6
Experimentos

> The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.
>
> John Von Neumann

Existem alguns critérios para se avaliar sistemas de classificação textuais (22), dentre eles estão:

- Precisão, a habilidade de prever a classe dos documentos. Isto é feito ao comparar rótulos atribuídos pelo classificador com rótulos atribuídos por um ser humano
- Velocidade e escalabilidade do treinamento
- Facilidade, velocidade e escalabilidade para inserção, deleção e alteração de documentos no corpus de treinamento
- Facilidade de diagnosticar, interpretar os resultados e adicionar julgamento humano para melhorar o classificador

Neste trabalho, nos restringimos à avaliar a precisão dos classificadores no *corpus* de teste. A precisão é dada por a/D, onde a é o número de classificações certas e D é o total de documentos classificados.

	positivos	negativos	total
Total de unigramas	685,069	609,846	1,294,915
Total de unigramas distintos	31,262	28,981	41,675
Total de bigramas distintos	256,287	228,869	419,586
Tamanho Médio dos Textos	685	610	647

Tabela 6.1: Estatísticas do *movie review data set*

6.1 Corpus

A palavra *corpus* significa corpo em latim. No contexto de Processamento de Linguagem Natural, *corpus* se refere a um conjunto de textos utilizados para experimentação e validação de modelos (15). O corpus com o qual trabalhamos é composto por 2000 avaliações de filmes (1000 positivas e 1000 negativas). Este corpus foi disponibilizado por Pang (19) e coletado do *IMDb*[1]. Alguns trabalhos que utilizam este corpus são (19, 18, 26, 12). Alguns critérios foram adotados para garantir a qualidade deste conjunto de dados. Uma destas restrições é que o limite de avaliações selecionadas de um mesmo autor para uma classe é de 20. O objetivo desta restrição é evitar que o estilo de um autor domine o conjunto de dados "viciando" o classificador. O *IMDb* utiliza um sistema de avaliação de 10 estrelas para os filmes. Avaliações de 8 ou mais estrelas foram consideradas positivas e as de 3 ou menos estrelas negativas. Algumas estatísticas do corpus estão descritas na Tabela (6.1).

6.2 Metodologia de Teste

A metodologia de testes adotada no projeto foi a Validação Cruzada com *K-folds* (*K-fold Cross Validation*). Neste método, a amostra original é particionada em k subamostras. Destas k subamostras, uma subamostra é retida para ser utilizada na validação do modelo, as $k - 1$ sub-amostras restantes são usadas como conjunto de treinamento. O processo de validação cruzada é então repetido k vezes, de modo que cada uma das k sub-amostras seja utilizada exatamente 1 vez como dado de teste para validação do modelo. O resultado final é o desempenho médio do classificador nos k testes. O objetivo de repetir os testes múltiplas vezes é aumentar a confiabilidade da estimativa da precisão do classificador. A teoria de validação cruzada foi primeiramente apresentada por Geisser em (08) onde foi demonstrado que a variância entre as precisões dos testes possui uma relação inversa com k.

[1] www.imdb.com

A Figura (6.2) ilustra uma validação cruzada com 3-folds. Todos os resultados descritos neste trabalho foram obtidos usando o *10-fold Cross Validation*. Como nosso *corpus* é composto por 1000 documentos de polaridade positivos e 1000 de polaridade negativa, em cada teste treinamos o classificador com 900 documentos de cada classe e o validamos com 100 documentos de cada classe.

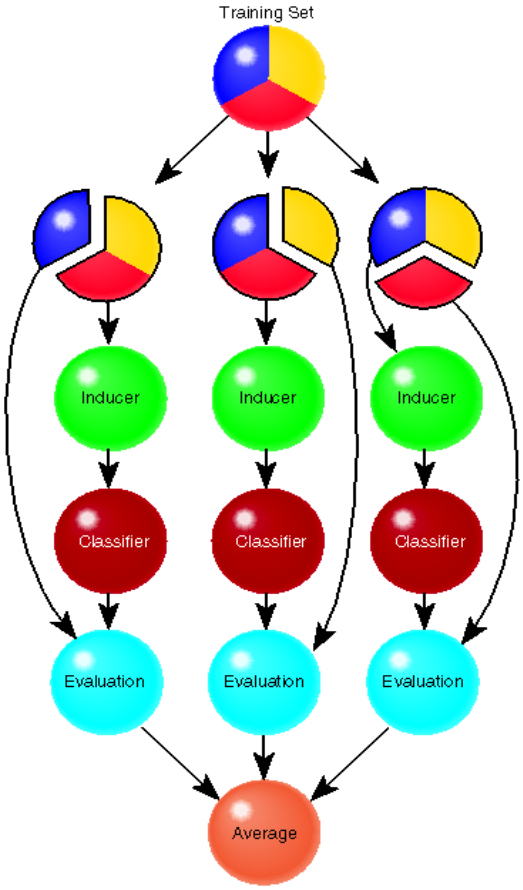


Figura 6.1: Validação Cruzada com 3 folds

6.3 Descrição dos Experimentos

Examinamos quatro modelos de representação de documento:

- unigrama (palavra)
- unigrama+*pos* (palavra e sua classe gramatical)
- bigrama (2 palavras consecutivas)
- bigrama+*pos* (2 palavras consecutivas e suas classes gramaticais)

	positivos	negativos	+ e -
Total de unigramas	428,544	406,682	835,226
Total de unigramas distintos	23177	22204	31,904
Total de bigramas distintos	256,287	228,869	419,586
Tamanho Médio dos Textos	429	407	418

Tabela 6.2: Estatísticas do *corpus* com subjetividade filtrada

Para cada um destes modelos, testamos os classificadores Naive Bayes e SVM em suas formas binária e multinomial. Aplicamos também um filtro de subjetividade nos documentos visando reproduzir os experimentos descritos em (17). Experimentos realizados neste *corpus* "subjetivo" estão rotulados como "subj". A geração deste conjunto de dados é descrito na próxima seção. O Apêndice (A) apresenta os resultados detalhados de cada experimento.

6.3.1
Corpus com Subjetividade Filtrada

Utilizamos como conjunto de treinamento para o classificador de subjetividade o *subjectivity data set* disponibilizados em Pang (17). O conjunto de exemplos de subjetividade é composto por 5000 "movie review snippets". Estes "snippets" foram extraídos de www.rottentomatoes.com. Um exemplo de um elemento deste conjunto é "*bold, imaginative, and impossible to resist*".

O conjunto de exemplos de objetividade (i.e. textos que descrevem dados factuais e não expressam opiniões) é composto por 5000 frases de resumos de roteiro extraídos do $IMDb$. Um exemplo é "The movie begins in the past where a young boy named Sam attempts to save celebi from a hunter".

Utilizando um classificador Naive Bayes multinomial e representando os documentos como a freqüência de unigramas, classificamos cada frase dos documentos como subjetiva ou objetiva. As frases classificadas como objetivas foram removidas do corpus.

Resultados dos experimentos realizados neste *corpus* "subjetivo" estão rotulados como "subj". A Tabela (6.3.1) apresenta as estatísticas do novo *corpus*.

6.4
Resultados

Esta seção descreve os resultados dos testes executados. Os gráficos ilustram o desempenho dos classificadores em função do número de *features* consideradas. A técnica adotada para seleção de *features* é baseado na

Informação Mútua Média conforme descrito na Seção (3.5.1). A Tabela (6.4) apresenta os melhores resultados obtidos para cada modelo.

Os experimentos envolvendo o classificador SVM foram realizados usando o pacote SVM^{light} disponibilizado por Joachims (10). Todos os resultados reportados foram obtidos usando o kernel linear para treinamento e o parâmetro $C = 1$.

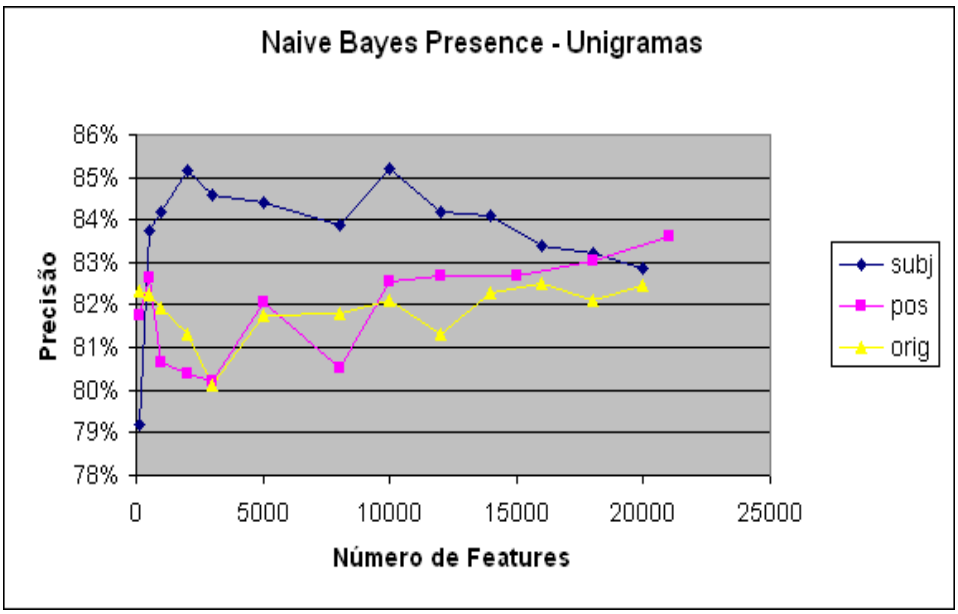
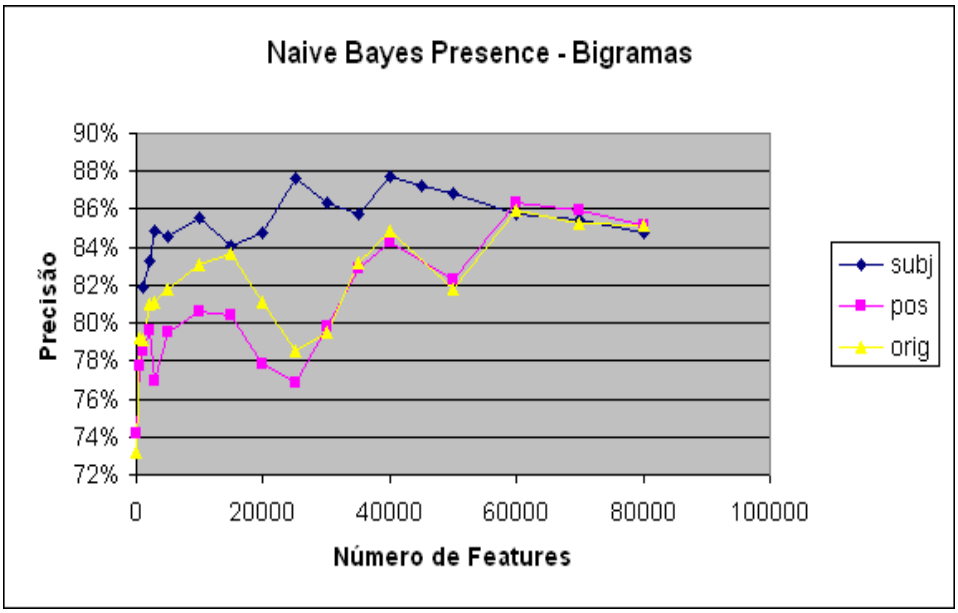
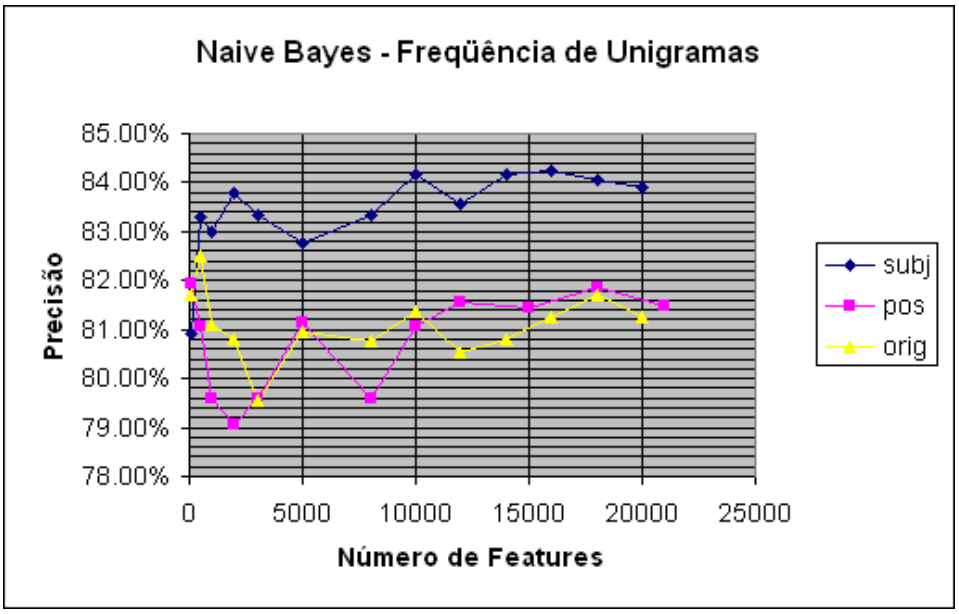
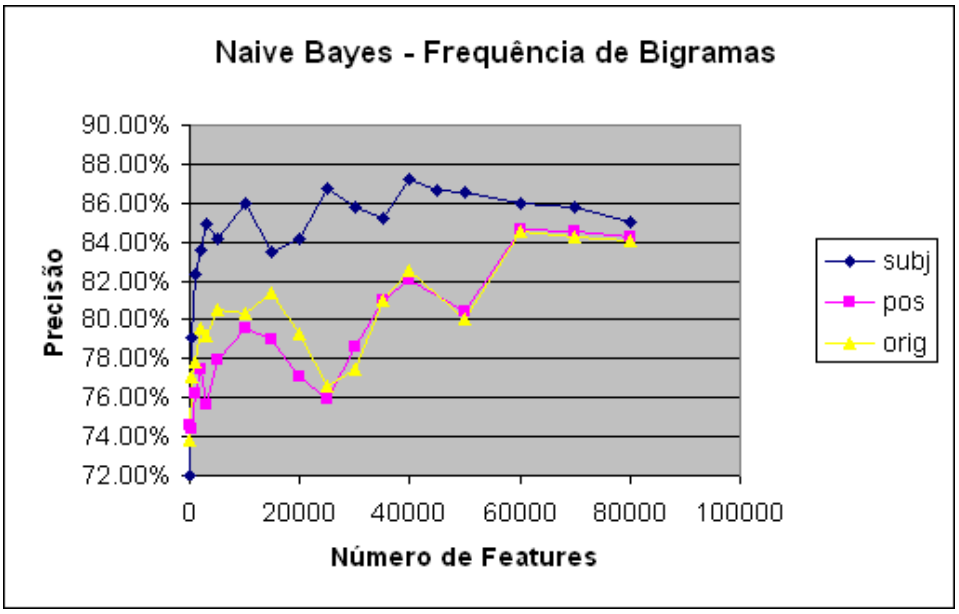


Figura 6.2: Desempenho do Classificador NB



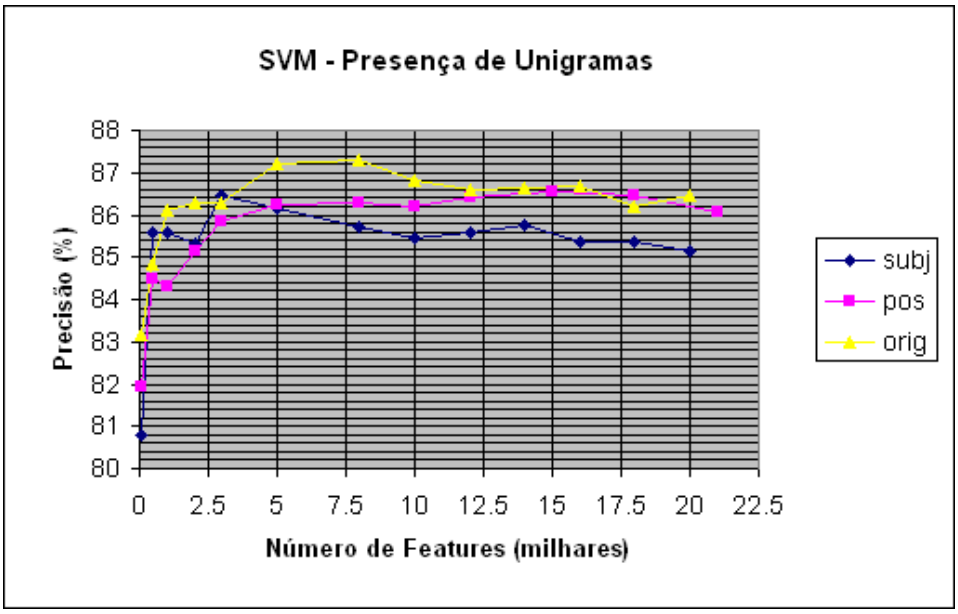


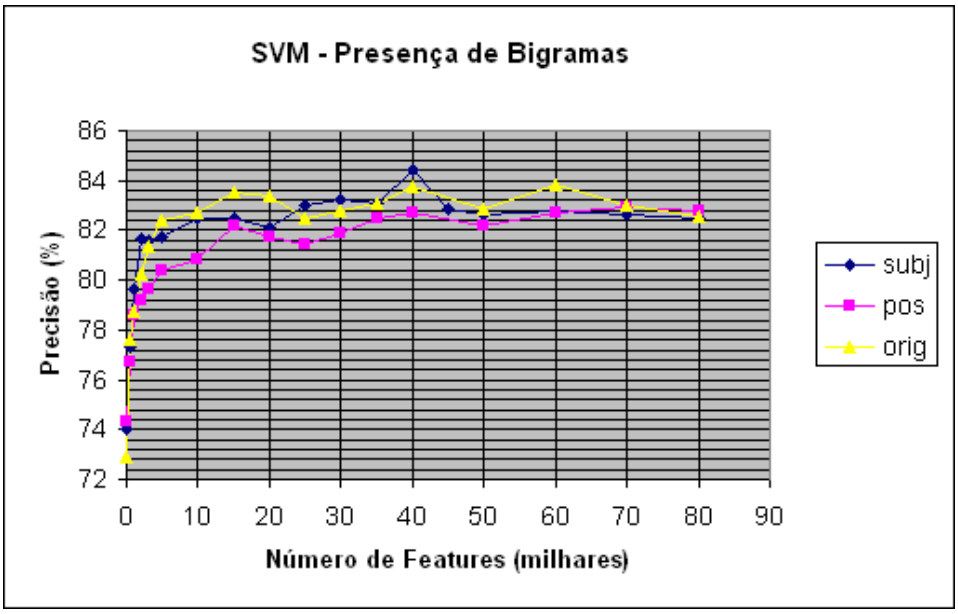


6.4.1 Discussão dos Resultados

O filtro de subjetividade se mostrou eficaz para o classificador Naive Bayes, porém não resultou em melhoria no desempenho do classificador SVM. Este resultado confirma experimentos descritos em (17). Note que ao usar o *corpus* de subjetividade filtrada, o classificador Naive Bayes melhora seu desempenho em todos os modelos de representação de documento.

O impacto do *POS tagging* em todos os experimentos é despresível. Apesar de melhorar a precisão na maioria dos modelos, o ganho não é estatisticamente significante. Note que as curvas de desempenho dos modelos

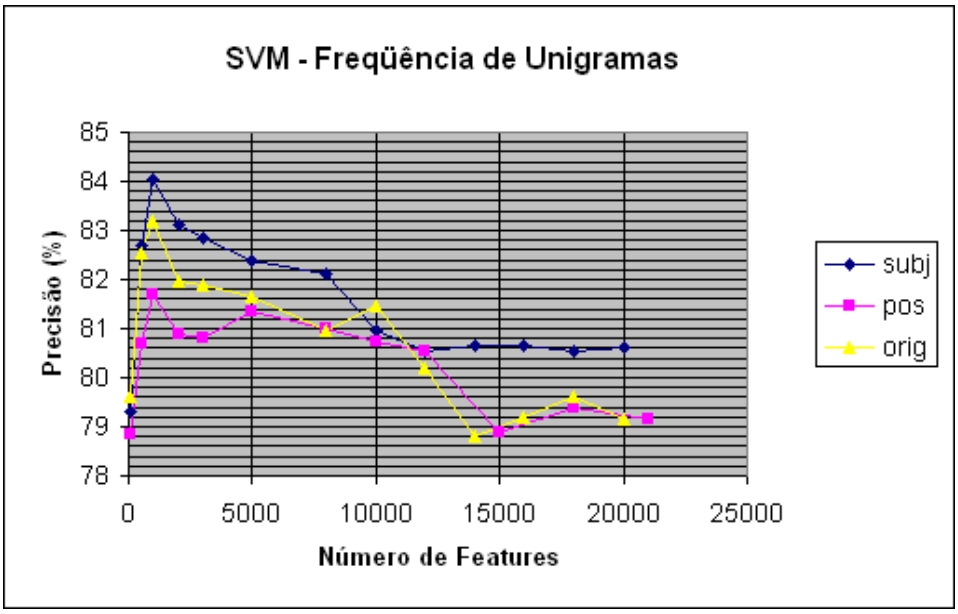


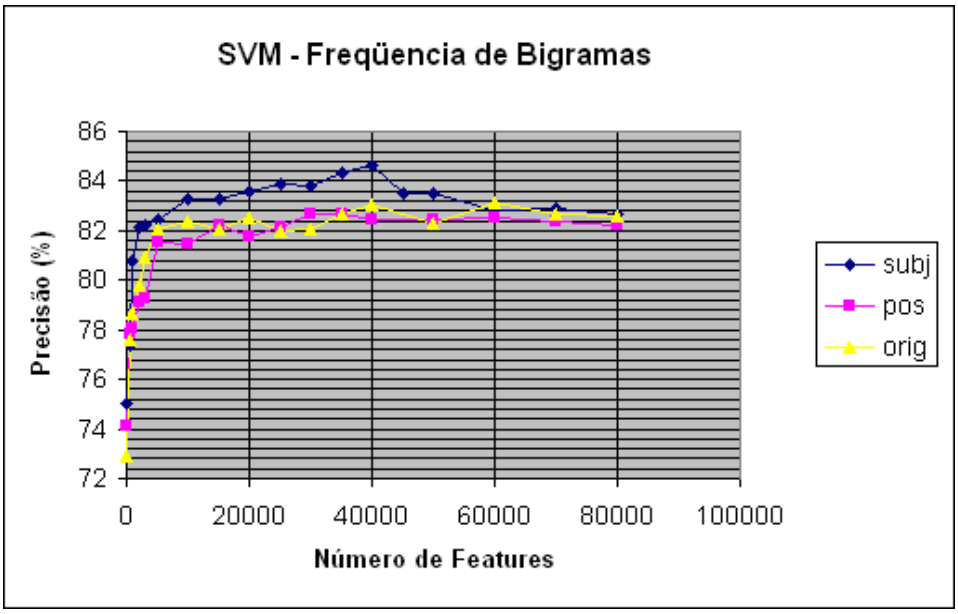


de documento com *POS tagging* ficam sempre muito próximas as curvas dos modelos N-grama do corpus original. Isto provavelmente se deve ao fato de que a grande maioria das palavras mantém sempre a mesma classe gramatical, logo a introdução dos *pos'* gera uma mínima variação nas crenças das máquinas.

Os classificadores binários, ou seja, aqueles que levam em consideração somente a presença das *features*, apresentam melhor desempenho do que os que consideram as freqüências. Isto está de acordo com os resultados dos experimentos em (19).

Representar documentos como ocorrência de bigramas quase sempre se mostrou superior a representação de unigramas com exceção dos resultados





reportados pelo classificador SVM binário.

A seleção de *features* não trouxe ganho de desempenho significativo para os classificadores. As curvas de desempenho em função do número de *features* apresenta uma alta variância não sendo possível inferir um número ótimo de *features*. Porém, é interessante notar que uma vez consideradas as 1000 *features* mais informativas (segundo o critério IMM), os classificadores não apresentam ganhos significativos ao incluir novas *features*.

Portanto, os experimentos sugerem que é possível trabalharmos com uma representação mais "limpa" do documentos (subjetividade filtrada e *features* mais significativas) sem perda de desempenho.

Os modelos e classificadores apresentaram desempenhos muito pare-

Modelo	NB Binário	NB Freq	SVM Binário	SVM Freq	média
unigrama	82.5	82.5	87.3	83.2	83.9
unigrama+POS	83.8	82.0	86.6	81.7	83.5
unigrama subj.	85.2	84.3	86.5	84.1	85.0
bigrama	86.0	84.5	83.2	83.2	84.2
bigrama+POS	86.3	84.7	83.0	82.7	84.2
bigrama subj.	**87.8**	87.2	84.4	84.7	**86.0**
média	**85.3**	84.2	85.2	83.3	

Tabela 6.3: Melhores precisões atingidas por cada classificador para cada modelo de representação de documentos

cidos (82% o pior e 87.8% o melhor). Portanto, não obtivemos conclusões definitivas sobre qual é o melhor modelo para a tarefa de classificação de sentimento. Além disso, o corpus utilizado é razoavelmente pequeno (2000 textos) e limitado a um domínio e fonte específicos. Portanto, os experimentos não trazem conclusões definitivas quanto aos melhores modelos para o problema.