-se útil a uso do aprendizado supervisionado.

Dentre os algoritmos supervisionados, alguns ganham destaque quando utilizados na classificação de documentos textuais. Um que apresenta uma boa empregabilidade e um bom desempenho é o classificador *bayesiano*, um método de aprendizado probabilístico que está baseado no *Teorema de Bayes*.

O objetivo do classificador Bayeasiano é verificar se uma amostra analisada pertence ou não a uma determinada classe. A obtenção desta resposta realiza-se através

de uma análise estatística das informações coletadas sobre as instâncias fornecidas. [Mitchell 1997] apresenta o seu funcionamento através da Equação 1

$$P\left(h|D\right) = \frac{P\left(D|h\right)P\left(h\right)}{P\left(D\right)} \tag{1}$$

O objetivo do classificador é encontrar a $P(h|D)$, que representa a probabilidade a *Posteriori* da ocorrência da classe h em razão do evento D. Para calculá-la, é necessário encontrar a probabilidade condicional $P(D|h)$, que representa a probabilidade de ocorrência do evento D dado a classe h. Também é necessário encontrar a probabilidade a *Priori* ($P(h)$) de ocorrência da classe a partir dos dados de treinamento. Também necessita-se da $P(D)$, o que representa a probabilidade do evento dentro do conjunto de treinamento. Em alguns casos onde são utilizadas mais de uma classe, a $P(D)$ pode ser desconsiderada do cálculo da probabilidade *bayesiana*, pois ela não altera a proporção entre os resultados. Neste caso é realizada uma comparação entre eles, e o evento é atribuído ao que obter o maior resultado [Mitchell 1997]. É importante ressaltar que o classificador assume que os eventos são independentes entre si, o que diminui a complexidade do cálculo probabilístico.

Uma análise do classificador *bayesiano* é realizada em [Koga 2011], onde é feita uma comparação com outros métodos de classificação existentes. O objetivo do trabalho envolveu a classificação automática do sujeito de frases. Para o treinamento foi utilizado um conjunto de atributos morfológicos e estruturais extraídos de frases pré-processadas. Os resultados obtidos demostraram um melhor desempenho do classificador *bayesiano*, o que foi justificado pelo fato de a maioria das informações analisadas realmente não possuírem dependência entre si.

Um problema clássico onde classificadores *bayesianos* são geralmente utilizados é a análise de mensagens do tipo *spam* [Silva and Vieira 2007]. Normalmente existem mensagens eletrônicas previamente classificadas como *spans*, o que permite que abordagens supervisionadas de classificação sejam utilizadas. O trabalho de [Rabelo et al. 2011] verificou que a resposta correta era retornada em 85% dos *emails* analisados, e que a precisão diminuía quando o conteúdo das mensagens ficava mais denso.

No problema de classificação textual, os eventos estudados costumam ser palavras contidas em documentos. Como o classificador *bayesiano* considera as variáveis independentes entre si, o contexto no qual uma palavra está inserida é descartado. Entretanto, muitas palavras podem assumir significados diferentes dependendo do contexto. [Peng et al. 2004] compara o algoritmo original, onde as variáveis são palavras independentes, e uma modificação, onde as variáveis são obtidas com a utilização de *n-gramas*, ou seja, n palavras adjacentes. Como resultado, verificou-se que o desempenho utilizando trigramas foi superior.

Nem todas informações em um documento podem estar relacionados ao assunto alvo da classificação. Assim, é necessário realizar a seleção das sentenças a serem usadas para treinar um classificador. Nessa linha, o trabalho de [Goldstein et al. 1999] mostra que a seleção de sentenças com poucas palavras melhora a classificação, assim como a remoção de *stop words*.

A seleção de sentenças tem outras aplicações além da tarefa de classificação.

Por exemplo, [Wang et al. 2012] procura por informações que consigam discriminar as principais informações contidas nos documentos, separando-os em grupos conforme suas semelhanças. Em seguida, é produzido um resumo com as principais diferenças encontradas nestes grupos de documentos. Também [Metzler and Kanungo 2008] busca selecionar sentenças para realizar a sumarização de documentos extraídos da *Web*.

3. Método de Classificação

Um classificador *bayesiano* divide o processamento em duas fases: treinamento e classificação. No treinamento os parâmetros do modelo são estimados, e na classificação eles são usados para retornar os documentos relevantes dado um determinado militar.

Para o classificador proposto neste trabalho, os passos necessários para a realização das duas fases são ilustrados na Figura 1. Ela esta dividida em duas partes. As etapas do lado esquerdo representam as atividades desempenhadas durante a fase de treinamento. Já as etapas do lado direito representam as atividades desempenhadas durante a fase de classificação.

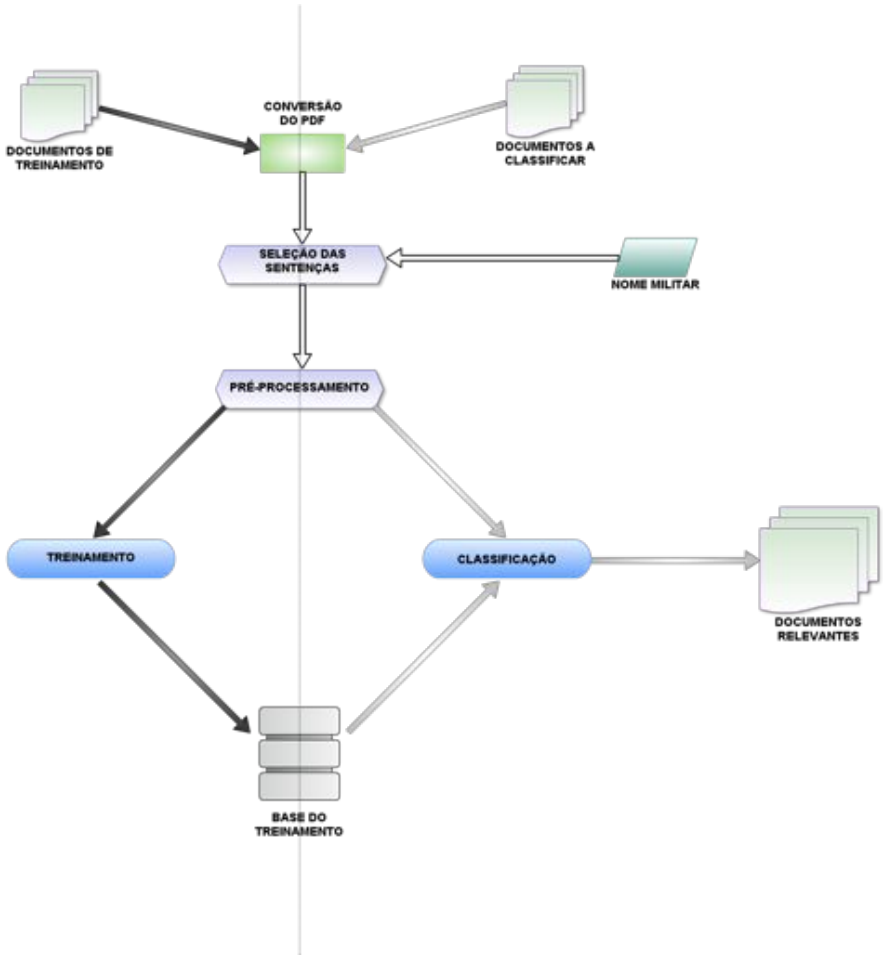

Figura 1. Arquitetura do classificador proposto.

Os *Documentos de Treinamento* são compostos pelas Folhas de Alterações e pelos BIs que deram origem a elas. Para montar a efetiva base de treinamento, são realizadas pesquisas pelo nome de um militar nas FAs. Caso sejam encontrados, os BIs selecionados

são atribuídos à classe dos documentos relevantes, enquanto os demais BIs são atribuídos à classe dos não relevantes. Essa base é utilizada pelo classificador para descobrir a probabilidade das evidências (palavras ou termos) estarem associadas às classes de interesse (relevante e não relevante).

Os *Documentos a Classificar* são um conjunto de BIs e FA que foram confeccionados em semestres anteriores, sendo estes os responsáveis pela formação das classes **relevantes** e **não relevantes**.

A fase chamada de *Conversão de PDF* é responsável pela transformação dos documentos, que encontram-se no formato PDF, em um formato textual que possibilite as próximas fases realizarem as suas atividades. Para realizar esta atividade, foi empregada a biblioteca *open source* para Java chamada *PDFBox*[1].

Na *Seleção de Sentenças*, de cada documento são escolhidos os trechos que serão processados. Inicialmente, são selecionados os documentos onde for encontrado o nome completo do militar (ex. Sander Pes Pivetta) ou a graduação mais o nome de guerra (ex. 3º Sgt Sander). Também é necessário encontra o trecho no texto que realmente referencie o militar em questão. Essa seleção é descrita na próxima seção.

Como a língua portuguesa possui uma grande quantidade de palavras distintas, porém com sentidos semelhantes entre si, ocorre um aumento na esparsidade dos dados, o que dificulta o processo de classificação. Na busca por diminuir esta diversidade, a etapa de *Pré-processamento* utiliza a remoção de *stop words* [Rigo et al. 2007], para remover palavras consideradas irrelevantes no processamento (ex. artigos, preposições e numerais), e o emprego de técnicas de *stemminhg* [Rezende 2005], para realizar uma normalização linguística das palavras.

Na etapa de *Treinamento*, para todas as sentenças selecionadas, são extraídos os eventos, que são representados por *n-gramas* de palavras. Estes eventos são então atribuídos às classes analisadas (relevante e não relevante), gerando tabelas contendo os termos e sua frequência ou incidência nos documentos de treinamento. Essas tabelas formam o *Modelo Treinado*, que depois é usado na etapa de Classificação.

Após a formação da base de conhecimento, torna-se possível realizar a *Classificação* dos documentos. Nesta etapa, através do *Teorema de Bayes* e das evidências coletadas anteriormente (Modelo Treinado), é analisada a probabilidade dos BIs pertencerem a uma das classes. Caso seja verificado que, em pelo menos um dos trechos analisados, a probabilidade de ser relevante for superior à de ser não relevante, o documento é considerado relevante para o militar em questão.

3.1. Seleção de Sentenças

Um dos passos mais importantes na execução do método proposto, tanto na fase de classificação quanto na de treinamento, é o da seleção de trechos dos documentos. Uma seleção inapropriada para um determinado caso implica na utilização de dados errados para o cálculo das probabilidades.

Neste trabalho são propostas duas técnicas para seleção dos trechos, chamadas de "Janela Fixa" e "Janela Deslizante". Ambas partem dos pontos no texto onde o nome

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

do militar, chamado pivô, aparece. Dado um pivô, cada técnica utiliza regras diferentes para selecionar o texto relacionado. A técnica da Janela Fixa considera como informações importantes aquelas que encontram-se mais próximas ao pivô. Para isso, a partir do pivô, é realizada a seleção dos κ caracteres anteriores e posteriores ao nome.

A técnica da Janela Deslizante considera que a informação importante possa estar afastada do pivô, principalmente quando ele estiver contido em uma tabela de nomes. Para fazer essa análise, primeiro deve-se verificar se o pivô encontra-se dentro de uma linha válida. Uma linha é válida se o número de palavras válidas da linha for igual ou superior a λ. Uma palavra é válida se possuir mais do que μ caracteres.

Em seguida, o texto anterior e o posterior ao pivô são selecionados até que símbolos de término de parágrafo sejam encontrados. Caso essa seleção possua $\lambda > 6$, ela é considerada válida e efetivamente selecionada. Caso contrário, ela é descartada e verificações sobre os parágrafos anteriores são realizadas, até que a condição de validação do texto seja satisfeita e esse texto efetivamente selecionado.

Ss Figuras 2 e 3 demonstram exemplos onde o texto "Sander Pes Pivetta" foi usado como pivô. Na Figura 2 foi utilizada a Janela Fixa com κ = 150, ou seja, foram selecionados os 300 caracteres mais próximos ao pivô. Na Figura 3 foi utilizada a Janela Deslizante com $\lambda = 6$ e μ = 3.

> 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 retormo 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. Seleção de texto por Janela Fixa com "Sander Pes Pivetta" como pivô.

> 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 retormo 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. Seleção de texto por Janela Deslizante com "Sander Pes Pivetta" como pivô.

Analisando as Figuras 4 e 5, verifica-se um exemplo onde o nome do militar encontra-se em uma tabela, juntamente com o nome de outras pessoas. Para esta situação, a sentença relevante encontra-se no parágrafo que aparece antes da tabela. Na Figura 4 foi utilizada a seleção por Janela Fixa com κ = 150. Como pode-se ver, a técnica seleciona

praticamente todas as informações contidas na tabela, que são irrelevantes para a análise. 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 desliza 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. Seleção de texto por Janela Fixa quando o pivô ("Sander Pes Pivetta") está em uma tabela.

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. Seleção de texto por Janela Deslizante quando o pivô ("Sander Pes Pivetta") está em uma tabela.

4. Experimentos e Resultados

Foram realizados diversos experimentos com o objetivo de avaliar o método de classificação proposto e as técnicas de seleção de sentenças por Janela Fixa e por Janela Deslizante. Para efeitos de comparação também foi verificada a qualidade de outros dois métodos, chamados "Pesquisa Nominal" e "Documento Inteiro".

O método da Pesquisa Nominal realiza uma categorização dos BIs através de uma pesquisa pelo nome do militar. Caso seja encontrada alguma referência a este militar, o documento é considerado relevante para a composição das suas FA. Neste método, não é empregado o aprendizado de máquina.

O método do Documento Inteiro realiza o treinamento dos BIs sem utilizar a seleção de sentenças. Ou seja, para cada documento utilizado para gerar a Folha de Alteração de um determinado militar, todas as palavras do documento são consideradas como eventos associados à classe de documentos relevantes para esse militar.

O método da Janela Fixa realiza o treinamento dos BIs utilizando a técnica de seleção de sentenças por Janela Fixa, com κ = 150. Assim, 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 à classe de documentos relevantes desse militar.

Similarmente, o método da Janela Deslizante realiza o treinamento dos BIs utilizando a técnica de seleção de sentenças por Janela Deslizante, com $\lambda = 6$ e $\mu = 3$. Caso um documento contenha informação relevantes a respeito de um militar, apenas as palavras que pertencem ao trecho selecionado por essa técnica são consideradas eventos associados à classe de documentos relevantes desse militar.

Os desempenhos dos experimentos foram avaliados usando as medidas de *precisão*, *cobertura* e *medida-f*. A precisão é calculada em função do número de BIs classificados como relevantes e que realmente o são. Já a cobertura é calculada em função do número de BIs relevantes que foram assim classificados. A medida-f é a média harmônica das outras duas medidas.

Para a base de treinamento foram utilizados 200 BIs, confeccionados durante o período de 1 ano, referentes a 64 militares. 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 109 BIs para um conjunto de 16 militares.

As Tabelas 1 e 2 apresentam os valores de precisão, cobertura e *medida-f* encontrados para cada um dos métodos de classificação descritos. Como pode ser observado, o método n-grama utiliza sequencia de 2 e 3 palavras (bigramas e trigramas) para a montagem de suas variáveis.

Observe que, dos quatro métodos, o que utiliza o "Documento Inteiro"é o que possui o pior desempenho, muito inferior aos demais. Isto ocorre porque o treinamento leva em consideração muitos parágrafos que não possuem nenhuma relação como os militares usados durante o treinamento. Com a execução da "Pesquisa Nominal", a medida-f ficou próxima a 50%. Apesar de encontrar todos os documentos relevantes, a precisão obtida foi baixa, pois foram retornados muitos BIs não importantes para a confecção das Folhas de Alterações do militar.

Os melhores resultados foram obtidos com a execução do classificador *bayseano* associado a uma técnica de seleção de sentenças, a Janela Fixa ou a Janela Deslizante. Dentre as duas, a seleção de sentenças através da Janela Deslizante conseguiu realizar uma melhor seleção dos trechos significativos para os militares. O resultado reflete o fato

Tabela 1. Resultados utilizando a frequência das *features*.

MÉTODO	N-GRAMA	PRECISÃO (%)	COBERTURA (%)	MEDIDA-F (%)
Pesquisa Nominal	-	33	100	49,6
Documento Inteiro	1	11	21	13,7
Janela Fixa	1	62	47	53
Janela Deslizante	1	88	51	64,5
Janela Fixa	2	65	48	55
Janela Deslizante	2	88	68	76,7
Janela Fixa	3	61	46	52
Janela Deslizante	3	86	60	69,7

Tabela 2. Resultados utilizando a ocorrência das *features*.

MÉTODO	N-GRAMA	PRECISÃO (%)	COBERTURA (%)	MEDIDA-F (%)
Pesquisa Nominal	-	33	100	49,6
Documento Inteiro	1	12	23	15,7
Janela Fixa	1	61	39	38
Janela Deslizante	1	85	46	59,6
Janela Fixa	2	55	44	49
Janela Deslizante	2	88	52	64,8
Janela Fixa	3	59	41	43
Janela Deslizante	3	77	45	56,8

de que, em muitos casos, as informações relevantes para o militar encontra-se distantes de seu nome, como por exemplo, dentro de listas de nomes. Nestes casos, a seleção com Janela Deslizante consegue percorrer o texto até encontrar a informação mais provável de ser a relevante.

Outra técnica explorada utiliza a frequência com que as palavras ocorrem dentro da base de treinamento, ou a incidência das palavras em cada documento. Nos resultados verificou-se que o uso da frequência apresenta um melhor desempenho comparado com a incidência das palavras, tanto com a Janela Fixa como com a Janela Deslizante. Este resultado mostra que palavras que ocorrem por mais vezes sao bons indicadores da classe.

A forma como as variáveis são selecionadas também merece destaque. Como muitas palavras apresentam sentidos distintos quando próximas a outras palavras, utilizar n-gramas para selecionar as variáveis torna-se oportuno. Nos resultados verifica-se que o emprego de bigramas e trigramas apresenta melhores resultados.

Apesar de modelos bayesianos no geral obterem bons resultados com relativamente poucos dados de treinamento, ainda assim é preciso atentar para que haja evidências suficientes para estimar os parâmetros do modelo. Para verificar como o tamanho da base de treinamento afeta a classificação no problema em questão, foram feitos experimentos variando-se o número de documentos utilizados no treinamento. As Figuras 7 e 6 mostram os resultados obtidos com Janela Fixa e Deslizante em comparação à Pesquisa Nominal.

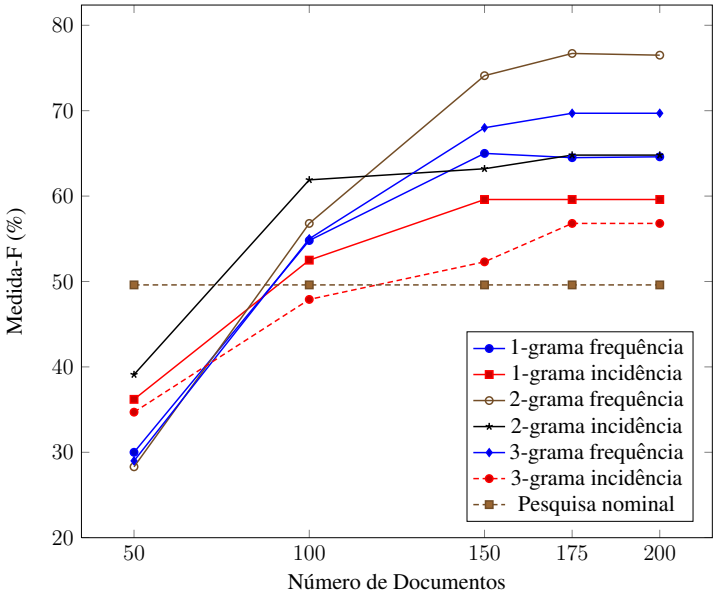

Figura 6. Variação dos resultados utilizando seleção por Janela Deslizante.

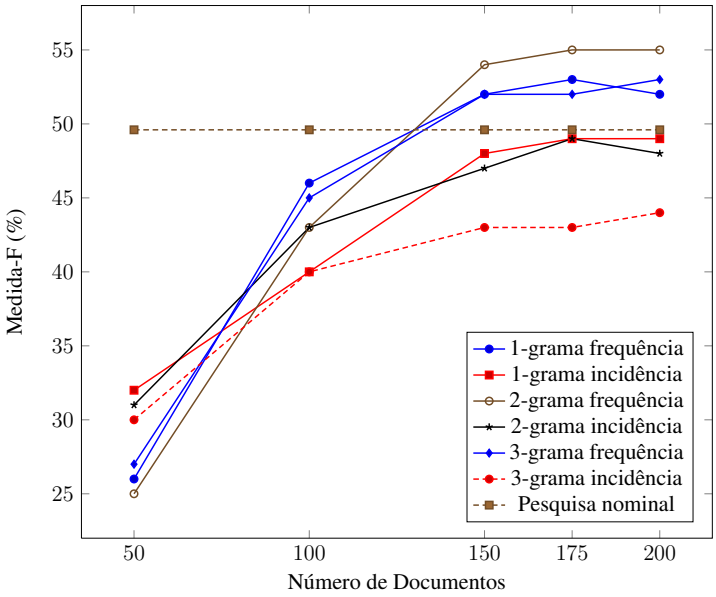

Figura 7. Variação dos resultados utilizando seleção por Janela Fixa.

Analisando os gráficos, é possível verificar que a classificação apresenta resultados baixos treinando a aplicação com um número pequeno de documentos. Conforme a quantidade de documentos utilizados aumenta, são obtidos melhores resultados, até o momento onde os resultados começam a estabilizar. Esta estabilidade é obtida a partir da utilização de 175 BIs, mantendo praticamente constante a partir deste valor.

De modo geral, os resultados são melhores quando é utilizada a seleção Janela Deslizante combinada com a frequência das palavras e com o emprego de bigramas. Esta combinação obteve um valor de medida-f igual a 76,7%, sendo seguido pelo uso de trigramas, que obteve uma medida-f igual a 69,7%.

5. Conclusões

Este artigo apresentou uma aplicação de classificação de documentos que emprega o método de aprendizado de máquina supervisionado *Naive Bayes*. O objetivo deste trabalho foi selecionar os Boletins Internos que deveriam compor as Folhas de Alterações de militares do Exército Brasileiro. Para testar a aplicação foram realizados experimentos com parâmetros e técnicas variados tais como: o uso da frequência ou incidência das palavras; a composição por unigramas, bigramas ou trigramas; e a seleção de sentenças por Janela Fixa ou Janela Deslizante.

A avaliação foi realizada através da classificação de um conjunto de Boletins Internos como relevantes para compor as Folhas de Alterações de militares. Na análise destes resultados, verificou-se que a combinação do algoritmo de *Naive Bayes* com a utilização de n-gramas maiores que um e a seleção adequada das sentenças são as técnicas mais influentes neste estudo. A seleção de sentenças por Janela Deslizante combinada com o uso de bigramas retornou os melhores resultados: 76,7% de medida-f.

O uso da Janela Deslizante retornou melhores resultados porque foi capaz de selecionar algumas sentenças relevantes ao militar que não estavam juntas com o seu nome. Já o uso de bigramas ajudou a identificar o contexto das palavras no texto. O uso de trigramas não obteve o mesmo desempenho possivelmente porque o número de sentenças analisadas não foram suficientes para estimar uma distribuição de probabilidade adequada.

Como trabalhos futuros pretende-se analisar o desempenho das técnicas apresentadas variando-se os valores de κ, λ e μ. Além disso, pretende-se utilizar o algoritmo SVM[2] a fim de verificar se esse método supera o classificador Bayesiano na classificação de BIs do Exército Brasileiro.

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.

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.

Peng, F., Schuurmans, D., and Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. *Inf. Retr.*, pages 317–345.

[2] *Suport Vector Machine*, em inglês.

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*.