Universidade Federal do Pampa

Sander Pes Pivetta

Classificação de Documentos do Exército Brasileiro Utilizando o Classificador *Naive Bayes* e Técnicas de Seleção de Sentenças

Alegrete

2013

Sander Pes Pivetta

Classificação de Documentos do Exército Brasileiro Utilizando o Classificador *Naive Bayes* e Técnicas de Seleção de Sentenças

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Dr. Sergio Luis Sardi Mergen

Alegrete

2013

Sander Pes Pivetta

Classificação de Documentos do Exército Brasileiro Utilizando o Classificador *Naive Bayes* e Técnicas de Seleção de Sentenças

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Trabalho de Conclusão de Curso defendido e aprovado em 05 de março de 2013

Dr. Sergio Luis Sardi Mergen
Orientador

Dr. Fabio Natanael Kepler
UNIPAMPA

Me. João Pablo Silva da Silva
UNIPAMPA

Alegrete
2013

Dedico este trabalho a todas as pessoas que estiveram ao meu lado durante este período. Aos amigos e colegas desta formação, aos meus pais que que muitas vezes escutaram minhas reclamações como: "estou cansado e com sono", "está muito difícil" e sem reclamar, ouviam e me motivavam, e a minha irmã Sanieli. Em especial a minha esposa Daniela que esteve ao meu lado durante este período, tendo compreensão durante os momentos de recolhimento aos estudos.

Resumo

Uma das necessidades do Exército Brasileiro é realizar a classificação dos documentos chamados Boletins Internos, os quais devem ser agrupados afim de gerar relatórios sumarizados a respeito dos militares. Para isto, é necessário encontrar referências relevantes à cada militar, dentro de um conjunto destes documentos confeccionados durante o período de um semestre. Para realizar esta classificação de forma automática, este trabalho utiliza o classificador *bayesiano*. O classificador emprega *n-gramas* como forma de selecionar os atributos de treinamento, recuperando a frequência/ocorrência das palavras nos documentos analisados. Também é necessário identificar quais as sentenças dos documentos são referentes ao militar analisado, para que apenas estas informações sejam empregadas pelo classificador. Este trabalho propõe duas heurísticas que selecionam sentenças relacionadas a cada militar. A aplicação proposta consegue atingir $78,5\%$ de medida-f na recuperação dos documentos relevantes. Além disso, constata-se que o uso dos *n-gramas* consegue realizar uma análise mais precisa das informações, e a seleção de sentenças influencia diretamente na classificação.

Palavras-chave:Classificação Textual, Naive Bayes, Seleção de Sentenças, *n-gramas*.

Abstract

One of the needs of the Brazilian Army is to perform the classification of documents called “Boletins Internos”, which must be grouped in order to generate summarized reports about the military. To accomplish this, it is necessary to find relevant references to each military inside a set of documents, elaborated during the period of one semester. To perform this classification automatically, this work uses the Bayes classifier. The classifier employs n-grams as a way to select the training attributes, identifying the frequency/occurrence of words inside the analyzed documents. It is also necessary to identify which sentences of the documents are related to the analyzed military. We propose two heuristics in order to better perform the selection of sentences that are related to each military. We can see that the proposed implementation can achieve 78.5% F-Measure in the recovery of relevant documents. Furthermore, the use of n-grams can perform a more accurate analysis of the information, and the sentence selection directly influences the classification.

Key-words: Text Classification, Naive Bayes, Sentences Selection, n-grams.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas

Cap Capitão

Cmt Gda Comandante da Guarda

I insatisfatória

Of Dia Oficial de Dia

S satisfatória

Sgt Sargento

TAF Teste de Aptidão Física

TAT Tiro de Ação Tática

Ten Tenente

Lista de siglas

BI Boletim Interno

EB Exército Brasileiro

OM Organização Militar

PDF *Portable Document Format*

SVM *Support Vector Machine*

TCC Trabalho de Conclusão de Curso

Sumário

1 Introdução

O desenvolvimento e a popularização 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 salvos em 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 qual visa separar documentos de acordo com algum critério, o que facilita a tomada de decisões sobre um conjunto de dados.

Muitas organizações necessitam realizar a classificação dos seus documentos, os quais podem ser úteis a várias atividades. O Exército Brasileiro (EB) é um exemplo destas organizações, onde documentos chamados de Boletim Interno (BI) são categorizados a fim de produzirem relatórios sumarizados com informações referentes aos militares. Os BIs são documentos confeccionados periodicamente que contém informações relacionadas às atividades realizadas pela instituição e pelos seus integrantes (EXERCITO, 2002).

A partir destes BIs são gerados documentos chamados de Folhas de Alterações, os quais são produzidas semestralmente e relatam o histórico de um militar referente as atividades por ele desempenhadas e sobre a sua vida pessoal (EXERCITO, 2001). Elas são confeccionadas para cada integrante da Organização Militar (OM), sendo uma atividade demorada que exige um grande trabalho, pois em média são produzidos 120 BIs e o número de Folhas de Alterações produzidas neste período pode ultrapassar 200 em várias OM. Além disso, as Folhas de Alterações devem ser produzidas e entregue aos interessados logo após o término do semestre, sendo um período considerado curto.

Conforme as normas vigentes para confecção das Folhas de Alterações, encontradas em Exercito (2001), nem todos BIs possuem informações relevantes para a sua confecção. Dessa forma, dado um militar, é preciso realizar pesquisas sobre o conteúdo de todos os BIs produzidos durante o período de um semestre, buscando todas as informações relativas ao militar. Estas são analisadas com o objetivo de verificar se elas devem ou não 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 trabalho propõe-se que a separação seja realizada com o auxílio de aprendizado de máquina. De modo geral pode se recorrer ao aprendizado supervisionado, ao aprendizado

semi-supervisionado e ao aprendizado não-supervisionado. Dentre estes, torna-se oportuno o emprego do **Aprendizado Supervisionado** (MITCHELL, 1997), pois dispõem-se de informações já classificadas em semestres anteriores, que podem ser utilizadas para a obtenção da base de conhecimento. Mais especificamente, a tarefa será realizada com o emprego do classificador *Naive Bayes* utilizando duas variações: o uso de *n-gramas* e a frequência/incidência das variáveis na montagem da base de treinamento.

Além disto, outro problema pesquisado envolve a escolha, para cada documento, das informações a serem utilizadas nesta tarefa de classificação. Devido os BIs serem compostos por um conjunto de pequenas informações, referentes a assuntos e pessoas distintas, torna-se necessário encontrar uma maneira de identificar quais sentenças são relevantes para cada militar. Esta escolha, caso seja realizada de forma equivocada, influencia diretamente no desempenho da aplicação, levando-o a realizar uma categorização menos precisa dos documentos. Assim, são analisadas maneiras de selecionar as sentenças a fim de encontrar a sentença que represente a correta informação referente ao militar.

Uma seleção correta das informações, por si só, pode não ser suficiente para a obtenção da melhor classificação das sentenças, pois elas podem apresentar palavras com variações morfológicas ou mesmo não relevantes para a análise. Uma forma de melhorar a apresentação das sentenças é através de um pré-processamento, ou seja, as palavras consideradas irrelevantes são excluídas e as palavras remanescentes são reduzidas ao seu radical. Estas atividades têm a finalidade de formatar a informação analisada e encontrar o melhor resultado.

O restante do trabalho encontra-se organizado da seguinte forma: As técnicas empregadas para realizar a classificação dos documentos textuais, como o aprendizado de máquina, seleção de sentenças e os trabalhos relacionados são apresentados no capítulo 2. O capítulo 3 apresenta a estrutura dos documentos confeccionados pelo EB. Uma explicação detalhada das soluções propostas neste trabalho estão presentes no capítulo 4. Já no capítulo 5 são apresentados os resultados obtidos e no capítulo 6 são tecidas as considerações finais.

2 Classificação Textual

Realizar a classificação de documentos é uma atividade que necessita a combinação de várias técnicas. Uma delas é o aprendizado de máquina, que é o responsável pela tomada de decisões sobre uma hipótese. Outra é a seleção das sentenças, que seleciona somente as informações necessárias para serem analisadas. Acoplada a elas, existem os processamentos das informações, que tem o objetivo de torná-las mais concisas e entendíveis para a análise da aplicação.

Nesta seção é feita uma contextualização a respeito das técnicas de aprendizado de máquina e seleção de sentenças. Além disso, são mencionados trabalhos relacionados que utilizam essas técnicas para resolver problemas semelhantes ao problema de classificação enfrentado pelo EB.

2.1 O Aprendizado de Máquina

Realizar a classificação de documentos é uma tarefa que exige uma análise das suas informações. Uma pessoa pode facilmente desempenhá-la, pois é capaz de entender o seu conteúdo e, com base nos seus conhecimentos, tomar decisões sobre a atividade. Entretanto, uma questão importante é como uma máquina consegue desempenhar essa mesma atividade. Para isto, é utilizada uma área da Ciência da Computação chamada de Aprendizado de Máquina.

O Aprendizado de Máquina destina-se a estudar formas de programar um computador para que este consiga entender padrões e a partir deles, tomar decisões automáticas sobre um assunto. Muitas são as aplicações que necessitam desempenhar esta atividade, por exemplo, a classificação automática de documentos, o reconhecimento de caracteres escritos manualmente, veículos autônomos, extrair conhecimentos de dados biológicos, dentre outras. Para que estas atividades possam ser realizadas, existe a necessidade da aplicação obter uma base de conhecimento sobre a qual são analisadas as hipóteses e, após a tomada de decisão, seja realizada uma ação sobre elas (NILSSON, 1996) (MITCHELL, 1997).

Este conhecimento é obtido através de um treinamento, que pode ser realizado conforme as formas descritas a seguir:

Aprendizado supervisionado: é uma maneira de treinar o classificador, onde é realizado o mapeamento de um expressivo conjunto de exemplos previamente rotulados

em classes. A partir destas informações, é extraído o conhecimento necessário para a execução da atividade (NILSSON, 1996).

Aprendizado não supervisionado: neste caso, os exemplos não possuem uma rotulação pré-definida, sendo o próprio algoritmo responsável por agrupá-los. Desta forma, a base de treinamento é formada através da similaridade entre as informações, sendo o modelo probabilístico o responsável por encontrar o resultado (NILSSON, 1996).

Aprendizado semi-supervisionado: é uma mesclagem realizada entre o aprendizado supervisionado e o não supervisionado. É utilizada quando a gama de exemplos rotulados utilizados no treinamento não é grande o suficiente para que a aplicação consiga analisar de forma correta as hipóteses (NILSSON, 1996).

Para este trabalho, dispõe-se de um grande conjunto de documentos previamente classificados, os quais podem ser utilizados no treinamento do classificador, tornando-se útil então, o emprego do aprendizado supervisionado. Dentre os algoritmos que utilizam este tipo de aprendizado para a obtenção da base de conhecimento, alguns merecem destaque quando empregados na classificação textual. A seguir são apresentados alguns destes algoritmos.

2.1.1 O Algoritmo *Naive Bayes*

Um bom algoritmo utilizado no aprendizado supervisionado é o classificador *Naive Bayes*. Ele é baseado no *Teorema de Bayes* (proposto por *Thomas Bayes*[1]) e tem o objetivo de encontrar a probabilidade a *Posteriori*. O classificador *bayesiano* apresenta uma maneira de calcular a probabilidade da ocorrência de um evento, baseando-se em probabilidades obtidas da análise dos eventos passados. Estas informações são utilizadas para construir a base de conhecimento da aplicação (conjunto de informações responsáveis pela classificação correta das hipóteses) (MITCHELL, 1997).

O modelo possui em seu nome o termo "*naive*", cuja tradução em português significa **simples** ou **ingênuo**. Esta denominação é a ele dada pois seus atributos são analisados independentes uns dos outros, o que permite o seu emprego na resolução de diversos problemas (RUSSELL; NORVIG, 2004).

Por trabalhar com as informações desta forma, o classificador *bayesiano* consegue obter resultados em um menor tempo de execução. Ele também torna-se mais fácil de ser programado, pois não considera o contexto que as informações estão inseridas. Neste caso, sua precisão é considerada máxima somente quando todas as informações forem independentes entre si (RUSSELL; NORVIG, 2004).

[1] 1702-1761

O objetivo do classificador *Naive Bayes* 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 2.1.

$$P\left(h|D\right) = \frac{P\left(D|h\right)P\left(h\right)}{P\left(D\right)} \tag{2.1}$$

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 $P(D)$, a qual 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).

Normalmente, deseja-se encontrar a classe mais provável do evento pertencer h_{MAP}, ou seja, a classe com o máximo valor da *Posteriori*. Assim, para realizar a classificação *bayesiana* é utilizada a Equação 2.2. A probabilidade $P(D)$ pode ser desconsiderada do cálculo da probabilidade do algoritmo de *Naive Bayes*, uma vez que esta é igual para todas as classes (MITCHELL, 1997).

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned} \tag{2.2}$$

Para encontrar a probabilidade da classe h ocorrer, a *Priori*, é necessário saber o número de exemplares de cada classe. Assim, divide-se o número de amostras associados à classe N_i pelo total de amostras utilizadas no treinamento da aplicação N (FERREIRA, 2005):

$$P(h) = \frac{N_i}{N} \tag{2.3}$$

Outro parâmetro necessário é a probabilidade do evento D em relação à classe h ($P(D|h)$). Este é calculado através da Equação 2.4, onde $t \in D$ e representa o atributo do evento e n é o total de atributos contidos no evento.

$$P(D|h) = \prod_{i=0}^{n} P(t_i|h) \tag{2.4}$$

Para encontrar a probabilidade de cada termo $P(t_i|h)$, é utilizada a Equação 2.5. Nela $T(t_i|h)$ representa a frequência que o atributo t_i é encontrado na classe h e n representa o total de atributos do evento. A ação de somar 1 a estes valores refere-se a suavização de *Laplace*, apresentada durante a seção 4.1.8. Ela tem o objetivo de evitar o encontro de uma probabilidade nula para o evento.

$$P(t|h) = \frac{T(t_i|h) + 1}{\sum_{i=0}^{n} T(t_i|h) + 1} \tag{2.5}$$

Uma empregabilidade do algoritmo *Naive Bayes* é a atividade da classificação textual. Neste caso, cada documento textual é um evento e as palavras dos documentos são os atributos. Normalmente os documentos podem ser categorizado em duas classes indicando documentos que sejam ou não relevantes para uma aplicação específica. O exemplo a seguir apresenta como é realizada esta operação pelo classificador *bayesiano*.

Um professor do curso de Ciência da Computação pediu aos seus alunos que realizassem uma avaliação da sua disciplina. Para isso, foi entregue um formulário onde o aluno deve caracterizar a disciplina como satisfatória (S) ou insatisfatória (I), além de ser escrito um comentário a respeito da disciplina. As respostas dos formulários encontram-se descritas na Tabela 1.

Tabela 1 – Resposta da avaliação da disciplina.

Nome Aluno	**Comentário**	**Avaliação**
aluno1	“bom.”	S
aluno2	“muito bom.”	S
aluno3	“ruim.”	I
aluno4	“muito ruim.”	I
aluno5	“muito ruim, muito ruim.”	I
aluno6	“bom? ruim! muito ruim.”	-

Como pode ser observado na tabela, todos os alunos completaram corretamente o formulário, com a exceção do aluno6. Ele somente preencheu o comentário sobre a disciplina, não realizando a sua avaliação. Com estas informações, o professor decidiu descobrir qual seria a avaliação deste aluno, a partir do comentário realizado. Para isto, utilizou o classificador *bayesiano* como meio de encontrar a resposta.

Assim, o primeiro passo realizado é encontrar a probabilidade a *Priori*, aplicando a equação 2.3.

$$P(\mathrm{S}) = \frac{N_{\mathrm{S}}}{N_{\mathrm{I}} + N_{\mathrm{S}}} = \frac{2}{2+3} = 0,40$$

$$P(\mathrm{I}) = \frac{N_{\mathrm{I}}}{N_{\mathrm{I}} + N_{\mathrm{S}}} = \frac{2}{3+2} = 0,60$$

Após é calculada a probabilidade de ocorrência dos eventos, ou seja, a probabilidade dos atributos (palavras) pertencerem às classes, utilizando a Equação 2.5.

$$P(bom/\mathrm{S}) = \frac{T_{bom/\mathrm{S}} + 1}{(T_{bom/\mathrm{S}} + 1) + (T_{muito/\mathrm{S}} + 1) + (T_{ruim/\mathrm{S}} + 1)}$$

$$P(bom/\mathrm{S}) = \frac{2+1}{(2+1) + (1+1) + (0+1)} = 0,50$$

$$P(muito/\mathrm{S}) = \frac{T_{muito/\mathrm{S}} + 1}{(T_{bom/\mathrm{S}} + 1) + (T_{muito/\mathrm{S}} + 1) + (T_{ruim/\mathrm{S}} + 1)}$$

$$P(muito/\mathrm{S}) = \frac{1+1}{(2+1) + (1+1) + (0+1)} = 0,33$$

$$P(ruim/\mathrm{S}) = \frac{T_{ruim/\mathrm{S}} + 1}{(T_{bom/\mathrm{S}} + 1) + (T_{muito/\mathrm{S}} + 1) + (T_{ruim/\mathrm{S}} + 1)}$$

$$P(ruim/\mathrm{S}) = \frac{0+1}{(2+1) + (1+1) + (0+1)} = 0,17$$

$$P(bom/\mathrm{I}) = \frac{T_{bom/\mathrm{I}} + 1}{(T_{bom/\mathrm{I}} + 1) + (T_{muito/\mathrm{I}} + 1) + (T_{ruim/\mathrm{I}} + 1)}$$

$$P(bom/\mathrm{I}) = \frac{0+1}{(0+1) + (3+1) + (4+1)} = 0,40$$

$$P(muito/\mathrm{I}) = \frac{T_{muito/\mathrm{I}} + 1}{(T_{bom/\mathrm{I}} + 1) + (T_{muito/\mathrm{I}} + 1) + (T_{ruim/\mathrm{I}} + 1)}$$

$$P(muito/\mathrm{I}) = \frac{3+1}{(0+1)+(3+1)+(4+1)} = 0,10$$

$$P(ruim/\mathrm{I}) = \frac{T_{ruim/\mathrm{I}}+1}{(T_{bom/\mathrm{I}}+1)+(T_{muito/\mathrm{I}}+1)+(T_{ruim/\mathrm{I}}+1)}$$

$$P(ruim/\mathrm{I}) = \frac{4+1}{(0+1)+(3+1)+(4+1)} = 0,50$$

Para encontrar a probabilidade a *Posteriori*, é necessário encontrar a probabilidade condicional, ou seja, a probabilidade da disciplina ser classificada como S e I, conforme a descrição do aluno6. Para encontrá-la é usada a Equação 2.4.

$$P(\mathrm{S}/aluno6) = P(bom/\mathrm{S}) \times P(ruim/\mathrm{S}) \times P(muito/\mathrm{S}) \times P(ruim/\mathrm{S})$$

$$P(\mathrm{S}/aluno6) = 0,50 \times 0,17 \times 0,33 \times 0,17$$

$$P(\mathrm{S}/aluno6) = 0,0048$$

$$P(\mathrm{I}/aluno6) = P(bom/\mathrm{I}) \times P(ruim/\mathrm{I}) \times P(muito/\mathrm{I}) \times P(ruim/\mathrm{I})$$

$$P(\mathrm{I}/aluno6) = 0,10 \times 0,50 \times 0,40 \times 0,50$$

$$P(\mathrm{I}/aluno6) = 0,010$$

Após obtidos estes resultados, torna possível o encontro da probabilidade a *Posteriori*, ou seja, a probabilidade do aluno6 ter classificado a disciplina como satisfatória ou insatisfatória. Neste momento é utilizada a Equação 2.1, que descreve o funcionamento do classificador *Naive Bayes*.

$$P(aluno6/\mathrm{S}) = P(\mathrm{S}/aluno6) \times P(\mathrm{S}) = 0,0048 \times 0,40 = 0,0019$$

$$P(aluno6/\mathrm{I}) = P(\mathrm{I}/aluno6) \times P(\mathrm{I}) = 0,010 \times 0,60 = 0,0060$$

Analisando os resultados, verifica-se qual das duas classes obteve a maior probabilidade e, de acordo com a Equação 2.2, verifica-se que $P(aluno6/S) < P(aluno6/I)$. Assim o professor concluiu que, conforme a descrição realizada, o aluno6 classificaria a disciplina como insatisfatória.

Utilizar o algoritmo de *Naive Bayes* na resolução de problemas deste âmbito obtém resultados satisfatórios, principalmente nas situações onde exista uma boa quantidade de exemplares a ser usados no treinamento. Observa-se também, que quanto maior for a abrangência do treinamento, maior será a confiabilidade dos resultados (MITCHELL, 1997).

2.1.2 O uso do algoritmo de *Naive Bayes* na classificação de textual

Como o classificador *bayseano* supõe uma independência entre as variáveis, torna-se fácil de ser treinado em termos de complexidade de tempo. Na classificação textual, devido os documentos serem compostos por várias palavras onde cada uma representa um evento, ele obtém o resultado em um menor tempo de execução. Este bom desempenho é verificado nos trabalhos que o utilizam na classificação de documentos.

O desempenho do classificador *Naive Bayes* é analisado em Koga (2011), sendo comparado com outros métodos de classificação existentes. Seu objetivo envolve a classificação automática dos sujeitos das frases. Em seu treinamento, foi utilizado um conjunto de atributos morfológicos e estruturais extraídos de frases pré-processadas. Os algoritmos foram implementados dento de *software Waikato Environment for Knowledge Analysis* (WEKA, 1996). Nos resultados, foi observado um melhor desempenho do classificador *bayesiano*, o que foi justificado pela maioria das informações analisadas possuírem independência entre si.

Um problema clássico onde classificadores *bayesianos* são geralmente utilizados é a análise de mensagens do tipo *spam*. Devido ao problema causado por eles aos provedores de *emails*, existe uma grande quantidade de mensagens eletrônicas já classificadas como *spams*. Isto possibilita que seja empregado o aprendizado supervisionado para realizar esta atividade. Rabelo, Filho e Oliveira (2011) verificaram que a resposta correta foi retornada em 85% dos *emails* analisados. Evidenciou também, que a precisão diminuía a medida que o conteúdo das mensagens apresentavam um maior número de palavras.

Como o classificador *Naive Bayes* não é o único algoritmo utilizado na classificação de documentos, Ting, Ip e Tsang (2011) apresentam uma comparação entre alguns algoritmos de aprendizado supervisionado. Dentre eles, o classificador *bayesiano* consegue obter o melhor resultado, superando até mesmo classificadores mais sofisticados, como o SVM. Também são analisadas formas de obter o melhor resultado, onde fica evidenciado o ganho provocado pelo pré-processamento das informações.

Devido os documentos serem compostos por várias palavras e estas podem estar repetidas no conjunto de treinamento, torna-se possível uma abordagem sobre o algoritmo de *Naive Bayes* de duas maneiras distintas: analisando a incidência ou a frequência das palavras (MCCALLUM; NIGAM, 1998).

No modelo de Bernoulli, que usa a incidência dos eventos, a base do treinamento é composta por um vetor binário que contém a palavra e a sua incidência dentro do conjunto de treinamento. Este vetor binário contem todas as palavras que foram encontradas pelo nenos uma vez durante o treinamento. Schneider (2004) mostra um estudo sobre as vantagens da utilização deste modelo na classificação de documentos, evidenciando que, em muitos casos, esta abordagem apresenta melhores resultados quando comparada com a outra. Tal superioridade também é constatado no trabalho realizado por Lewis (1998).

A outra abordagem analisa a frequência das palavras nos arquivos de treinamento. Diferente da técnica anterior, onde somente é verificada a presença das palavras, neste modelo são atribuídos pesos a elas. Estes pesos são calculados conforme o número de vezes em que elas são encontradas durante a fase de treinamento, sendo guardada a sua frequência em uma tabela, juntamente com a palavra propriamente dita. McCallum e Nigam (1998) fazem uma comparação entre os dois processos, e diferente dos trabalhos acima, verifica que este modelo consegue realizar uma classificação mais eficiente dos documentos analisados. Nota-se também que, um aumento no vocabulário, obtido com o treinamento, proporciona uma maior precisão na classificação dos documentos.

Na fase de classificação, ao ser realizada um busca por uma palavra, esta pode não ser encontrada na base de treinamento, o que pode mascarar o resultado do classificador. Assim, é necessário realizar uma normalização sobre estas sentenças. Grau et al. (2004) apresenta em seu trabalho formas de realizar esta normalização, onde tem destaque o método de *Laplace*. Este método propõem a adição de um valor λ a todos os vocábulos pesquisados, evitando assim o retorno de uma probabilidade nula para este vocábulo, e uma inconsistência no resultado da aplicação.

Como o algoritmo de *Naive Bayes* analisa as sentenças como atividades isoladas, ele é considerado *ingênuo*, pois a maioria dos eventos possui uma dependência com os outros. Mesmo assim ele consegue realizar uma boa classificação (RUSSELL; NORVIG, 2004). Por este motivo, ele torna-se um algoritmo que não apresenta uma alta complexidade de programação, tornando-se muito utilizado no aprendizado de máquina.

Assim, na classificação de documentos, o significado das palavras no texto acaba sendo descartado. Sabe-se que elas podem assumir sentidos diferentes, conforme a sua associação com outras palavras. Isto possibilita realizar outra abordagem, a qual objetiva analisar este significado. Peng, Schuurmans e Wang (2004) também fazem uma abordagem sobre o problema clássico de classificação de *emails* como *spams*, onde é realizada uma comparação entre o algoritmo *Naive Bayes* sem mudanças, onde as variáveis são

palavras independentes, com uma modificação, onde as variáveis são obtidas com a uso de *n-gramas*, ou seja, n palavras adjacentes. Na comparação dos resultados verificou-se que com o uso de trigramas foi realizada uma melhor categorização das sentenças.

Para exemplificar esta atividade, considere a seguinte frase: "Sander está caminhando". Com o uso de uni-gramas, cada palavra irá representar um evento. Utilizando bigramas, serão formados dois eventos, "Sander está" e "está caminhando" e com trigramas irá formar apenas um evento, "Sander está caminhando". Assim é possível verificar como são selecionados os eventos através de *n-gramas*.

Apesar do uso desta técnica não empregar o real significado que estas palavras apresentam no texto, esta prática consegue, em muitos casos, melhorar o desempenho do classificador. Hartmann et al. (2011) compara o uso de uni-gramas, bigramas e trigramas, verificando que os bigramas apresentaram os melhores resultados, superando os trigramas.

De forma semelhante, Braga, Monard e Matsubara (2009) analisaram o desempenho de bigramas, verificando que ele apresenta um desempenho superior aos uni-gramas. Também foi analisado o desempenho da aplicação ao ser empregado bigramas e uni-gramas simultaneamente, ou seja, quando não for encontrado o bigrama são usados os uni-gramas que o compõe, não resultado em uma melhor classificação das sentenças.

2.1.3 O Algoritmo *Support Vector Machine*

Outro algoritmo de aprendizado de máquina que pode ser empregado na classificação textual é o *Support Vector Machine* (SVM). Ele é uma técnica de aprendizado supervisionado estudada e desenvolvida por Vladimir Vapnik que pode ser aplicada em várias áreas de estudo como bioinformática, classificação de textos e imagens, identificação de *spams* e reconhecimentos de assinaturas. É baseado no princípio da minimização do risco estrutural, ou seja, procura diminuir os erros verdadeiros[2] retornados da análise de uma hipótese (VAPNIK, 2006; KECMAN, 1948).

É um método linear (podendo ser adaptado para ser não-linear) que tem por objetivo reconhecer os padrões de problemas (encontrados em treinamentos realizados), defini-los sobre um espaço vetorial e separá-los em duas classes distintas. Seu problema está exatamente em encontrar o hiperplano ótimo, ou seja, realizar o processo de categorização dos espaços vetoriais distintos, positivos e negativos, e maximizar a margem de separação das classes. O hiperplano é uma superfície de decisão composta por margens (vetores que demarcam os limites das classes), onde as suas dimensões alteram a generalização dos padrões analisados (margens largas elevam a generalização, enquanto margens estreitas a diminuem) (VAPNIK, 2006; JUNIOR, 2003).

[2] Um erro verdadeiro é obtido quando uma hipótese classificada como 'verdadeira' para uma classe é retornada como 'falsa' pela aplicação

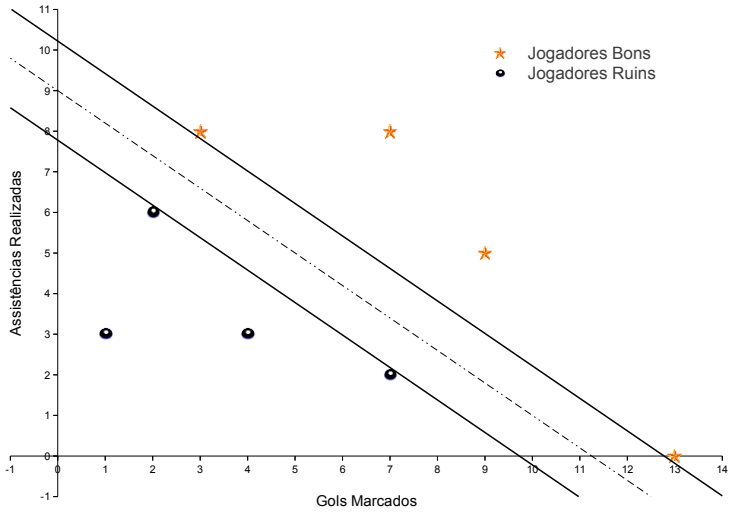
Para demonstrar o funcionamento do SVM, considere o exemplo apresentado na Tabela 2. Nele é analisado a qualidade dos atacantes em um campeonato de futebol, sendo estes classificados em jogadores "bons" ou "ruins". A análise leva em consideração dois parâmetros: o número de gols marcados e a quantidade de assistências realizadas.

Tabela 2 – Desempenho obtido pelos atacantes durante o campeonato.

Nome	**Gols Marcados**	**Assistências**	**Classificação**
Jogador A	13	0	Bom
Jogador B	4	3	Ruim
Jogador C	7	2	Ruim
Jogador D	9	5	Bom
Jogador E	1	3	Ruim
Jogador F	2	6	Ruim
Jogador G	7	8	Bom
Jogador H	3	8	Bom

Analisando o desempenho dos jogadores mostrado na Tabela 2, é possível plotar estes valores em um gráfico e realizar a separação através de hiperplanos. Observe no gráfico da Figura 1 como é realizada a separação entre as classes de jogadores e a obtenção do hiperplano ótimo, representado pelo hiperplano tracejado.

Figura 1 – Representação do hiperplano ótimo de separação entre as classes



Com base no exposto acima, verifica-se que a largura do hiperplano interfere diretamente na probabilidade de ocorrer um erro na classificação. Quanto maior for a distância entre as margens de separação, mais preciso tende a ser o algoritmo de classificação. Isto mostra que o principal objetivo do SVM é encontrar a maior separação entre as margens das classes analisadas (DUDA; HART; STORK, 2001).

A obtenção deste hiperplano ótimo está definido em Vapnik (2006), sendo utilizados vetores de suportes, encontrados através da minimização da função quadrática descrita conforme a a Equação 2.6

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i (h(x), h(x_i)) + \beta_0 \tag{2.6}$$

Observação: α e β representam os parâmetros encontrados durante o treinamento, $(h(x), h(x_i))$ constituem os vetores de características e y é a classificação das classes.

Apesar do SVM suportar somente a separação entre duas classes de forma linear, nem sempre isto é possível, por exemplo, quando aplicado na bioinformática ou em algumas situações da classificação textual. Nesses casos, surge a possibilidade de utilizar classificadores não-lineares, onde é utilizado a estrutura de *kernels*. Para encontrar o *kernel*, função de separação entre pontos não-lineares de um conjunto de treinamento, deve-se dividir os pontos encontrados com o treinamento em pequenos conjuntos. Estes conjuntos são tratados de forma linear onde esta divisão pode ser realizada através da decomposição multi-classe **Um-contra-todos** (uma classe é comparada com todas as outras) ou **Todos-contra-todos** (todas as classes são testadas, duas a duas, com as restantes) (PIMENTA, 2004).

Para realizar esta classificação, algumas regras são necessárias para enquadrá-las na função *kernel*. Uma delas é a necessidade de que estejam contidas dentro da definição do *Teorema de Mercer* definida no trabalho de Mercer (1909), ou seja, sua matriz de representação deve ser uma matriz positiva (SMOLA; BERNHARD, 1999).

Apesar de existir a necessidade das funções *kernel* estarem definidas no *Teorema de Mercer*, nem sempre há esta possibilidade. Nestas situações, surge a necessidade do emprego de outras técnicas para a obtenção do hiperplano ótimo. As mais comuns podem ser encontradas na Tabela 3 (SMOLA; BERNHARD, 1999; KECMAN, 1948).

Tabela 3 – Funções de Kernels mais utilizados pelo classificador SVM.

Tipo	**Função**
Polinomial A	$[(x_i \times X_j) + 1]^p$
Gaussiano B	$exp\left(-\frac{\parallel X_1 - X_j \parallel^2}{2\sigma^2}\right)$
Sigmoidal C	$tanh\left[\beta_0 \left(X_i \times X_j\right)\right] + \beta_1$

Fonte: (SMOLA; BERNHARD, 1999; KECMAN, 1948)

2.1.4 O uso do *Support Vector Machine* na classificação de textual

O SVM é um algoritmo que possui uma maior complexidade quando comparado com o *Naive Bayes*. Assim, Kim, Howland e Park (2005) realizaram um estudo da complexidade computacional deste classificador, onde é analisado o seu desempenho com uma diminuição de seus eventos. Além deste tema, vários autores apresentam exemplos de seu uso na classificação de documentos, com a adição de funções, visando uma melhoria em seus resultados.

Em Silva e Vieira (2007) é realizada a classificação automática de artigos extraídos do jornal Folha de São Paulo. São realizadas comparações entre algoritmos de classificação, onde verifica-se que o SVM possui um melhor desempenho quando na presença de uma maior quantidade de exemplos utilizados no treinamento. De forma semelhante, Joachimis (1998) utiliza 5 algoritmos de aprendizado supervisionado para realizar a classificação de dois conjuntos de documentos textuais. Ele observa um desempenho significativamente melhor do algoritmo SVM comparado aos demais em ambos os conjuntos de testes. Basu, Watters e Shepherd (2003) mostram que o SVM apresenta um melhor desempenho comparado com o algoritmo de Redes Neurais Artificiais, principalmente na presença uma grande quantidade de informações a serem usadas no treinamento.

Já Tong e Koller (2001) apresentam em seu estudo o desempenho obtido pelo SVM, ao realizar a classificação de documentos extraídos dos conjuntos de dados *Reuters-21578* e *Newsgroups* utilizando três produções: o *Simple Margin*, o *MaxMin Margin* e o *Ratio Margin*. É observado que a produção que utiliza *Ratio Margin* consegue obter resultados mais consistentes em um período de tempo não muito longo.

Em Carpineto, Michini e Nicolussi (2009) é realizada uma comparação entre os resultados obtidos com o uso do textitkernel linear e do *kernel gaussian* associados aos métodos baseados em *kernel* e com o modelo tradicional bag-of-words. Observou-se que o método baseado em *kernel* obteve uma performance superior.

Outra forma de avaliar o desempenho do SVM é apresentada em Alves (2010). Nele é verificado o desempenho obtido com a utilização de *stop-words* e *stemminig* na etapa de pré-processamento das informações. Também são associados os modelos *Bag-of-words*, *n-grams* e *Pos-Tag* ao SVM no fase de seleção das variáveis para verificar qual combinação de técnicas apresenta o melhor desempenho. Ficou evidenciado que *n-gramas* não apresentaram melhoras na classificação das sentenças, mas a técnica de *stemming* obteve bons resultados.

Em outra tentativa de melhorar seu desempenho, Silic et al. (2007) também realiza uma comparação com *n-gramas* como forma de selecionar os eventos a serem analisadas. Ao realizar a classificação de artigos produzidos na Croácia em diversas classes, percebe que o uso de *n-gramas* apresenta resultados inferiores ao classificador original. Carpineto,

Michini e Nicolussi (2009) realizam experimentos com o *kernel linear* e com o *kernel gaussian* mesclando-os com métodos de seleção de eventos: o método tradicional de *bag-of-words* e com o método baseado em *kernels*. Verificou-se que o *kernel linear* apresenta os melhores resultados, independente da técnica de seleção de eventos usada.

Seguindo esta ideia, Pang, Lee e Vaithyanathan (2002) realizam uma aplicação que tem o objetivo de classificar sentimentos existentes em resenhas de filmes, extraídas de sites da *Web*, onde as palavras foram separadas em uni-gramas e em bigramas. Associados a elas foram utilizados os classificadores de *Naive Bayes*, SVM e Máxima Entropia. Nos resultados percebe-se que o SVM realiza uma melhor classificação, mas os resultados são semelhantes aos obtidos pelo *Naive Bayes*.

2.2 O Uso da Seleção de Sentenças

Os documentos textuais normalmente são compostos por várias sentenças (parte de um texto capaz de apresentar informações pertinentes a uma informação). Estas sentenças podem estar associadas com o objetivo de relatar informações sobre um único assunto, assim como apresentarem sentidos distintos e apresentarem novas informações.

Por este motivo, algumas aplicações necessitam separar estas informações para conseguir realizar a sua atividade. Um exemplo são os documentos que podem ser compostos tanto por informações relevantes como não relevantes. Nestes casos, existe a necessidade de percorrê-los e encontrar as informações pertinentes para a realização da atividade. Este processo recebe o nome de **seleção de sentenças**.

Em uma forma mais simples, **seleção de sentenças** é a atividade que seleciona informações pertinentes, dentro de um documentos, possibilitando a realização de uma atividade, como é o caso da classificação de documentos.

Schönhofen e Benczúr (2005) apresentam a grande quantidade de tempo gasto para realizar uma classificação de um documento onde não é aplicada a seleção de sentenças. Com o intuito de melhorar o seu desempenho temporal, ele propõe uma técnica baseada na correlação entre as palavras, como forma de selecionar partes do texto que consigam representar o seu conteúdo com o menor número de palavras possível.

Para realizar uma boa classificação de documentos, é necessário utilizar somente as informações pertinentes, sendo necessário selecionar estas sentenças. Ko, Park e Seo (2004) analisam a importância das sentenças verificando a semelhança com o título e a sua importância no contexto do documento. Em Hachey e Grover (2004) é realizada uma classificação de documentos jurídicos conforme o seu tema onde a seleção de sentenças é usada para encontrar a informação que melhor defina o tema abordado pelo documento. Esta seleção resulta em uma melhor classificação, visto que são ignoradas informações que

não representam o conteúdo destes documentos.

Em Khoo, Marom e Albrecht (2006) é apresentado como a seleção de sentenças melhora os resultados de um classificador textual. Verifica-se que o emprego das técnicas de seleção de sentenças produzem uma melhor classificação dos documentos. Segundo ele, este resultado está vinculado a grande gama de palavras existentes nos documentos, onde a seleção acaba em uma redução neste número melhorando o resultado.

Também é importante mencionar que, até mesmo a seleção das sentenças pode utilizar os algoritmos de aprendizado de máquina, como apresentado no trabalho de Metzler e Kanungo (2008), que utiliza o aprendizado supervisionado. Já García-Hernández et al. (2008) utilizada o aprendizado não supervisionado, onde na comparação com as outras técnicas de seleção de informações, percebe-se uma diminuição no número de sentenças redundantes selecionadas.

Outra atividade onde a seleção de sentenças possui empregabilidade é no processo de sumarização de textos[3]. Sumarizar documentos significa encontrar partes dele que melhor representem o seu conteúdo e, a partir delas, compor um resumo que apresente o conteúdo do arquivo. Um exemplo é a ferramenta TXTRACTOR desenvolvida por McDonald e Chen (2002). Ela sumariza os textos, realizando a seleção das informações através de uma segmentação do texto, onde estas são separadas conforme as semelhanças apresentadas.

Já no trabalho apresentado por Jorge e Pardo (2011), é realizado um ranqueamento entre as informações selecionadas e, aquelas que apresentarem um melhor resultado, são utilizadas na confecção do resumo. Outra técnica de selecionar as informações relevantes considera os parágrafos como sentenças. Estes são atribuídos a três categorias (*essenciais*, *complementares* ou *supérfluas*), sendo somente selecionados para o processo de sumarização aqueles que pertencerem as classes *essenciais* e *complementares* (PARDO; RINO; NUNES, 2003)

2.3 O Processamento das Palavras Aplicado ao Classificador

A classificação de documentos é uma atividade que trabalha com uma grande quantidade de variáveis, as quais são representadas por palavras. Concomitantemente, na língua portuguesa muitas destas palavras apresentam significados semelhantes. Isto significa que, ao utilizar as sentenças em seu formato original, pode ocorrer uma influência negativa na obtenção do resultado.

Com o objetivo de melhorar a forma como estas informações se apresentam, são utilizadas técnicas que visam diminuir o espectro de palavras utilizadas pela aplicação.

[3] versão mais curta do documentos que contém as informações mais relevantes (JORGE; PARDO, 2011)

Neste sentido, ganha destaque o emprego de *stop-word* e *stemming*.

Stop-word consiste em uma lista de palavras (*stoplist*) irrelevantes para a classificação de textos, ou seja, são palavras que ao serem removidas não alteram o sentido das informação. A *stoplist* é composta por pronomes, preposições, numerais, artigos, dentre outras palavras (SURESH et al., 2011; WAJEET; ADILAKSHMI, 2009; REZENDE, 2005). Goldstein et al. (1999) mostra que a remoção de *stop-words* melhora o desempenho, visto que diminui o número de palavras, ocorrendo uma diminuição dos eventos a serem analisados.

Em Wajeet e Adilakshmi (2009) é verificado que a remoção de *stop-words* facilita a seleção dos eventos do documento. Em Suresh et al. (2011) é apresentado o desempenho de seu classificador utilizando *stop-words*, sendo que este superou em 10% os resultados do classificador que não removeu os *stop-words* dos documentos analisados.

Já *stemming* é a atividade que busca reduzir as palavras ao seu radical sem realizar alterações em seu sentido (Normalização Morfológica). Nesta atividade palavras são passadas para o singular, verbos são colocados no infinitivo, afixos são removidos, dentre outros (IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005; REZENDE, 2005). O trabalho desenvolvido por Alvares (2005) propôs um algoritmo de *stemming* para a língua portuguesa, visto que a maioria dos algoritmos existentes eram para a língua inglesa.

3 Documentos do Exército Brasileiro

Para que a classificação automática dos documentos possa ser programada, é necessário ter o entendimento sobre a estrutura dos documentos e quais os tipos de informações estão presentes neles. Além disso, e com uma maior importância, é necessário entender o processo de construção do histórico militar. A partir destes conhecimentos, é possível delimitar as estratégias que utilizem, da melhor forma, os documentos disponíveis e conseguir aumentar a eficácia do classificador.

Focando esta ideia, as seções a seguir realizam uma descrição dos Boletim Interno (BI) e das Folhas de Alterações, além da forma como eles são confeccionados.

3.1 Boletim Interno

Diariamente são produzidos diversos tipos de documentos pelo Exército Brasileiro (EB), tendo destaque os BIs. Estes documentos são definidos como o "(...) instrumento pelo qual o comandante, chefe ou diretor divulga suas ordens, as ordens das autoridades superiores e os fatos que devam ser do conhecimento da Organização Militar (OM) (...)" (EXERCITO, 2002, p. 16). Cada OM é responsável pela confecção deste documentos e, neles relatar todas as atividades a ela importantes e às pessoas que nela trabalham.

Segundo Exercito (2004), os BIs são documentos que devem ser publicados periodicamente, conforme as necessidades apresentadas pela OM. Ele é estruturado em quatro seções, **Serviços Diários**, **Instrução**, **Assuntos Gerais e Administrativos** e **Justiça e Disciplina**, as quais apresentam informações referentes aos seguintes assuntos:

Serviços Diários : Esta é a primeira seção, onde está apresentada a relação dos militares escalados para compor a equipe do serviço diário, como por exemplo, Oficial de Dia (Of Dia), Comandante da Guarda (Cmt Gda), dentre outros. Estas são as pessoas responsáveis por realizar a segurança das instalações e a defesa da OM (EXERCITO, 2004).

Instrução : Nesta segunda seção, estão publicadas informações referentes as instruções (atividades de ensino realizadas durante o ano que visam formar e preparar os militares para a realização das atividades afins), que serão desempenhadas pela OM, assim como os militares responsáveis por executá-las (EXERCITO, 2004).

Assuntos Gerais e Administrativos : Esta parte do documento é a que apresenta o maior quantitativo de informações, pois relata as atividades a serem realizadas pela organização, ordens emitidas pelo Comandante e por autoridades superiores, alterações ocorridas com o material e todas as informações referentes aos seus integrantes, excetuando-se as específicas de outras seções (EXERCITO, 2004).

Justiça e Disciplina : A última parte do BI apresenta assuntos relacionados a vida disciplinar dos militares, como por exemplo, informações sobre punições de militares e referências elogiosas (EXERCITO, 2004).

Figura 2 – Exemplo de um Boletim Interno

ORIGINAL

Folha 001

MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
Comando Militar de Área – Divisão de Exército - Brigada
Organização Militar
Nome Histórico da Organização Militar

Quartel em CIDADE, ESTADO, **01 Jan 01**

BOLETIM INTERNO Nr 001

PARA CONHECIMENTO DESTE ORGANIZAÇÃO MILITAR E DEVIDA EXECUÇÃO, PUBLICO O SEGUINTE:

1ª PARTE - SERVIÇOS DIÁRIOS

1. **SERVIÇOS PARA O DIA 02 Jan 01**

Of Dia	2º Ten **Fulano**
Cmt Gda	3° Sgt **Beltrando**
Sgt Dia	3º Sgt **Ciclano**
Cb Gda	Cb **Fulano**
Cb Gda	Cb **Beltrano**
Gda	Sd **Ciclano**
Gda	Sd **Fulano**

2ª PARTE - INSTRUÇÃO

2. **INSTRUÇÃO DO DIA 02 Jan 01**
a. Determino que todos soldados do efetivo variável participem da instrução de defesa do aquartelamento, que se realizará amanha no campo de futebol, sob responsabilidade do 1º Ten Fulano.

3ª PARTE – ASSUNTOS GERAIS E ADMINISTRATIVOS

3. **APRESENTAÇÃO DE MILITARES**
a. O 3º Sgt Mnt Com **SANDER** PES PIVETTA, Da sub unidade alfa, apresentou-se no dia 01 Jan 01, por término de férias relativas ao 1° semestre do corrente ano, estando pronta para o serviço.

4. **DESIGNAÇÃO DE MILITAR PARA REALIZAR REFEIÇÕES NO RANCHO DOS CABOS E SOLDADOS**
a. Determino que o 3º Sgt Mnt Com **SANDER** PES PIVETTA, Da SU alfa, realize as refeições do café da manhã e do almoço, no dia 02 Jan 01, no rancho dos cabos e soldados da Organização Militar.

4ª PARTE – JUSTIÇA E DISCIPLINA

5. **DISPENSA DO SERVIÇO COMO RECOMPENSA**
a. Concedo ao 1° Ten **FULANO** BELTRANDO DA SILVA, da sub unidade alfa, 02 (dois) dias de dispensa total do serviço como recompensa, a contar do dia 04 Jan 11, de acordo com o inciso Nr do Art. Nr do REGULAMENTO. Apresentação em 06 Jan 01.

NOME DO COMANDANTE OM – POSTO/GRADUAÇÃO
Cmt Organização Militar

Na Figura 2 é possível verificar uma representação simplificada de um BI, visto que um documento original possui em média de cinco folhas, sendo inviável sua apresentação. Nela é possível verificar como estão dispostas as suas seções e exemplos de informações que cada uma contém.

Também é possível verificar que existem duas informações distintas destacadas, uma com a cor amarela ("apresentação de militar") e a outra com a cor azul ("designação de militar para realizar refeições no rancho dos cabos e soldados"), as quais serão utilizadas na próxima seção, onde é explicada a forma como as Folhas de Alterações são produzidas.

Ressalta-se que as informações contidas neste exemplo são meramente figurativas, devido a questões de sigilo e por possuírem informações pessoais de interesse somente interno à organização, fato que impossibilita a publicação de informações reais.

3.2 Folhas de Alterações

Outro documento produzido pelo Exército são as Folhas de Alterações, documento produzido semestralmente que tem a finalidade de registrar os fatos mais significativos ocorridos com um militar durante este período. Estas informações são extraídas dos BIs e cada exemplar é referente a um único militar (EXERCITO, 2002).

No documento Exercito (2002) são encontradas informações que normalizam a estrutura das Folhas de Alterações. Elas são compostas por um cabeçalho e um corpo. O cabeçalho possui informações referentes à OM que o militar está vinculado, o semestre de referência e os dados de identificação do militar. Já o corpo é dividido em duas seções, a **Primeira Parte** e a **Segunda Parte**.

Primeira Parte : Esta seção apresenta informações extraídas dos BIs sobre a vida funcional do militar ocorridas durante o semestre. Ela compreende diversos assuntos, como por exemplo, funções exercidas, referências elogiosas, punições, resultados de Teste de Aptidão Física (TAF) e Tiro de Ação Tática (TAT), movimentações, promoções, entre outros. Estas referências são expostas em ordem cronológica dos acontecimentos sendo separados pelo mês de sua ocorrência. As Folhas de Alterações não contemplam todos os assuntos presentes nos BIs, sendo necessário realizar uma seleção dos assuntos afim de decidir quais informações serão utilizadas em sua confecção (EXERCITO, 2002).

Segunda Parte : Nesta seção consta o tempo de serviço do militar, como por exemplo, o tempo de serviço realizado com o transcorrer do semestre, o tempo total de serviço computado para a aposentadoria, tempo de serviço computado para medalha militar, entre outros (EXERCITO, 2002).

Também em Exercito (2002) são encontradas as normas que regulamentam quais informações são relevantes para a sua elaboração. Através delas percebe-se que nem todas as informações referentes a um militar são importantes para a composição das suas Folhas de Alterações. Assim, existe a necessidade de realizar uma busca sobre todos os BIs e selecionar quais informações são relevantes.

Figura 3 – Exemplo de uma Folha de Alterações

MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
Comando Militar de Área – Região Militar
Organização Militar – Codon OM — Guarnição da Cidade – Estado
FOLHA Nº 01

NOME: Sander Pes Pivetta
POSTO OU GRADUAÇÃO: 3º Sgt
ARMA (SERVIÇO OU QUALIFICAÇÃO): Manutenção de Comunicações
IDENTIDADE: 01234567-89 — 1º SEMESTRE DE 2001
CP: 010203-012 — PERÍODO DE 01 JAN À 30 JUN

1ª PARTE:

JANEIRO:

APRESENTAÇÃO DE MILITARES

- a 01, BI Nr 001- Apresentou-se no dia 01 Jan 01, por término de férias relativas ao 1° semestre do corrente ano, estando pronta para o serviço.

FEVEREIRO:

MARÇO:

ABRIL:

MAIO:

DISPENSA DO SERVIÇO COMO RECOMPENSA

- a 22, BI Nr 099- Foi concedido 02 (dois) dias de dispensa total do serviço como recompensa, a contar de 01 Mai 01 de acordo com o inciso Nr do Art. Nr do Regulamento. Apresentação em 06 Mai 01.

JUNHO:

2ª PARTE:

1. TEMPO COMPUTADO DE EFETIVO SERVIÇO (TC):.....00a 06m 01d
   a. Arregimentado.....00a 06m 01d
      - de 01 Jan a 30 Jun.....Pronto na OM

2. TEMPO NÃO COMPUTADO (TNC).....00a 00m 00d

3 . TEMPO DE SERVIÇO COMPUTAVEL PARA MEDALHA MILITAR
   - até 30 Jun 01 (TSCMM).....00a 06m 01d

4. TEMPO DE SERVIÇO NACIONAL RELEVANTE (TSNR).....00a 00m 00d

5. TEMPO TOTAL EFETIVO SERVIÇO (TTES).....01a 06m 01d

Cidade – Estado, 21 de Junho de 2001.

NOME DO COMANDANTE SUB UNIDADE ALFA – POSTO/GRADUAÇÃO
Cmt SU Alfa

Na Figura 3 é possível observar um exemplo de Folha de Alterações, que devido aos motivos apresentados na seção anterior, possui informações figurativas. É possível ver exemplos de informações que devem estar em sua composição.

Analisando as Figuras 2 e 3, consegue-se ter uma noção de como o processo de seleção de informações ocorre nos BIs. Na Figura 2 percebe-se a existência de duas informações destacadas que referenciam o 3º Sargento (Sgt) Sander Pes Pivetta, e na Figura 3 apenas uma destas está presente. A informação destacada com a cor amarela, "apresentação de militar", é encontrada nos dois documentos, pois é considerada importante para compor as Folhas de Alterações. Já a destacada com a cor azul, "designação de militar para realizar refeições no rancho dos cabos e soldados", representa uma atividade cotidiana desenvolvida pelo militar, não sendo importante para a confecção das Folhas de Alterações deste militar.

Isto mostra que nem todas as informações contidas nos BIs são relevantes para as Folhas de Alterações, sendo necessário a realização de uma busca sobre todos os BIs produzidos durante o semestre, por informações referentes ao militar. Ao ser encontrada, esta é analisada para identificar se deve ser usada na produção das Folhas de Alterações.

4 Método de Classificação Proposto

O principal objetivo deste trabalho é desenvolver uma aplicação capaz de classificar os BIs como relevantes para a produção das Folhas de Alterações. Para isto, é utilizado o aprendizado de máquina, uma área que possibilita a realização de uma atividade de maneira independente por uma aplicação. Dentre os tipos de aprendizado de máquina existentes, optou-se pelo aprendizado supervisionado, visto o bom número de Folhas de Alterações existentes que podem ser empregadas no treinamento e na validação da aplicação. Associado a ele, e visando um melhor desempenho, são utilizadas técnicas de seleção de sentenças e de processamento de palavras.

Dentre os algoritmos de aprendizado supervisionado utilizando na classificação textual, foram analisados os algoritmos *Naive Bayes* e SVM. Estes algoritmos foram propostos durante o Projeto de Trabalho de Conclusão de Curso (TCC), onde seria realizada uma comparação entre eles. Para iniciar a comparação entre as técnicas foi escolhido o algoritmo *Naive Bayes*, devido a estrutura simples, quando comparado aos outros algoritmos, e pela sua maior facilidade em ser programado.

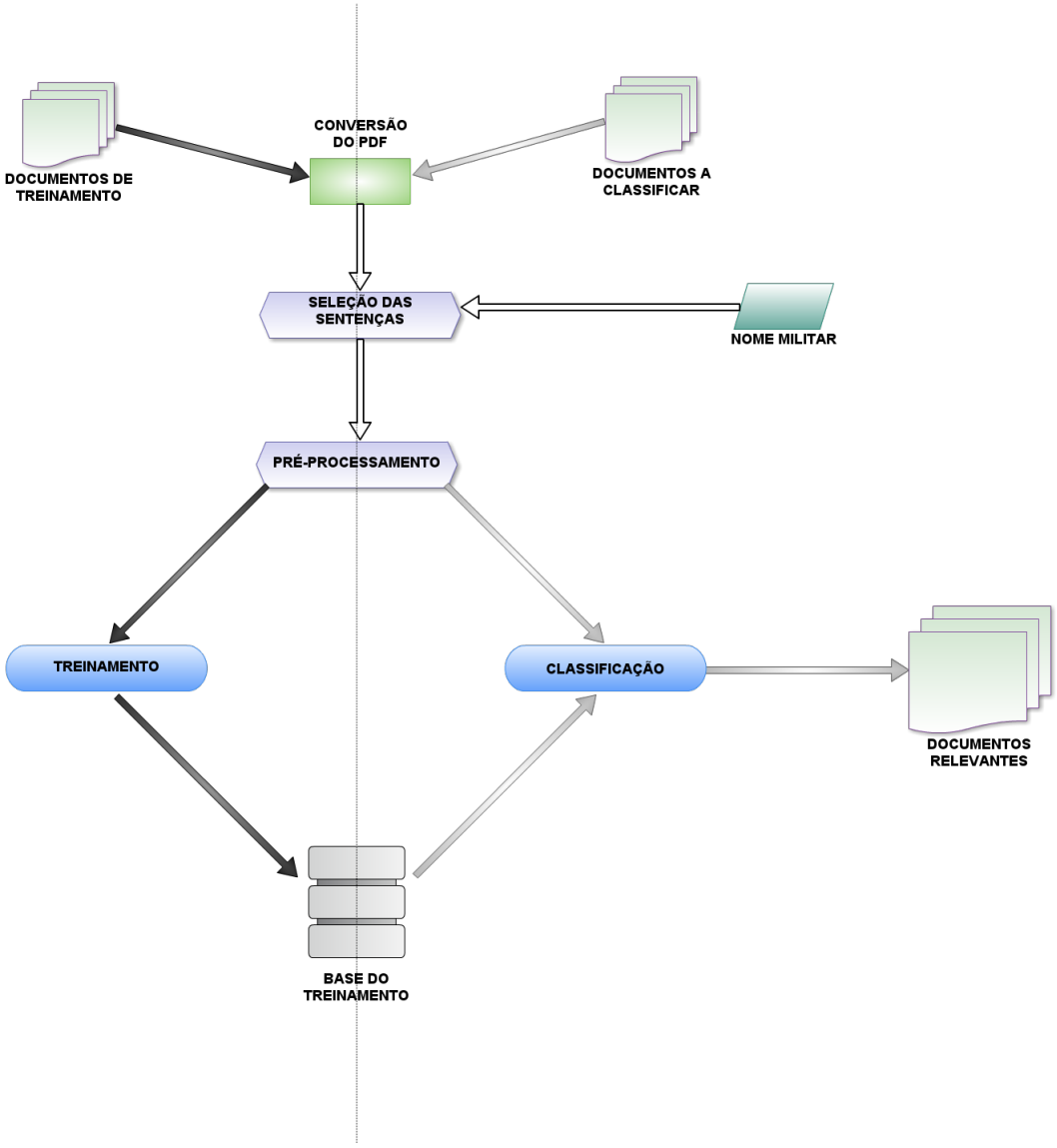
Com a implementação do classificador, verificou-se a possibilidade realizar uma associação de outras técnicas a este algoritmo com o objetivo de melhorar o seu desempenho. Desta forma, comparar o desempenho do classificador *bayesiano* com o das suas variações proporciona um maior ganho científico, visto que mostra formas de extrair o maior potencial deste algoritmo.

4.1 Organização do Classificador *Bayseano*

Para que uma aplicação consiga realizar uma atividade de forma independente, deve existir uma base de conhecimento, sobre a qual são extraídas informações para a análise das novas suposições. Para isto, o algoritmo de *Naive Bayes* divide-se em duas fases: a fase de **treinamento** e a fase de **classificação**.

Durante a fase de **treinamento**, os parâmetros são estimados, ou seja, os documentos de treinamento são analisados e deles extraídas todos os eventos. No caso em questão os eventos se referem às palavras existentes nos documentos. Para cada palavra do vocabulário pode-se usar como parâmetro a sua frequência nos documentos ou a sua incidência. Já na fase de **classificação** estes valores obtidos são usados afim de retornar se um BI é relevante para um determinado militar.

Figura 4 – Fases Desenvolvidas pelo Classificador Textual



A execução dessas atividades está representada pela Figura 4. Através dela é possível verificar como ocorrem as duas fases do classificador ocorrem, onde as setas brancas apresentam o caminho percorrido por ambas as fases, as setas pretas o caminho percorrido durante o treinamento e as setas cinzas, o caminho percorrido durante a classificação. Em uma maneira mais simples, a Figura apresenta em seu lado direito a sequência realizada pelo treinamento, já analisando o lado esquerdo é observada a sequência seguida durante a classificação. Nas seções que seguem, as partes que compõem esta arquitetura são detalhadas.

4.1.1 Documentos de Treinamento

A utilização do aprendizado supervisionado necessita da existência de documentos pré-classificados. Nesta aplicação, os documentos de treinamento são compostos por

Folhas de Alterações de militares, as quais foram confeccionadas durante semestres anteriores e pelos BIs que deram origem a elas. Sintetizando, esta base de documentos é utilizada para descobrir a probabilidade das evidências (palavras ou termos) estarem associadas a uma das classes de interesse (relevantes e não relevantes). Estas bases são abordadas na seção 4.1.7.

A montagem da efetiva base de treinamento é realizada através das seleção de todos os BIs de um semestre, onde o nome de um militar específico apareça. A consulta pelo nome do militar é feita através do seu 'nome completo' (e.g., Sander Pes Pivetta) ou da sua 'graduação mais o nome de guerra' (e.g., 3º Sgt Sander). Para cada um dos BIs selecionados, é verificado se ele está contido na Folha de Alterações confeccionada para o militar em questão. Esta etapa é realizada verificando se o número dos BIs selecionados estão descritos dentro do documento de Folha de Alterações. Para os BIs que forem empregados na confecção da Folha de Alterações, é feita uma seleção das sentenças que estejam relacionadas ao militar. Estas sentenças são atribuídas à classe dos relevantes. A seleção de sentenças também é feita sobre os BIs que não foram utilizando na confecção das Folhas de Alterações. Nesse caso as sentenças selecionadas são atribuídas à classe dos não relevantes. Na seção 4.2 encontra-se a explicação de como é realizada a seleção de sentenças.

4.1.2 Documentos a Classificar

O outro conjunto de documentos é composto pelos BIs não treinados, ou seja, àqueles que se deseja classificar automaticamente. Eles são agrupados pelo semestre que foram gerados. Com essas informações, dado o nome de um militar e o semestre, é realizada uma busca dentro desse conjunto de documentos para localizar os que serão enviados para a etapa de classificação. Assim como ocorre nas seleção dos documentos para treinamento, a seleção dos documentos a classificar é feita através de uma busca pelo 'nome completo' do militar ou pela sua 'graduação mais o nome de guerra'.

4.1.3 Conversão do *Portable Document Format*

Para realizar a classificação de documentos, é necessário que o conteúdo deles possa ser entendido pela aplicação. Como as duas bases de documentos encontram-se no formato *Portable Document Format* (PDF), há uma necessidade em transformá-los em um formato textual que possibilite a manipulação de suas informações.

Para isto é utilizada uma biblioteca *open source* para java chamada **PDFBox**[1], que possibilita a criação de manipulação de arquivos neste formato.

[1] pdfbox.apache.org

4.1.4 Seleção de Sentenças

Conforme explicado na seção 4.1.1, é necessário utilizar um mecanismo que selecione as sentenças a serem rotuladas como relevantes ou não relevantes. Este mesmo mecanismo também é utilizado durante a classificação para selecionar as sentenças representativas de um militar que, de fato, deverão ser classificadas. Ele é explicado em maiores detalhes na seção 4.2.

4.1.5 Pré-processamento

Devido a língua portuguesa possuir uma grande variação morfológica, existe uma grande quantidade de palavras que apresentam sentidos semelhantes entre si. Isso acarreta no problema da esparsidade dos dados, aumentando o número de eventos a serem analisadas e dificultando o processo de classificação. Na tentativa por diminuir a gama de palavras que irá compor a base de conhecimento, são utilizadas técnicas de *stemming* e remoção de *stop words* (REZENDE, 2005).

Para a criação da *stoplist* foram utilizados os pronomes, artigos, numerais, preposições, conforme pode ser encontrado em Rezende (2005). Adicionadas a estes conjuntos de palavras, também foram incluídas palavras do contexto militar que não alteram o sentido do texto, como Sgt, Capitão (Cap), Tenente (Ten), Diex, dentre outras.

Após a remoção das *stop words*, é realizada uma análise do restante das palavras, sendo retiradas todas aquelas que possuírem menos de três caracteres. Esta medida foi adotada uma vez que a maioria das palavras com tamanho inferior a este número de caracteres não atribuírem nenhuma característica especial ao texto. A remoção de palavras com um número superior de caracteres se tornou inviável, pois removem palavras relevantes e os resultados foram piores.

Mesmo com a remoção dos *stop words*, ainda há a necessidade de reduzir as outras palavras. O objetivo é diminuir ainda mais o número de palavras que irá compor as classes da base de treinamento. Para isto, são reduzidas as palavras ao mesmo radical. Este processo é chamado de *stemming* (REZENDE, 2005). Neste trabalho foram removidos os sufixos de pluralidade e os artigos de gênero, como por exemplo a palavra **interessados**, a qual gera a palavra *interessad*, transformando as sentenças em um formato mais apropriado para o algoritmo realizar a sua classificação.

4.1.6 Treinamento

Nessa etapa, todas as sentenças estão selecionadas e processadas pelas etapas anteriores, sendo possível realizar o treinamento da aplicação. Ao serem selecionados, os eventos são atribuídos às classes analisadas (relevantes e não relevantes), gerando uma tabela que contém o evento e a sua frequência ou incidência nas sentenças de treinamento.

Esses eventos são representados por *n-gramas* de palavras, onde n deve ser superior ou igual a 1. O uso dos *n-gramas* é uma variação do classificador *bayesiano*, onde o objetivo do seu uso é tentar analisar o contexto em que as palavras estão inseridas nas sentenças.

4.1.7 Base do Treinamento

Ao ser realizada uma atribuição de um evento a uma das classes, é requisitada uma atualização da tabela. As tabelas são as responsáveis pelo "conhecimento" da aplicação, ou seja, elas possuem as informações necessárias para que a aplicação realize sua função afim. Estas tabelas são o resultado obtido pela fase de treinamento, e podem ser produzidas por duas formas distintas:

Analisando a incidência dos eventos nos documentos utilizados durante o treinamento. Neste caso, a tabela é composta apenas pelos eventos encontrados nos documentos de treinamento, onde, sempre que for requisitada uma atualização, é feita uma busca pelo evento analisando na tabela, com o objetivo de verificar se ele já está nela presente. Caso não seja encontrado o evento é inserido.

Analisando a frequência que estes eventos ocorrem nos documentos de treinamento, sendo a tabela composta pelos eventos e por sua frequência. No instante que for requisitada um atualização da tabela, é verificado o evento analisado esta presente nela. Caso esteja, é realizada uma atualização de sua frequência, porém em caso contrário, este evento é inserido com a frequência igual a 1.

Estes eventos são produzidos com o emprego de *n-gramas* ou, conjunto de n palavras. Desta forma os eventos podem ser compostos por mais de uma palavra, $n > 1$, havendo nestes casos a necessidade de realizar uma inserção recursiva nas tabelas. Esta inserção é realizada inicialmente com os eventos compostos por n palavras. Após, estes são separados em "sub-eventos" compostos por $n - 1$ palavras, sendo inseridos na tabela, repetindo esta operação até o instante que eles sejam compostos por uma única palavra.

4.1.8 Classificação

A outra fase do algoritmo de *Naive Bayes* é a de **classificação**, onde são realizados os cálculos probabilísticos de uma sentença pertencer a cada uma das classes. Para encontrar este valor é utilizado o *Teorema de Bayes*, onde a probabilidade é calculada com base nas evidências coletadas durante a fase de treinamento.

Para evitar inconsistências nos resultados, é necessário realizar uma suavização destes, ou seja, evitar que a probabilidade de ocorrência de um evento seja nula, ocasionado no retorno de uma probabilidade nula também na análise da sentença. Para isto é utilizada a suavização de *Laplace*, a qual consiste em adicionar 1 ao número da frequência ou da incidência de todos os evento (NEY; MARTIN; WESSEL, 1997; JUAN; NEY, 2002).

Nos casos em que os eventos são compostos por *n-gramas*, com $n > 1$, é usada a recursividade de *back-off*[2] (KNESER; NEY, 1995; DUPONT; BARBE, 2006). Caso a consulta não encontre este evento, mesmo com o emprego de uni-gramas, é aplicada a suavização de *Laplace*.

Ao final da classificação, se em pelo menos uma das sentenças selecionadas a probabilidade *bayesiana* do evento ser relevante for superior ao de ser não relevante, o documento é considerado relevante para o militar em questão.

4.1.9 Documentos Relevantes

O objetivo deste classificador é encontrar os BIs relevantes para compor as Folhas de Alterações dos militares. Para isto, são desenvolvidas duas fases, primeiro é realizado o treinamento e, depois a classificação. Após realizada a classificação dos BIs, esta etapa é a responsável por agrupar os documentos relevantes retornados.

4.2 Seleção de Sentenças Propostas

Uma das atividades mais importantes na execução do método de classificação, tanto na fase de treinamento como na fase de classificação, é a seleção dos trechos dos documentos referentes ao militar. Uma seleção inapropriada das informações acarreta na utilização de informações errôneas para o cálculo da probabilidade *bayesiana*, provocando uma classificação equivocada dos BIs.

Para mostrar esta necessidade foi construída uma técnica que considera o documento inteiro como uma sentença. Aos ser encontrado o "nome completo" ou a "graduação mais o nome de guerra" do militar (pivô), todas as informações contidas no documento são tratadas como uma única informação e consideradas referentes exclusivamente a este militar. O desempenho obtido é encontrado na seção 5.1, mostrando a ineficácia desta técnica.

Partindo para técnicas mais aprimoradas, foram propostas outras duas técnicas para selecionar as sentenças, a **Janela Fixa** e a **Janela Deslizante**. Ambas partem do ponto no texto onde o pivô é encontrado, sendo que cada técnica utiliza regras distintas para selecionar o texto que se refere ao militar.

Na **Janela Fixa** são consideradas importantes todas as informações que encontram-se próximas ao pivô. Assim, a partir do índice do pivô, são selecionados os k caracteres anteriores e posteriores a este índice.

[2] A recursividade de *back-off* consiste em realizar uma busca recursiva dos eventos na base de treinamento. Caso o *n-grama* não seja encontrado é realizada uma busca com "sub-eventos", compostos por *(n-1)-grama* (KNESER; NEY, 1995; DUPONT; BARBE, 2006)

A técnica da **Janela Deslizante** pressupõe que a informação possa estar afastada do pivô, ou mesmo que esta não possua um tamanho fixo de caracteres. Isto ocorre, por exemplo, quando um nome é encontrado em uma lista ou em uma tabela.

Nesta seleção, o primeiro passo é analisar se o pivô se encontra em um trecho válido. Um trecho possui o texto compreendido entre dois símbolos de término de parágrafo. Para que ele seja considerado válido, deve possuir um número superior a λ palavras válidas. Para que a palavra seja considerada válida, ela não deve apresentar menos de μ caracteres e não estar contida na *stoplist*.

Caso este trecho selecionado possua mais do que λ, ela é efetivamente selecionada e transformada em uma sentença a ser analisada pelo classificador. Caso contrário, este trecho é descartado e verificações são realizadas sobre os parágrafos anteriores, até que a condição de validação do trecho seja satisfeita e este texto efetivamente selecionado para compor a sentença.

Figura 5 – 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 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 6 – Seleção de texto por Janela Deslizante 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 retorno estão previstos para os dias 13 Jul 11 e 15 Jul 11, respectivamente (solução à nota Nr 040-STA, de 06 Jul 11);

Para encontrar o número de parâmetros ideal para λ e μ, foram realizados testes afim de obter a melhor categorização das sentenças. Nos exemplos a seguir, é verificada a forma como as técnicas de seleção de sentenças por Janela Fixa e Janela Deslizante se comportam.

Exemplificando as duas técnicas de seleção, considere as Figuras 5 e 6, que selecionam o texto usando “Sander Pes Pivetta” como pivô. Na Figura 5 foi utilizada a Janela Fixa com $\kappa = 150$, ou seja, foram selecionados os 300 caracteres mais próximos ao pivô. Na Figura 6 foi utilizada a Janela Deslizante com $\lambda = 6$ e $\mu = 3$.

As Figuras 7 e 8 mostram um exemplo no qual o nome do militar está em uma tabela juntamente com o nome de outras pessoas. Nesse caso, a informação referentes a esses militares encontra-se no parágrafo que aparece antes da tabela. Na Figura 7 foi utilizada a seleção por Janela Fixa com $\kappa = 150$. Como pode ser verificado, a técnica seleciona praticamente todas as informações contidas na tabela e ignora a sentença anterior a ela, que possui a informação realmente pertinente. Já na Figura 8 foi utilizada a Janela Deslizante com $\lambda = 6$ e $\mu = 3$. Observa-se que, nesse caso, como o trecho onde

Figura 7 – 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 8 – Seleção de texto por Janela Deslizante 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

ocorre o nome do militar não é considerada válida, a janela de texto desliza para cima até o encontro de um trecho válido.

5 Resultados Obtidos

Este capítulo tem o objetivo de avaliar a classificação automática dos BIs como relevantes para a confecção das Folhas de Alterações. Os resultados são obtidos com a execução do classificador *Naive Bayes* associado aos métodos de seleção de sentenças "Janela Fixa" e "Janela Deslizante". E para efeitos de comparação, também são apresentados os resultados obtidos com a execução de dois métodos básicos chamados de "Pesquisa Nominal" e "Documentos Inteiro". Além destas técnicas, será apresentado o desempenho das variações do classificador *bayesiano* usando *n-gramas* como forma de seleção dos eventos, sendo que será analisada tanto a frequência como a incidência dos eventos.

Para avaliar o desempenho dos experimentos, são usadas as métricas de **Precisão**, **Cobertura** e medida-f. A precisão é calculada em função do número de BIs classificados como relevantes e que realmente são. Já a cobertura é calculada em função do número de BIs relevantes que assim foram classificados. A medida-f apresenta a média harmônica das duas métricas acima. Obter um valor de medida-f próximo a zero indicam que tanto a precisão quanto a cobertura foram pobres, enquanto valores próximos a 1 indica que tanto a precisão quanto a cobertura obtiveram resultados satisfatórios.

Durante a etapa de treinamento foram realizados testes com quantidades variadas de BIs, os quais foram confeccionados durante o período de um ano, utilizando o quantitativo de 64 militares selecionados aleatoriamente. A marcação dos documentos em "relevantes" e "não relevantes" foi realizada automaticamente, com o auxílio das Folhas de Alterações desses militares, conforme descrito na seção 3.2.

5.1 Resultados do Classificador *Bayesiano* associado as Técnica de Seleção de Sentenças

Para encontrar um parâmetro básico de avaliação, sobre o qual se calculado o desempenho obtido pelos classificadores, foi realizada uma aplicação simples de classificação que não envolve o algoritmo *bayesiano* e nem as técnicas de seleção de sentenças, chamado de **Pesquisa Nominal**. Nesta técnica, para um militar, são considerados relevantes todos os BIs onde for encontrada alguma referência ao pivô, sem a análise do seu conteúdo.

O outro parâmetro básico é o método do **Documento Inteiro**. Diferente do anterior, ele utiliza o classificador *bayesiano* para verificar se o BI que referencia o militar é relevante, mas não utiliza as técnicas de seleção de sentenças. Para cada documento

utilizado para gerar a Folha de Alterações de um determinado militar, todas as suas palavras são consideradas como eventos e são associadas à classe de documentos relevantes para esse militar. Esta técnica tem o objetivo de mostrar o comportamento do classificador *bayesiano* quando são usadas informações incoerentes e justificar a necessidade de selecionar corretamente as sentenças.

Partindo para o emprego das técnicas de seleção de sentenças, foram desenvolvidos dois métodos: **Janela Fixa** e **Janela Deslizante**.

O método da **Janela Fixa** realiza o treinamento dos BIs utilizando a técnica de seleção de sentenças por Janela Fixa, com $\kappa = 150$. Assim, caso um documento contenha informações referentes a um militar, apenas as palavras próximas ao pivô, são selecionado por ela e, consideradas eventos associados à uma das duas classe do classificador.

De maneira similar, 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 possua referência ao pivô, apenas as palavras que pertencem ao trecho selecionado por essa técnica são consideradas eventos associados a uma das classes do classificador.

Para encontrar os melhores valores de κ, λ e μ, juntamente com as suas técnicas de seleção de sentenças, foram realizados testes variando o valor destes parâmetros. Foram executados testes com κ variando entre 100 e 250, λ variando entre 2 e 10 e μ variando entre 2 e 5. Os melhores resultados foram obtidos com $\kappa = 150$, $\lambda = 6$ e $\mu = 3$.

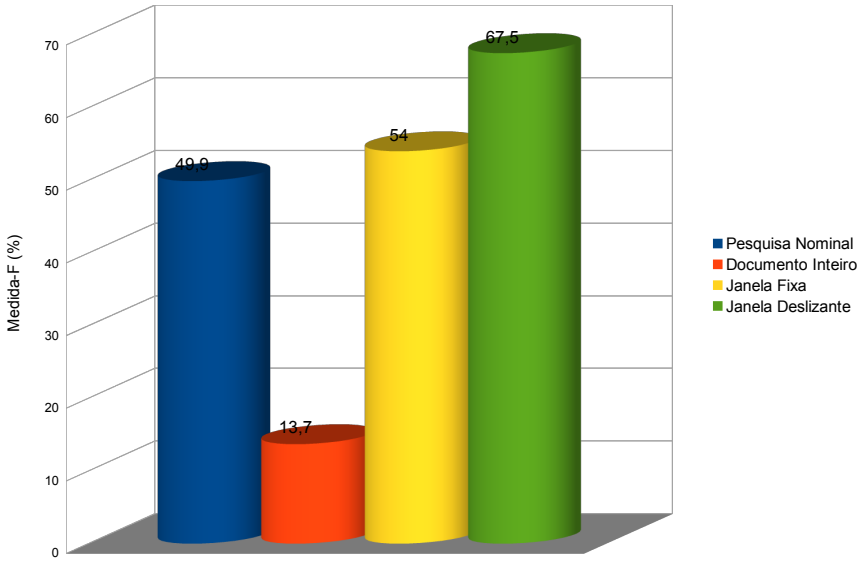
A geração da base de treinamento, depende do método de classificação empregado. No método Documento Inteiro, a lista de documentos relevantes e não relevantes é composta por todas as palavras existentes no BIs. Por exemplo, se um documento for usado na confecção de uma Folha de Alterações de um militar, todo este documento é incorporado à base de documentos relevantes. Já nos métodos de Janela Fixa e Janela Deslizante, a lista de relevantes e não relevantes é composta pelos trechos dos arquivos referentes aos militares. Para todo documento onde for encontrada uma referência ao pivô, é utilizada um método de seleção de sentenças e, conforme explicado anteriormente, é incorporado a classe da base de treinamento mais adequada.

Após o treinamento do classificador e obtido o conhecimento necessário para a categorização das sentenças, é executada a classificação dos novos documentos. Estes arquivos são constituídos por um novo conjunto de documentos, diferente dos utilizados na fase de treinamento. Nesta etapa, foram realizados testes sobre um conjunto de 228 BIs e para 49 militares selecionados aleatoriamente.

No gráfico da Figura 9 é possível observar o resultado obtido entre os métodos de classificação propostos.

Observando os resultados, percebe-se que a técnica que utiliza o Documento Inteiro

Figura 9 – Resultados do classificador *bayesiano* com a seleção de sentenças



apresenta o pior desempenho, muito inferior aos demais. Este fraco desempenho deve-se ao motivo dos BIs serem compostos por varias informações, tanto relevantes como não relevantes, além de possuírem referências a outros militares. No instante que todo o documento é considerado uma sentença, são usadas informações relevantes e não relevantes ao mesmo tempo. Isto provoca a formação de classes inconsistentes e ao final, em uma classificação errônea. Assim, fica evidenciada a necessidade de realizar uma melhor seleção destas sentenças.

Com a execução da “Pesquisa Nominal", a medida-f ficou próxima a 50%. Apesar de encontrar todos os documentos relevantes, foram retornados muitos BIs não importantes para a confecção das Folhas de Alterações do militar, ocasionando em uma precisão média de 33%.

Com a execução dos métodos Janela Fixa e Janela Deslizante, foram obtidos os melhores resultados. Dentre estes, o método que realiza a seleção de sentenças através da Janela Deslizante conseguiu realizar uma melhor seleção dos trechos significativos para os militares e realizar uma melhor classificação. O resultado reflete o fato de que, em muitos casos, as informações relevantes encontram-se distantes de seu nome, como por exemplo, dentro de listas ou de uma tabela de nomes. Nestes casos, a seleção com Janela Deslizante consegue percorrer o texto até encontrar a informação mais provável de ser relevante.

5.2 Uso de ocorrência dos eventos e *n-gramas*

Na tentativa de encontrar melhores maneiras de realizar a classificação dos documentos do EB, foram desenvolvidas variações do algoritmo de *Naive Bayes*. Uma delas

propõe o uso da incidência dos eventos, ou seja, somente é analisada a presença ou não no conjunto de treinamento.

A outra variação está relacionada a maneira como são selecionados os eventos. O classificador *bayesiano* utiliza palavras isoladas para compor cada evento, não analisando o seu sentido dentro do texto. Na tentativa de analisar o sentido destas palavras, foram utilizados bigramas e trigramas para selecionar os eventos.

O desempenho obtido por estas variações do algoritmo *Naive bayes* estão apresentadas nas tabelas abaixo. A Tabela 4 apresenta os resultados quando é utilizada a incidência dos eventos. Na Tabela 5, são apresentados os resultados quando é utilizada a frequência dos eventos.

Tabela 4 – Resultados utilizando a frequência dos eventos.

Método	n-grama	Precisão (%)	Cobertura (%)	Medida-F (%)
Pesquisa Nominal	-	33	100	49,6
Janela Fixa	1	62	46,3	53
Janela Deslizante	1	89	54,4	67,5
Janela Fixa	2	67	48	56
Janela Deslizante	2	90	69,6	**78,5**
Janela Fixa	3	63	47	54
Janela Deslizante	3	86	63,7	72,2

Tabela 5 – Resultados utilizando a incidência dos eventos.

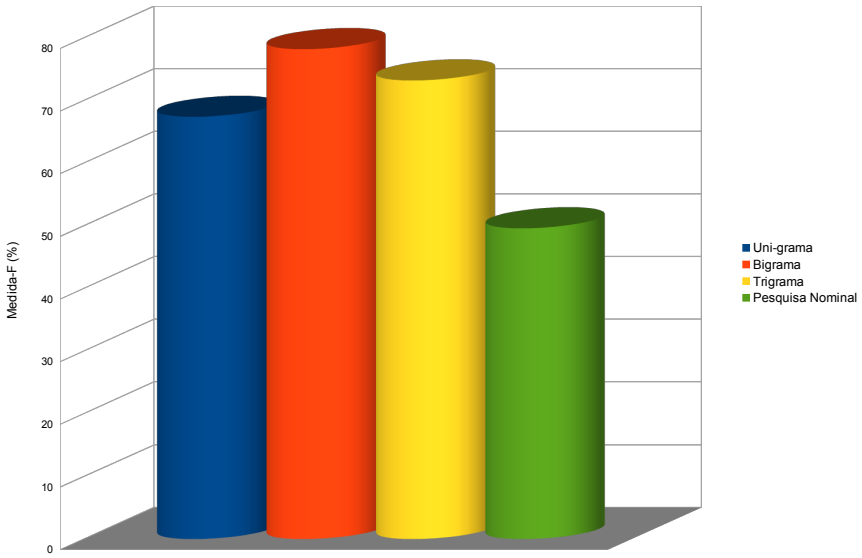
Método	n-grama	Precisão (%)	Cobertura (%)	Medida-F (%)
Pesquisa Nominal	-	33	100	49,6
Janela Fixa	1	61	42,5	50
Janela Deslizante	1	85	46	59,6
Janela Fixa	2	57	46	51
Janela Deslizante	2	88	52	**65**
Janela Fixa	3	56	38	45
Janela Deslizante	3	80	46,5	58,8

Comparando as duas tabelas, verifica-se que o uso da frequência dos eventos apresenta um desempenho superior ao obtido com a incidência das palavras, tanto com a Janela Fixa como com a Janela Deslizante. Este resultado mostra que palavras encontradas mais vezes, tem uma maior relevância para a classe e deve possuir um maior peso no cálculo da probabilidade.

Analisando a forma como os eventos foram selecionados, percebe-se que o uso de *n-gramas* consegue realizar uma melhor classificação dos BIs, tornando oportuno o seu uso. Como muitas palavras possuem sentidos distintos quando próxima as outras, a utilização mais de uma palavra para formar os eventos consegue dar uma maior importância para aquelas que ocorrem juntas. Analisando o gráfico da Figura 10, percebe-se que dentre os *n-gramas* analisados, os **bigramas** realizam uma melhor classificação dos documentos,

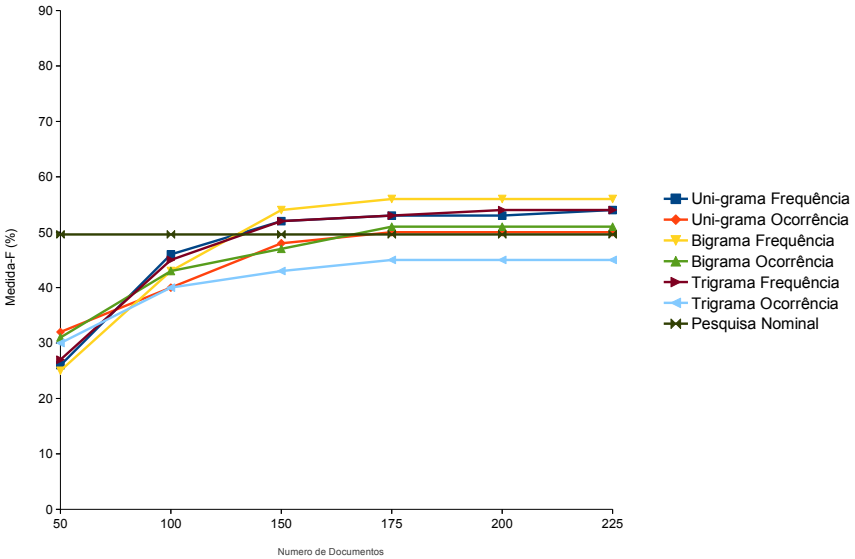
sendo superior aos trigramas. Este resultado se deve as sentenças serem compostas por poucas palavras, tornando mais raras as vezes em que trigramas iguais são encontrados.

Figura 10 – Desempenho do classificador utilizando Janela Deslizante e a Frequência dos termos.



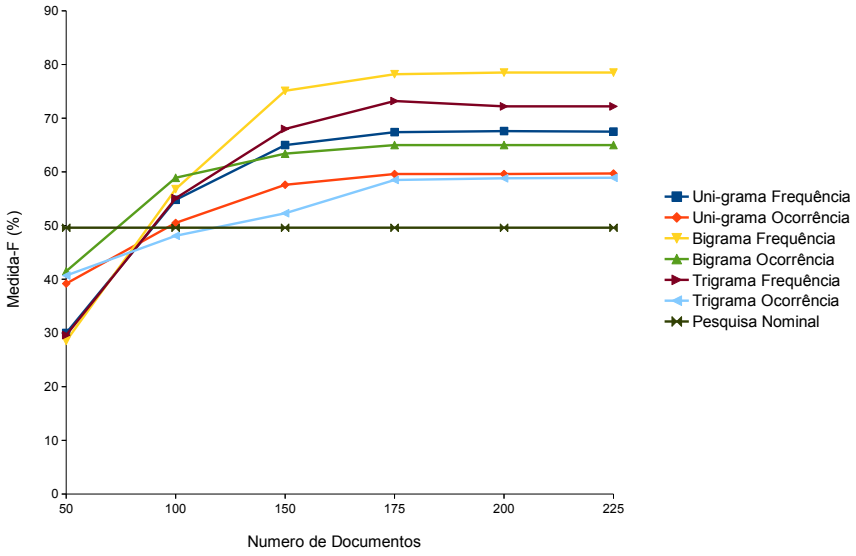
5.3 Número de documentos usados no treinamento

Figura 11 – Variação dos resultados utilizando seleção por Janela Fixa.



Apesar de modelos *bayesianos* no geral obterem bons resultados com relativamente poucos dados de treinamento, ainda assim é preciso atentar para que hajam evidências suficientes para estimar os parâmetros do modelo. Para verificar como o tamanho da base

Figura 12 – Variação dos resultados utilizando seleção por Janela Deslizante.



de treinamento afeta a classificação no problema em questão, foram feitos experimentos variando-se o número de documentos utilizados no treinamento. As Figuras 11 e 12 mostram os resultados obtidos com o uso dos métodos de Janela Fixa e Janela Deslizante. Também é possível realizar uma comparação com o método de Pesquisa Nominal.

Analisando os gráficos, é possível verificar que a classificação apresenta resultados baixos, quando o treinamento da aplicação usa um número pequeno de documentos. Conforme a quantidade de documentos 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-se praticamente constante a partir deste valor.

Na combinação de todas as técnicas ao classificador *bayseano*, os resultados se apresentam 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 **78,5%**, sendo seguido pelo uso de trigramas, que obteve uma medida igual a **72,2%**.

6 Conclusão

Este trabalho apresentou um classificador de documentos que emprega o algoritmo de *Naive Bayes*. O objetivo deste trabalho foi desenvolver uma aplicação que classificasse de forma automática os BIs em relevantes para comporem as Folhas de Alterações de militares do EB. Na busca pelos melhores resultados foram realizados testes com variações do algoritmo, utilizando a frequência e a incidência das palavras e o uso de uni-gramas, bigramas e trigramas como forma de compor os eventos. Além disso foram propostas duas técnicas de seleção de sentenças, a Janela Fixa e a Janela Deslizante, como forma de realizar a seleção das informações a serem categorizadas pelo algoritmo. Estas técnicas tem a função de escolher a melhor base de conhecimento, sobre a qual são extraídas as informações necessárias para classificar os BIs, além de encontrar a correta sentença a ser analisada pela aplicação.

A validação dos resultados realizou-se através de um conjunto de BIs, os quais deveriam ser categorizados como relevantes para compor as Folhas de Alterações dos militares. A realização da classificação considerando relevante todos os documentos onde o nome do militar for encontrado não obtém bons resultados, pois são retornados muitos BIs que não são relevantes para o militar. Já a técnica que realiza o treinamento sobre o documento inteiro, os resultados mostram-se extremamente ruins, comprovam a necessidade de realizar uma correta seleção das sentenças.

Com o uso das técnicas de seleção de sentenças, a seleção com Janela Deslizante apresentou os melhores resultados. Isto porque as informações relevantes para um militar podem estar distantes do seu nome, e esta técnica consegue deslizar sobre o texto e encontrar com mais facilidade a sentença correta.

Combinada a esta técnica, foram realizados testes com variações do algoritmo de *Naive Bayes*. Na análise entre a frequência e a incidência dos eventos, percebe-se uma melhor classificação com o uso da frequência das palavras para montar a base de conhecimento. Este resultado mostra a necessidade de atribuir um peso maior para os eventos que possuem uma maior ocorrência na base de treinamento.

A outra variação testada foi o uso de *n-gramas*. Verificou-se que os bigramas conseguiram realizar uma melhor classificação das sentenças, superando os resultados obtidos pelos trigramas. Este desempenho inferior com o uso dos trigramas pode estar associado a quantidade de informações empregadas durante o treinamento do classificador. Um conjunto maior de documentos usados durante o treinamento poderia possibilitar o

encontro de uma maior quantidade de termos termos, e também em uma maior frequência destes. Mesmo assim, fica comprovado que usar mais de uma palavra para compor um evento, na tentativa de analisar o contexto que estas estão inseridas nos documentos, realiza uma classificação mais eficiente das sentenças.

Em uma maneira geral, o uso dos bigramas para formar os eventos, a frequência que estes são encontradas durante o treinamento e a seleção com Janela Deslizante apresenta a melhor classificação dos BIs. Esta combinação foi superior em aproximadamente 63% superior ao método base que emprega somente a Pesquisa Nominal.

Analisando os pontos positivos de cada método, percebe-se que a Pesquisa Nominal, apesar de realizar uma classificação incorreta de muitos BI, pode ser útil nos casos que se deseja diminuir o número de documentos a serem analisados. Como ele encontra todos os arquivos que possuem pelo menos uma referência ao militar, não é descartado nenhum documento relevante. O método de Janela Fixa torna-se ineficaz, pois mesmo que a informação esta perto do nome, em algumas sentenças são utilizadas informações referentes a outras sentenças. O método Janela Deslisante consegue alcançar bons resultados, sendo útil para os casos onde se deseja obter os resultados prontos, podendo haver a exclusão de alguma sentença relevante.

Baseado nestes resultados, conclui-se que a técnica proposta para realizar a classificação dos BIs consegue realizar uma eficiente classificação dos BIs. Como trabalhos futuros, pretende-se estender este estudo através do uso de outros algoritmos de aprendizado de máquina, como por exemplo, o uso do algoritmo de SVM. Caso este venha apresentar melhorias no desempenho, será realizada a confecção automática da Folhas de Alterações, porém em caso contrário, será retornado uma lista com todos os BIs em ordem de relevância.

Também com o intuito de proporcionar as outras OM utilizarem esta aplicação em seu favor, será disponibilizada a aplicação e seu código fonte, sendo estes no formato *Open Source*. Isto possibilitará que, outras pessoas possam reaproveitem este código na solução deste problema por um foco diferente ainda não pensado, além de utilizarem para confeccionar as Folhas de Alterações de seus militares.

Referências

ALVARES, R. V. *Investigação do Processo de Stemming na Língua Portuguesa*. Dissertação (Mestrado) — Universidade Federal Fluminense - UFF, Niteroi- RJ, Março 2005. Citado na página 37.

ALVES, A. I. M. *Modelo de representação de texto mais adequado à classificação*. Dissertação (Mestrado) — Instituto Superior de Engenharia do Porto, Porto, 2010. Citado na página 34.

BASU, A.; WATTERS, C.; SHEPHERD, M. Support vector machines for text categorization. In: *Proceedings of the 36th Hawaii International Conference on System Sciences*. Hawai: [s.n.], 2003. Citado na página 34.

BRAGA, I. A.; MONARD, M. C.; MATSUBARA, E. T. Combining unigrams and bigrams in semi-supervised text classification. In: *14th Portuguese Conference on Artificial Intelligence*. Aveiro(Portugal): [s.n.], 2009. Citado na página 31.

CARPINETO, C.; MICHINI, C.; NICOLUSSI, R. A concept lattice-based kernel for svm text classification. In: *Proceedings of the 7th International Conference on Formal Concept Analysis*. Berlin: Springer-Verlag, 2009. Citado 2 vezes nas páginas 34 e 35.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. Segunda. Nova Iorque: Wiley Interscience, 2001. Citado na página 32.

DUPONT, P.; BARBE, P. S. Noisy sequence classification with smoothed markov chains. In: *In Conférence francophone sur l?apprentissage automatique 2006, (CAp 2006*. [S.l.: s.n.], 2006. p. 187–201. Citado na página 50.

EXERCITO. *Boletim do Exército Número 02*. Brasilia, 2001. Citado na página 21.

EXERCITO. *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)*. Brasilia, 2002. 16 p. Citado 4 vezes nas páginas 21, 39, 41 e 42.

EXERCITO. *Regulamento Interno e dos Serviços Gerais (RISG)*. Brasilia, 2004. Citado 2 vezes nas páginas 39 e 40.

FERREIRA, E. *Estudo de um Algoritmo de Mineração de Dados Aplicado à Avaliação de Curvas de Consumo de Energia Elétrica*. Dissertação (Mestrado) — Universidade Federal de Itajubá., Itajubá, 2005. Citado na página 25.

GARCíA-HERNáNDEZ, R. A. et al. Text summarization by sentence extraction using unsupervised learning. In: *Proceedings of the 7th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer-Verlag, 2008. Citado na página 36.

GOLDSTEIN, J. et al. 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, 1999. Citado na página 37.

GRAU, S. et al. Dialogue act classification using a bayesian approach. In: *Proceedings of the Ninth International Conference Speech and Computer (SPECOM 2004)*. Petersburg(Russia): [s.n.], 2004. Citado na página 30.

HACHEY, B.; GROVER, C. Sentence classification experiments for legal text summarisation. In: *In Proceedings of the 17th Annual Conference on Legal Knowledge and Information Systems (Jurix)*. Amsterdam: [s.n.], 2004. Citado na página 35.

HARTMANN, T. et al. Sentiment detection with character n-grams. In: *Procedings of the 2011 International Conference on Data Mining*. Las Vegas: CSREA Press, 2011. Citado na página 31.

IKONOMAKIS, M.; KOTSIANTIS, S.; TAMPAKAS, V. Text classification using machine learning techniques. *WSEAS TRANSACTIONS on COMPUTERS*, v. 4, n. 8, p. 966–974, August 2005. Citado na página 37.

JOACHIMIS, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Lehrstuhl VII Künstliche Intelligenz*. Dortmund: [s.n.], 1998. Citado na página 34.

JORGE, M. L. R. C.; PARDO, T. A. S. Sumarização automática multidocumento: Seleção de conteúdo com base no modelo cst (cross-document structure theory). In: *Anais do XXIV Concurso de Teses e Dissertações da Sociedade Brasileira de Computação*. Natal/RN, Brazil: [s.n.], 2011. Citado na página 36.

JUAN, A.; NEY, H. Reversing and smoothing the multinomial naive bayes text classifer. In: *In Proceedings of the 2nd Int. Workshop on Pattern Recognition in Information Systems (PRIS 2002*. [S.l.: s.n.], 2002. p. 200–212. Citado na página 49.

JUNIOR, J. M. *Classificação de Páginas na Internet*. Dissertação (Mestrado) — Universidade de São Paulo., São Paulo, 2003. Citado na página 31.

KECMAN, V. *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. Londres: Massachusetts Institute of Technology, 1948. Citado 2 vezes nas páginas 31 e 33.

KHOO, A.; MAROM, Y.; ALBRECHT, D. Experiments with Sentence Classification. In: *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*. [S.l.: s.n.], 2006. Citado na página 36.

KIM, H.; HOWLAND, P.; PARK, H. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 2005. Citado na página 34.

KNESER, R.; NEY, H. Improved backing-off for m-gram language modeling. In: *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. [S.l.: s.n.], 1995. Citado na página 50.

KO, Y.; PARK, J.; SEO, J. Improving text categorization using the importance of sentences. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, 2004. Citado na página 35.

KOGA, M. L. Classificadores Bayesianos: Aplicados a análise sintática da língua portuguesa. In: *Escola Politécnica da Universidade de São Paulo*. São Paulo: [s.n.], 2011. Citado na página 29.

LEWIS, D. D. Naive (bayes) at forty: The independence assumption in information retrieval. In: *Proceedings of the 10th European Conference on Machine Learning*. Londres: Springer-Verlag, 1998. Citado na página 30.

MCCALLUM, A.; NIGAM, K. A comparison of event models for naive bayes text classification. In: *IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION*. [S.l.]: AAAI Press, 1998. Citado na página 30.

MCDONALD, D.; CHEN, H. Using sentence-selection heuristics to rank text segments in txtractor. In: *Joint Conference on Digital Libraries - JCDL*. New York: [s.n.], 2002. Citado na página 36.

MERCER, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Royal Soc. (A)*, Londres, 1909. Citado na página 33.

METZLER, D.; KANUNGO, T. Machine learned sentence selection strategies for query-biased summarization. In: *SIGIR Learning to Rank Workshop*. [S.l.: s.n.], 2008. Citado na página 36.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill Science/Engineering/Math, 1997. Citado 5 vezes nas páginas 22, 23, 24, 25 e 29.

NEY, H.; MARTIN, S. C.; WESSEL, F. Statistical language modeling using leaving one-out. In: *Corpus-Based Methods in Language and Speech Processing*. [S.l.]: Kluwer Academic Publishers, 1997. p. 174–207. Citado na página 49.

NILSSON, N. J. *Introduction to Machine Learning: An Early Draft of a Prospected Textbook*. Stanford: Robotics Laboratory Departament of Computer Science Stanford University, 1996. Citado 2 vezes nas páginas 23 e 24.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia: [s.n.], 2002. Citado na página 35.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. G. V. NeuralSumm: Uma abordagem conexionista para a sumarização automática de textos. In: *Anais do IV Encontro Nacional de Inteligência Artificial - ENIA*. Campinas-SP, Brasil: [s.n.], 2003. Citado na página 36.

PENG, F.; SCHUURMANS, D.; WANG, S. Augmenting naive bayes classifiers with statistical language models. *Inf. Retr.*, p. 317–345, 2004. Citado na página 30.

PIMENTA, E. M. C. *Abordagens para decomposição de problemas multi-classe: Os códigos de correcção de erros de saída*. Dissertação (Mestrado) — Faculdade de Ciências da Universidade do Porto, Porto, 2004. Citado na página 33.

RABELO, J. P.; FILHO, M. A.; OLIVEIRA, T. Mineração de Textos Através do Algoritmo de Classificação. In: *Instituto de Matemática. Universidade Federal da Bahia (UFBA)*. Salvador: [s.n.], 2011. Citado na página 29.

REZENDE, S. O. *Sistemas Inteligentes. Fundamentos e Aplicação*. Barueri: Editora Manole Ltda, 2005. Citado 2 vezes nas páginas 37 e 48.

RUSSELL, S. J.; NORVIG, P. *Inteligência Artificial*. Tradução da segunda. Rio de Janeiro: Massachusetts Institute of Technology, 2004. Citado 2 vezes nas páginas 24 e 30.

SCHNEIDER, K.-M. On word frequency information and negative evidence in naive bayes text classification. In: . [S.l.: s.n.], 2004. Citado na página 30.

SCHöNHOFEN, P.; BENCZúR, A. A. Feature selection based on word?sentence relation. In: *Fourth International Conference on Machine Learning and Applications*. Budapest: [s.n.], 2005. Citado na página 35.

SILIC, A. et al. N-grams and morphological normalization in text classification: A comparison on a croatian-english parallel corpus. In: *Progress in Artificial Intelligence, 13th Portuguese Conference on Aritficial Intelligence*. [S.l.]: Springer, 2007. Citado na página 34.

SILVA, C. F.; VIEIRA, R. 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: [s.n.], 2007. Citado na página 34.

SMOLA, A. J.; BERNHARD, S. *Learning With Kernels: Support Vector Machine, Regularization, Optimization and Beyond*. Londres: Massachusetts Institute of Technology, 1999. Citado na página 33.

SURESH, V. et al. A non-syntactic approach for text sentiment classification with stopwords. In: *WWW (Companion Volume)*. [S.l.]: ACM, 2011. p. 137–138. Citado na página 37.

TING, S.; IP, W.; TSANG, A. H. Is naïve bayes a good classifier for document classification? *International Journal of Software Engineering and Its Applications*, v. 5, n. 3, 2011. Citado na página 29.

TONG, S.; KOLLER, D. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2001. Citado na página 34.

VAPNIK, V. *Estimation of Dependences Based on Empirical Data*. New York: Springer Science+Business Media, 2006. Citado 2 vezes nas páginas 31 e 33.

WAJEET, M. A.; ADILAKSHMI, T. Text classification using machine learning. *Journal of Theoretical and Applied Information Technology*, p. 119–123, 2009. Citado na página 37.

WEKA. *Waikato Environment for Knowledge Analysis: Data Mining with Open Source Machine Learning in Java*. 1996. Disponível em: www.cs.waikato.ac.nz/ml/weka. Acessado em 08/02/2013. Citado na página 29.