5. Estudo de Casos

Neste capítulo apresentaremos os experimentos feitos com a ferramenta SQLLOMining objetivando a obtenção de Objetos de Aprendizagem. Descreveremos também alguns procedimentos adotados para a geração dos dados que foram utilizados nos experimentos e finalmente apresentaremos os resultados obtidos.

5.1. Geração do Corpus

O processo de especificação do Corpus foi um trabalho que partiu praticamente do zero. Em alguns domínios de problemas mais comuns, já existem Corpus disponibilizados em um formato padrão, mas no estudo de caso que nos propomos trabalhar, encontramos um material disponível extremamente amplo e rico, mas totalmente desprovido de qualquer tratamento. O nosso material consiste de qualquer documento que possa ser encontrado na Internet ou em uma biblioteca digital que possua o conhecimento que queremos capturar em forma de texto.

Para restringir um pouco esta vasta opção de recursos, optamos por nos concentrar em textos de um determinado domínio permitindo assim o desenvolvimento de um Corpus especializado e por isso mais rico em informação para o aprendizado de máquina.

Apesar de estarmos restringindo o nosso domínio, estamos falando sempre de extração de conhecimento através do aprendizado a partir de exemplos. Isso nos leva a uma necessidade de exemplos extremamente precisos e ricos em informação. Eles precisam espelhar claramente o objeto de interesse representando da melhor forma as classes de texto que podem ser encontradas. Desta forma estaremos procurando manter o modelo o mais próximo possível das premissas adotadas na sua formulação e assim reduzir erros de classificação. Para isso utilizaremos a ontologia de tipos de ALOs descrita anteriormente que visa descrever o domínio que estamos abordando.

Outro resultado positivo que podemos esperar obter com a restrição de um domínio é a restrição também do léxico e da forma de escrita. O uso da linguagem natural pode ser extremamente variado e esta variação ocorre por diversas causas seja pelo lado de quem está escrevendo como pelo lado de para quem o texto está sendo escrito.

Um texto escrito por um romancista é totalmente diferente do que um escrito por um mestre ou doutor. E da mesma forma, um texto escrito para alunos do ensino médio tem uma abordagem totalmente deferente de um escrito para aqueles mesmos mestres ou doutores.

Partimos então das seguintes premissas: trabalharemos com um tipo de texto específico, voltado para o ensino médio, que apresentem conceitos de áreas específicas como física e química e procuraremos nestes textos parágrafos que possam ser classificados na ontologia de tipos de ALOs e utilizados na criação de Objetos de Aprendizagem (LOs).

5.1.1.Extração automática de exemplos

O processo de coleta de exemplos é extremamente trabalhoso. A quantidade mínima requerida para começarmos a alcançar resultados razoavelmente satisfatórios é grande, e fazer uma etiquetagem manual é praticamente impossível quando se quer alcançar este número. Partimos então para uma abordagem mais simplista que nos permitisse desenvolver um Corpus razoavelmente satisfatório.

Uma das abordagens objetivou a busca de textos que contivessem apenas um tipo de sentença. Um exemplo muito fácil de encontrar são textos de definição. Um glossário, que pode ser facilmente encontrado na Internet, possui apenas textos da classe Definição e nos permite colecionar de uma vez centenas de exemplos válidos.

Por outro lado, textos romanceados não possuem definições e podem ser considerados como contendo apenas sentenças da classe negativa com uma margem bastante pequena de erro.

Partindo destas abordagens colecionamos mais de 300 exemplos de definições do domínio de física extraindo glossários de termos de física e química disponíveis na Internet e utilizamos como exemplos da classe negativa, textos de livros que falavam sobre o mesmo assunto, mas de uma forma romanceada.

5.1.2.Pré-etiquetagem com conferência manual

Outra abordagem válida e que nos permitiu gerar uma quantidade de exemplos bastante satisfatória é a pré-etiquetagem com conferência manual. Este processo necessita de um modelo já criado a partir de exemplos já coletados. Este modelo é utilizado para etiquetar textos ainda não classificados e sobre esta etiquetagem é feita uma conferência revisando apenas os textos etiquetados positivamente.

Este método nasceu naturalmente da técnica de aprendizado semi-supervisionado utilizada pelo algoritmo existente no sistema SQLLOMining onde na segunda fase do aprendizado o algoritmo classifica as amostras a partir do modelo pré gerado e aprimora seus parâmetros estatísticos aumentando sua precisão.

5.2.Descrição dos experimentos

A seguir apresentaremos os dois Corpus que foram desenvolvidos utilizando a ferramenta SQLLOMining. Ambos utilizaram textos orientados ao ensino médio de ciências exatas como, por exemplo, a física e a química.

5.2.1.Corpus Definição

O primeiro Corpus que desenvolvemos possui apenas dois tipos sendo um positivo e um negativo. Estamos objetivando encontrar apenas textos relativos a definições de conceitos separando-os de todo o resto. Este Corpus utilizou a técnica descrita anteriormente de extração automática de exemplos através de glossários de termos de física encontrados na Internet e em livros da área.

Para exemplos de texto da classe não-definição, utilizamos livros romanceados relativos ao mesmo assunto como, por exemplo, “Uma Breve

História do Tempo" de Stephen Hawking (Hawking, 1988). Criamos então o Corpus Inicial.

Numa segunda etapa fizemos uma revisão manual das sentenças dos livros que foram consideradas como da classe positiva pelo algoritmo e encontramos diversas delas que eram realmente definições. Um exemplo é a seguinte sentença: "Com efeito, o metro é definido como a distância percorrida pela luz em 0,000000003335640952 segundos medidos por um relógio de césio." Ela está no livro "Uma Breve História do Tempo" e inicialmente estava definida como sendo da classe negativa. Corrigimos então a classificação inicial delas as incluindo como exemplos da classe positiva.

Acrescentamos o livro "A Dança do Universo" de Marcelo Gleiser (Gleiser, 2006) e refizemos este procedimento utilizando todos os tipos de modelagens possíveis variando o parâmetro de N-Grama e *stemming*. Aumentamos com isso a quantidade de exemplos de definição que tínhamos inicialmente. Alcançamos então o número de 628 exemplos de definição e 2558 exemplos de não-definição para o livro "Uma Breve História do Tempo" e "A Dança do Universo". Chamamos este Corpus de Corpus Aumentado.

Fizemos os experimentos de aprendizado com o Corpus desenvolvido nas etapas anteriores com o objetivo de colher as métricas de desempenho do aprendizado que serão apresentadas a seguir. Para executar os cálculos destas métricas consideramos todo o resto de amostras provenientes dos livros como sendo da classe negativa. Esta é uma premissa simplista, mas consideramos uma aproximação razoável devido ao fato de realmente serem muito reduzidos os casos de definição existentes no texto dos livros.

Devido a quantidade reduzida de exemplos utilizamos apenas dois grupos de teste e por isso executamos os experimentos fazendo uma validação cruzada reduzida que utilizou estes dois grupos de exemplos.

5.2.2.Corpus Estendido

Como mencionado em (Chakrabarti, 2002) o modelo Bayesiano de misturas parte do pressuposto que cada texto deve estar relacionado a uma e apenas uma classe assim como cada componente da mistura. Esta mistura

representará a geração de todo o documento existente. Partindo destas premissas podemos supor que, para que possamos obter um modelo que possa classificar qualquer texto, ele deve estar munido de informação suficiente para conseguir classificá-lo corretamente em uma das classes que ele contempla. Por isso seria importante que este modelo estivesse "informado" em relação a todos os tipos de texto que existem.

Sob este ponto de vista procuramos sempre trabalhar com um conjunto de classes que satisfaça essas premissas e para definir este conjunto utilizamos a ontologia apresentada anteriormente. Como a tarefa de classificar qualquer texto que existe é bastante complexa, usamos continuamente a classe negativa. Esta classe engloba todos os outros tipos de texto que não têm uma classe especificada. Por exemplo, nos primeiros experimentos que fizemos, trabalhamos com a classe de Definição e a classe de Não-definição dividindo claramente todos os textos existentes.

Como é amplamente discutido em (Nigam, Maccallum, Thrun e Mitchell, 2000) quando temos uma classe negativa que engloba diversos tipos de texto, vamos de encontro às premissas do modelo e por isso o Aprendizado de Maquina sofre muito, levando à perda de desempenho no processo de aprendizado com textos não etiquetados. Ainda neste artigo são sugeridos dois aprimoramentos para o algoritmo com o objetivo de minimizar esta perda de desempenho: o primeiro define uma diferenciação de peso para as informações relativas às amostras diminuindo o efeito que elas têm na definição do modelo e o segundo calcula de outra forma as componentes do modelo de mistura permitindo que elas, por sua vez, sejam também compostas de uma mistura de distribuições. Desta forma relaxamos a premissa que diz que cada componente tem uma relação de um para um com as classes. Dentro de uma classe podem existir diversas subclasses e esta relação passa a ser de muitos para um.

Com o objetivo de explorar este raciocínio de uma outra forma fizemos um segundo grupo de testes utilizando um “Corpus Estendido”. Neste Corpus separamos da classe negativa um dos subtipos que antes lhe pertencia. Para nos apoiar na criação deste Corpus utilizamos a ontologia de Tipos de ALOs apresentada no primeiro capítulo visando obter com isso uma melhora no desempenho do classificador e do aprendizado semi-supervisionado.

5.3.Resultados

A seguir apresentaremos os resultados obtidos nos experimentos feitos separando-os em tópicos que abordarão cada uma das avaliações propostas.

5.3.1.Corpus Definição

No primeiro Corpus de Definição, fizemos diversos experimentos com objetivos distintos. O primeiro levantou as diferenças encontradas com a variação das *features* (quantidade de atributos) utilizadas no modelo. No segundo avaliamos especificamente o processo do aprendizado semi-supervisionado e finalmente avaliamos a utilização do Corpus desenvolvido para a classificação de arquivos com origens totalmente distintas dos textos utilizados no Corpus.

5.3.1.1.Seleção de *features* do modelo

O primeiro experimento utilizou quatro casos distintos. Dois deles utilizaram o caso mais simples, onde cada palavra é considerada separadamente, e os outros dois agruparam as palavras duas a duas o que traria maiores informações quanto à ordenação delas.

Modelagem	**Qtd**	**Accuracy**	***Recall***	***Precision***
1 Grama	7797	90,55%	96,007%	84,8866%
1 Grama com *stemming*	7110	91,42%	95,946%	86,3241%
2 Grama	17784	75,62%	86,969%	60,9113%
2 Grama com *Stemming*	17601	76,15%	87,458%	61,4078%

Tabela 3 – Resultados com o Corpus inicial avaliação de *features*

Podemos observar que quando utilizamos o recurso de *stemming* as classificações apresentam um desempenho um pouco melhor, mas podemos concluir que o uso do recurso de 2-grama diminui a performance da classificação.

A partir destes resultados resolvemos reduzir nossos experimentos as *features* que apresentaram maior desempenho. Passamos a utilizar então sempre unigrama e o algoritmo de *stemming* nos experimentos seguintes.

5.3.1.2. Aprendizado Semi-Supervisionado

Objetivando apresentar o aprendizado semi-supervisionado ocorrendo em um experimento fizemos um gráfico com o valor da Accuracy para todas as amostras de ambas as classes.

Nestes gráficos podemos observar que a Accuracy aumenta à medida que a quantidade de amostras é acrescentada aos experimentos. Cada uma das curvas dos gráficos abaixo equivale a uma quantidade diferente de exemplos utilizados no cálculo do modelo básico. Todas as curvas de aprendizado do primeiro gráfico estabilizaram-se em um patamar e a sua maior parte em torno de 90%. Podemos então observar que a partir de um número específico de exemplos, as curvas de aprendizado não são muito afetadas mantendo-se no mesmo nível ao acrescentar amostras.

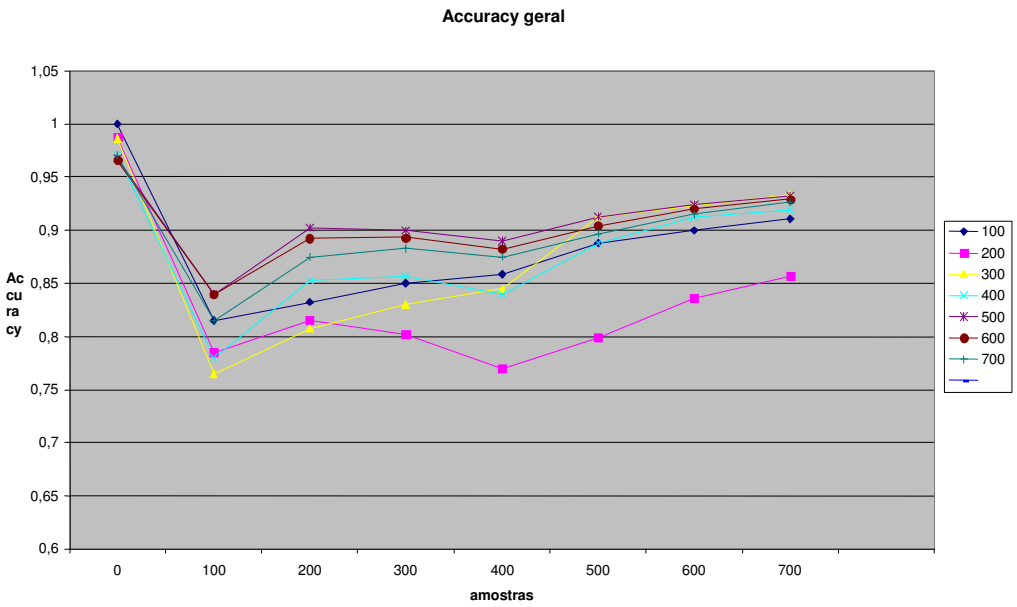

Figura 244 – Accuracy Corpus Definição

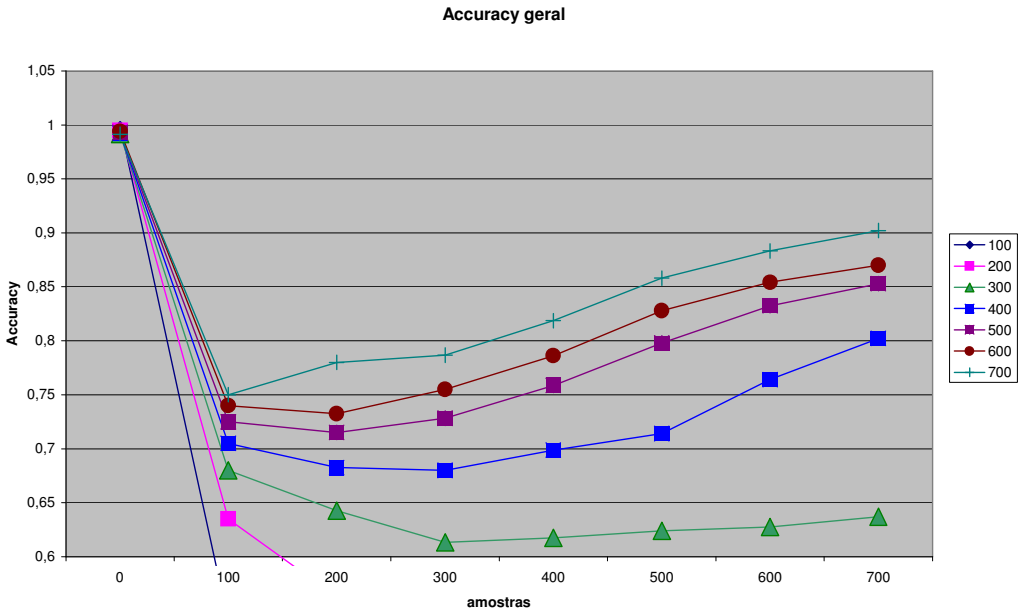

Figura 255- Accuracy Corpus Aumentado

Note que as curvas no primeiro gráfico são bem diferentes da do segundo. Cada um destes experimentos utilizou um corpus diferente e esta variação nos mostra como o aprendizado semi-supervisionado é sensível à forma como os exemplos são divididos. No Corpus Aumentado, apesar de não termos obtido nenhum ganho para a quantidade de exemplos menor que 400, tivemos uma melhora bem maior para as curvas a partir de 400 exemplos, chegando a ter uma melhora de 17%. Este ganho praticamente se estabiliza a partir de 500 exemplos sugerindo ser a melhor escolha.

Abaixo segue tabela com o cálculo dos ganhos:

Exemplos	**100**	**200**	**300**	**400**	**500**	**600**	**700**
100 amostras	54%	64%	68%	71%	73%	74%	75%
700 amostras	46%	48%	64%	80%	85%	87%	90%
Ganho	-17%	-31%	-7%	12%	15%	15%	17%

Tabela 4 – Ganhos de desempenho com o Aprendizado Semi-Supervisionado

5.3.1.3.Classificação de arquivos

Criamos então, manualmente, alguns arquivos com sentenças de um tipo específico para serem utilizados em experimentos e em coleta de métricas para

avaliação da classificação. Definimos ser necessário utilizar sentenças parecidas com as utilizadas nos exemplos e também utilizar outras sentenças originárias de textos bem diferentes. Desta forma seria possível avaliar corretamente a classificação, pois no arquivos de teste não teríamos os mesmos textos que os encontrados nos textos utilizados no treinamento. Desenvolvemos finalmente os últimos dois arquivos de testes como casos extremamente difíceis, pois se tratavam de textos próprios de livros de ensino médio contendo diversos tipos de sentenças misturados.

Arquivos de teste	Classe	*Recall*	*Precision*	F1
Definições retiradas de livros de ensino médio	Definição	93%	100%	96%
Não definições retiradas de livros de ensino médio	Não Definição	30%	100%	46%
Livros de física do ensino médio (Aula sobre medidas)	Definição	100%	08,6%	21%
	Não Definição	36%	100%	53%
Livros de física do ensino médio (Aula sobre Energia)	Definição	100%	12%	16%
	Não Definição	27%	100%	42%

Tabela 5 – Resultados com o Corpus Definição

Podemos observar por estes resultados que o valor de *Recall* se manteve sempre elevado para a classe de Definição indicando que a grande maioria das definições existentes nos textos foram corretamente indicadas pela classificação. Mas o valor da métrica *Precision* se apresentou bastante baixo em casos mais complexos indicando que diversas partes dos textos que não eram definição foram mal classificadas.

5.3.2.Corpus Estendido

O Corpus Estendido tem mais uma classe que diferencia outro subconjunto da classe negativa. Para escolher esta nova classe utilizamos a ontologia de LOs descrita em capítulos anteriores. Escolhemos a classe Lei, pois este tipo de texto era encontrado com muita freqüência e também foi, por diversas vezes, classificado como definição.

Inicialmente fizemos um Corpus com duas classes: Lei e Não-Lei e depois trabalhamos com um Corpus com três classes: Definição, Lei e Outros.

5.3.2.1.Desempenho da Classificação

Para mantermos um domínio comum entre os experimentos refizemos o Corpus de Definição mantendo os mesmos exemplos nos três casos. Utilizamos um arquivo com exemplos de Lei, três arquivos com exemplos de Definição e dois arquivos com exemplos de Não-Definição que eram textos retirados dos dois livros mencionados anteriormente. Para o Corpus de Lei, apenas o arquivo de exemplos de Lei foi classificado como Lei, os outros ficaram na classe negativa. Para o Corpus de Definição o arquivo de Lei e os livros foram incluídos na classe negativa e os três arquivos de definição ficaram na classe Definição. E finalmente no Corpus Estendido definimos três classes, o arquivo de Lei ficou na classe Lei, os três arquivos de Definição na Classe de Definição e os dois livros na classe negativa.

Para testar igualmente os três Corpus, utilizamos a mesma quantidade de exemplos, limitado à quantidade reduzida que tínhamos disponível para a classe de Lei. Neste caso utilizamos apenas 100 exemplos em todos os experimentos. A partir do modelo criado com estes 100 exemplos, classificamos 1100 amostras. Abaixo seguem os resultados:

	Corpus de Definição	Corpus de Lei	Corpus de Definição e Lei
Precision Definição	97%		87%
Precision Lei		16%	41%
Precision Negativa	73%	99%	97%
Recall Definição	46%		83%
Recall Lei		76%	72%
Recall Negativa	99%	90%	93%
F1 Definição	62%		85%
F1 Lei		26%	46%
F1 Negativa	84%	94%	95%
Accuracy	77,7273%	89,2273%	89,2332%

Tabela 6 – Comparação do Corpus Definição com o Corpus Estendido

É importante salientar que os exemplos de Lei foram utilizados nos três experimentos sendo que no primeiro eles fizeram parte dos exemplos da classe negativa e, no segundo e no terceiro, da classe positiva de Lei.

Ao classificar as 27 amostras de Lei a *precision* no Corpus Estendido ficou bem maior e o *Recall* um pouco menor. Obtivemos então uma melhora considerável na medida F1 para estas amostras. Para as amostras de definição tivemos uma queda na precisão quando utilizamos o Corpus Estendido, mas o *Recall* praticamente dobrou. Obtivemos novamente uma melhora considerável na medida F1 para estas amostras. Novamente é importante frisar que os exemplos da classe negativa foram diferentes nos três casos.

Resolvemos então repetir os experimentos utilizando sempre os mesmos exemplos para termos outra avaliação. A quantidade de 99 exemplos foi dividida da seguinte forma:

	Corpus de Definição	Corpus de Lei	Corpus de Definição e Lei
Classe Definição	33		33
Classe Lei		27	27
Classe Negativa	27 + 39	33 + 39	39

Tabela 7 – Distribuição dos exemplos nos Corpus

Utilizando as mesmas 1000 amostras nos três experimentos, obtivemos os seguintes resultados:

	Corpus de Definição	Corpus de Lei	Corpus de Definição e Lei
Precision Definição	98%		89%
Precision Lei		75%	64%
Precision Negativa	70%	98%	95%
Recall Definição	19%		89%
Recall Lei		22%	67%
Recall Negativa	99%	99%	95%
F1 Definição	32%		89%
F1 Lei		34%	65%
F1 Negativa	82%	99%	95%
Accuracy	72%	98%	93%

Tabela 8 – Comparação do Corpus Definição com o Corpus Estendido

Novamente podemos observar que o valor de *Recall* tanto para a classe de definição quanto para a classe de Lei aumentou no Corpus Estendido. A precisão caiu em ambos os casos, mas a medida de F1 aumentou.

Podemos concluir que o Corpus Estendido parece apresentar maior desempenho no reconhecimento das classes. A variação nos resultados ocorreu devido à variação nos exemplos que foram utilizados e como para a classe de lei não tínhamos um Corpus muito extenso para trabalhar tivemos que reduzir a opção de testes e consequentemente de resultados, o que não é o ideal neste tipo de experimento. Mas a tendência a melhora de desempenho para o Corpus Estendido parece bastante clara.

5.3.2.2. Desempenho do Aprendizado Semi-Supervisionado

Objetivando observar o que acontece com o desempenho do aprendizado semi-supervisionado quando utilizamos o Corpus Estendido fizemos um gráfico com o valor da Accuracy para todas as amostras de ambas as classes em cada um dos três Corpus utilizados anteriormente. Os experimentos utilizaram os mesmos 100 exemplos utilizados no experimento anterior na mesma distribuição entre as classes. Variamos então a quantidade de amostras acrescidas ao modelo e obtivemos os resultados a seguir:

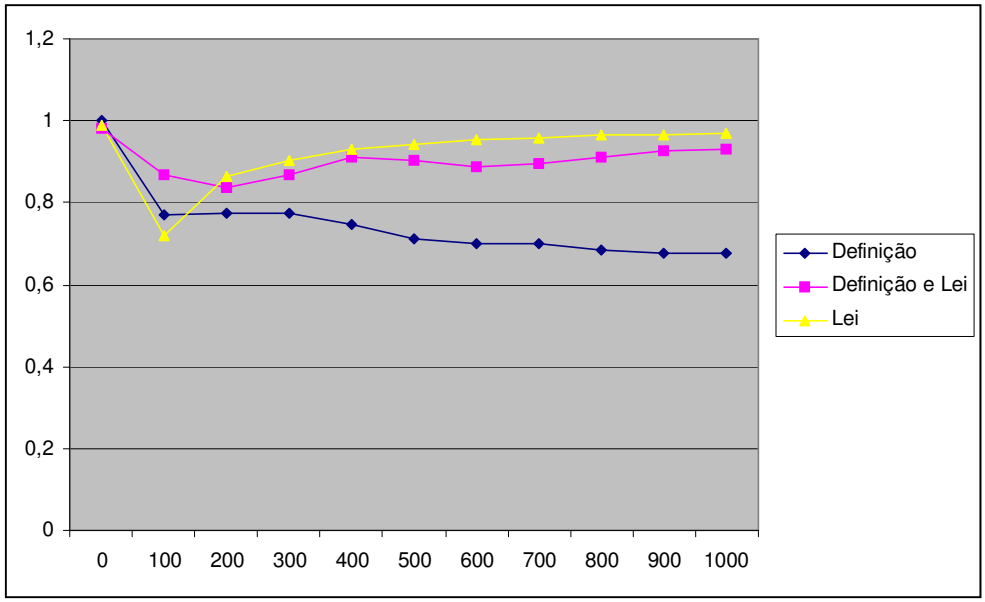

Figura 266 - Comparação da curva de Aprendizado - Valores de Accuracy

Mais uma vez podemos comprovar que, quanto mais as premissas usadas pelo modelo Bayesiano de misturas forem verídicas para o experimento, melhor é a curva de aprendizado. Quando utilizamos o Corpus de Definição perdemos desempenho do classificador com o uso de amostras. Neste Corpus a classe negativa possuía textos de Lei e Outros misturados. Quando utilizamos o Corpus de Lei o aprendizado também foi ruim. E finalmente para o Corpus Estendido a curva de aprendizado foi bastante boa. A seguir apresentamos os gráficos.

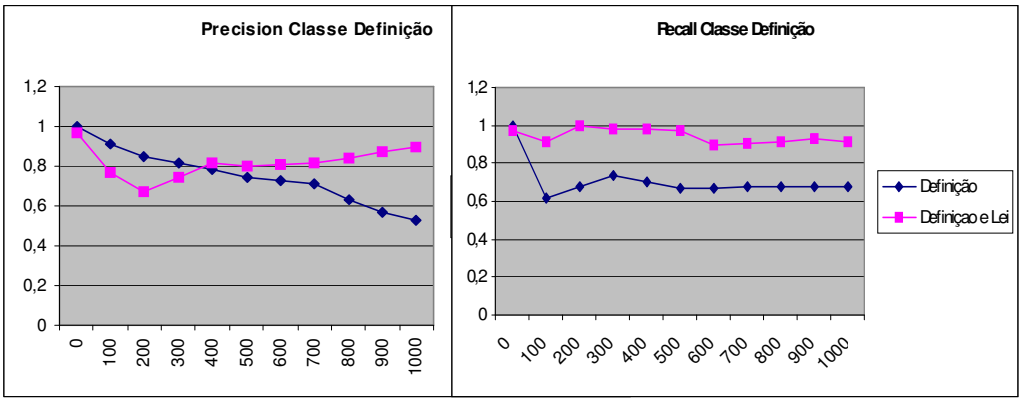

Figura 277 – *Precision* e *Recall* para Classe Definição

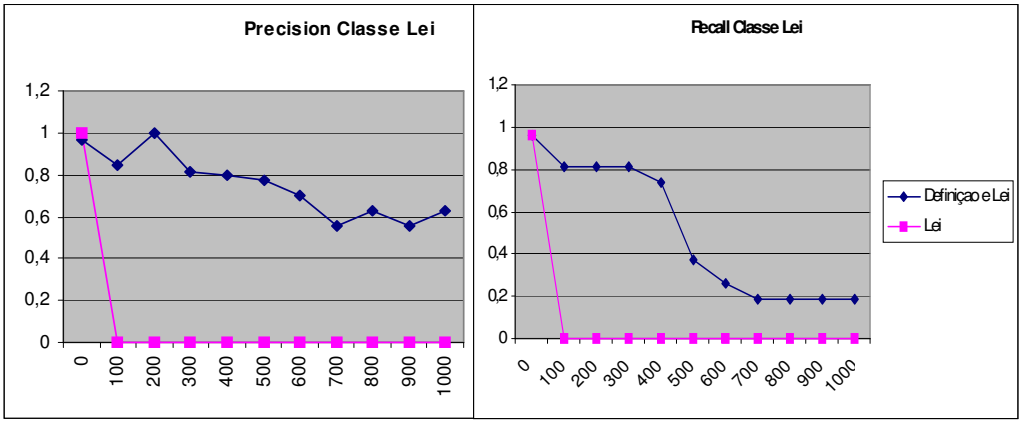

Figura 288– *Precision* e *Recall* para Classe Lei

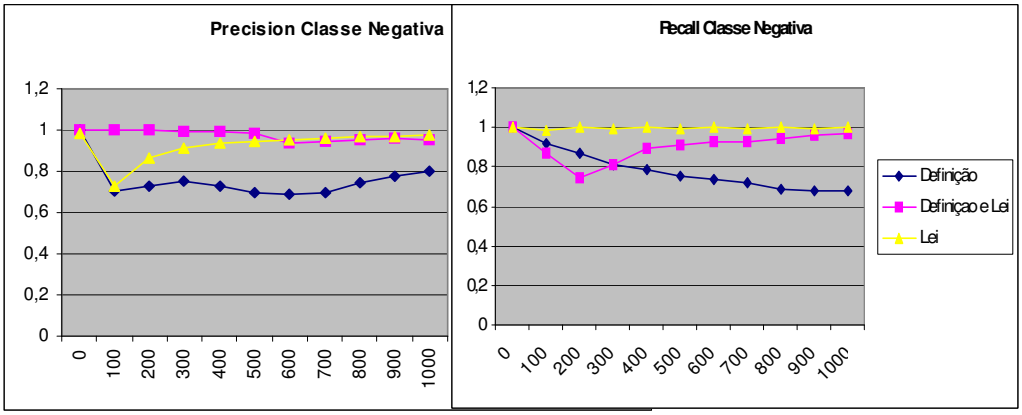

Figura 299 – *Precision* e *Recall* para Classe Negativa