



Programa Interdisciplinar de Pós-Graduação em
Computação Aplicada
Mestrado Acadêmico

ALEXANDRE NUNES BARBOSA

Descoberta de Conhecimento Aplicado à Base de
Dados Textual de Saúde

São Leopoldo, 2012

ALEXANDRE NUNES BARBOSA

**DESCOBERTA DE CONHECIMENTO APLICADO
À BASE DE DADOS TEXTUAL DE SAÚDE**

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre, pelo Programa Interdisciplinar de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos – UNISINOS.

Professor orientador: Dr. João Francisco Valiati

São Leopoldo

2012

Dedico essa dissertação a três pessoas da minha maior estima:

A minha mãe (*in memóriam*), pela pessoa especial que foi, e que sua falta vai ser sentida sempre dentro de mim.

A minha filha Natália, que meu esforço sirva de exemplo para moldar a sua personalidade, te amo Nat.

A minha namorada Helena, que merece esta homenagem e muito mais, te amo Lena.

AGRADECIMENTOS

Analisando estes dois anos que passaram e lembrando todas as pessoas que estiveram ao meu lado e contribuíram de alguma forma para que esse momento fosse concretizado, me dou conta que tenho muitos a agradecer.

À UNISINOS, em especial ao Programa Interdisciplinar de Pós-graduação em Computação Aplicada (PIPCA), o qual me acolheu e me forneceu subsídios para a realização deste mestrado.

Ao professor Arthur Tórgo Gómez, não tenho palavras para agradecer, pois na ocasião do meu ingresso na pós-graduação, era o coordenador do PIPCA, e junto com meu orientador soube entender o momento difícil que eu estava enfrentando em função da doença de minha mãe.

Ao meu orientador João Francisco Valiati são inúmeros agradecimentos, pela compreensão das minhas dificuldades, pelo conhecimento e experiência transmitidos e por ter acreditado em mim, principalmente pela orientação durante este importante processo de aprendizagem.

Ao corpo docente do PIPCA que sempre ajudou dirimir dúvidas.

Ao colega Rodrigo Moraes muitíssimo obrigado, pela ajuda com sua experiência, dicas, conselhos, conhecimento, debates que contribuíram para o enriquecimento desta dissertação, sempre solícito quando precisei.

Ao Mestre em Informática em Saúde pela Universidade Federal de São Paulo Amilton Martha, que viabilizou este trabalho com a disponibilização da base de dados utilizada.

A minha filha Natália, que teve um pai que não pode lhe dedicar a devida atenção por mais de dois anos, e que com o seu jeito meigo e carinhoso compreendeu a importância do que eu estava fazendo.

Agradeço, in memoriam, com saudades e ternura a meus pais, Vilson e Cesarina que foram a base da minha formação como pessoa, em especial a minha mãe que sempre acreditou em mim e esteve a meu lado.

A minha namorada Helena, pelo apoio incondicional, abdicando de muito para me apoiar durante essa jornada, sobretudo pelo amor e carinho. Sem ela não sei se teria conseguido.

Aos meus colegas de curso, Fabiane Penteado, Lucas Graebin, Lucas Monteiro, Michele Lermen, Toni Wickert, Tassia Serrao, Cristiano Galafassi, que foram importantes durante a realização das disciplinas.

Ao Dr. Claiton Brenol, Dra. Aline Defaveri do Prado e ao acadêmico de Medicina Cristiano Köhler Silva, muito obrigado pela análise e explicações quanto as regras extraídas.

A minha família que sempre acreditou e me apoiou durante a realização do curso.

A todos meus amigos que torceram por mim.

Aos funcionários do PIPCA que sempre deram o apoio necessário.

A Deus que nos dotou de intelecto e capacidade de aprendizagem.

Ao banco Santander pelo apoio financeiro em forma de bolsa de estudos, que viabilizou a realização deste curso.

RESUMO

Este trabalho propõe um processo de investigação do conteúdo de uma base de dados, composta por dados descritivos e pré-estruturados do domínio da saúde, mais especificamente da área da Reumatologia. Para a investigação da base de dados, foram compostos 3 conjuntos de interesse. O primeiro composto por uma classe com conteúdo descritivo relativo somente a área da Reumatologia em geral, e outra cujo seu conteúdo pertence a outras áreas da medicina. O segundo e o terceiro conjunto, foram constituídos após análises estatísticas na base de dados. Um formado pelo conteúdo descritivo associado as 5 maiores frequências de códigos CID, e outro formado por conteúdo descritivo associado as 3 maiores frequências de códigos CID relacionados exclusivamente à área da Reumatologia. Estes conjuntos foram pré-processados com técnicas clássicas de Pré-processamento tais como remoção de *Stopwords* e *Stemmer*. Com o objetivo de extrair padrões que através de sua interpretação resultem na produção de conhecimento, foram aplicados aos conjuntos de interesse técnicas de classificação e associação, visando à relação entre o conteúdo textual que descreve sintomas de doenças com o conteúdo pré-estruturado, que define o diagnóstico destas doenças. A execução destas técnicas foi realizada através da aplicação do algoritmo de classificação *Support Vector Machines* e do algoritmo para extração de Regras de Associação *Apriori*. Para o desenvolvimento deste processo foi pesquisado referencial teórico relativo à mineração de dados, bem como levantamento e estudo de trabalhos científicos produzidos no domínio da mineração textual e relacionados a Prontuário Médico Eletrônico, focando o conteúdo das bases de dados utilizadas, técnicas de pré-processamento e mineração empregados na literatura, bem como os resultados relatados. A técnica de classificação empregada neste trabalho obteve resultados acima de 80% de Acurácia, demonstrando capacidade do algoritmo de rotular dados da saúde relacionados ao domínio de interesse corretamente. Também foram descobertas associações entre conteúdo textual e conteúdo pré-estruturado, que segundo a análise de especialistas, podem conduzir a questionamentos quanto à utilização de determinados CIDs no local de origem dos dados.

Palavras-chave: Prontuário Médico Eletrônico. Mineração Textual. Descoberta de Conhecimento em Textos. Classificação. Associação.

ABSTRACT

This study suggests a process of investigation of the content of a database, comprising descriptive and pre-structured data related to the health domain, more particularly in the area of Rheumatology. For the investigation of the database, three sets of interest were composed. The first one formed by a class of descriptive content related only to the area of Rheumatology in general, and another whose content belongs to other areas of medicine. The second and third sets were constituted after statistical analysis in the database. One of them formed by the descriptive content associated to the five highest frequencies of ICD codes, and another formed by descriptive content associated with the three highest frequencies of ICD codes related exclusively to the area of Rheumatology. These sets were pre-processed with classic Pre-processing techniques such as *Stopword Removal* and *Stemming*. In order to extract patterns that, through their interpretation, result in knowledge production, association and classification techniques were applied to the sets of interest, aiming at to relate the textual content that describes symptoms of diseases with pre-structured content, which defines the diagnosis of these diseases. The implementation of these techniques was carried out by applying the classification algorithm Support Vector Machines and the Association Rules Apriori Algorithm. For the development of this process, theoretical references concerning data mining were researched, including selection and review of scientific publications produced on text mining and related to Electronic Medical Record, focusing on the content of the databases used, techniques for pre-processing and mining used in the literature, as well as the reported results. The classification technique used in this study reached over 80% accurate results, demonstrating the capacity the algorithm has to correctly label health data related to the field of interest. Associations between text content and pre-structured content were also found, which, according to expert analysis, may be questioned as for the use of certain ICDs in the place of origin of the data.

Keywords: Electronic Medical Record. Text Mining. Knowledge Discovery in texts. Classification. Association.

LISTA DE FIGURAS

Figura 1 - Sub-divisões do CID	28
Figura 2 - Matriz de Confusão.....	39
Figura 3 - Espaço dimensional com documentos mapeados	43
Figura 4 - Mapeamento no espaço dimensional original e espaço característico	45
Figura 5 - Descrição gráfica do processo	63
Figura 6 - Exemplo de conteúdo a ser pré-processado	71
Figura 7 - Acurácia por técnica de transformação	80
Figura 8 - Gráfico de taxa média de FP.....	81
Figura 9 - Gráfico de Acurácia conjunto 5 CIDs mais Frequentes.....	85
Figura 10 - Gráfico de Acurácia entre técnicas de transformação.....	90
Figura 11 - Gráfico de média de FP	91

LISTA DE TABELAS

Tabela 1 - Representação Binária.....	36
Tabela 2- Matriz de Confusão n classes.....	39
Tabela 3 - Distribuição dos documentos	42
Tabela 4 - Frequências de CID	68
Tabela 5 - Maiores frequências individuais e associadas	68
Tabela 6 - Composição dos conjuntos de interesse com base na frequência	69
Tabela 7 - Exemplo de termos excluídos.....	72
Tabela 8 - Composição dos conjuntos de interesse	73
Tabela 9 - Percentuais de Acurácia conjunto não balanceado.....	77
Tabela 10 - Resultados para as métricas obtidas com conjunto não balanceado	77
Tabela 11 - Percentuais de Acurácia conjunto balanceado	78
Tabela 12 - Resultados para as métricas de dados balanceados	78
Tabela 13 - Matriz de confusão conjunto <i>M Não M</i>	80
Tabela 14 - Resultados de classificação conjunto não balanceado	82
Tabela 15 - Lista de métricas para o conjunto 5 CIDs mais Frequentes não balanceado frequência 5.....	83
Tabela 16 - Resultados de classificação para classes balanceadas	84
Tabela 17 - Resultados de classificação para o conjunto 5 CIDs mais Frequentes balanceado frequência 5	84
Tabela 18 - Matrizes de confusão 5 CIDs mais Frequentes.....	85
Tabela 19 - Quantidade média de FP por classe	86
Tabela 20 - Percentuais de Acurácia para conjunto não balanceado.....	87
Tabela 21 - Resultados de classificação para conjunto não balanceado frequência 5.....	87
Tabela 22 - Percentuais de Acurácia conjunto balanceado	88
Tabela 23 - Métricas obtidas para conjunto balanceado frequência 5.....	88
Tabela 24 - Matriz de Confusão geradas.....	89

Tabela 25 - Quantidade de regras extraídas	97
Tabela 26 - Lista de regras extraídas para o conjunto dos 5 CID mais Frequentes	98
Tabela 27 - Quantidade de regras extraídas conjunto 3 Ms mais Frequentes....	99
Tabela 28 - Lista de regras extraídas para o conjunto dos 3 CID mais frequentes	101

LISTA DE QUADROS

Quadro 1 - Exemplo de transações.....	47
Quadro 2 - Síntese dos trabalhos.....	60
Quadro 3 - Trecho da base de dados original.....	66

LISTA DE EQUAÇÕES

Equação 1	34
Equação 2	35
Equação 3	35
Equação 4	36
Equação 5	36
Equação 6	40
Equação 7	40
Equação 8	40
Equação 9	41
Equação 10	42
Equação 11	45
Equação 12	45
Equação 13	45
Equação 14	46
Equação 15	46

LISTA DE SIGLAS E ABREVIATURAS

BD	Banco de Dados
CAR	<i>Class Associations Rules</i>
CFM	Conselho Federal de Medicina
CID	Classificação Internacional de Doenças
DATASUS	Departamento de Informática do Sistema Único de Saúde
DCBD	Descoberta de Conhecimento em Bases de Dados
DCT	Descoberta de Conhecimento em Texto
DECS	Descritores em Ciências da Saúde
FN	<i>False Negatives</i>
FP	<i>False Positives</i>
HTML	<i>Hyper Text Markup Language</i>
IDF	<i>Inverse Document Frequency</i>
MT	Mineração Textual
NER	<i>Named Entity Recognition</i>
PIPCA	Programa Interdisciplinar de Pós-Graduação em Computação Aplicada
PLN	Processamento de Linguagem Natural
PME	Prontuário Médico Eletrônico
PMP	Prontuários Médicos de Pacientes
RI	Recuperação de Informações
ROC	<i>Receiver Operating Characteristic</i>
SQL	<i>Structure Query Language</i>
TF	<i>Term Frequency</i>
TI	Tecnologia da Informação
TN	<i>True Negatives</i>
TP	<i>True Positive</i>
WHO	<i>World Health Organization</i>

SUMÁRIO

1 INTRODUÇÃO	16
1.1 MOTIVAÇÃO	16
1.2 OBJETIVOS GERAIS E ESPECÍFICOS DO TRABALHO.....	17
1.3 ESTRUTURA DO TRABALHO	18
2 PRONTUÁRIO MÉDICO ELETRÔNICO	20
2.1 PRONTUÁRIO MÉDICO DO PACIENTE	20
2.2 PRONTUÁRIO MÉDICO ELETRÔNICO	22
2.3 CLASSIFICAÇÃO INTERNACIONAL DE DOENÇAS	26
3 FUNDAMENTOS TEÓRICOS	29
3.1 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS.....	29
3.2 DESCOBERTA DE CONHECIMENTO EM TEXTOS	31
3.2.1 Pré-Processamento	31
3.2.1.1 Remoção de <i>Stopwords</i>	32
3.2.1.2 Tokenização.....	32
3.2.1.3 <i>Stemming</i>	33
3.2.1.4 Seleção de Termos Relevantes.....	34
3.2.2 Transformação dos Dados	35
3.2.3 Mineração Textual	37
3.2.3.1 <i>Support Vector Machines</i>	41
3.2.3.2 Algoritmo <i>Apriori</i>	46
4 TRABALHOS RELACIONADOS	49
4.1 RECUPERAÇÃO DE INFORMAÇÕES EM TEXTOS LIVRES DE PME POR MEIO DE SEMELHANÇA	49
4.2 DESCOBERTA DE CONHECIMENTO EM PRONTUÁRIO ELETRÔNICOS	51

4.3 CLASSIFICAÇÃO DE CÓDIGOS MÉDICOS PARA DESCOBERTA DE RELACIONAMENTOS	54
4.4 EXTRAÇÃO DE REGRAS EM CONTEÚDO TEXTUAL COM ALGORITMO <i>APRIORI</i>	56
4.5 CONSIDERAÇÕES SOBRE OS TRABALHOS	59
5 DESCOBERTA DE CONHECIMENTO APLICADA A DADOS DE SAÚDE.....	63
5.1 BASE DE DADOS.....	64
5.1.1 Aquisição da Base de Dados.....	64
5.1.2 Descrição da Base de Dados.....	65
5.2 SELEÇÃO DOS DADOS	65
5.3 COMPOSIÇÃO DOS CONJUNTOS DE INTERESSE	68
5.4 PRÉ-PROCESSAMENTO	70
5.4.1 Tokenização	72
5.4.2 Remoção de <i>Stopwords</i>.....	72
5.4.3 <i>Stemming</i>	72
5.4.4 Redução de Dimensionalidade.....	73
5.4.5 Seleção de Termos Relevantes	74
5.5 EXPERIMENTOS REALIZADOS.....	74
5.5.1 Classificação.....	75
5.5.1.1 Experimentos sobre o conjunto <i>M Não M</i>	76
5.5.1.2 Experimentos sobre o conjunto dos 5 CIDs mais Frequentes	81
5.5.1.3 Experimentos sobre o conjunto 3 <i>Ms</i> mais Frequentes.....	86
5.5.1.4 Discussão dos Resultados da Classificação.....	91
5.5.2 Regras de Associação	95
5.5.2.1 Experimentos sobre o conjunto 5 CIDs mais Frequentes.....	96
5.5.2.2 Experimentos sobre o conjunto dos 3 <i>Ms</i> mais Frequentes	99
5.5.2.3 Discussão dos Resultados de Regras de Associação.....	102

6 CONCLUSÃO	106
6.1 TRABALHOS FUTUROS.....	108
REFERÊNCIAS	110
 ANEXOS	
ANEXO A – RELAÇÃO DE <i>STOPWORDS</i>	115
ANEXO B – DECLARAÇÃO DA ESPECIALISTA.....	118

1 INTRODUÇÃO

Os avanços obtidos na área de Tecnologia da Informação (TI), aliados ao baixo custo de aquisição de equipamentos de última geração, propiciam às instituições das mais diversas áreas da sociedade o armazenamento de um expressivo volume de informações de variados tipos. A área da saúde não foge a esse contexto.

As instituições da área da saúde, mais especificamente os hospitais, armazenam informações diversas a respeito da vida clínica de cada paciente no formato de Prontuário Médico de Paciente (PMP). Um problema enfrentado pelos hospitais, sejam eles da esfera pública ou privada, é o acúmulo dos PMP gerados, mas que pode ser equacionado pela utilização de sua versão informatizada, o Prontuário Médico Eletrônico (PME).

Assim como a versão física, o PME armazena toda informação clínica a respeito do paciente, e levando-se em conta a quantidade de pacientes que buscam tratamentos nos hospitais, estamos diante de um expressivo volume de informações. Parte dessas informações está armazenada sobre formato estruturado (ou predefinido) e sob a forma de textos livres (ou descritivos). Buscando investigar informações armazenadas em bases de dados, como as informações que compõem um PME, foram desenvolvidas técnicas visando à Descoberta de Conhecimento em Bases de Dados (DCBD). A DCBD tem por objetivo a extração e análise de padrões obtidos em banco de dados (BD), para uma possível produção de conhecimento (FELDMAN, 1995). Esses padrões são extraídos através da aplicação de técnicas computacionais em campos que possuem uma pré-formatação dos valores inseridos. Similarmente existem técnicas desenvolvidas para a Descoberta de Conhecimento em Textos (DCT), onde não existe nenhum controle dos dados armazenados (TAN, 1999), tal como os campos livres existentes no PME.

1.1 MOTIVAÇÃO

A saúde é um segmento crítico da sociedade, que possui uma grande concentração de estudos buscando soluções computacionais viáveis para os mais diversos tipos de problemas relativos ao gerenciamento de informações.

Uma dessas soluções é o PME, que é a evolução natural do PMP (meio físico) para o meio digital. Dessa forma, toda a informação relativa à vida clínica do paciente fica armazenada no PME e, por esse motivo, possui uma grande variedade em seu teor (MASSAD; MARIN; AZEVEDO NETO, 2003).

Para o armazenamento dos dados, são utilizados campos com conteúdo pré-estruturado e descritivo com relacionamento direto entre estes. Com a intenção de conhecer melhor o conteúdo de bases de dados que compõem o PME, através da análise de possíveis padrões a serem descobertos, bem como a produção de conhecimento a partir destes, os dados sob a forma de texto livre oriundos de sistemas de PME são factíveis de investigação através da aplicação de técnicas de DCT. Por ser um processo não trivial, há uma ausência de mecanismos para exploração desse conteúdo, e também se deve considerar que uma base de dados com volume expressivo de informações descritivas pode servir para avaliar o poder do classificador na distinção de classes.

Consultada a literatura, é visível o aumento no número de trabalhos que abordam DCT em PME. Alguns trabalhos, como o de Martha, Campos e Sigulem (2010), propõem melhorar os resultados de recuperação de textos da área da biomédica. Já o trabalho realizado com base em conteúdo descritivo de um sistema de PME associado a campos estruturados (YAN et al., 2010) propõe uma variação de uma técnica clássica para a classificação de doenças.

Existe um déficit no que tange a trabalhos com foco na exploração de conteúdo descritivo, através de associações entre termos, na tentativa de extrair padrões pouco comuns ou desconhecidos no domínio pesquisado, e que por consequência conduza a produção de conhecimento. Os resultados alcançados podem justificar a extensão das técnicas abordadas neste trabalho, a sistemas de PME, tendo foco na validação do conteúdo textual relacionado a diagnósticos, busca de inconsistências ou classificação de pacientes quanto a doenças da área da Reumatologia (no contexto deste trabalho).

1.2 OBJETIVOS GERAIS E ESPECÍFICOS DO TRABALHO

O objetivo geral deste trabalho é propor um processo investigativo e exploratório de dados descritivos de um sistema de PME concentrados no

domínio da Reumatologia, visando à realização de experimentos computacionais para avaliar a capacidade de técnicas de mineração de dados textuais relacionados a classes específicas da Classificação Internacional de Doenças (CID) no que se refere ao domínio de interesse, bem como a aplicação de regras associativas para investigar possíveis relacionamentos de conteúdo descritivos com a ocorrência de determinadas CIDs.

Para atingir tal objetivo foram investigadas técnicas clássicas de mineração textual com vistas à adequação das técnicas existentes, na tentativa de explorar, via experimentação computacional, a descoberta de conhecimento assim como avaliar e sugerir técnicas mais promissoras no tratamento de dados de PMEs.

Como objetivos específicos do trabalho, destacam-se:

- Composição de conjuntos de dados baseados nos códigos CID de maior frequência exclusiva;
- Observar o desempenho do classificador na distinção de diferentes classes relacionadas ao domínio de interesse com base no conteúdo textual;
- Extração de padrões através de regras associativas que possam conduzir a produção de conhecimento;
- Realização de experimentos visando à validação do processo exploratório executado, com o emprego de técnicas de validação cruzada;
- Sugerir um novo conhecimento extraído da base, com a finalidade de apoiar o profissional de saúde na identificação de doenças relacionadas à Reumatologia, assim como a identificação de pacientes com sintomas relacionados.

1.3 ESTRUTURA DO TRABALHO

O trabalho segue com a descrição do universo do problema no capítulo 2, descrevendo o PMP, PME e a CID.

A sequência do trabalho apresenta uma visão geral do processo de mineração textual, no capítulo 3, abordando os principais conceitos para a compreensão do processo.

Os trabalhos correlatos, ou seja, aqueles trabalhos que aplicam tais técnicas de mineração textual para investigação de problemas relacionados aos PMEs, são descritos no capítulo 4.

O capítulo 5 apresenta os experimentos e resultados obtidos, descrevendo antes a base de dados, estatísticas, técnicas de Pré-Processamento e Mineração de Dados utilizadas, finalizando com uma discussão dos resultados.

A apresentação de um relato das conclusões do trabalho e suas perspectivas de possíveis extensões são descritas no capítulo 6.

2 PRONTUÁRIO MÉDICO ELETRÔNICO

Para uma maior compreensão do domínio do problema abordado, este capítulo apresenta a descrição dos principais conceitos associados a Prontuário Médico do Paciente e Prontuário Médico Eletrônico, abordando o surgimento de ambos, fatores positivos e negativos relacionados a cada um, fornecendo uma visão sobre o gerenciamento de informações clínicas de paciente em instituições de saúde. A última seção deste capítulo é reservada para a descrição da CID, relatando sua origem e evolução até o padrão atualmente usado.

2.1 PRONTUÁRIO MÉDICO DO PACIENTE

O primeiro relato de registro de informação médica sobre um paciente, segundo o “Prontuário Médico do Paciente: Guia para uso Prático” (CRM, 2006), é no período de 3.000 a 2.500 a.C.. Eram informações cirúrgicas descritas por um médico egípcio em papiros. Após foram encontradas anotações feitas por Hipócrates sobre doentes, datadas de 460 a.C. O próximo registro é no ano 1137, são anotações relativas aos pacientes realizadas no Hospital São Bartolomeu em Londres. Em 1897 o Hospital Geral de Massachussets é o primeiro a organizar um serviço de arquivo médico (CRM, 2006).

Segundo o Conselho Regional de Medicina (CRM) de Brasília, o PMP foi desenvolvido por médicos e enfermeiros para garantir a memória dos fatos e eventos na vida clínica de cada paciente, de forma sistemática, onde todos os profissionais envolvidos no tratamento pudessem ter acesso às informações (CRM, 2006).

Em 1944, o PMP foi introduzido no Brasil após estudos de especialização em sistemas de arquivo e classificação de observações e foi implantado no hospital de Clínicas de São Paulo (CRM, 2006).

O PMP é uma importante ferramenta utilizada pelos profissionais da área de saúde no decorrer do tratamento de cada paciente. Conforme a Resolução nº1.638/2002 do Conselho Federal de Medicina (CFM) (CFM, 2002), o PMP é um documento único; é nele que ficam registradas todas as informações relativas a procedimentos, prescrições de medicações, dietas, evoluções, exames, dados

cadastrais, estabelecendo-se, dessa forma, um meio único de registro e acesso a informações, utilizadas pelos profissionais envolvidos no tratamento (CRM, 2006). Além da utilização pela equipe clínica, o PMP pode ser utilizado para rotinas de ordem administrativas, legais, ensino, pesquisa e estatística. O grau de importância do PMP fica evidente pela obrigatoriedade de toda instituição de saúde possuir uma Comissão Permanente de Revisão de Prontuário implantado pelo CFM (CFM, 2002).

O CRM, através do Processo Consulta nº 1.401/02 - Conselho Federal de Medicina (30/02) de 21 de junho de 2002, conceitua o PMP como um documento único, composto por várias informações registradas. Tem sua origem em fatos e situações diversas sobre o tratamento e/ou assistência prestados ao paciente e serve como fator facilitador na troca de informações entre os profissionais da área da saúde, devendo ser de uso reservado e sigiloso (CRM, 2006).

Uma vez que as informações estão armazenadas apenas em meio físico, alguns pontos negativos podem ser constatados. É possível citar como exemplo:

- Por ser meio físico, somente pode estar disposto para análise por apenas um profissional, ou ainda indisponível, pois pode estar em trânsito entre arquivo médico e demais setores dentro da instituição;
- Por sua natureza, pode ser retirado de seu lugar, intencionalmente ou não, e, caso não haja uma definição e execução de regras que componham um protocolo de utilização, pode ser extraviado;
- Por ser preenchido manualmente através de escrita, é possível que haja rasuras na caligrafia do profissional, podendo gerar dificuldade na interpretação das informações registradas ou conduzir a interpretações ambíguas;
- A necessidade de espaço para o arquivamento e guarda do PMP, uma vez que o mesmo deve ser guardado por 20 anos conforme o art. 8º da Resolução CFM nº 1.821/07 (CFM, 2007). Para mensurar o problema, em consulta realizada junto ao Serviço de Arquivo Médico do Hospital Materno Infantil Presidente Vargas, estima-se a quantidade aproximada de 12.600 novos prontuários gerados ao ano;
- É possível também mencionar os riscos de incêndios, vazamento de canos, etc., que podem causar grande prejuízo caso não exista uma

rotina de microfilmagem (*backup*) para prever possíveis perdas de informações;

- Por não possuir um dispositivo eletrônico que gerencie, não é possível criar dispositivos de alertas para diferentes situações.

2.2 PRONTUÁRIO MÉDICO ELETRÔNICO

O Prontuário Médico Eletrônico é um sistema informatizado, que através da utilização de recursos computacionais, permite que sejam armazenadas informações relativas à saúde e tratamentos de um determinado paciente. Em outras palavras, o PME é uma versão digital do prontuário tradicional, ficando todos os registros realizados pelo profissional da área da saúde armazenados em um banco de dados (BD), e gerenciado por um sistema que, através de sua interface, proporciona a inserção, alteração, consulta de dados, estatísticas, etc.

A utilização do PME está amparada pela resolução 1.639/2002 do Conselho Federal de Medicina que aprova as normas técnicas para o uso de sistemas informatizados para a guarda e manuseio do prontuário médico.

No PME, da mesma forma que o PMP, devem constar obrigatoriamente segundo a resolução nº 1.638/2002 do CFM, os seguintes dados:

- Dados de identificação do paciente – Nome completo, data de nascimento, sexo, nome da mãe, naturalidade, endereço completo;
- Anamnese, exame físico, exames complementares solicitados, resultados, laudos, hipóteses diagnosticadas, diagnóstico definitivo e tratamento efetuado;
- Evolução diária do paciente, com data e hora, discriminação de todos os procedimentos aos quais o mesmo foi submetido e identificação dos profissionais (enfermeiros, fisioterapeutas, fonoaudiólogos) que os realizaram, assinados eletronicamente quando elaborados e/ou armazenados em meio eletrônico;
- Nos casos emergenciais, nos quais seja impossível a coleta do histórico clínico do paciente, deverá constar relato médico completo de todos os procedimentos realizados e que tenham possibilitado o diagnóstico e/ou a remoção para outra unidade;

- Formulários e fichas de avaliação diversos, como de descrição cirúrgica, interconsultas, ficha de anestesia, etc.

Os dados a serem inseridos no PME são armazenados em campos estruturados, ou seja, campos com valores predefinidos, que garantem uma uniformidade dos dados, porém impondo limites ao profissional que utiliza o sistema. Para suprir essa deficiência, proporcionando uma maior liberdade de expressão (similar ao prontuário físico) para compor a informação que será inserida no sistema, são disponibilizados campos de texto livre ou descritivos. Esses campos são espaços onde o profissional de saúde que manuseia o sistema pode, na forma de texto livre, fazer registros com maior quantidade de detalhes, tendo como restrição apenas o grau de abstração do profissional, algo que os campos estruturados não permitem. Pela série de avaliações, exames, observações e questionamentos realizados pelo médico durante uma consulta ou atendimento, e como cada um desses itens influenciam aos outros, pode-se afirmar que é muito difícil para o profissional da área da saúde realizar um registro que contemple todos os itens citados, de uma forma clara e precisa através de campos com formatação predefinida, justificando assim a utilização dos campos de textos livre. Porém é possível elencar como pontos negativos em campos de textos livres, principalmente no que tange a DCT, conforme segue (CRM, 2006):

- A diversidade de profissionais que utilizam o prontuário eletrônico (médicos, enfermeiros, fisioterapeutas), cada um com suas expressões próprias;
- A linguagem médica é rica em expressões próprias, abreviações, etc., que causam problemas que devem ser pré-processados;
- Geralmente esses campos não possuem um corretor ortográfico, que ajudaria a evitar erros de ortografia (por exemplo, corrigiria sefaleia para cefaleia) no momento da inserção dos dados (MARTHA; CAMPOS; SIGULEM, 2010);
- Muitas palavras presentes nos campos de texto livre possuem um grande índice de repetição e não possuem nenhum valor representativo, por isso devem ser retiradas do texto (WIVES, 2004);

- Por sua natureza, esses campos não obedecem à estrutura de dados alguma, e as informações vão sendo armazenados conforme o profissional que opera o sistema vai alimentando o mesmo, portanto impossibilitando a criação de gráficos e/ou estatísticas de uma forma automática (MARTHA; CAMPOS; SIGULEM, 2010);
- Em situações de emergência, pode ocorrer que o PME seja criado com ausência de informações.

Como todo sistema eletrônico, o PME possui uma série de quesitos de segurança que devem ser observados. O PME pode servir de prova judicial, por exemplo, em processos de erro médico, devendo assim possuir um mecanismo que possibilite a identificação de forma incontestável de qual profissional é responsável por qual informação. Essa identificação de quem registrou qual(is) informação(ões) ou qual(is) prescrição(ões) inseriu no PMP se dá através da assinatura do profissional responsável pelo registro acompanhado do carimbo. No PME de forma análoga, a identificação do profissional se dá pela assinatura digital, com padrão de Infraestrutura de Chaves Públicas do Brasil (ICP), garantindo dessa forma a validade jurídica do PME (CFM, 2007).

Especificamente os quesitos de segurança que devem ser observados são:

- *Integridade*: o sistema deve assegurar que os dados não sejam alterados por pessoas não autorizadas;
- *Confidencialidade*: o sigilo médico deve ser preservado. Dessa forma, as informações não devem ser acessadas por pessoas não autorizadas;
- *Autenticação*: o PME deve possuir recursos para identificar quem está acessando os dados;
- *Autorização*: cada profissional cadastrado deve estar associado a uma lista de privilégios de acesso e/ou restrições;
- *Auditoria*: rotina pela qual o sistema assegura que a atividade do profissional possa ser registrada para possíveis investigações sobre ações no PME.

Com auxílio da Tecnologia da Informação (TI), o PME deixa de ser um armazenador de informações e começa a interagir com o profissional. Exemplificando, por possuírem diversos campos pré-estruturados, é possível estabelecer limites para resultados de exames, assim quando um valor de um resultado fica fora dos limites previstos, o sistema pode acionar uma função de alerta para o profissional.

O PME possui vantagens que são inerentes a todo sistema computacional (MASSAD; MARIN; AZEVEDO NETO, 2003):

- Os dados inseridos nunca terão problema de serem ilegíveis;
- A forma de localização de um determinado paciente é rápida, dependendo somente das funções de busca do PME utilizado;
- O acesso ao PME é realizado por um computador conectado a uma rede de dados, não dependendo dessa forma de: solicitação, localização e transporte;
- O acesso às informações pode ocorrer por mais de um profissional ao mesmo tempo;
- Elimina a redundância de informações e solicitação de exames;
- Possibilidade de integração com outros sistemas da área da saúde;
- Facilidade de pesquisa a grandes grupos a partir de determinadas características;
- Ocupa pouco espaço físico.

Apesar das vantagens, alguns obstáculos devem ser vencidos quando se trata da implantação ou aquisição de um PME. Muitas vezes, a dificuldade de compreender as soluções, o tempo necessário para implantação e treinamento de pessoal, inclusive do médico, é escasso. Outras vezes, é a quebra do paradigma, mudando a forma de trabalho de muitos profissionais (SABBATINI, 1998).

Em contraponto às vantagens, também podem ser elencados alguns pontos negativos no PME:

- Não conseguir acesso às informações por falha de equipamento;
- Não conseguir acesso aos sistemas por falta de energia elétrica;

- Resistência de alguns profissionais ao uso por falta de conhecimento;
- Investimentos de *hardware*, *software*, treinamento.

Atualmente não existe uma obrigatoriedade ou uma previsão para as instituições de saúde abolirem o PMP e migrarem para o PME, ficando isso a critério de cada uma.

Apesar dos fatos supracitados, é uma tendência as diversas instituições em saúde migrarem para o prontuário eletrônico uma vez que as vantagens obtidas com sua implantação permitem maior agilidade dos procedimentos realizados, acesso às informações com uma velocidade expressiva, principalmente se tratando de uma área onde dependendo da situação, segundos são preciosos e podem fazer a diferença na vida de um paciente.

É crescente a quantidade de estudos e produção científica que envolve o PME. Um estudo realizado envolvendo 174 trabalhos a partir do ano de 1995 até 2007, fazendo uma revisão na produção científica sobre DCT em bases textuais de PME, comprova isso (MEYSTRE et al., 2008).

Conforme exposto nos objetivos, este trabalho visa tratar a informação textual relacionada com códigos CID, que compõem a base de dados. Dessa forma, é apresentada na próxima seção uma breve descrição conceitual da CID, bem como aspectos históricos da sua constituição e organização.

2.3 CLASSIFICAÇÃO INTERNACIONAL DE DOENÇAS

A Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde, também conhecida simplesmente como CID, é um sistema que visa padronizar e fornecer códigos para diagnósticos de doenças, sintomas, queixas e procedimentos (DATASUS, 2008).

O CID teve sua origem no século XVII com Jonh Graunt, em um estudo estatístico sobre diversas causas de mortes. No século posterior, François Bosieer de Lacroix tentou uma classificação sistemática de doenças. Nos séculos posteriores, continuaram a surgir diversos relatos de estudos para a formulação de uma classificação, até que um evento em destaque ocorreu em 1853, em Bruxelas, o primeiro Congresso Internacional de Estatística, onde se

discutiu a necessidade de uma classificação de causa de mortes por doenças de uma forma padronizada. Em 1858, no segundo Congresso Internacional de Doenças, foi apresentado por Farr uma classificação que dividia as doenças causadoras de mortes em cinco grupos. Essa classificação foi adotada e passou por revisões nos anos de 1874, 1880 e 1886. Farr reconhecia que o sistema de classificação deveria ser ampliado para as demais doenças (WHO, 2004). Somente em 1898, na reunião da Associação de Saúde Pública, em Ottawa (Canadá), foi recomendado que Canadá, México e Estados Unidos adotassem uma classificação para causa de mortes apresentada por Jacques Bertillon do Instituto Internacional de Estatística e que fosse revisada a cada dez anos. As três primeiras revisões ocorreram em 1900, 1909 e 1920, paralelamente a uma classificação de morbidades. Em 1946, uma comissão da World Health Organization (WHO) revisou as listas internacionais de causas de mortes aceitando a classificação dos Estados Unidos. A partir desse momento, surgiu a Classificação Internacional de doenças, traumatismos e causas de morte. Porém, a partir da sexta revisão em 1948, começaram a se estabelecer comitês nacionais de estatísticas vitais e de saúde, trabalhando de forma articulada com a WHO no formato atual. A décima revisão (atual) foi realizada em 1999 (WHO, 2004).

Alguns fatores influenciaram no atraso das revisões, como a expansão do uso da classificação CID para diversos países e a necessidade de mais tempo para avaliar relatórios e críticas de diversos centros colaboradores da WHO (ENSP, 2006).

Neste trabalho sempre que for realizada uma referência à Classificação Internacional de Doenças, será feita pela sua abreviação, precedido do número de sua revisão: CID-8, CID-9, CID-10.

Basicamente, a CID é dividida em três volumes, conforme segue (DATASUS, 2008):

- *Volume I*: contém a classificação propriamente dita, composta por três caracteres, uma letra seguido de 2 algarismos; ainda pode se ter uma subcategoria, nesse caso a subcategoria é um algarismo antecedido por um ponto;

- *Volume II*: contém orientações, regras, guias para os usuários;
- *Volume III*: composto por um índice alfabético.

O volume I é dividido em 22 capítulos, sendo que as doenças são agrupadas em uma ramificação principal denominada agrupamento. Cada agrupamento é dividido em n categorias mais específicas de doenças. Por último, cada categoria é dividida em subcategorias finais, que correspondem ao código da doença. A Figura 1, retirada do site do Departamento de Informática do Sistema Único de Saúde (DATASUS) (DATASUS, 2008), demonstra essas divisões. É possível observar na Figura 1, na coluna à esquerda, a lista dos 22 capítulos, sendo selecionado o capítulo XIII- Doenças do Sistema Osteomuscular e do tecido conjuntivo. Localizado na parte superior mais centralizado na figura, um dos diversos agrupamentos de doenças, nesse caso foi selecionado o agrupamento M15-M19 Artroses.

Figura 1 - Sub-divisões do CID (DATASUS, 2008)

Fonte: DATASUS (2008)

Centralizado na Figura 1, estão diversas categorias, com destaque para a categoria M.15.0. A subcategoria é o valor ao lado de cada categoria separado por um ponto.

3 FUNDAMENTOS TEÓRICOS

Este capítulo descreve a fundamentação teórica envolvida neste trabalho, detalhando o processo de Descoberta de Conhecimento em Bases de Dados e Descoberta de Conhecimento em Textos. O segundo processo é enfatizado por ser o foco deste trabalho, onde são descritas algumas técnicas de Pré-Processamento de Textos e técnicas de Mineração Textual que podem ser empregadas.

3.1 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

É denominado de Descoberta de Conhecimento em Bases de Dados o processo não trivial, pelo qual é possível a extração de padrões que estão implícitos em um expressivo volume de informação armazenado em meio digital, passando por uma análise, focada em um problema específico e com o objetivo de produção de conhecimento (FAYYAD, 1996).

Para Feldman e Sanger (2006), DCBD é um paradigma que se concentra na exploração informatizada de grandes quantidades de dados, ocorrência de padrões nesses dados e suas causas, devido à falta de capacidade dos métodos tradicionais de manipulação de informação operar grandes volumes de dados. Dessa forma, o processo de DCBD faz uso de ferramentas robustas para indução de dados. O conhecimento adquirido é utilizado como sustentação no processo de tomada de decisão.

Como se trata de um processo, a DCBD tem uma série de etapas que devem ser cumpridas, e são caracterizadas por possuírem uma natureza iterativa (apesar de serem etapas sequenciais, pode-se retornar ao começo e reiniciar o processo, ou apenas retornar a etapas anteriores) e interativa (porque o usuário pode optar por retomar qualquer uma dessas etapas).

Para iniciar o processo, é necessário ter um entendimento do domínio dos dados a serem investigados, bem como do objetivo que se deseja atingir. Uma vez de posse dessas informações, inicia-se o processo, com as seguintes etapas conforme sugerido por Han e Kamber (2006):

- *Seleção dos dados*: etapa em que os dados a serem utilizados no processo são selecionados de acordo com os objetivos estabelecidos, e ainda é possível realizar uma nova seleção nesses dados formando-se subconjuntos;
- *Pré-processamento*: etapa responsável pela limpeza da base de dados. Junto com a etapa de mineração, é uma das etapas mais importantes do processo, pois é nessa etapa que será garantida a qualidade dos dados que serão submetidos às etapas posteriores. Portanto, se os dados seguem no processo com a presença de ruídos, redundâncias, inconsistências, o processo não conseguirá obter sucesso na extração de padrões, obtendo como resultados valores que não condizem com a realidade;
- *Transformação*: essa etapa tem o objetivo de consolidar os dados em um formato apropriado para a etapa de mineração; isso pode envolver remoção de ruídos, generalização dos dados, criação de novos atributos;
- *Mineração de dados*: etapa em que serão aplicadas as técnicas para mineração dos padrões, levando em consideração o objetivo a ser atingido como critério de escolha da mesma;
- *Pós-processamento*: etapa em que os padrões obtidos são interpretados para produção do conhecimento.

A visualização ou representação do conhecimento aparece como uma última etapa do processo DCBD e diz respeito à forma de como será realizada a representação do conhecimento gerado (HAN; KAMBER, 2006).

O processo de DCBD descrito nos parágrafos anteriores é aplicado sempre em bases de dados que possuem as informações armazenadas em campos com formatação previamente estabelecida, ou seja, campos estruturados. Porém existe uma grande quantidade de informação, como notícias, artigos, páginas WEB, laudos de exames e diagnósticos médicos que estão armazenados no formato de textos, e com potencial de possuir conhecimento implícito em seus registros (HAN; KAMBER 2006). Por serem informações que variam em seu conteúdo e volume, sem possuir um formato fixo, a preparação dos dados deve ser mais aprimorada. O processo de extração

de conhecimento desse tipo de informação é denominado de Descoberta de Conhecimento em Textos e será descrito na próxima seção.

3.2 DESCOBERTA DE CONHECIMENTO EM TEXTOS

A exemplo da DCBD, o objetivo do processo de DCT é o mesmo: a descoberta de conhecimento gerado a partir da análise de padrões implícitos obtidos em grandes volumes de informação. A diferença entre esses processos é que o segundo está concebido para inferir em dados estruturados, enquanto, no primeiro, os dados não possuem estruturação senão as normas gramaticais. Essas informações podem ter origem em notícias, trabalhos de pesquisa, livros, bibliotecas digitais, *e-mails*, páginas da WEB, e mais especificamente como foco deste trabalho, campos de texto livre (ou descritivos) em sistemas de Prontuário Médico Eletrônico. Há estimativas que 85% dos dados armazenados hoje são em formato texto (SAYED, 2008), e potencialmente esses campos possuem uma grande quantidade de conhecimento implícito, entretanto de difícil análise (JOHN, 2002).

3.2.1 Pré-Processamento

A etapa de Pré-processamento é considerada uma das etapas mais importantes do processo de DCT. Assim como no processo de DCBD, de forma análoga, no processo DCT o pré-processamento visa à preparação dos textos que formam a base de dados textual para as etapas seguintes, transformando uma base de dados sem estrutura em uma coleção de dados tratáveis agregando qualidade (COHEN; HERSH, 2005).

Essa preparação passa por eliminação de palavras com pouco ou nenhum significado (*Stopwords*), remoção de variações morfológicas (*Stemming*) e seleção das palavras que melhor representam o texto (termos relevantes).

As etapas do pré-processamento de bases textuais são abordadas a seguir.

3.2.1.1 Remoção de *Stopwords*

Denomina-se Remoção de *Stopwords* a retirada de palavras ou termos sem valor representativo e com grande frequência em todos os textos que compõem a base textual. A lista de palavras a serem removidas, também conhecida como *stop list*, é composta geralmente por pronomes, advérbios, preposições, conjunções e artigos, etc., e a sua remoção da base de dados minimiza substancialmente o tamanho do texto. As palavras que compõem a lista de *Stopwords* podem ser diferentes dependendo do contexto que está sendo pesquisado (HAN; KAMBER 2006). As palavras que farão parte da lista de exclusão, também conhecidas como *stop list*, podem ser selecionadas de uma forma manual, sendo que o pesquisador avalia uma a uma e decide quais palavras farão parte da mesma.

3.2.1.2 Tokenização

Essa técnica tem o objetivo de reduzir um texto em unidades mínimas. O resultado da aplicação pode ser: uma única palavra (denominada unigrama), duas palavras (bigrama) e três ou mais palavras (multigramas).

A pontuação, bem como os espaços em branco presentes no texto, agem como um facilitador na identificação dos *tokens*, pois eles efetuam a separação das palavras, assumindo a forma nesse caso de delimitadores. Porém se tratando de pontuação, deve se observar que um ponto utilizado para término de uma sentença, também é utilizado para abreviações (SOARES, 2008). A fase de tokenização, segundo Konchady (2006), é composta por seis passos:

- Geração de *tokens* tendo como base os delimitadores;
- Identificação de abreviações;
- Identificação de palavras combinadas;
- Identificação de símbolos da internet;
- Identificação de números;
- Identificação de *tokens* multi-vocabulares.

3.2.1.3 Stemming

Uma das abordagens utilizadas, denominada aplicação de *Stemming*, visa eliminar variações morfológicas que uma palavra possua (ou promover a redução a seu radical), através da remoção de seu sufixo. Com a eliminação do sufixo, palavras que antes possuíam diferentes representações, passam a ter somente uma, aumentando sua representatividade no contexto pesquisado (BAEZA; RIBEIRO, 1999). Após a aplicação da técnica de *Stemming*, palavras como *caminhar*, *caminhando* e *caminha* passam a ser representadas pelo termo *caminh*.

O algoritmo de Orengo e Huyck (2001) faz uso de 199 regras para a remoção de sufixos. Essas regras possuem exceções que são tratadas com o uso de um dicionário com 32.000 termos. Esse algoritmo apresenta uma sequência de 8 passos que executam o processo de remoção de sufixo:

- *Redução do plural*: eliminação e ou substituição da letra 's' ou sufixo que represente o plural;
- *Conversão de palavras de gênero feminino para masculino*: conversão do gênero masculino para o feminino removendo 'a' letra a do final da palavra, nesse passo somente as palavras com sufixos mais comuns são removidos;
- *Eliminação de sufixos que caracterizem advérbios*: remoção do sufixo 'mente' desde que não façam parte de uma lista de exceção;
- *Redução de aumentativos e diminutivos*: remoção dos sufixos de aumentativos e diminutivos mais comuns;
- *Remoção dos sufixos dos substantivos mais comuns* (devido à grande quantidade de variações da língua portuguesa): remoção de 61 possíveis sufixos para adjetivos e substantivos;
- *Redução de formas verbos a sua raiz*: redução da forma verbal ao seu radical;
- *Testes de vogais*: remoção das vogais: 'a', 'e', 'o' das palavras que não foram tratadas pelos passos anteriores;
- *Remoção de acentos fonéticos*: remoção e ou substituição dos sufixos que denotam plural.

Já o algoritmo de Porter (1980), desenvolvido para a língua Inglesa e muito utilizado na literatura, efetua a remoção de sufixos através de uma lista de regras que são aplicadas sobre as palavras. A aplicação desse algoritmo consiste na utilização de uma lista de sufixos associados a regras, para remoção ou não desses sufixos.

A adaptação desse algoritmo para sua aplicação na língua Portuguesa segue 5 passos (VIEIRA; VIRGIL, 2007):

- Remoção de sufixos;
- Se o primeiro passo não executou nenhuma alteração são removidos os sufixos verbais;
- Remoção do sufixo i, se precedido de c;
- Remoção dos sufixos residuais os, a, i, o, á, í, ó;
- Substituição da cedilha e remoção dos sufixos e, é, ê e.

3.2.1.4 Seleção de Termos Relevantes

Essa etapa possui o objetivo de selecionar termos com maior valor representativo, ou seja, selecionar as palavras que melhor representem a base textual. As formas mais utilizadas para selecionar as palavras em relação ao texto são baseadas na frequência do termo (WIVES, 2004).

A técnica de Frequência Absoluta dos termos considera a frequência de ocorrência de um determinado termo em toda a base textual, ou seja, a quantidade de vezes que a palavra é encontrada ao longo da base de dados, conforme demonstra Equação 1.

$$FA = \sum_{i=1}^m (A_i), \quad (\text{equação 1})$$

onde A_i é a quantidade de documentos na base de dados textual que contém o termo em questão.

A Frequência Relativa dos termos é uma técnica que utiliza o valor da frequência absoluta dividida pela quantidade de palavras da base textual, conforme Equação 2;

$$FR = \frac{FA}{\text{qntPalavras}}, \quad (\text{equação 2})$$

onde FA representa a Frequencia Absoluta descrita na técnica anterior.

Informação Mútua é uma técnica estatística que mede a associação de um determinado termo t com uma determinada classe c , levando em consideração a quantidade de vezes que t e c ocorrem ao mesmo tempo, quando t ocorre sem c ; quando c ocorre sem t , conforme Equação 3.

$$IM = \log \frac{A_i \times N_{\text{all}}}{(A_i + C_i)(A_i + B_i)}, \quad (\text{equação 3})$$

onde:

- A_i é a quantidade de documentos na base de dados textual que contém o termo em questão;
- N_{all} é a quantidade total de documentos da base de treinamento;
- C_i é a quantidade de documentos que não contém o termo em questão, mas pertencem a uma determinada classe C ;
- B_i é a quantidade de documentos que contém o termo em questão, mas não pertencem a classe C .

3.2.2 Transformação dos Dados

Após o pré-processamento, os textos não podem ser submetidos à etapa de MT diretamente, em sua forma original. Antes é necessária a execução de uma técnica, que transformará os termos selecionados em valores numéricos, que serão a representatividade de cada termo em relação à base textual.

A técnica de representação por frequência Binária consiste em representar a ocorrência de um termo relevante em relação aos textos que compõem a base 1, e sua ausência com 0. Por exemplo, se uma determinada base de dados textual for representada por n palavras, cada texto que compõe a base possuirá um vetor de n posições que assumirá o valor 1 para a presença ou 0 para ausência de cada termo no texto correspondente, conforme demonstra Tabela 1.

Tabela 1 - Representação Binária

Termos Textos	Termo 1	Termo 2	Termo 3	.	.	Termo n
Texto 1	1	1	0			1
Texto 2	1	0	0			0
Texto 3	1	0	1			0
.						
.						
.						
Texto n	0	0	1			1

Fonte: Elaborado pelo autor

Outra técnica utilizada, similar a anterior, que utiliza o valor estatístico em vez do valor binário, é denominada TF-IDF (*Term Frequency–Inverse Document Frequency*). Essa técnica calcula um peso de representatividade do termo em relação à base de dados (FELDMAN; SANGER, 2006). Para a realização do cálculo, é necessário obter o valor de TF, que é a frequência do termo t no documento d , conforme representado pela expressão $TF_{(d,t)}$.

O valor de IDF (*Inverse Document Frequency*), que representa a importância de um determinado termo para a base de dados é obtido pelo logaritmo da divisão da quantidade de documentos da coleção, pela quantidade de documentos com a presença do termo, conforme demonstra a Equação 4. O valor de IDF tem o objetivo de reduzir a importância de termos que possuem uma grande frequência, mas sem valor representativo proporcional. Como forma de garantir que não aconteça uma divisão por zero, é somado ao denominador 1 da equação.

$$IDF_t = \log\left(\frac{N}{DF_t + 1}\right), \quad (\text{equação 4})$$

onde N é a quantidade total de documentos e DF é a quantidade de documentos em que o termo está presente.

O produto da frequência do termo (TF) pela representatividade geral do termo (IDF) resulta no valor final denominado TF-IDF, conforme demonstra a Equação 5.

$$TF - IDF = TF_{(d,t)} \times IDF_t \quad (\text{equação 5})$$

Considere uma coleção de 5 documentos, onde um termo t consta em 3 documentos desta coleção, o valor de TF-IDF é calculado da seguinte forma:

- Frequência do termo t na coleção = 3;
- $\text{IDF} = \log\left(\frac{5}{3}\right) = 0,221$;
- $\text{TF-IDF} = 3 \times 0,221 = 0,663$.

Após a aplicação das técnicas de transformação na base textual, ela está pronta para ser submetida à etapa de Mineração Textual, que é descrita na próxima seção.

3.2.3 Mineração Textual

Mineração Textual (MT) pode ser descrita como um processo de extração de padrões não triviais ou de extração de conhecimentos em textos, e que pode ser considerado uma extensão do processo de DCBD (TAN, 1999). É importante diferenciar Mineração Textual de Recuperação de Informações (RI). A Mineração Textual busca a descoberta de padrões implícitos, enquanto RI visa à recuperação de forma automática de documentos que satisfaçam as necessidades de informação do usuário (BAEZA; RIBEIRO, 1999).

Feldman e Sanger (2006) definem MT como um processo no qual o usuário com auxílio de ferramentas computacionais promove a análise e a extração de padrões em uma base textual.

Essas ferramentas utilizam técnicas de aprendizagem de máquina, processamento de linguagem natural, recuperação de informação e gestão do conhecimento para extração de padrões que conduzam à produção de conhecimento.

Segundo Fayyad (1996), os métodos de mineração com maior destaque são Sumarização, Análise de Agrupamento, Regras de Associação e Classificação.

- *Sumarização*: seleciona as informações mais importantes do texto, diminuindo a dispersão da descrição e tornando-o mais compacto, mas sem perder o significado;

- *Análise de agrupamento*: também conhecida por clusterização, associa um conjunto de dados similares a um ou mais clusters;
- *Classificação*: classifica um item a uma ou a várias classes categóricas predefinidas;
- *Associação*: busca descobrir associações entre dados que de alguma maneira devem possuir algum relacionamento, através de elementos que implicam a presença de outros fazendo uso de métricas de suporte e confiança.

Segundo Han e Kamber (2006), a tarefa de Classificação é o processo que busca um modelo que descreva dados de forma distinta em conceitos ou classes. Diversas são as técnicas utilizadas para esse fim, dentre as quais é possível citar Árvore de Decisão, Redes Neurais Artificiais e *Support Vector Machines* (SVM) (SEMOLINI, 2002). O modelo é construído fazendo uso de uma técnica que utiliza conjuntos de dados para treinamento e outro de teste, ou ainda utilizando um método denominado Validação Cruzada.

A técnica de Validação Cruzada divide a base de dados em n subconjuntos com a mesma quantidade de amostras. O algoritmo efetua o treinamento de classificação com $n-1$ subconjuntos, e realiza o teste sobre o subconjunto que não foi utilizado durante o treinamento. Esse processo é repetido até que todos os n subconjuntos sejam testados (HAN; KAMBER, 2006).

Para avaliar os resultados obtidos em problemas de classificação, pode-se fazer uso de uma ferramenta como Matriz de Confusão e métricas como Acurácia, Precisão e Abrangência.

A Matriz de Confusão é uma forma de analisar o quanto o classificador consegue distinguir diferentes classes, pois nela é demonstrado o número de classificações corretas em oposição às classificações preditas para cada classe. O ideal de uma Matriz de Confusão é que a maioria das quantidades de amostras esteja representada ao longo da diagonal principal, com o restante das entradas próximas a zero. É possível reconhecer na Matriz de Confusão os seguintes termos (WITTEN; FRANK, 2005):

- *Verdadeiros Positivos (TP)*: são aquelas amostras que pertencem à classe X e foram classificadas na classe X ;

- *Falsos Negativos (FN)*: são aquelas amostras que pertencem à classe X , mas foram classificadas como sendo da classe Y ;
- *Falsos Positivos (FP)*: são aquelas amostras que pertencem à classe Y , mas erroneamente foram classificadas na classe X ;
- *Verdadeiros Negativos (TN)*: são as amostras da classe Y , que foram corretamente classificadas na classe Y .

Os termos vistos anteriormente podem ser visualizados na Figura 2.

Figura 2 - Matriz de Confusão

		Classificadas	
		X	Y
Classes	X	TP	FN
	Y	FP	TN

Fonte: Elaborado pelo autor

A coluna da esquerda lista a distribuição das amostras que uma classe possui em todas as classes listadas no cabeçalho da Matriz de Confusão. A quantidade de TP está descrita no cruzamento da linha da classe X com a coluna também da classe X .

Para uma Matriz de Confusão de n classes, o valor de FP para qualquer classe é dado pelo somatório da coluna cuja classe está sendo analisada, excluindo-se o valor de TP. Já o valor de FN é obtido somatório dos valores presentes na linha relativa a classe em questão, menos o valor TP desta classe, conforme é demonstrado na Tabela 2.

Tabela 2- Matriz de Confusão n classes

	Classe 1	Classe 2	Classe 3	Classe n
Classe 1			FP ₁₃	
Classe 2			FP ₂₃	
Classe 3	FN ₃₁	FN ₃₂	TP	FN _{3n}
....
Classe n			FP _{n3}	

Fonte: Elaborado pelo autor

Tendo como exemplo a classe 3 representada na Tabela 2, o valor de FP será dado pela soma de FP_{13} , FP_{23} , ... , FP_{n3} , e o valor de FN será o resultado da soma de FN_{31} , FN_{32} , ..., FN_{3n} .

A Acurácia é uma medida de desempenho que mede a quantidade de acerto do classificador, ou seja, a quantidade de amostras corretamente classificadas. Também conhecida com taxa de reconhecimento global, por tratar da capacidade do classificador em distinguir diferentes classes, conforme demonstra a Equação 6 (HAN; KAMBER 2006).

$$\text{Acurácia} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}. \quad (\text{equação 6})$$

O numerador da Equação 6 é composto pelo resultado da soma da quantidade de TP com a quantidade de TN, e o denominador pelo total de amostras.

A métrica de Precisão fornece o percentual de acertos atribuído à classe. É calculada pela divisão de amostras classificadas corretamente, pelo total de amostras da classe, como é demonstrado na Equação 7 (WITTEN; FRANK, 2005).

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (\text{equação 7})$$

Caso todos os documentos sejam classificados corretamente, será atingido o valor máximo para essa métrica que é 1.

A métrica de Abrangência é calculada pela razão entre a quantidade de amostras corretamente classificadas, pela quantidade total de amostras atribuídas à classe, conforme é demonstrado na Equação 8 (WITTEN; FRANK, 2005).

$$\text{Abrangência} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (\text{equação 8})$$

O valor máximo que essa métrica atinge, da mesma forma que a métrica Precisão é 1, que acontece quando todas as amostras são classificadas corretamente. Como o nome da métrica mesmo diz, ela informa o percentual do

que foi classificado corretamente em relação à quantidade total de amostras que poderiam ser atribuídas à classe.

Já as métricas como suporte e confiança, utilizadas para Regras de Associação, devido a seu contexto, serão descritas junto com o algoritmo *Apriori* na subseção 3.2.3.2.

Em consonância com os objetivos relacionados no capítulo 1 deste trabalho, as duas próximas subseções apresentam as técnicas de SVM e de Regras de Associação, mais especificamente o algoritmo *Apriori*.

3.2.3.1 *Support Vector Machines*

O *Support Vector Machines* é um algoritmo rápido e eficiente para classificação de dados textuais, fazendo uso de um hiperplano para a classificação das amostras em um espaço característico. Através de uma etapa de aprendizagem (ou treino) com dados previamente rotulados, o SVM efetua treino para classificação posteriormente na etapa de teste. Essa aprendizagem sobre os dados rotulados é aplicada em amostras cuja sua classificação é desconhecida (FELDMAN; SANGER, 2006).

Dado um conjunto de treinamento com textos classificados nas classes *A* e *B*, a técnica consegue retornar os dados de teste submetidos, quais pertencem à classe *A* ou *B*. Para isso são utilizados os termos relevantes que representam a base de dados (no caso deste trabalho, com conteúdo textual) como dimensões, para executar um mapeamento dos dados de treino, em um espaço dimensional. Os documentos são posicionados em relação a cada dimensão fazendo uso de seus pesos representativos, conforme demonstra Tabela 3 (HAN; KAMBER, 2006).

Tabela 3 - Distribuição dos documentos

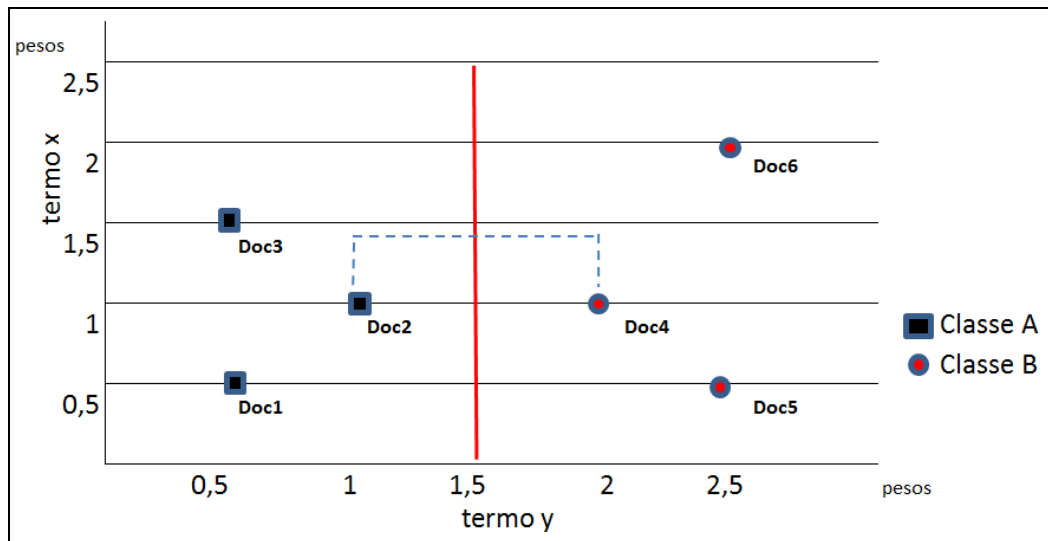
DOCUMENTOS	DIMENSÕES		CLASSES
	termo x	termo y	
Doc1	0,5	0,5	A
Doc2	1	1	A
Doc3	1,5	0,5	A
Doc4	1	2	B
Doc5	0,5	2,5	B
Doc6	2	2,5	B

Fonte: Elaborado pelo autor

Após o posicionamento das amostras, o algoritmo busca o traçado de uma linha denominada de hiperplano que separe as duas classes. Observando a Figura 3, é possível constatar que diversas linhas podem ser traçadas para separar as duas classes, entretanto o traçado da linha visa não somente separar as 2 classes, mas maximizar a distância entre essa linha e a amostra mais próximo de cada classe do conjunto de treinamento, criando, dessa forma, uma fronteira de decisão, que posteriormente é utilizada para classificar os dados de teste cujas classes são desconhecidas (HAN; KAMBER, 2006). Essa maior distância entre as margens de cada classe impede que alguma possível variação que ocorra no momento de classificação dos dados de treino venha a afetar os resultados do classificador (LOVELL; WALDER, 2006).

A Figura 3 demonstra a distribuição das amostras de cada classe, no espaço dimensional, segundo os pesos de cada dimensão. A linha tracejada representa o hiperplano separando as duas classes.

Figura 3 - Espaço dimensional com documentos mapeados



Fonte: Elaborado pelo autor

A linha traço-ponto, perpendicular ao hiperplano, indica a distância dos documentos de cada classe que estão mais próximos. Esses documentos recebem a denominação de Vetores de suporte (*Support Vectors*) e são fundamentais no processo classificatório (LOVELL; WALDER, 2006).

A definição dos Vetores de suporte é realizada na fase de treinamento, através da maximização da função W , em relação a um vetor de variáveis multiplicadoras α de Lagrange, conforme demonstra Equação 9 (SMOLA; SCHÖLKOPF, 2002).

$$\text{Max } W(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j}^N y_i y_j \alpha_i \alpha_j (x_i x_j), \quad (\text{equação 9})$$

onde:

- α : vetor de variáveis multiplicadoras;
- N : quantidade de elementos do conjunto de treinamento;
- x : elemento do conjunto de treinamento;
- y : classe a qual o elemento do conjunto de treinamento está rotulado (+1 ou -1).

O valor de α visto na Equação 9 para ser válido deve, para cada elemento do conjunto de treino, ser um valor maior ou igual a zero, e menor ou igual ao parâmetro $Cost$ (ou C). O valor desse parâmetro é escolhido empiricamente, baseado no desempenho que o classificador apresenta,

funcionado como um fator de regularização. Efetuada a maximização através da Equação 9, fazendo uso dos dados de treino (\mathbf{x}_i), todo α_i que apresentar seu valor maior do que zero é considerado um Vetor de suporte (SEMOLINI, 2002).

Concluída a etapa de treino, o algoritmo SVM utiliza os Vetores de suporte gerados para efetuar comparações com os dados do conjunto de teste a serem classificados. O resultado dessa comparação será de forma numérica e simétrica, +1 ou -1. Utilizando o exemplo anterior, o algoritmo retornará o valor de +1 caso o texto pertença à classe A; e -1 caso não pertença. A Equação 10 demonstra a função de decisão utilizada pelo algoritmo SVM. O valor obtido no cálculo é arredondado para +1 ou -1 pela função *sgn*.

$$d(X^t) = \text{sgn}\left[\sum_{i=1}^{N_{sv}} Y_i \alpha_i (X_i, X^t) + b_0\right], \quad (\text{equação 10})$$

onde:

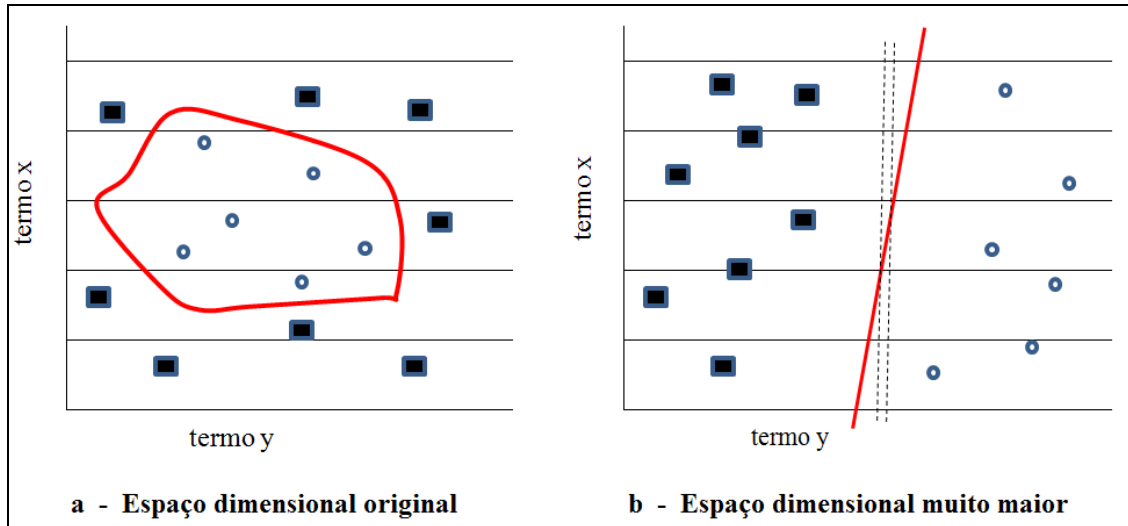
- Y é a classe;
- α é um fator multiplicador Lagrange (para máximos e mínimos);
- \mathbf{X}_i é o Support Vector;
- \mathbf{X}^t é o dado a ser classificado;
- b_0 é a constante de deslocamento da reta.

Os dados a serem classificados por esse algoritmo podem não ser linearmente separáveis. Para esse tipo de problema, o SVM substitui o produto interno das equações por uma função de *Kernel* (LOVELL; WALDER, 2006).

A função de *Kernel* objetiva o mapeamento desses dados em um espaço muito maior do que o original denominado espaço característico, cujos dados são mapeados em posições diferentes das originais, e devido à maior dimensão é possível realizar a separação linear. A Figura 4 demonstra esse processo, com os dados mapeados primeiro no espaço dimensional original (a), e após os dados são mapeados novamente em um espaço característico muito maior do que o original (b). Por questões de tamanho da figura, o espaço característico foi representado graficamente com o mesmo tamanho do espaço original, as 2 linhas paralelas tracejadas verticais sobre o espaço característico representam

uma descontinuidade do traço para indicar que o mesmo é maior do que a representação.

Figura 4 - Mapeamento no espaço dimensional original e espaço característico



Fonte: Elaborado pelo autor

Como exemplo de funções de *Kernel*, podem-se listar (SMOLA; SCHÖLKOPF, 2002):

- Função Polinomial, onde o parâmetro d é especificado *a priori* pelo usuário, conforme demonstra a Equação 11;

$$\mathcal{K}(X_i, X_j) = (X_j^T X + 1)^d, \quad (\text{equação 11})$$

- Função *Perceptron*, onde β_0 e β_1 são definidos pelo usuário, conforme demonstra a Equação 12;

$$\mathcal{K}(X_i, X_j) = \tanh(\beta_0 X_j^T X + \beta_1), \quad (\text{equação 12})$$

- Função de Base Radial (RBF), onde o parâmetro σ^2 é definido pelo usuário, conforme demonstra a Equação 13;

$$\mathcal{K}(X_i, X_j) = \left(\frac{-\|X_j - X_i\|^2}{2\sigma^2} \right). \quad (\text{equação 13})$$

Além das 3 apresentadas, existem outras funções de *Kernel* que podem ser usadas para implementar diferentes mapeamentos não lineares. Não foi encontrada referência na literatura que indicasse qual a função mais apropriada, pois o problema de classificação é dependente do domínio, cada um com distribuição de classes. Segundo Witten e Frank (2005), as mais utilizadas são a RBF e a *Perceptron*, e os resultados de cada uma dependerá da sua aplicação.

Testes realizados com as 3 funções de *Kernel* descritas, com o conjunto *M Não M*, revelaram resultados promissores com a função RBF em comparação com as outras 2, motivo pelo qual foi escolhida para a realização dos experimentos.

3.2.3.2 Algoritmo *Apriori*

O algoritmo *Apriori*, proposto por Agrawal e Srikant (1994), visa buscar a identificação de padrões que ocorrem em uma base de dados sob a forma de associações entre esses dados, e que satisfaçam os valores preestabelecidos para as métricas de suporte e confiança. Para um conjunto de dados, busca-se a probabilidade de ocorrência de um determinado valor (ou item) uma vez constatada a presença de outro valor em uma mesma transação. Uma transação é um registro de um BD onde ficam armazenados dados (itens) que co-ocorrem com uma determinada frequência. A transação também é conhecida como obtenção de *itemset* (WITTEN; FRANK, 2005).

Seja *D* um conjunto de transações onde *a* e *b* são itens frequentes, pode-se inferir que a ocorrência do item *a* implica a possibilidade de ocorrência do item *b* ($a \rightarrow b$). O suporte dessa regra é o resultado da quantidade de vezes que a regra é verdadeira dividida pela quantidade de transações do conjunto *D*, conforme demonstra a Equação 14 (AGRAWAL; SRIKANT, 1994).

$$\text{suporte}(a \rightarrow b) = \frac{a \cup b}{D}. \quad (\text{equação 14})$$

O limiar de confiança mede o grau de certeza da regra, que é calculado pela quantidade de vezes que os itens *a* e *b* ocorrem juntos, divididos pela quantidade de vezes que *a* ocorre no conjunto de transações, conforme demonstra a Equação 15.

$$\text{confiança}(a \rightarrow b) = \frac{a \cup b}{a}. \quad (\text{equação 15})$$

Seja o Quadro 1, o conjunto de transações *D* (as linhas em destaque são as transações em que *a* e *b* co-ocorrem). O valor de suporte para a regra ($a \rightarrow b$) é 0,5 e o valor de confiança para a mesma regra é 0,62.

Quadro 1 - Exemplo de transações

TRANSAÇÕES	CAMPO 1	CAMPO 2	CAMPO 3	CAMPO 4
T1	a	c	d	e
T2	a	d	f	b
T3	a	b	-	-
T4	e	f	-	-
T5	c	b	f	-
T6	a	b	-	-
T7	a	-	-	-
T8	a	f	-	-
T9	a	d	b	-
T10	a	b	-	-

Fonte: Elaborado pelo autor

O algoritmo *Apriori* implementa uma sequência de passos com o objetivo de identificar as regras de associação. Essas regras devem possuir valores iguais ou maiores do que os valores mínimos preestabelecidos para suporte e confiança, que são denominados respectivamente de suporte mínimo (*supmim*) e confiança mínima (*confmim*):

- Identificar a frequência com que os itens ocorrem nas transações, gerando C_1 ;
- Executar a contagem dos itens frequentes com os itens de C_1 que atingiram o limiar de *supmim*, gerando L_1 ;
- Identificar em L_1 os conjuntos frequentes de dois itens, estes se tornam pares candidatos em C_2 . Os pares candidatos que atingirem o critério de *supmim* serão pares frequentes em L_2 ;
- Identificar em L_2 as transações compostas por 3 itens, estas serão as trincas candidatas em C_3 . As trincas geradas devem possuir como subconjuntos de 2 itens, duplas que devem estar presentes em L_2 , caso contrário, estas devem ser podadas. As trincas que atingirem o limiar de *supmim* serão trincas frequentes em L_2 ;
- Esse procedimento deve ser executado até que o conjunto de regras candidatas não possua transações que satisfaça o critério de *supmim* em C_n , e por consequência, o conjunto L_n seja vazio.

O algoritmo *Apriori* pode também ser aplicado através de uma abordagem que objetiva a extração de regras considerando somente associações dentro de cada uma das classes que formam o conjunto de dados investigado. Para isso, é necessário identificar associações de características, que possuam valores iguais ou superiores aos mínimos preestabelecidos para as métricas de suporte e confiança, e dessa forma possam distinguir cada uma dessas classes. Nessa abordagem denominada de *Class Associations Rules* (CARs), o consequente é o atributo classe.

O CAR representa um conjunto de regras de associação do tipo $X \rightarrow y$, onde X pode ser uma ou mais características que implicam uma classe y . O valor de confiança dessa regra é dado pela quantidade de vezes que a associação X é verdadeira dentro da classe y , dividido pela quantidade de vezes que essa associação ocorre em todo conjunto. O valor de suporte é a quantidade de vezes que o antecedente da regra ocorre no conjunto de dados inteiro, dividido pela quantidade de registros (LIU; HSU; MA, 1998).

4 TRABALHOS RELACIONADOS

Este capítulo aborda trabalhos correlatos que possuem como tema de pesquisa a Descoberta de Conhecimento em Textos, utilizando o conteúdo de bases textuais relacionados à área da saúde, mais especificamente em Prontuários Médicos Eletrônicos.

O primeiro trabalho correlato apresentado na seção 4.1 aborda a Recuperação de Informação utilizando semelhança semântica aplicada a textos de um PME. A seção 4.2 apresenta um trabalho que visa à extração de informação em um PME através da utilização de conceitos. A seção 4.3 foca a classificação automática de códigos CID e seus relacionamentos. O último trabalho correlato apresentado na seção 4.4 aborda a mineração textual com Regras de Associação a partir de opiniões sobre restaurantes. Por fim, a seção 4.5 apresenta uma visão geral e crítica dos trabalhos correlatos.

4.1 RECUPERAÇÃO DE INFORMAÇÕES EM TEXTOS LIVRES DE PME POR MEIO DE SEMELHANÇA

Este trabalho aborda a Recuperação de Informações em campos de texto livre de um PME, através de um sistema desenvolvido e denominado por Martha, Campos e Sigulem (2010) como SIRIMED. Esse sistema utiliza a semelhança semântica e ortográfica para RI. Um estudo comparativo foi realizado utilizando um sistema de busca tradicional, o *Clinic Manager* (SIGULEM, 1994), desenvolvido pelo Departamento de Informática em Saúde (DIS) da Universidade Federal de São Paulo, utilizado em um sistema de PME. Os resultados obtidos no SIRIMED foram comparados aos resultados obtidos no *Clinic Manager* para averiguar a quantidade de informações que deixam de ser recuperadas pela falta da aplicação de uma técnica mais apurada na busca.

A RI estuda técnicas para a automação da recuperação de textos, com conteúdo relevante associado aos termos utilizados pelo usuário como parâmetros de pesquisa. Um dos métodos utilizados para esse fim é a criação de um conjunto de palavras com significado relevante para representar o conteúdo do texto. Esse conjunto de palavras é denominado índice, sendo a base dos

sistemas de RI. A premissa do SIRIMED é o desenvolvimento de um índice, compacto e eficiente, para a representação dos textos.

Nesse estudo, o conteúdo dos campos pesquisados em ambos os sistemas foram textos livres (que compreendem a descrição de sintomas, queixas do paciente, respostas a questões apresentadas na consulta e demais observações que o médico julgue importante). Foram utilizadas duas bases de dados: a primeira oriunda de uma clínica especializada em neurologia e psiquiatria, com 6.732 registros; e uma segunda de uma clínica especializada em nefrologia e clínica médica, com 26.072 registros.

O trabalho foi dividido em duas partes distintas. A primeira chamada de indexação automática consiste na criação do índice invertido de pesquisa que será utilizado para comparação com os termos usados na consulta pelo usuário. Ainda nessa etapa, ocorre a realização do pré-processamento das bases, conforme descrição a seguir:

- Substituição de letras maiúsculas por minúsculas;
- Remoção de caracteres especiais, tais como: [{ ? ! \$ # % ; , . &;
- Substituição de letras acentuadas pela mesma letra sem a acentuação ortográfica;
- Substituição de “ç” por “c”;
- Remoção de *Stopwords*;
- Aplicação de técnicas de *Stemming*;
- Cálculo do peso de cada palavra (valor representativo de cada palavra em relação ao texto) através do padrão Termo-Frequência.

A segunda parte foi a recuperação dos textos tendo como base uma pergunta elaborada. A pergunta feita pelo usuário também deve passar pelo pré-processamento. Essa recuperação faz uso da semelhança semântica, semelhança ortográfica, inserção de sinônimos e preservação da ordem dos termos. Os termos utilizados dentro da área de saúde são inúmeros, em virtude disso foi feito uso de uma padronização dinâmica do vocabulário médico, denominada Descritores em Ciências da Saúde (DECS¹) (BIREME, 2011). Essa padronização foi utilizada pelo sistema SIRIMED para implementar a

¹ <http://decs.bvs.br/>

recuperação semântica dos textos. Ainda nessa etapa foi realizada uma validação das palavras utilizadas com o dicionário br.ispell² do português falado no Brasil (KARPISCHEK, 1999). Para analisar a quantidade de informações que deixaram de ser recuperadas, foram selecionados 34 termos médicos nas duas bases de dados e pesquisados através do *Clinic Manager* e do SIRIMED.

Foi constatado que na base que contém dados textuais de neurologia e psiquiatria (base de dados 1), formado por 6.732 textos com 830.471 palavras, somente 58% destas estão presentes no dicionário da língua portuguesa utilizado, portanto 42% das palavras não são passíveis de correção ortográfica. Ainda na análise dessa base de dados, somente 8,8% desse conjunto de palavras estava presente vocabulário DECS. A base textual (base de dados 2) referente à nefrologia que contém 26.072 textos com 3.990.900 palavras, possui percentuais inferiores à primeira. Somente 43% das palavras são possíveis de correção ortográfica fazendo uso do dicionário de língua portuguesa Ispell, e 7,3% estavam presentes no vocabulário DECS.

Foi constatado que a remoção de *Stopwords* reduziu a quantidade de palavras nas duas bases de dados, em 32,6% para a base 1 e 37,9% para a base 2. A aplicação de *Stemming* obteve uma redução 1,5% das palavras que compõem o índice que representam os textos.

Em casos específicos, como “ronco” e “tontura”, apenas com a aplicação de *Stemming*, foram obtidos percentuais de melhoria na RI de 80,1% e 73,7% respectivamente, em comparação aos resultados obtidos com o sistema *Clinic Manager*. Com a aplicação das técnicas de semelhança ortográfica e semelhança semântica para esses casos específicos, não foram observadas melhorias expressivas no que tange à RI, mas em outros casos como em “desmaio” com a aplicação de semelhança ortográfica e semântica, foi obtido um percentual de melhoria da RI de 193%.

4.2 DESCOBERTA DE CONHECIMENTO EM PRONTUÁRIO ELETRÔNICOS

Este trabalho visa à apresentação de uma estratégia para Extração de Informação (conhecimento sobre doenças, padrões de preenchimento de

² <http://www.ime.usp.br/~ueda/br.ispell/>

prontuários, comportamentos dos pacientes, perfil dos pacientes, etc.) em uma base textual que compõe um PME de uma clínica psiquiátrica (LOH et al., 2002).

A base de dados empregada neste trabalho foi coletada em uma clínica psiquiátrica privada e corresponde ao período de quatro meses, o que permitiu a obtenção de 400 registros. Esses registros contêm informações relativas a laudos de internação e evoluções médicas diversas, na forma de textos livres, que foram cadastradas pelos profissionais da área da saúde (médicos, enfermeiros, psicólogos) do quadro funcional da clínica. Para cada prontuário (paciente), é associado um diagnóstico através de codificação do CID-10.

A estratégia utilizada para Extração de Informação foi dividida na Análise Qualitativa e Análise Quantitativa de conceitos presentes na base textual do PME. Segundo Loh et al. (2002), os conceitos são uma forma de representação de objetos, eventos, opiniões e ideias do mundo real. É possível citar, como exemplo, em uma base textual psiquiátrica o conceito “alcoolismo”, relacionado ao uso de substâncias psicoativas. Foram selecionados para as quatro classes 97 conceitos (65 referentes a características dos pacientes e 32 referentes a remédios).

A Análise Qualitativa visou à identificação dos conceitos presentes nos textos, através da existência de termos que confirmavam a presença desses conceitos. A identificação dos conceitos foi realizada utilizando SVM e um modelo contextual. No SVM, cada conceito foi representado por um vetor de termos simples, e atribuído um peso a cada um desses termos, que descreve o grau de importância do termo sobre o conceito. Já o papel do modelo contextual foi minimizar interpretações erradas, pois esse modelo considera que a relação entre os termos pode influenciar na representação dos conceitos.

A Análise Quantitativa foi caracterizada pela aplicação de técnicas estatísticas nos conceitos encontrados na primeira etapa. Nessa etapa, foram aplicadas as técnicas de análise de distribuição e técnica associativa.

O papel da técnica de análise de distribuição foi verificar a frequência com que os conceitos ocorriam na base textual, gerando como resultado uma lista de cada conceito com sua frequência. Enquanto a técnica associativa descobria associações entre os conceitos através de uma probabilidade

condicional, ou seja, se existia um ou mais conceitos específicos no texto, a presença de outro determinado conceito também estaria confirmada com uma probabilidade de certeza.

A combinação dessas duas etapas teve como resultado o conhecimento sobre o domínio. A base de dados foi dividida em duas partes, um grupo de 200 registros para treinamento, e outro com os 200 restantes para teste. Para os experimentos realizados, foram selecionadas quatro classes que representavam os diagnósticos mais frequentes na clínica, e a distribuição dos registros do conjunto de treino seguem os seguintes percentuais relativos a uma dessas classes:

- *Transtornos afetivos*: 27 textos (13,5%);
- *Esquizofrenia*: 103 textos (51,5%);
- *Orgânicos*: 18 textos (9%);
- *Substâncias Psicoativas*: 52 textos (26%).

O primeiro grupo de dados (treino) foi dividido em quatro partes, correspondendo às quatro classes especificadas. A fração da base textual destinada a cada uma das classes continha os laudos de internação com diagnósticos referentes à classe associada.

Os resultados foram medidos usando as métricas de Abrangência, Precisão e *F-measure*. Por haver uma distribuição em quatro classes diferentes, foram utilizadas as medidas de *Microaveraging* (considera a coleção toda como uma única classe então avalia a Abrangência e a Precisão) e *Macroaveraging* (calcula a Precisão e Abrangência em cada classe e então extrai os valores médios para a coleção toda). A média entre os valores de *Microaveraging F-measure* e *Macroaveraging F-measure* foi usada para medir o melhor desempenho, e são apresentados na forma de percentual.

O experimento realizado usou a primeira parte dos dados para o treinamento, identificando as características das quatro classes especificadas, o conhecimento descoberto foi acoplado em um sistema de classificação que se baseia em espaço de vetores. Quando utilizados todos os conceitos de cada lista, para caracterizar cada classe de diagnóstico, foi alcançado um percentual de 44%. O maior percentual médio para métricas de desempenho foi obtido

fazendo uso apenas dos conceitos menos frequentes (probabilidade < 50%) de cada lista, resultando em 62%; com a utilização de apenas pares de conceitos presentes nas associações, o percentual foi de 51%; fazendo uso dos pares de conceitos presentes nas associações, porém somente para os pares exclusivos de cada classe, foi obtido o percentual de 57%; usando todos os conceitos de cada lista mais os pares de conceitos extraídos nas associações, o percentual alcançado foi de 52%.

O melhor método de caracterização utilizado obteve uma média de 62% de acertos entre os valores de *Microaveraging F-measure* e *Macroaveraging F-measure*. Esse método usa apenas os conceitos menos frequentes (probabilidade < 50%) de cada lista de conceitos de cada classe. O trabalho fez uso da análise de especialistas e, segundo eles, o resultado é satisfatório, uma vez que desempenhos maiores do que 60% são melhores do que algumas decisões de especialistas humanos.

4.3 CLASSIFICAÇÃO DE CÓDIGOS MÉDICOS PARA DESCOBERTA DE RELACIONAMENTOS

Este trabalho aborda a classificação de múltiplos códigos CID-9 utilizados durante a realização de tratamentos ou consultas médicas. Dependendo do(s) motivo(s) da consulta, um ou mais códigos CID podem ser atribuídos ao paciente, ficando de forma implícita, relacionados às anotações realizadas pelo médico (YAN et al., 2010).

Além de codificar a(s) doença(s) dos pacientes de uma forma padronizada, também, o CID é utilizado na identificação de procedimentos e ou exames médicos realizados. Tal identificação é utilizada por seguradoras, operadoras de planos de saúde e poder público para transações financeiras, como reembolsos, quantificação de custo de tratamentos, repasses do governo e cobranças. As instituições de saúde do poder público também utilizam a ocorrência de códigos CID no processo de Acreditação Hospitalar (programa de certificação de qualidade hospitalar) (BRASIL, 2002), elaboração de relatórios estatísticos, e controle de determinados tipos de doenças, que por suas características possam gerar uma epidemia.

Tendo como base registros do tipo texto, retirados do PME (anotações médicas ou de enfermagem, laudos laboratoriais, avaliações diversas), é abordado o problema de atribuição automatizada de códigos CID a esses registros. Conforme citado, dependendo do motivo que levou o paciente a procura de atendimento, um ou mais códigos CID podem ser atribuídos ao prontuário do paciente. Essa ação pode ser executada na hora da consulta, mas na maioria das vezes, especialmente quando o paciente necessita de internação, esse ato é postergado. Um especialista da área médica realiza uma revisão de todo o material registrado durante a consulta, e com embasamento nas legislações, normas, e sua experiência profissional efetua a atribuição dos referidos códigos.

A estrutura do CID é composta por ramificações de categorias que se dividem em diversas subcategorias. Existe no domínio médico conhecimento de relacionamentos de ocorrência entre códigos CID. Por exemplo, ocorrência de “dor no peito” implica a ocorrência de “congestiva insuficiência cardíaca”, em uma mesma consulta (YAN et al., 2010). Esses relacionamentos, até certo ponto, são conhecidos dentro da área médica e devem ser considerados sempre que possível.

Diante do exposto, o desafio deste trabalho foi o desenvolvimento de um método que classificava múltiplos códigos CID (associados a um conteúdo textual), a estrutura que o compõe bem como o código CID específico atribuído, agregando o prévio conhecimento médico sobre os relacionamentos entre códigos existentes.

O problema poderia ter uma abordagem de classificação binária, mas isso excluiria da análise os relacionamentos entre classes subjacentes. As relações entre os códigos CID podem ser tratadas como uma nova classe, mas devido à quantidade de combinações entre cada classe e a falta de registros suficientes para uma avaliação relevante, tornam essa abordagem inviável. A principal característica do problema é o relacionamento do conteúdo textual com um ou mais códigos CID, e os relacionamentos desses códigos entre si. Isso dificulta a utilização de algumas técnicas, caso se deseje uma exploração total, tanto da base textual quanto dos dados estruturados relacionados com esta, de forma que cubra a estrutura dos inter-relacionamentos dos códigos.

A base de dados textual utilizada foi composta por textos livres e gerada pelos diversos profissionais da área da saúde (enfermeiros, médicos, fisioterapeutas, etc.). Os textos cadastrados foram oriundos de observações médicas, exames de laboratórios, avaliações físicas dos pacientes, totalizando 978 registros. Sobre esses registros, foi aplicada a representação unigrama com cada característica indicando presença ou ausência de uma palavra. Após foi executada a remoção de *Stopwords*, e aplicação de técnicas de *Stemming*, obtendo como resultado 1.931 palavras. Foram escolhidas como palavras relevantes aquelas que ocorreram em pelo menos 5% da base textual, chegando a um total de 1.155 palavras. A base de dados possui 140 códigos de CID diferentes relativos a sintomas ou doenças. Devido à maioria das classes terem poucos dados para treinamento, foram utilizados apenas os códigos que ocorreram no mínimo em 5% na base textual, totalizando 20 CIDs.

A técnica desenvolvida denominada de SVM-sim, uma extensão do SVM clássico, utiliza o conhecimento existente sobre os relacionamentos entre códigos para suprir a falta de dados para treinar o classificador. A comparação foi realizada com três tipos diferentes de classificadores de múltiplos códigos:

- SVM-rank;
- Método do vizinho mais próximo – múltiplos códigos (ML-KNN);
- Rede Neural (BP-MLL).

Os resultados alcançados em termos de acurácia na classificação dos códigos CID, pela técnica SVM-sim, alcançaram os melhores percentuais em 18 classes, tendo como percentual mínimo alcançado 95,3%, ficando com o mesmo índice a técnica SVM-rank. O maior percentual de acurácia alcançado pela técnica SVM-sim foi de 99,8%, sendo que o 2º maior percentual de classificação na mesma classe atingiu o valor máximo de 99,0% também na técnica SVM-rank.

4.4 EXTRAÇÃO DE REGRAS EM CONTEÚDO TEXTUAL COM ALGORITMO *APRIORI*

Investigada a literatura sobre o uso de Regras de Associação, em dados descritivos relacionados à PME, não foi encontrado nenhum trabalho

relacionado. Desta forma, é apresentado um trabalho que trata do tema de mineração de conteúdo textual e Regras de Associação em conteúdo textual, porém em outro domínio.

O trabalho realizado por Gupta, Tennesi e Gupta (2009), embora não utilize uma base de dados do domínio da saúde e sim uma base formada com opiniões sobre restaurantes, aborda a temática da extração de Regras de Associação com algoritmo *Apriori* em conteúdo textual.

A técnica foi aplicada em opiniões cadastradas sobre 120 pizzarias que foram extraídas do portal www.yelp.com, que é especializado em cadastrar e disponibilizar opiniões diversas sobre restaurantes.

O objetivo do trabalho foi poder identificar e extrair determinadas características presentes nas opiniões cadastradas, classificando-as em classes predefinidas, bem como quanto a sua polaridade. É necessário destacar que é expressivo o volume de opiniões referentes a um determinado restaurante cadastrado, o que torna inviável para uma pessoa a leitura de todas as opiniões cadastradas de diversos restaurantes, dificultando a comparação entre estabelecimentos para formalização de uma escolha (tomada de decisão).

As características predefinidas para a classificação das opiniões foram relativas a ambiente, comida e serviço, e as polaridades positiva, negativa e neutra. As opiniões foram coletadas fazendo uso de um *WEB crawler*³, coletando opiniões sobre 120 pizzarias. Para a identificação de adjetivos e frases nominais, foi utilizada a ferramenta Stanford POS, que faz uso de *tags* para a identificação gramatical das palavras.

Durante a leitura do trabalho, em nenhum momento é mencionada a utilização de alguma técnica de pré-processamento. Porém, Gupta, Tennesi e Gupta (2009) citam que na seleção de palavras foram utilizados substantivos comuns (no plural e singular), excluindo-se dessa forma nomes próprios.

O algoritmo *Apriori* gerou os conjuntos de itens frequentes a serem classificados nas classes preestabelecidas como características, entretanto sua utilização acarreta problemas como:

³ Ferramentas de busca utilizada para coletar informações em páginas WEB de forma refinada, retornando os resultados mais relevantes (MAGALHÃES, 2008).

- O algoritmo *Apriori* na geração dos conjuntos frequentes não faz nenhuma consideração quanto à ordenação ou mapeamento das palavras, dessa forma um conjunto frequente, tal como [Palo; Alto; Pizza] ficaria [Alto; Palo; Pizza], conduzindo a uma avaliação do conjunto em uma ordem equivocada;
- Todas as palavras (características) da frase são colocadas dentro do mesmo conjunto, às vezes compondo estes com palavras que estão muito distantes umas das outras, pois o algoritmo como dito no item anterior não leva em consideração a posição da palavra;
- A ocorrência de forma individual de algumas palavras frequentes, que podem ser consideradas características, mas que individualmente não possuem grande significado, por exemplo, a característica vinho francês, a palavra francês pode ocorrer de forma isolada em um conjunto, mas não representa por si só uma característica, a não ser que esteja acompanhada da palavra vinho.

Para a correção desses problemas, foram aplicados 2 filtros. O primeiro levando em consideração a ordem que as características ocorrem e também a proximidade das mesmas, sempre considerando características que estejam a uma distância máxima de 2 palavras. O segundo filtro tem por objetivo a eliminação de palavras que ocorrem sem a presença de uma segunda palavra e, dessa forma, não constituem uma característica.

Para a classificação das características, foram levados em consideração o contexto da palavra na frase e os hiperônimos⁴ entre outras formas linguísticas que possam estar associadas à palavra, conforme os passos a seguir:

- Obtenção da orientação semântica de cada palavra utilizando o *Wordnet*⁵ (PRINCETON, 2011). Foram obtidas 400 características (conjunto de treino) a partir das opiniões coletadas e rotuladas seguindo as classes preestabelecidas (comida, ambiente e serviço);

⁴ Palavra do mesmo campo semântico, mas de forma mais abrangente (CEGALLA, 2008).

⁵ Banco de dados léxico do inglês, que agrupa as palavras por sua classe gramatical e seus relacionamentos (PRINCETON, 2011).

- Obtenção dos conjuntos de sinônimos de cada característica através da ferramenta *Wordnet*, para cada conjunto de características foi computado um hiperônimo;
- Para treinamento, foi utilizada a estimativa de probabilidade máxima, para calcular se uma determinada palavra pertence a uma classe;
- Para execução dos testes, foi multiplicada a probabilidade de a palavra pertencer a uma determinada classe, pela probabilidade de todos os seus hiperônimos também pertencerem a essa classe. A classe com maior probabilidade então é atribuída a palavras.

Nos resultados apresentados, é afirmado que o limiar foi estipulado um limiar de suporte de 0,04% para determinar os itens frequentes, que embora muito baixo, qualquer aumento nesse valor resultaria em apenas 3 conjuntos de itens frequentes, desperdiçando características com potencial representativo, seguindo de uma apresentação das características selecionadas para o limiar de 0,04%. Quanto à polarização das características, são apresentadas as métricas de Precisão e Abrangência para cada uma das polaridades:

- Positiva, Precisão de 100% e recall de 98%;
- Negativa, Precisão de 100% e recall de 95%;
- Neutra, Precisão de 96% e recall de 100%.

4.5 CONSIDERAÇÕES SOBRE OS TRABALHOS

Entre os trabalhos apresentados nas seções anteriores, todos possuem a DCT como área de pesquisa, 3 destes trabalhos utilizam dados descritivos relativos a PME. Um destes trabalhos aplica a técnica de Regras de Associação sobre opiniões relativas a restaurantes, cadastradas em um portal. A metodologia empregada varia de acordo com o objetivo de cada um. Uma síntese dos trabalhos referidos é apresentada no Quadro 2.

Quadro 2 - Síntese dos trabalhos

REFERÊNCIA	Martha, Campos e Sigulem (2010)	Loh et al. (2002)	Yan et al. (2010)	Gupta, Tenneti e Gupta (2009)
PESQUISA	RI em textos livres de PME por meio de semelhança	Descoberta de Conhecimento em PME	Classificação de doenças e relacionamentos entre códigos médicos	MT em Opiniões através de regras de associação
OBJETIVOS	RI por meio de semelhança semântica em PME	Estratégia para classificação de PME através de conceitos extraídos de campos descritivos	Classificação multi-classes de PME	Classificação de opiniões e descoberta de polaridade
PRÉ-PROCESSAMENTO	Stemming; Stopwords; Correção Ortográfica	Stopwords	Stemming; Stopwords	Sem referência
TÉCNICA EMPREGADA	Indexação automática com auxílio de dicionários	SVM	SVM	Regras de associação: algoritmo <i>Apriori</i>
QNT. DE REGISTROS	Base 1: 6.732 reg. Base 2: 27.072 reg.	Base de dados com 400 registros	Base de dados com 978 registros	Não divulgado

Fonte: Elaborado pela autor

Os resultados obtidos pelo sistema desenvolvido por Martha, Campos e Sigulem (2010) em comparação com os resultados do sistema SIRIMED, que executa a busca pelo termo exato, apresentaram aumento de 83% na recuperação de textos em determinados casos. Porém não existe no trabalho referência às quantidades de textos relevantes armazenados para cada termo, impossibilitando a realização de uma comparação dos resultados dos dois sistemas com uma possível solução ótima, ou mesmo o grau de relevância dos textos recuperados. Foi afirmado no trabalho que somente a remoção de *Stopwords* e a aplicação de *Stemming* alavancaram de forma expressiva os percentuais de recuperação de informação. Porém comparando-se o percentual de RI, obtido somente com a etapa de pré-processamento, com o percentual obtido posteriormente a aplicação da semelhança ortográfica e semântica, torna-

se evidente que os resultados são menos expressivos, destacando-se dessa forma a importância da etapa de pré-processamento.

No trabalho investigado por Loh et al. (2002), é pequeno o volume de registros da base de dados investigada, com 400 registros, e essa quantidade de registros ainda é dividida em conjunto de treino e teste. Apesar de os registros associados a cada uma das classes pertencerem realmente à classe, essa distribuição de registros no conjunto de treino não é uniforme. A distribuição dos textos no conjunto de teste é muito semelhante, tendo pouca variação. Uma vez que um dos objetivos do trabalho é aplicação de técnicas estatísticas (Análise Quantitativa), pode-se observar que existe uma pequena amostra de dados relativos a cada classe, sendo que a grande maioria está concentrada na classe referente à Esquizofrenia (51,5%). Isso pode comprometer a qualidade da classificação favorecendo os resultados para uma classe em detrimento de outras. Foram selecionados 32 conceitos referentes a remédios e 65 conceitos referentes a características do paciente, mas não existe referência quanto à distribuição desses conceitos nas classes. A única técnica de pré-processamento mencionada durante o trabalho foi a remoção de *Stopwords* composta por uma lista de 220 termos.

Embora o problema abordado por Yan et al. (2010) seja de multiclassificação, ou seja, um determinado texto possuir uma ou mais classes associadas a ele, os resultados (acurácia) são apresentados de forma isolada por classe. Os valores de acurácia obtidos pela técnica SVM-sim variam de 95,6% (classe 1) até seu valor máximo em 99,8% (para a classe 7 e classe 10), não seguindo uma ordem crescente uniforme, como é demonstrado nos gráficos de acurácia e desvio padrão das 20 classes. Foram utilizadas 1.155 palavras para a identificação das 20 classes. Através de um gráfico apresentado, é possível identificar 177 palavras para a classe Tosse (*Cough*) que obteve uma acurácia de 95,9%, mas para diversas classes, inclusive as 2 classes com melhor percentual (classe 7 e 10 com 99,8%), possuem menos de 5 palavras para sua identificação.

O trabalho de Gupta, Tenneti e Gupta (2009) destaca que é inviável a leitura de todas as opiniões por ser expressiva a quantidade de registros e que foram coletadas opiniões sobre 120 restaurantes, mas em momento algum o

autor faz referência à quantidade de opiniões ou características que compõem a base de teste. Não é relatado se o conjunto de treino possui uma distribuição homogênea dos registros nas 3 classes investigadas, sendo dessa forma um conjunto balanceado. Durante a apresentação dos resultados, não é mencionado o nível de acerto da técnica utilizada nas classes preestabelecidas, deixando de fornecer um parâmetro para análise de desempenho do algoritmo.

Durante o estudo, ficou evidente que a aplicação de técnicas de pré-processamento, como *Stemming* e remoção de *Stopwords*, foram fatores fundamentais nos resultados alcançados, devendo ser consideradas no decorrer desta pesquisa, bem como a seleção de termos por frequência. Não foi mencionado o uso de técnicas de transformação utilizadas para a representação do conteúdo diante dos algoritmos utilizados.

Analisando-se quantitativamente cada base de dados (volume de registros), sem levar em conta a distinção por classe utilizada em cada trabalho, em relação à base de dados da área da Reumatologia inferida nesse trabalho, chega-se a um valor intermediário, ficando um pouco abaixo da base de dados utilizada por Martha, Campos e Sigulem (2010), mas expressivamente superior às demais bases de dados.

A base de dados utilizada nessa proposta e descrita no capítulo 5 é composta por campos de conteúdo de texto livre, associados a campos pré-estruturados onde estão armazenados os códigos CID. Similarmente o trabalho desenvolvido por Yan et al. (2010) infere sobre os dados o algoritmo SVM, demonstrando que este é viável de ser utilizado na investigação de dados da área da Reumatologia.

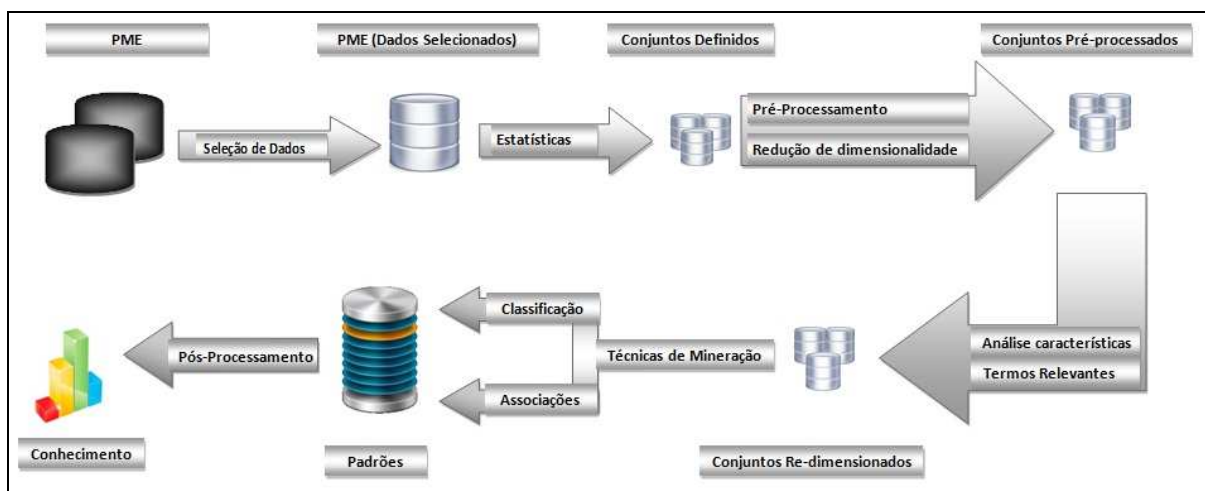
Assim como o trabalho de Gupta, Tenneti e Gupta (2009), este trabalho visa explorar uma base com conteúdo descritivo utilizando a técnica de Regras de Associação. Embora as bases pertençam a domínios diferentes, o trabalho correlato buscou a identificação de características preestabelecidas no conteúdo textual e sua polaridade, enquanto neste trabalho se buscam associações de termos que melhor representem cada uma das doenças que compõem as classes investigadas. Buscam-se ainda associações entre termos que são desconhecidas no domínio médico que possam constituir conhecimento válido ou passível de ser criticado.

5 DESCOBERTA DE CONHECIMENTO APLICADA A DADOS DE SAÚDE

Este capítulo, em consonância com os objetivos do trabalho, apresenta os experimentos realizados e seus resultados. Para isso, é realizada uma contextualização do problema de Classificação e extração de Regras de Associação, em dados textuais de um sistema de PME, relativos à área da Reumatologia. As etapas do processo são descritas, bem como a base de dados e a seleção dos conjuntos de interesse. Também são detalhadas as técnicas de Pré-processamento aplicadas, para então descrever os experimentos com seus resultados, encerrando com uma discussão sobre estes.

Para a realização do processo descrito, é executada uma sequência de sucessivas e distintas etapas, que podem ser visualizadas graficamente através da Figura 5, que representa a estrutura de funcionamento do processo como um todo, bem como seus elementos componentes e o fluxo que o processo deve seguir.

Figura 5 - Descrição gráfica do processo



Fonte: Elaborado pelo autor

O processo de investigação tem início na aquisição da base de dados composta por dados clínicos relativos à área de Reumatologia, cadastrados em um sistema de PME. Os dados estão em sua forma íntegra, com características próprias (presença de linguagem de hipertexto, por exemplo), com todos os registros e campos que a compõem, sem restrição alguma. O próximo elemento destacado é a seleção dos dados da área de interesse. Essa seleção visa à

composição de um conjunto principal de dados, que seja composto pelos registros que possuem os códigos CID com maior frequência.

A etapa seguinte é um levantamento estatístico de ocorrências dos códigos CID para então efetuar a composição dos conjuntos de interesse. A partir dessa definição, é possível executar a redução de dimensionalidade, sendo que o conteúdo descritivo com poucos caracteres que prejudicariam a etapa de mineração é removido. Nessa etapa do processo, é iniciado o pré-processamento (limpeza) dos conjuntos de interesse, através da aplicação de técnicas específicas para esse fim. Com os conjuntos pré-processados, a atividade a ser executada é a análise de características para seleção de termos relevantes.

A partir desse momento, inicia-se a realização de experimentos com a aplicação das técnicas de mineração de dados, desdobrando-se em Classificação e Associação, e como resultado a produção de padrões e regras, respectivamente.

A última etapa é o Pós-processamento, ou seja, análise dos padrões extraídos para a produção de conhecimento. Nessa etapa, principalmente para a análise das Regras de Associação, contou-se como auxílio de 2 especialistas da área da Reumatologia.

5.1 BASE DE DADOS

Esta seção descreve como ocorreu a aquisição da base de dados e como foram constituídos os conjuntos de interesse.

5.1.1 Aquisição da Base de Dados

Durante a leitura de trabalhos correlatos, chegou-se a um autor que possuía um trabalho da área de interesse (dados da área da saúde oriundos de PME). Foi realizado contato com esse autor questionando a viabilidade da utilização da mesma base de dados. A utilização da mesma base não foi possível, mas em razão de ele trabalhar no desenvolvimento de sistemas para a área da saúde, conseguiu a disponibilização de uma base de dados clínicos

relativos à área da Reumatologia. Devido à natureza da especialidade da clínica, os dados são relativos a pacientes de uma faixa etária predominante de idosos.

5.1.2 Descrição da Base de Dados

A base de dados cedida possuía originalmente 42.537 registros. Após a execução de um filtro, foram removidos 9.534 registros (22,4%), que por não possuírem registros com conteúdo descritivo, não contribuíam para o desenvolvimento deste trabalho. Após a aplicação do filtro, restaram 33.003 registros armazenados.

Os dados investigados estão distribuídos em 3 campos com conteúdo pré-estruturado e 1 com conteúdo descritivo, não possuindo qualquer dado que conduza a identificação do paciente ou caracterização do mesmo (sexo, idade, cor, etc.), o que não caracteriza um problema, pois para o objetivo deste trabalho não são necessárias tais informações. Os campos que compõem a base de dados são os seguintes:

- *CODCLI*: campo com formato pré-estruturado, alfanumérico, destinado que armazenava o código identificador de cada paciente, usado para individualizar os registros de um mesmo paciente;
- *TEXTO*: campo com conteúdo descritivo, destinado a armazenar as observações, descrições, prescrições, diagnósticos, que o profissional da área da saúde emite sobre o paciente;
- *CID*: campo com formato pré-estruturado, alfanumérico, destinado a armazenar o código de identificação da doença diagnosticada;
- *DESCRIÇÃO*: campo com formato pré-estruturado, texto, para armazenar a descrição de cada código CID relacionado ao diagnóstico da doença.

5.2 SELEÇÃO DOS DADOS

Essa base de dados, originalmente de uma plataforma WEB, continha em seus registros descritivos, marcações da linguagem de hipertexto (*HyperText Markup Language* - HTML), que para a realização deste trabalho, poluíam e

aumentavam o volume do conteúdo textual de forma expressiva, além de impossibilitar uma avaliação da base de dados para estabelecer o conjunto principal de dados.

Devido à quantidade de registros, foi inviável executar uma remoção manual da marcação referida. Para equacionar esse problema, foi utilizada uma série de expressões regulares maximizando a quantidade e o volume de texto a ser removido. As expressões regulares foram executadas através do editor de textos para scripts Notepad++⁶.

Para cada consulta realizada, o profissional da área da saúde cadastrava informações descritivas no sistema, associando a essas informações um ou mais códigos CIDs. Cada código CID cadastrado correspondia a um registro gerado na base de dados, com a repetição do conteúdo descritivo associado. Hipoteticamente, em uma consulta que foram diagnosticados 5 códigos CIDs para um paciente, essa consulta tem como resultado na base de dados 5 registros com os 5 códigos CIDs diferentes, mas todos os registros com o mesmo conteúdo textual, conforme mostra o Quadro 3.

Quadro 3 - Trecho da base de dados original

Descrição	CID
ALÉRGICA: TEGRETOL	M32.9
ALÉRGICA: TEGRETOL	E66.0
no intervalo teve outra candidíase, usou medicamento tópico e nizoral . resultado de exames: EFP= DISCR	M79.0
no intervalo teve outra candidíase, usou medicamento tópico e nizoral . resultado de exames: EFP= DISCR	D56.9
No mês de junho a Vó teve queimadura extensa, teve que ser internada, rebaixou nível de consciência, t	M82.8
No mês de junho a Vó teve queimadura extensa, teve que ser internada, rebaixou nível de consciência, t	M79.0
No mês de junho a Vó teve queimadura extensa, teve que ser internada, rebaixou nível de consciência, t	K77.0
No mês de junho a Vó teve queimadura extensa, teve que ser internada, rebaixou nível de consciência, t	E78.2
t4livre: 1,19; TSH= 2,82; ANTITIREO E TPO +.-> Dr. Hyungmeus= TSH= 5,01; GLIC= 85; U= 37; C= 0,9; CPK=9	M75.1
t4livre: 1,19; TSH= 2,82; ANTITIREO E TPO +.-> Dr. Hyungmeus= TSH= 5,01; GLIC= 85; U= 37; C= 0,9; CPK=9	M15.0
t4livre: 1,19; TSH= 2,82; ANTITIREO E TPO +.-> Dr. Hyungmeus= TSH= 5,01; GLIC= 85; U= 37; C= 0,9; CPK=9	I10
t4livre: 1,19; TSH= 2,82; ANTITIREO E TPO +.-> Dr. Hyungmeus= TSH= 5,01; GLIC= 85; U= 37; C= 0,9; CPK=9	E78.5
ALERGIA A PIROXICAM	E78.2
ALÉRGICA: UNASYN E AVALOX	M79.0
ALÉRGICA: UNASYN E AVALOX	M15.0
ALÉRGICA: UNASYN E AVALOX	I10

Fonte: Base de dados pesquisada

A partir de um levantamento realizado na base de dados, foi estimada a quantidade máxima de até 14 códigos CIDs diferentes em uma mesma consulta, relacionados a um mesmo conteúdo descritivo.

⁶ <http://notepad-plus-plus.org/>

Os diferentes registros que compunham a consulta de um paciente foram agrupados de maneira que formassem apenas um registro, sem a repetição do conteúdo descritivo, e com cada código CID registrado em um campo diferente. Essa operação reduziu a base de dados a 19.958 registros, ou seja, uma redução de 39,53% nos 33.003 registros. Analisando os campos que armazenavam os códigos CID, constatou-se que a base de dados se caracteriza pela grande incidência de códigos CID começando com a letra *M* (M00-M99). Esses códigos no CID-10 estão presentes no capítulo XIII (DATASUS, 2008), que correspondem a doenças do sistema Osteomuscular e do Tecido Conjuntivo (DATASUS, 2008). A primeira tarefa realizada foi uma análise minuciosa do volume de códigos CID que compõem a base. Foi constatado que dos 19.958 registros:

- Quantidade de registros com a ocorrência de pelo ao menos 1 código CID começando por *M*, associados a outros códigos CIDs no mesmo registro era de 13.345 registros;
- Quantidade de registros com a ocorrência de somente códigos CIDs começando por *M* era de 9.349 registro;
- Quantidade de registros sem a presença de código CID começando pela letra *M* era de 6.613 registros;
- Código CID com maior incidência era o M15.0 com 2.199 registros, podendo haver ocorrência de outros códigos CID no mesmo registro;
- Quantidade de registros isolados para o código M15.0 era de 659 registros (sem a presença de outros códigos CID).

Após uma limpeza preliminar (remoção de marcadores HTML) e uma reorganização da base de dados, pode-se afirmar que 67% dos registros da base de dados possuem código que iniciam pela letra *M*, que caracteriza doenças relativas à área de Reumatologia. Para formação dos conjuntos de dados visando à realização dos experimentos, foi realizado um levantamento estatístico de frequências e co-ocorrências dos códigos CID. As 5 maiores frequências constatadas são relacionadas na Tabela 4.

Tabela 4 - Frequências de CID

CID	DESCRIÇÃO CID	QNT REG
M54.2	Cervicalgia	931
E78.2	Hiperlipidemia mista	1474
I10	Hipertensão essencial (primária)	2095
M79.1	Mialgia	2119
M15.0	(Osteo)artrose primária generalizada	2199

Fonte: Elaborado pelo autor

Porém é necessário ressaltar que as frequências de CID listadas na Tabela 4, são as quantidades totais por CID, podendo ter a presença de outros códigos CIDs no mesmo registro. Desta forma, foi realizado um levantamento estatístico onde foram computadas somente as ocorrências de forma individual de cada CID, ou seja, sem a presença de outros CIDs, conforme é demonstrado na Tabela 5.

Tabela 5 - Maiores frequências individuais e associadas

CID	M15.0	M79.1	M54.2	I10	E78.2	OUTROS	TOTAL
M15.0	659	289	66	404	223	558	2199
M79.1	289	899	107	147	82	595	2119
M54.2	66	107	529	27	30	172	931
I10	404	147	27	425	522	570	2095
E78.2	223	82	30	522	390	227	1474

Fonte: Elaborado pelo autor

Analisando-se a Tabela 5, é possível notar a presença de 2 CIDs que não pertencem a área da Reumatologia listados no grupo das 5 maiores frequências da base de dados.

5.3 COMPOSIÇÃO DOS CONJUNTOS DE INTERESSE

A seleção dos conjuntos de interesse investigados ocorreu tendo em foco características da base de dados e estatísticas realizadas.

A principal característica da base de dados é que seu conteúdo armazenado é composto, em sua grande maioria, por dados da área da Reumatologia. Dessa forma, é possível separar os dados em duas classes: Uma

composta por conteúdo textual associado unicamente a doenças reumatológicas (classe *M*) com 13.345 registros; e outra classe composta por conteúdo textual que não está associado a doenças da área citada na classe anterior (*Não M* ou *Não M*) com 6.613 registros, assim compondo o primeiro conjunto, nomeado *M Não M*.

Os conjuntos de interesse posteriores foram definidos através de um levantamento estatístico realizado tendo como base as frequências. O segundo conjunto é composto pelos 5 códigos CIDs de maiores frequências (nomeado como 5 CIDs mais Frequentes). O terceiro conjunto é um subconjunto do segundo, sendo formado pelos 3 códigos CIDs com maiores frequências, mas somente relacionados à Reumatologia e titulado como 3 Ms mais Frequentes.

Conforme é demonstrado na Tabela 6, é possível identificar para os 3 conjuntos a quantidade geral e a quantidade exclusiva de registros para cada classe. Tomando como exemplo o conjunto *M Não M*, a quantidade geral é quantidade de registros com a presença de códigos CIDs começando com a letra *M*, em conjunto com CIDs de outros domínios. A coluna de quantidade exclusiva lista as quantidades de ocorrências dos CIDs, segundo um critério de exclusividade. Esse critério adotado permite a presença de somente códigos CIDs que comecem com a letra *M*, para classe *M*, e o contrário para a classe *Não M*.

Tabela 6 - Composição dos conjuntos de interesse com base na frequência

CONJUNTOS	CLASSES	QNT GERAL	QNT EXCLUSIVA
M Não M	CIDs com a letra M	13.345	9.349
	CIDs sem a letra M	6.613	6.613
5 mais frequentes	M54.2	931	529
	M79.1	2.119	899
	M15.0	2.199	659
	I10	2.095	425
	E78.2	1.474	390
3 M mais frequentes	M54.2	931	529
	M79.1	2.119	899
	M15.0	2.199	659

Fonte: Elaborado pelo autor

A partir do exposto, e focando os objetivos deste trabalho, deve ser aplicado nos 3 conjuntos o algoritmo de classificação escolhido. Com isso, é possível observar o comportamento do algoritmo em conjuntos com diferentes quantidades de registros, tendo em foco sua capacidade de classificação.

Durante a consulta, todas as observações e os demais fatos que o médico julgue importantes são transcritos no PME, em seu campo descritivo. O diagnóstico é realizado, através do cadastro de um código CID específico, que está vinculado ao conteúdo textual cadastrado. Tendo isso como base, a coerência do conteúdo descritivo com o código CID correspondente deve ser explorada, bem como verificar a existência de associações entre termos que descrevem sintomas, que sejam desconhecidas no domínio da reumatologia.

5.4 PRÉ-PROCESSAMENTO

O conteúdo descritivo possui dados clínicos sobre pacientes que buscaram alguma forma de tratamento. Esse conteúdo é muito variável, sendo possível listar alguns exemplos:

- Indicadores de exames e seus valores;
- Nomes de medicamentos com suas posologias;
- Queixas diversas sobre dores em locais pontuais;
- Informações sócio-econômicas sobre o paciente;
- Termos médicos diversos.

Através da Figura 6, são demonstrados alguns exemplos, retirados da base de dados, dos itens citados anteriormente.

Figura 6 - Exemplo de conteúdo a ser pré-processado

Conteúdo Descritivo	
	EFP= NL; P= 4,2; CA= 1,34; PCR= 0,23; PTH= 50; CA= 9,5; C= 0,65; TSH= 1,7; CALCIURIA= 182 MG/24h; Trouxe exames de ...
e	No intervalo foi a viagem para Munique a para Praga, gostou muito. Trouxe vitamina D = 22.60. Teve episódio de dor em ...
	Colonoscopia deu erticulose e colite segmentar em colon. A biopsia deu um processo inflamatório crônico inespecífico, ...
c	pneumo pediu PPD, negativo. Pediu nova TC após 3 meses. Pneumo achou que não tinha piorado, eu acho que houve ...
	Nasofibro: Rinite, hipertrofia de cornetos. EDA= esofagite distal leve. Hernia de hiato por deslizamento grau II. Melhora ...
	Paciente está tendo dor de estômago (suspeita de úlcera) irá realizar os exames esta semana. Disse que emagreceu (...
d	Ha cerca de 3 meses comprou videogame para filho e ficou com dor em 1ª MCF D. A seguir n trabalho notou algumas dores ...
	Trouxe resultado de exames: aC. uRICO= 6,2; FR= -; U= 41; CT= 193; TG= 237; GLIC= 95; C= 0,9; LDL= 108; HDL= 38; HMG= NL; ...
	Hoje me contou que e membra da I. Batista do Morumbi. Acha que melhorou com a fisioterapia e a medicação. Raramente ...
b	Uso Interno: Ciclobenzaprina 5 mg Famotidina 40 mg Paracetamol 250 mg formulário 60 cápsulas Diacereina 50 mg Meloxicam...

Fonte: Elaborado pelo autor

A utilização de dados em formato abreviado também é muito frequente, bem como a utilização de pontuação e caracteres especiais tais como ° & % # =, entre outros.

A quantidade de palavras armazenada em cada registro também é muito aleatória, variando de somente 2 palavras até conteúdos com mais de 120 palavras e com uma série de erros de ortografia.

Objetivando executar a limpeza dos conjuntos de dados selecionados e tentar aproximá-los de uma forma estruturada para submetê-los as técnicas de MT, foram aplicadas aos conjuntos de interesse os processos de Tokenização, *Stopwords*, *Stemming* e Redução da dimensionalidade.

As técnicas de Pré-Processamento foram executadas fazendo uso parcial de uma ferramenta desenvolvida por um aluno do programa de Iniciação Científica da UNISINOS, em um projeto que visava à consolidação de um ferramental para pré-processamento de textos. Essa ferramenta foi utilizada, por estar preparada para pré-processar dados descritivos oriundos da língua Portuguesa, enquanto outras ferramentas não possuem suporte para caracteres com acentos, gerando dados truncados no conteúdo textual. Complementar ao uso dessa ferramenta foi utilizado o *software WEKA*⁷ (*Waikato Environment for Knowledge Analysis*).

⁷ <http://www.cs.waikato.ac.nz/ml/weka/>

5.4.1 Tokenização

A Tokenização foi realizada usando a técnica *AlphabeticTokenizer*, implementada na ferramenta WEKA. A aplicação dessa técnica resulta na remoção de caracteres tais como . , ; : () e dígitos numéricos, restando somente unidades mínimas (*tokens*), formados apenas a partir de sequências alfabéticas contínuas.

5.4.2 Remoção de *Stopwords*

A técnica de remoção de *Stopwords* foi realizada fazendo uso de uma *stoplist* geral, formada por 382 termos que também são utilizados em outros domínios, e adaptada para área da saúde, uma vez que não existe uma lista específica para a área da medicina. A Tabela 7 lista alguns exemplos de *Stopwords* removidas bem como suas quantidades. A lista completa de *Stopwords* está presente no anexo I.

Tabela 7 - Exemplo de termos excluídos

TERMO REMOVIDO	QNT. REMOVIDA
Agora	818
Algumas	236
Alguns	391
Cada	219
Esta	4.014
Mesmo	713
Pelo	822

Fonte: Elaborado pelo autor

5.4.3 *Stemming*

O conteúdo descritivo possui uma grande incidência de termos com variação na sua grafia. Essa variação ocorre no seu sufixo, através de aumentativos e diminutivos entre outros, conforme relacionado a seguir:

- Sufixos: o termo varia no fim da palavra, tal como nos termos abdômen, abdominal, articulação e articular;
- Aumentativos como barrigão e vermelhão;

- Diminutivos como em dorzinhas e pedrinhas;
- A variação também pode ocorrer no plural como nos casos de náuseas e acidentes.

Após testes realizados com as 2 técnicas, o algoritmo de Porter (1980) mostrou melhores resultados que o algoritmo de Orengo e Huyck (2001), motivo pelo qual foi escolhido neste trabalho. Para a sua execução foi utilizada a ferramenta de Pré-Processamento de conteúdo textual desenvolvida na UNISINOS.

5.4.4 Redução de Dimensionalidade

Uma análise do conteúdo da base de dados mostrou que uma quantidade expressiva de palavras e ou termos compostas por 3 ou menos caracteres. Esse volume de palavras, além de dificultar o processamento elevando o custo computacional, também não possui valor significativo para a descrição da base de dados.

Em outra análise realizada, similar a anterior, foi constatada a presença de diversos textos com poucas palavras e de pouca contribuição, tanto para a tarefa de Classificação quanto para a extração de Regras de Associação. Esses textos foram medidos através da quantidade total de caracteres presentes em cada um, excluídos dessa forma todos os textos menores ou iguais a 30 a caracteres. A Tabela 8 demonstra as quantidades de textos excluídos nos conjuntos de interesse.

Tabela 8 - Composição dos conjuntos de interesse

CONJUNTO	QNT TEXTOS ORIGINAL	QNT TEXTOS EXCLUÍDAS	QNT TEXTOS RESULTANTES
<i>M Não M</i>	15.962	825	15.137
5 mais frequentes	2.902	173	2.729
3 mais frequentes	2.087	105	1.982

Fonte: Elaborado pelo autor

O percentual médio de textos excluídos com menos de 30 caracteres foi de 6%.

5.4.5 Seleção de Termos Relevantes

A aplicação das técnicas de Pré-Processamento, ao removerem termos sem valor representativo, tem como resultado uma prévia seleção de termos relevantes.

A execução de testes com frequências diversas, para definir a frequência mínima que um termo deveria possuir para ser representativo, demonstrou que a aplicação de filtros com valores de ocorrências acima de 10 geravam grandes quantidades de amostras, com poucos ou nenhum termo representativo. Levando tal fato em consideração, foi aplicado um filtro considerando as frequências 5, 8 e 10. Sendo assim, somente termos que ocorrem no mínimo em 5, 8 ou 10 documentos foram considerados para a representação dos textos. Esses valores foram escolhidos visando ao cuidado que alguns textos não ficassem sem representatividade. Também é possível citar como consequência da aplicação desse filtro a diminuição de ruído.

5.5 EXPERIMENTOS REALIZADOS

Nesta seção, são relatados os experimentos computacionais executados na base de dados, objetivando a produção de conhecimento, bem como a observação do comportamento dos algoritmos utilizados nos experimentos.

A realização dos experimentos computacionais nos conjuntos de interesse, descritos neste capítulo, focam na tarefa de Classificação e na extração de Regras de Associação. Para a realização destes, algumas operações específicas foram realizadas sobre o conteúdo descritivo, em virtude da técnica de mineração empregada ou do algoritmo utilizado.

Para a realização dos experimentos, foram utilizadas as frequências mínimas citadas associadas a 3 técnicas de transformação de dados clássicas e amplamente utilizadas na literatura: Binária, TF-IDF e TF.

O WEKA foi o *software* escolhido para a execução dos algoritmos de mineração aplicados nos experimentos deste trabalho. O mesmo foi desenvolvido pela universidade de Waikato na Nova Zelândia, é uma interface gráfica, que implementa técnicas Mineração de Dados tais como Árvores de

Decisão, Redes Neurais Artificiais, classificação com SVM e Regras de Associação, bem como outras etapas do processo de descoberta de conhecimento (WITTEN; FRANK, 2005).

A opção pelo uso dessa ferramenta foi devido ao fato dela implementar as técnicas e os algoritmos utilizados neste trabalho, a sua ampla utilização em trabalhos que abordam mineração textual, bem como o fato de ser um *software* livre, sobre licença *GNU - General Public License (GPL)*.

5.5.1 Classificação

Os experimentos foram realizados visando à tarefa de classificação dos registros, através de seu conteúdo textual. Contudo ao levar em conta as quantidades de registros de cada classe que compõe cada um dos conjuntos de interesse, destaca-se disparidade entre estas. Para eliminar essa diferença foi realizado um balanceamento nas classes que compõem cada conjunto.

5.5.1.1 Balanceamento das Classes

Efetuando uma comparação entre as quantidades de registros que compõem cada uma das classes em um mesmo conjunto de interesse, foi constatada uma diferença, em termos de percentual, que variam de 14% até 60% da classe com menor quantidade de registros em relação à classe majoritária, no conjunto dos 5 CIDs mais Frequentes. Para o conjunto dos 3 *Ms* mais Frequentes, a diferença foi entre 23% e 44%, considerando o conjunto *M Não M*, que possui somente 2 classes, a diferença percentual foi de 31%.

Os experimentos realizados com o conjunto *M Não M* obtiveram resultados pouco expressivos, que conduziram a equiparar o volume de dados das classes dentro de cada conjunto. No sentido de evitar um possível favorecimento de uma classe específica em detrimento de outra durante o processo classificatório, foi selecionada a mesma quantidade de registros da classe minoritária nas classes majoritárias.

Tomando como exemplo o conjuntos dos 5 CIDs mais Frequentes onde a classe E78.2 possui 390 registros, foram removidos de forma aleatória nas 4

demais classes deste conjunto, quantidades de registros que resultasse em igualar todas as 5 classes com 390 registros, formando um conjunto com 1.950 registros.

5.5.1.2 Escolha do Algoritmo

O algoritmo SVM alcançou significativos índices de Acurácia na classificação de conteúdo textual da área da saúde. A base de dados utilizada por Yan et al. (2010), composta por conteúdo descritivo relacionado a sintomas de doenças, e associados ao CID-9 obtiveram percentuais de Acurácia entre 95,3% e 99,8%, o que motivou a escolha deste algoritmo para a realização dos experimentos deste trabalho.

Os resultados obtidos na classificação das amostras foram avaliados segundo as métricas de Precisão e Abrangência. Isso se deve ao fato de ambas fornecerem uma visão do desempenho do algoritmo, levando em conta diferentes aspectos na classificação dos conjuntos. Também foi levada em consideração a ampla utilização dessas métricas na literatura.

5.5.1.3 Experimentos sobre o conjunto *M Não M*

O primeiro experimento realizado foi a classificação deste conjunto, com o algoritmo SVM aplicado as 3 técnicas de transformação (Binária, TF-IDF e TF), e nas 3 frequências mínimas de 5, 8 e 10 com respectivamente 4.947, 3.671 e 3.214 termos para o conjunto não balanceado e 4.811, 3.575 e 3.141 termos para o conjunto balanceado em cada frequência. A aplicação do algoritmo foi executada por diversas vezes, com diferentes valores para o parâmetro *Cost*, objetivando melhorar o percentual de Acurácia. Este parâmetro teve uma variação entre 3 e 20, atingindo o percentual máximo de Acurácia com o valor 5.

A Tabela 9 lista os percentuais de Acurácia obtida durante a realização dos experimentos para o conjunto *M Não M* não balanceado.

Tabela 9 - Percentuais de Acurácia conjunto não balanceado

TÉCNICA	ACURÁCIA (%) POR FREQUENCIA		
	5	8	10
Binário	76,27	76,44	76,40
TF-IDF	76,60	76,67	76,71
TF	76,38	76,49	76,47

Fonte: Elaborado pelo autor

Observando a Tabela 9, é possível constatar que os resultados de Acurácia para esse conjunto obtiveram os melhores percentuais na técnica de transformação TF-IDF, com a frequência 10 atingindo o valor máximo, se comparados aos percentuais obtidos pela técnica Binária e TF nas 3 frequências.

A Tabela 10 lista os resultados das métricas de Precisão, Abrangência e taxa de Falsos Positivos (FP) obtidos para cada técnica em cada frequência.

Tabela 10 - Resultados para as métricas obtidas com conjunto não balanceado

BINÁRIO			TF-IDF			TF								
		FP %	Abrang %	Prec %			FP %	Abrang %	Prec %			FP %	Abrang %	Prec %
5	M	42,10	75,40	88,90	5	M	38,00	76,80	86,70	5	M	41,40	75,70	88,60
	Não M	11,10	78,20	57,90			Não M	13,30	76,20			62,00	Não M	11,40
8	M	41,30	75,70	88,70	8	M	40,50	76,10	88,50	8	M	39,50	76,30	87,50
	Não M	11,30	78,00	58,70			Não M	11,50	78,00			59,50	Não M	12,50
10	M	40,80	75,90	88,20	10	M	40,00	76,20	88,20	10	M	39,70	76,20	87,60
	Não M	11,80	77,60	59,20			Não M	11,80	77,70			60,00	Não M	12,40

Fonte: Elaborado pelo autor

Na Tabela 10, é possível observar que os melhores percentuais individuais por classe foram alcançados na técnica de transformação Binária com frequência 5, com destaque para a classe M que obteve o maior valor de Precisão e FP de todos os experimentos, sendo que este último atingiu mais do que o dobro conforme destaque na Tabela 10.

O valor máximo para a métrica de Abrangência pertence à classe *Não M*, também listado na técnica Binária com frequência 5, porém com valores próximos em outras frequências e técnicas. A classe *Não M* se caracterizou em todos os experimentos pelos menores percentuais de FP.

As técnicas TF-IDF e TF também revelaram resultados próximos ao máximo obtido para a métrica de Precisão. Porém não houve diferenças expressivas, para cada classe, com a mudança da frequência.

Após a execução dos experimentos descritos anteriormente, o conjunto *M Não M* foi balanceado e uma nova bateria de testes foi iniciada com o algoritmo SVM. Os resultados de Acurácia estão listados na Tabela 11.

Tabela 11 - Percentuais de Acurácia conjunto balanceado

TÉCNICA	ACURÁCIA POR FREQUENCIA (%)		
	5	8	10
Binário	80,74	80,78	80,81
TF-IDF	80,98	81,10	80,96
TF	79,74	80,80	80,78

Fonte: Elaborado pelo autor

Observando os resultados de Acurácia apresentados na Tabela 11, é possível notar que novamente os melhores percentuais estão presentes quando foi utilizado a técnica TF-IDF, principalmente na frequência 8.

A Tabela 12 apresenta um resumo dos resultados das métricas obtidas para cada técnica em cada frequência para o conjunto balanceado.

Tabela 12 - Resultados para as métricas de dados balanceados

BINÁRIO		TF-IDF			TF									
		FP %	Abrang %	Prec %										
5	M	17,7 0	81,70	79,20	5	M	17,9 0	81,70	79,90	5	M	18,5 0	80,80	77,90
	Não M	20,8 0	79,80	82,30			Não M	20,1 0	80,30			82,10	Não M	22,1 0
		FP %	Abrang %	Prec %			FP %	Abrang %	Prec %			FP %	Abrang %	Prec %
8	M	17,8 0	81,70	79,40	8	M	17,6 0	81,90	79,80	8	M	17,7 0	81,80	79,30
	Não M	20,6 0	79,90	82,20			Não M	20,2 0	80,30			82,40	Não M	20,7 0
		FP %	Abrang %	Prec %			FP %	Abrang %	Prec %			FP %	Abrang %	Prec %

		%	%			%	%			%	%			
1	M	17,8	81,70	79,40	1	M	17,6	81,90	79,50	1	M	17,8	81,70	79,30
0	Não M	20,6	80,00	82,20	0	Não M	20,5	80,10	82,40	0	Não M	20,7	79,90	82,20

Fonte: Elaborado pelo autor

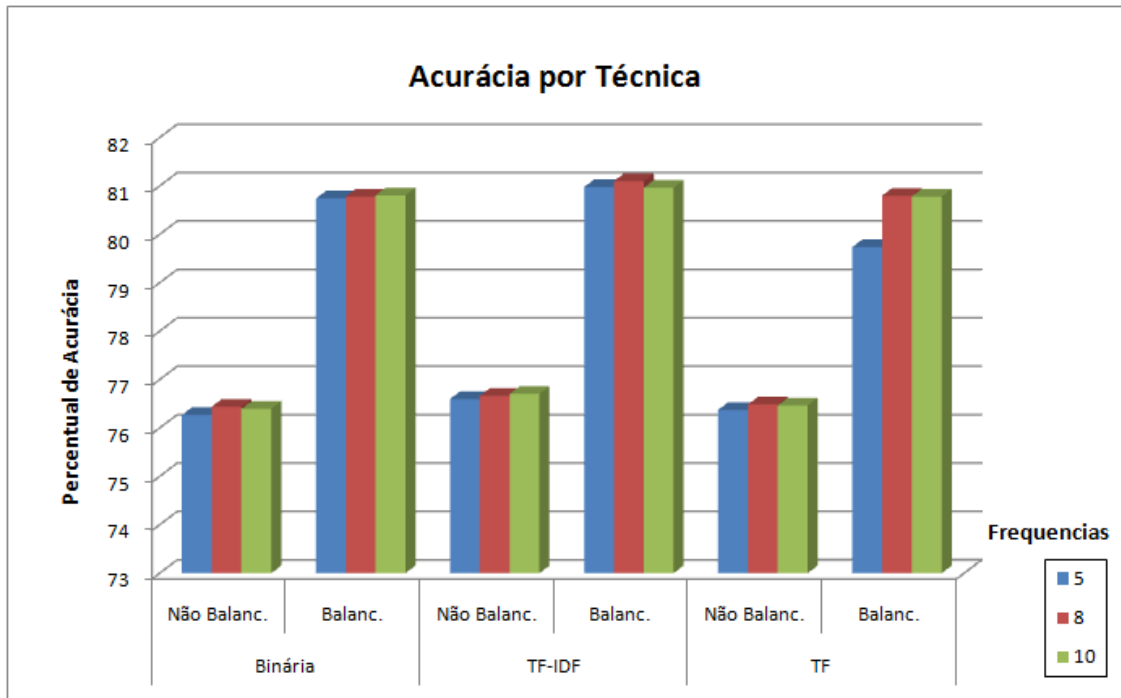
Observando a Tabela 12 de um modo geral, é possível destacar que os experimentos com o conjunto balanceado obtiveram em todos os experimentos valores de Precisão acima de 79%, com exceção para a técnica TF com frequência 5. Esta métrica na classe *M* teve uma redução de aproximadamente 9% em relação aos experimentos com conjunto não balanceado, porém com aumento de aproximadamente 20% para a classe *Não M*, mostrando equilíbrio nos valores alcançados para a métrica de Precisão.

O comportamento verificado quanto à distribuição de FP, relatado nos experimentos com o conjunto *M Não M* com classes não balanceadas, não foi novamente verificado, uma vez que os percentuais de FP estão mais próximos, mas com predominância de maiores valores para a classe *Não M*. Também é observado na Tabela 12 valores de Precisão e Abrangência próximos, com a menor diferença em 1,4% na técnica TF-IDF com frequência 5.

A diferença de quantidade de termos selecionados para o conjunto balanceado e não balanceado dentro de cada uma das 3 frequências utilizadas, variou em torno de 2,5%. A média de acerto obtida na classificação com o primeiro conjunto foi de 76,49%, com um valor médio de FP de 26,13%.

A classificação média do conjunto *M Não M* balanceado mostrou-se promissora se comparado com os resultados obtidos com o mesmo conjunto não balanceado. As médias de acerto e de FP obtidas com experimentos no primeiro conjunto foi de 80,75% e 19,26%, respectivamente. Analisando-se a técnica de transformação TF-IDF, de forma isolada, o conjunto balanceado com frequência 8 obteve 81,10% de Acurácia contra 76,66% do não balanceado. A Figura 7 mostra um gráfico dos percentuais de Acurácia para os 2 conjuntos, de cada uma das 3 técnicas de transformação com cada umas das frequências.

Figura 7 - Acurácia por técnica de transformação



Fonte: Elaborado pelo autor

O gráfico da Figura 7 mostra que os experimentos realizados com conjunto balanceado alcançaram os melhores percentuais de Acurácia em comparação com o conjunto não balanceado, principalmente na técnica de TF-IDF com frequência 8.

Analisando a Matriz de Confusão apresentada na Tabela 13, é possível observar que existe um maior equilíbrio nas quantidades de registros classificados corretamente com o conjunto balanceado. No experimento com o conjunto não balanceado, a classe *M* obteve mais do que o dobro em amostras reconhecidas corretamente que a classe *Não M* (valores em destaque na Tabela 13).

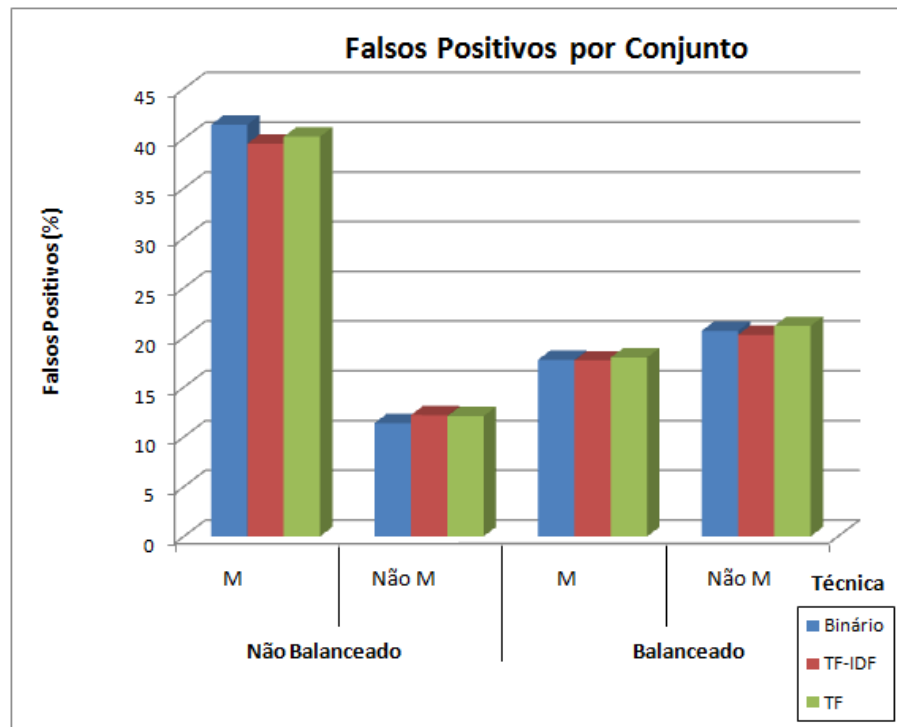
Tabela 13 - Matriz de confusão conjunto *M Não M*

TF-IDF					
Não Balanceada			Balanceada		
Matriz de Confusão			Matriz de Confusão		
M	Não M		M	Não M	
88,48%	11,52%	M	79,78%	20,22%	M
40,52%	59,48%	Não M	17,58%	82,42%	Não M

Fonte: Elaborado pelo autor

A classificação do conjunto balanceado também obteve menores índices de FP na comparação com o conjunto não balanceado conforme demonstra o gráfico de taxa média de FP na Figura 8.

Figura 8 - Gráfico de taxa média de FP



Fonte: Elaborado pelo autor

O conjunto não balanceado concentrou a maior média de FP na classe *M* (mais que o dobro da classe *Não M*), já os experimentos com o conjunto balanceado tiveram as médias de FP com valores próximos nas 2 classes.

5.5.1.4 Experimentos sobre o conjunto dos 5 CIDs mais Frequentes

Os experimentos realizados no conjunto *M Não M*, apresentaram melhores resultados no conjunto balanceado, dessa forma os experimentos executados no conjunto dos 5 CIDs mais Frequentes seguiram a mesma sistemática, iniciando os experimentos pelo conjunto não balanceado, para depois efetuar uma comparação com os resultados obtidos com o conjunto balanceado.

O conjunto dos 5 CIDs mais Frequentes não balanceado, totaliza 2.729 amostras, onde foram aplicadas as técnicas de transformação Binária, TF-IDF e TF, com as frequências 5, 8 e 10 para 1.239, 798 e 663 termos respectivamente.

A versão do mesmo conjunto, porém balanceado com 1.725 amostras utilizou as mesmas técnicas e também as frequências 5, 8 e 10 para 1.072, 649 e 524 termos respectivamente.

A aplicação do algoritmo SVM foi executada por diversas vezes, variando o valor para o parâmetro *Cost*, com o objetivo de melhorar o percentual de Acurácia. O valor desse parâmetro variou entre 5 e 25, atingindo o percentual máximo de Acurácia com valores próximos a 12. A Tabela 14 lista os percentuais de Acurácia obtidos durante a realização dos experimentos para o conjunto dos 5 CIDs mais Frequentes não balanceados.

Tabela 14 - Resultados de classificação conjunto não balanceado

TÉCNICA	ACURÁCIA POR FREQUENCIA (%)		
	5	8	10
Binário	56,83	57,75	57,38
TF-IDF	59,00	58,92	58,26
TF	57,60	57,86	57,82

Fonte: Elaborado pelo autor

Os resultados apresentados na Tabela 14, embora tenham ficado abaixo dos resultados do conjunto anterior, novamente concentram os melhores resultados na técnica de transformação TF-IDF, com a melhor Acurácia na frequência 5 (em destaque na Tabela 14).

Os experimentos realizados com o conjunto dos 5 CIDs mais Frequentes, obtiveram menores índices de FP na utilização da técnica Binária que teve valores menores que as outras 2 técnicas. As classes com maiores percentuais de FP foram as classes M15.0 e M79.1, com destaque para a classe M79.1, que obteve percentual superior frente as outras 4 classes, conforme destaque na Tabela 15 que apresentando os resultados das 3 técnicas para a frequência 5.

Tabela 15 - Lista de métricas para o conjunto 5 CIDs mais Frequentes não balanceado frequência 5

	BINÁRIA			TF-IDF			TF					
	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %			
5	E78.2	3,70	61,30	40,90	E78.2	3,90	62,20	44,30	E78.2	4,20	59,60	43,20
	I10	2,90	61,10	26,60	I10	4,20	59,40	35,30	I10	4,20	57,80	33,30
	M15.0	13,00	57,60	58,60	M15.0	13,50	57,30	60,20	M15.0	14,10	55,50	58,90
	M54.2	5,90	62,80	46,10	M54.2	6,70	63,10	53,30	M54.2	6,90	60,90	49,80
	M79.1	33,20	53,50	81,90	M79.1	26,60	57,80	78,10	M79.1	27,20	57,20	78,00

Fonte: Elaborado pelo autor

Entre as 3 classes relacionadas à área da Reumatologia, somente a classe M54.2 obteve melhor percentual de Abrangência do que as classes E78.2 e I10, isso pode ser observado nas 3 técnicas usadas (conforme destaques na Tabela 15). O experimento com a técnica Binária na frequência 5 obteve os melhores valores de Abrangência em comparação com as outras 2 técnicas.

A métrica de Precisão obteve os melhores resultados na aplicação da técnica TF-IDF, com exceção para a classe M79.1 que alcançou o melhor percentual de todos os experimentos na técnica Binária.

Os valores de Abrangência das classes E78.2 e I10 na técnica Binária, obtiveram valores muito próximos, com uma diferença de 0,20%.

Após a realização dos experimentos com o conjunto dos 5 CIDs mais Frequentes não balanceados, foram realizados os experimentos de classificação com o mesmo conjunto, porém balanceado.

Os experimentos com o conjunto dos 5 CIDs mais Frequentes obteve, similarmente ao conjunto *M Não M*, melhores resultados de Acurácia com o conjunto balanceado, embora em menores índices que os atingidos pelo conjunto *M Não M*, conforme é demonstrado na Tabela 16.

Tabela 16 - Resultados de classificação para classes balanceadas

TÉCNICA	ACURÁCIA POR FREQUENCIA (%)		
	5	8	10
Binário	58,90	58,26	58,43
TF-IDF	60,99	59,42	59,01
TF	58,49	58,38	58,67

Fonte: Elaborado pelo autor

A exemplo dos experimentos anteriores, os resultados de classificação obtiveram melhores índices de Acurácia nas 3 frequências, quando foi utilizada a técnica de transformação TF-IDF.

Analisando os resultados dos experimentos com o conjunto balanceado apresentados na Tabela 17 em comparação com os resultados do conjunto não balanceado, nota-se maiores valores de FP nas classes, porém sem a concentração nas classes M79.1 e M15.0, constatada no conjunto não balanceado. Nos experimentos com o conjunto balanceado a classe M79.1 obteve os menores valores de FP.

Tabela 17 - Resultados de classificação para o conjunto 5 CIDs mais Frequentes balanceado frequência 5

	BINÁRIA			TF-IDF			TF					
	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %			
5	E78.2	13,20	53,90	61,70	E78.2	12,90	55,40	64,10	E78.2	14,60	52,60	64,60
	I10	9,40	52,90	42,30	I10	8,30	57,30	44,30	I10	8,60	53,10	39,10
	M15.0	11,40	57,50	62,00	M15.0	10,70	59,10	62,00	M15.0	11,20	58,30	62,90
	M54.2	10,70	61,30	67,80	M54.2	10,70	62,40	70,70	M54.2	11,70	59,30	68,10
	M79.1	6,60	69,70	60,60	M79.1	6,20	71,90	63,80	M79.1	5,80	71,30	57,70

Fonte: Elaborado pelo autor

Com o conjunto balanceado os percentuais de Precisão principalmente na técnica TF-IDF, foram maiores na comparação com experimentos anteriores. Já a métrica de Abrangência teve comportamento inverso, havendo uma redução em seus valores. Os resultados de Precisão na técnica Binária para as classes E78.2 e M15.0, foram próximos com uma diferença de 0,30%, fato que não foi verificado nas demais técnicas.

A Matriz de Confusão apresentada na Tabela 18 demonstra um aumento nas quantidades de acertos nas 2 classes compostas por registros não ligados a

Reumatologia (E78.2 e I10), quando o experimento foi realizado com o conjunto balanceado, porém com maiores quantidades de FP.

Tabela 18 - Matrizes de confusão 5 CIDs mais Frequentes

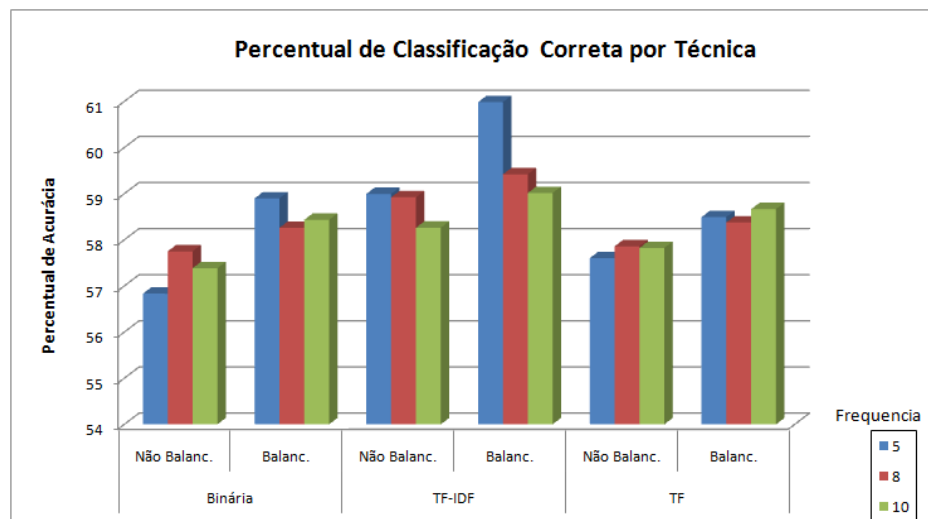
		TF-IDF										
Freq.		Não Balanceada					Balanceada					
		Matriz de Confusão					Matriz de Confusão					
		E78.2	I10	M15.0	M54.2	M79.1	E78.2	I10	M15.0	M54.2	M79.1	
5	E78.2	44,35%	11,30%	16,81%	4,64%	22,90%	E78.2	64,06%	12,46%	9,56%	9,86%	4,06%
	I10	10,70%	35,32%	25,62%	3,98%	24,38%	I10	25,51%	44,35%	17,39%	8,69%	4,06%
	M15.0	3,97%	5,40%	60,16%	7,30%	23,17%	M15.0	11,30%	9,86%	62,03%	8,98%	7,83%
	M54.2	1,03%	0,82%	9,30%	53,31%	35,54%	M54.2	10,14%	3,77%	6,38%	70,72%	8,99%
	M79.1	2,30%	2,30%	8,88%	8,41%	78,11%	M79.1	4,64%	6,96%	9,56%	15,07%	63,77%

Fonte: Elaborado pelo autor

Já as classes relacionadas à área da Reumatologia apresentaram uma redução na quantidade de acertos com o conjunto balanceado, com destaque para a classe M79.1 que sofreu a maior redução.

Embora os resultados dos experimentos de classificação do conjunto dos 5 CIDs mais Frequentes tenha retornado resultados abaixo dos apresentados no conjunto *M Não M*, novamente os melhores resultado são retornados nas classes com dados balanceados, principalmente na técnica de transformação TF-IDF. A única variação foi quanto a frequência de melhor desempenho, que neste experimento foi a frequência 5, conforme pode ser constatado no gráfico da Figura 9.

Figura 9 - Gráfico de Acurácia conjunto 5 CIDs mais Frequentes



Fonte: Elaborado pelo autor

A realização de experimentos com o conjunto balanceado retornaram uma distribuição equilibrada de FP nas 5 classes que compõem este conjunto. A Tabela 19 lista as médias de FP para os conjuntos balanceado e não balanceado por técnica de transformação.

Tabela 19 - Quantidade média de FP por classe

Classe	BINÁRIA					TF-IDF					TF				
	E78.2	I10	M15.0	M54.2	M79.1	E78.2	I10	M15.0	M54.2	M79.1	E78.2	I10	M15.0	M54.2	M79.1
Conjunto não balanceado (Qnt)	26	25	74	38	128	23	27	72	40	120	28	31	77	43	109
Conjunto Balanceado (Qnt)	48	31	39	37	23	48	30	38	35	23	49	31	39	38	22

Fonte: Elaborado pelo autor

Ao observar a Tabela 19, é possível constatar que as classes relacionadas à área da Reumatologia do conjunto balanceado sofreram uma redução na quantidade de FP, com destaque para a classe M79.1 que possuía uma grande concentração de FP no conjunto não balanceado. A classe E78.2 alcançou a maior média de FP na aplicação das 3 técnicas de transformação.

5.5.1.5 Experimentos sobre o conjunto 3 Ms mais Frequentes

O conjunto formado pelos 3 códigos CIDs relacionados exclusivamente à Reumatologia com o conjunto não balanceado contém 1.982 textos e balanceado possui 1.452 textos. Como nos outros experimentos, foram utilizadas as técnicas de transformação Binária, TF-IDF e TF com as frequências mínimas de 5, 8 e 10 com 1.077, 714 e 595 termos para o conjunto não balanceado e 986, 638 e 534 com o conjunto balanceado.

A aplicação do algoritmo SVM teve uma variação do parâmetro *Cost* para esses experimentos entre 3 e 20, obtendo o valor máximo de Acurácia para o conjunto próximo a 15.

A Tabela 20 lista os percentuais de Acurácia obtidos nas 3 técnicas de transformação em todas as frequências definidas para classes não balanceadas.

Tabela 20 - Percentuais de Acurácia para conjunto não balanceado

TÉCNICA	ACURÁCIA POR FREQUENCIA (%)		
	5	8	10
Binário	68,82	68,52	68,87
TF-IDF	69,17	68,87	68,77
TF	68,42	68,72	68,87

Fonte: Elaborado pelo autor

Observando a Tabela 20 é possível constatar novamente que os melhores índices de Acurácia foram obtidos quando foi utilizada a técnica de transformação TF-IDF, alcançando seu máximo na frequência 5, com exceção para a frequência 10 que obteve percentuais abaixo das demais técnicas.

A classe M79.1 da mesma forma que ocorreu no conjunto dos 5 CIDs mais Frequentes, obteve uma concentração expressiva de FP nos experimento com o conjunto não balanceado, enquanto que nas outras 2 classes a distribuição é equilibrada. A classe M79.1 também despontou no valor de Precisão com valores acima de 80% enquanto as outras classes obtiveram valores menos expressivos, tal comportamento se repete nas 3 técnicas, conforme mostra os valores em destaque na Tabela 21.

Tabela 21 - Resultados de classificação para conjunto não balanceado
frequência 5

	Binária			TF-IDF			TF					
	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %			
5	M15.0	8,50	77,80	64,10	M15.0	8,90	77,00	64,10	M15.0	10,40	74,70	65,60
	M54.2	7,90	66,00	47,70	M54.2	7,90	68,10	52,10	M54.2	8,90	64,50	50,00
	M79.1	34,50	65,50	84,00	M79.1	33,40	65,80	82,40	M79.1	31,70	66,50	80,08

Fonte: Elaborado pelo autor

A Tabela 21 mostra que a classe M15.0 obteve os maiores percentuais de Abrangência nos 3 experimentos com valores acima dos 70%. Na técnica Binária as classes M54.2 e M79.1 obtiveram percentuais de Abrangência muito próximos.

O experimento realizado com a aplicação do classificador no conjunto balanceado dos 3 Ms mais Frequentes, obtém os percentuais de Acurácia listados na Tabela 22.

Tabela 22 - Percentuais de Acurácia conjunto balanceado

TÉCNICA	ACURÁCIA POR FREQUENCIA (%)		
	5	8	10
Binário	75,14	74,86	74,31
TF-IDF	75,00	74,59	74,86
TF	74,72	74,59	74,45

Fonte: Elaborado pelo autor

Os melhores percentuais de Acurácia, nas 3 técnicas de transformação, estão concentrados na frequência 5 e 8 da técnica Binária, sendo que o melhor valor obtido para a frequência 10 foi na técnica TF-IDF, conforme destaques na Tabela 22.

Com comportamento similar aos demais experimentos realizados com conjuntos balanceados, o percentual de FP teve uma distribuição mais equilibrada, com destaque da classe M54.2, cujo valor despontou dos demais. A classe M79.1, que antes possuía o maior percentual de FP, passou a ter o menor, conforme destaque na Tabela 23.

Tabela 23 - Métricas obtidas para conjunto balanceado frequência 5.

	BINÁRIA			TF-IDF			TF					
	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %	FP %	Abrang %	Prec %			
5	M15.0	10,60	77,70	74,20	M15.0	10,00	78,20	71,90	M15.0	11,50	76,50	74,80
	M54.2	16,20	71,30	80,60	M54.2	17,50	70,10	82,00	M54.2	16,70	71,00	81,80
	M79.1	10,40	77,20	70,70	M79.1	10,00	78,00	71,10	M79.1	9,70	77,70	67,60

Fonte: Elaborado pelo autor

As classe M79.1 e M15.0 alcançaram os melhores percentuais de Abrangência dos experimentos com a diferença máxima de 8,1% em relação a classe M54.2. Na técnica TF-IDF, onde os maiores percentuais de Abrangência foram alcançados, a diferença entre as classes M79.1 e M15.0 foi de 0,20%. Uma pequena diferença de percentual também pode ser observado na técnica Binária para estas 2 classes.

É possível constatar um aumento nos percentuais de Precisão para as classes M15.0 e M54.2, o mesmo não pode ser verificado para a classe M79.1 que sofreu uma redução para essa métrica, com seu mínimo atingido na técnica TF. As classes M79.1 e M15.0 apresentam valores próximos para a métrica de

Precisão na técnica TF-IDF, com 0,80% de diferença. A classe M54.2 obteve aumentos expressivos para esta métrica, atingindo a 32,9% na técnica Binária.

Analisando a Matriz de Confusão dos melhores resultados obtidos (ambos na mesma frequência, porém em técnicas de transformação diferentes), as maiores classificações corretas com conjunto não balanceado, foram obtidas nas classes com maiores quantidades de amostras (M15.0 e M79.1), comportamento que não ocorreu no experimento com conjunto balanceado, onde houve um equilíbrio na classificação, conforme é demonstrado na Tabela 24.

Tabela 24 - Matriz de Confusão geradas

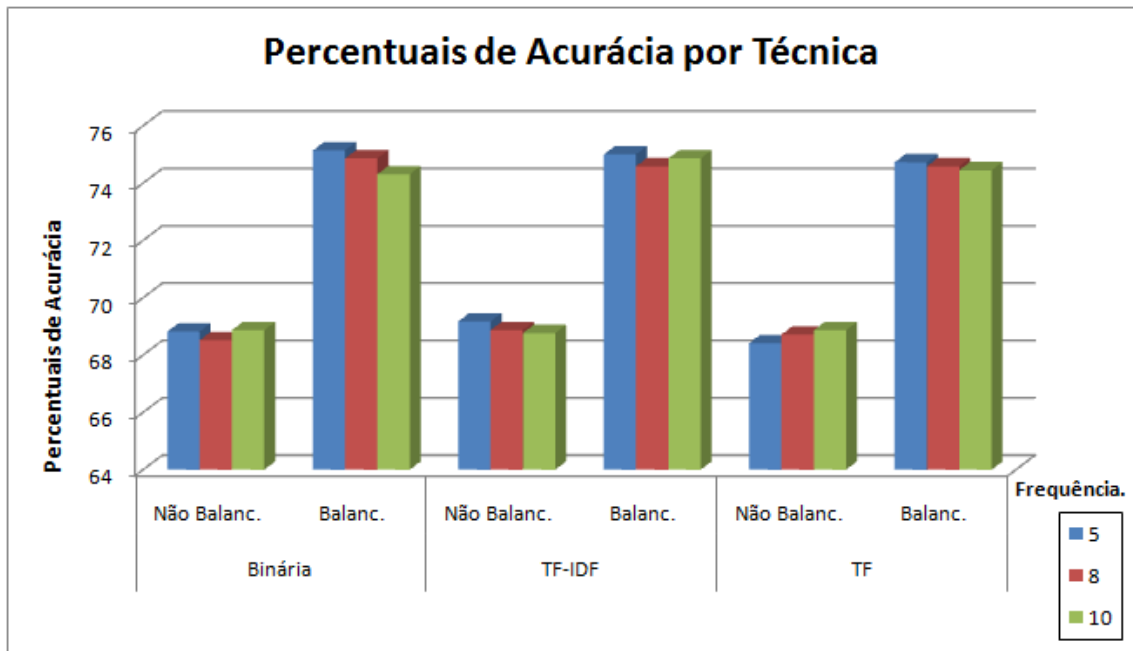
Freq	TF-IDF Não Balanceada			Binária Balanceada				
	Matriz de Confusão			Matriz de Confusão				
	M15.0	M54.2	M79.1		M15.0	M54.2	M79.1	
5	64,13%	7,46%	28,41%	M15.0	74,17%	16,73%	9,10%	M15.0
	8,06%	52,07%	39,87%	M54.2	7,64%	80,58%	11,78%	M54.2
	9,45%	8,18%	82,37%	M79.1	13,63%	15,70%	70,67%	M79.1

Fonte: Elaborado pelo autor

Também é possível observar na Tabela 24 uma redução na quantidade de FP para as classes M15.0 e M79.1 e um aumento na classe M54.2.

Igual aos experimentos anteriores, os resultados alcançados com o conjunto dos 5 CIDs mais Frequentes balanceado, revelaram melhores resultados na comparação com o mesmo conjunto não balanceado, conforme demonstra o gráfico na Figura 10.

Figura 10 - Gráfico de Acurácia entre técnicas de transformação

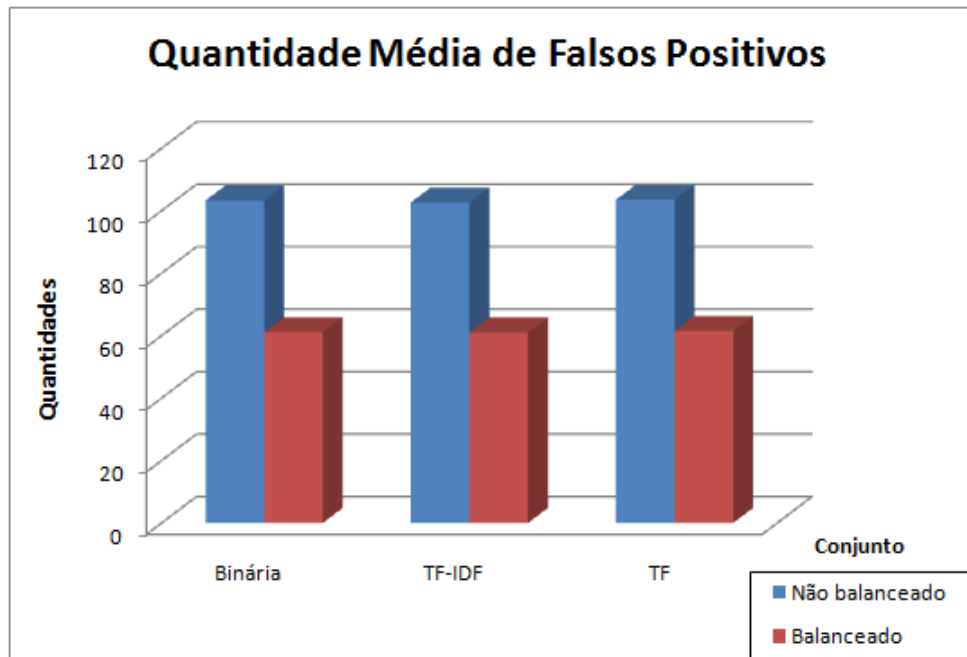


Fonte: Elaborado pelo autor

Os resultados nas 3 técnicas de transformação no conjunto balanceado obtiveram percentuais próximos e com Acurácia acima dos 74%, com destaque para a técnica de transformação Binária com frequência mínima 5 que obteve 75,14%, enquanto o melhor resultado obtido com os experimentos com o conjunto não balanceado foi de 69,17% na técnica TF-IDF também com frequência 5. Os valores obtidos com conjunto não balanceado nas demais frequências em todas as 3 técnicas foram próximos, tendo atingido seu mínimo em 68,42% na técnica TF com frequência 5.

O conjunto balanceado apresentou menores quantidades de FP, conforme demonstram as médias apresentadas pelo gráfico da Figura 11.

Figura 11 - Gráfico de média de FP



Fonte: Elaborado pelo autor

A técnica TF-IDF apresentou a média de 60,94, a menor nos experimentos com conjunto balanceado, e a maior foi de 61,50 atingida pela técnica TF. Os experimentos com conjunto não balanceado obtiveram a média mínima de 102,61 na técnica TF-IDF, e a máxima de 103,5 na técnica TF.

5.5.1.6 Discussão dos Resultados da Classificação

Os resultados apresentados nos experimentos realizados no conjunto *M Não M*, evidenciam um melhor desempenho do classificador, quando este conjunto de dados teve suas classes balanceadas. Embora os resultados mínimos de Acurácia dos experimentos com as classes não balanceadas, não fossem menos de 76%, com as classes balanceadas, estes percentuais atingiram um valor mínimo de 79,74% e máxima de 81,10%.

A taxa de FP do experimento com melhor resultado do conjunto não balanceado, ficou concentrada principalmente na classe *M*, com uma diferença de 29% entre as duas classes. Com a realização dos experimentos com o conjunto balanceado esta diferença foi reduzida a 3%, porém com a classe *Não M* com maior valor.

Os experimentos realizados com o conjunto não balanceado revelam um favorecimento do classificador para a classe *M*, em detrimento da classe *Não M*. O classificador obteve, na frequência de melhor percentual de Acurácia, reconhecimento de 88,20% das amostras da classe *M* e 60% da classe *Não M*. Já a realização dos experimentos com o conjunto balanceado, obteve Precisão de 79,80% de reconhecimento para a classe *M*, e 82,40% para a classe *Não M*. A diferença em prol da classe *Não M* é de 3% e foi uma constante nos experimentos com o conjunto balanceado.

Para o conjunto não balanceado a elevada quantidade de registros que a classe *M* possui quando comparada com a classe *Não M*, aliada a uma provável linguagem melhor consolidada para descrever sintomas e prescrição de tratamentos, para doenças da área da Reumatologia (que compõem a classe *M*), podem ter influenciado neste resultado.

A classe *M* do conjunto não balanceado sofreu uma redução de 31,25% na sua quantidade de registros para ficar equiparada a classe *Não M*, o que resultou em perda de representatividade e uma redução de 8,40% de acerto para esta classe. Já para a classe *Não M* o balanceamento do conjunto elevou a Precisão em 22,40%. Mesmo a classe *M* sendo composta por registros relativos à especialidade da instituição de saúde geradora dos dados, a classe *Não M*, apresenta melhores características representativas quando o conjunto é balanceado. Isso pode ocorrer devido a existência de termos que estão presentes nas 2 classes com maior quantidade de ocorrências para a classe *M*, devido a sua maior quantidade de registros. Com a equiparação da classe *M*, a quantidade destes termos representativos também é reduzida, passando a classe *Não M* a ter a maior quantidade de ocorrências, elevando a sua representatividade e conseqüente melhora no percentual de Precisão.

A realização de experimentos de classificação com o conjunto de dados composto pelos 5 CIDs mais Frequentes, com classes não balanceadas obteve percentual máximo de Precisão de 58,99%, e mínimo de 56,83%, índices inferiores aos obtidos com o conjunto *M Não M*. Nas Matrizes de Confusão geradas para os experimentos, a classe M79.1 é a que possui maiores quantidades de FP, com maior contribuição das classes M54.2 e M15.0, o que faz sentido pois são todas classes ligadas a área da Reumatologia.

Após a realização de experimentos com classes balanceadas, a métrica de FP apresentou valores mais equilibrados em todas as classes, com destaque para a redução que obteve a classe M79.1. Os experimentos com conjunto balanceado também obtiveram maiores valores de Abrangência com as classes *M* em relação ao não balanceado. Com exceção da classe da área M79.1 que sofreu uma redução no percentual de Precisão, as demais classes tiveram acréscimo em seus percentuais.

Novamente os maiores percentuais de Acurácia foram obtidos com as classes minoritárias com a aplicação da técnica de transformação TF-IDF, mas variando a frequência que para esse conjunto o melhor resultado foi obtido na frequência 5.

Os experimentos realizados com as classes não balanceadas obtiveram uma grande quantidade (o dobro ou mais dependendo da classe que se compara) de amostras reconhecidas pela classe M79.1, fato que não ocorreu com a realização dos experimentos com classes balanceadas, ficando a classe M79.1 equiparada às demais classes. O CID (classe) M79.1, correspondente a Mialgia, segundo especialistas consultados, é o diagnóstico do sintoma de dores musculares. Este CID é o que possui maior predominância na quantidade de registros deste conjunto, que pode ter ocasionado uma maior representatividade dos termos a ele relacionados perante as classes minoritárias, representatividade que ficou equiparada no balanceamento do conjunto, pela redução de registros e pela possibilidade de características relacionadas a este CID, serem encontradas em outros CIDs deste conjunto.

Os experimentos realizados no conjunto dos 3 *Ms* mais Frequentes obtiveram valores de classificação superiores aos experimentos do conjunto dos 5 CIDs mais Frequentes. Os experimentos com classes não balanceadas obtiveram resultados de Acurácia acima dos 68%, com seu valor máximo em 69,17% atingido na frequência 5 da técnica de transformação TF-IDF. Aplicando o classificador nos dados balanceados o percentual de Acurácia mínimo foi de 74,31%, e seu máximo de 75,13% atingido com a técnica transformação Binária com frequência mínima 5, valores também superiores aos alcançados pelo conjunto dos 5 CIDs mais Frequentes.

Analisando as métricas geradas para o conjunto não balanceado, novamente a classe M79.1 desponta com a maior taxa de FP das 3 classes que compõem o conjunto de interesse. Após executar o balanceamento a taxa FP para a classe M79.1 diminui e as outras classes aumentam um pouco, ficando os valores equilibrados nas 3 classes.

A classe M79.1 também apresentou o maior valor de Precisão no conjunto não balanceado, com uma diferença de 40% dependendo da classe comparada. Com a realização do experimento no conjunto balanceado. A classe M79.1 teve uma redução nesta métrica, apresentando valores compatíveis com as demais classes, porém destacando-se agora a classe M54.2 com melhor percentual de Precisão.

Observando as Matrizes de Confusão geradas (Tabela 24), destaca-se novamente a grande quantidade de amostras reconhecidas da classe M79.1 com classes não balanceadas, e esta quantidade sendo reduzida com a realização dos experimentos em classes balanceadas.

Os melhores resultados nos 3 experimentos de classificação realizados, foram obtidos com a utilização da técnica de transformação TF-IDF, principalmente quando foi utilizado o valor intermediário das 3 frequências (8). As técnicas TF-IDF e Binária mostraram maior equilíbrio nos resultados quando se utilizou diferentes valores para frequência. O mesmo equilíbrio nos resultados não foi observado na técnica de transformação TF, que teve o valor de classificação das frequências 8 e 10 se destacando em relação a frequência 5. Como esperado, o classificador mostrou melhor desempenho quando o conjunto possuía maior quantidade de amostras e menor quantidade de classes, caso do conjunto *M Não M*. Nota-se também que os resultados são equilibrados quando o conjunto é homogêneo quanto a suas classes, como foram os resultados dos conjuntos *M Não M* e dos 3 *Ms* mais Frequentes. Quando o conjunto de dados possui diversidade em suas classes (caso do conjunto dos 5 CIDs mais Frequentes que possui 2 classes fora da área da Reumatologia), ocorre uma disparidade nos resultados com máximos e mínimos se destacando, o que indica um melhor aproveitamento do classificador para conjuntos homogêneos.

5.5.2 Regras de Associação

Os experimentos realizados nessa etapa do trabalho visaram à descoberta de padrões, aplicando a técnica de Regras de Associação no conteúdo textual de um sistema de PME. O conteúdo descritivo da base de dados estudada, como visto na seção 5.1.2, é composto por sintomas, resultados de exames entre outros, que o médico observa na hora da consulta.

O conteúdo descritivo está sempre associado a 1 ou mais códigos CIDs, que compõem o diagnóstico da(s) doença(s). Os experimentos com a técnica de Regras de Associação focam na extração de associações entre termos que descrevem esses diagnósticos, na tentativa de descobrir associações que não são peculiares a cada doença, bem como a validação do conteúdo textual com o código CID associado.

Os experimentos de extração de Regras de Associação foram realizados primeiro nos conjuntos de dados não balanceados dos 5 CIDs mais Frequentes e 3 Ms mais Frequentes. Entretanto, alguns experimentos exigiram que o limiar de suporte fosse muito reduzido e mesmo assim, retornaram regras que contemplavam somente uma classe. Em busca da extração de regras com percentual de suporte maior, foram executados experimentos também com o conjunto de dados balanceados.

Para a realização dos experimentos, os conjuntos foram pré-processados com aplicação da técnica de Tokenização, remoção de *Stopwords*, finalizando com a aplicação da técnica de transformação Binária. Visando obter regras mais claras e facilitar a interpretação das mesmas, não foi utilizada a técnica de redução da palavra ao seu radical (*Stemming*).

Todos os experimentos foram iniciados começando com os valores de 0,1 para a métrica de suporte e 0,7 para a métrica de confiança. A partir dos resultados retornados para esses parâmetros iniciais, como baixas quantidades de regras, o experimento era realizado novamente, porém diminuindo os limiares de suporte e de confiança de forma gradativa, visando à preservação dos maiores valores para essas métricas.

Os experimentos foram realizados com foco nas classes de cada conjunto, com a utilização do CAR (descrito na seção 3.2.3.2) e com a forma clássica do algoritmo *Apriori*.

Para análise e validação das regras extraídas dos conjuntos de interesse, foram consultados 2 especialistas da área da Reumatologia.

5.5.2.1 Experimentos sobre o conjunto 5 CIDs mais Frequentes

O conjunto dos 5 CIDs mais Frequentes é composto por 2.729 textos (1.725 balanceado) associados as 5 classes que compõem este conjunto com 3.786 termos selecionados. O primeiro experimento realizado foi a extração de regras com foco nas 5 classes que compõem o conjunto, com os valores iniciais para as métricas de suporte e confiança de 0,1 e 0,7, respectivamente. Esta configuração inicial não retornou nenhum tipo de regra, o que ocasionou a redução dos valores iniciais para realizar os experimentos seguintes. Os valores em questão foram reduzidos gradativamente, visando à extração de regras com os valores mais elevados possíveis para os limiares de suporte e confiança. Esta sistemática foi executada para a extração de regras fazendo uso do algoritmo *Apriori* clássico.

Foram extraída 6 regras relacionadas a classe M79.1 com 0,04 e 0,60 para os limiares de suporte e confiança. Uma vez que as 6 regras eram relacionadas a apenas uma classe, buscou-se uma alternativa que extraísse regras que contemplassem as outras classes do conjunto.

A Tabela 25 lista os experimentos realizados a cada redução nos valores de suporte e confiança (alternando a extração de regras com e sem a utilização do CAR) e as quantidades de regras extraídas para esses valores.

Tabela 25 - Quantidade de regras extraídas

SUPORTE	CONFIANÇA	CAR	QNT. DE REGRAS
0,10	0,70	<i>true</i>	0
0,04	0,70	<i>true</i>	1
0,03	0,70	<i>true</i>	8
0,10	0,70	<i>false</i>	0
0,06	0,70	<i>false</i>	1
0,04	0,70	<i>false</i>	10
0,04	0,60	<i>false</i>	19

Fonte: Elaborado pelo autor

Os experimentos realizados com as classes balanceadas retornaram 8 regras, 6 associadas à classe M79.1 e 2 associadas à classe M54.2.

A extração de regras sem a utilização de CAR no conjunto não balanceado, foi possível somente com os limiares de suporte e confiança estabelecidos em 0,05 e 0,70 (com este limiares em 0,06 e 0,70 foi extraída somente uma regra), respectivamente, o que gerou a extração de 4 regras. Repetindo o experimento com as classes balanceadas configurando os limiares em 0,04 e 0,60, foi possível a extração de 19 regras com valores de confiança acima de 0,70. A lista das regras obtidas nos experimentos com classes balanceadas estão listadas na Tabela 26.

Tabela 26 - Lista de regras extraídas para o conjunto dos 5 CID mais Frequentes

Parâmetros Apriori			Regra	Confiança
Suporte	Confiança	CAR		
0,03	0,7	<i>true</i>	dores=yes points=yes 64 ==> class=M79.1 54	0,84
			dores=yes tender=yes 62 ==> class=M79.1 52	0,84
			fibromialgia=yes 92 ==> class=M79.1 76	0,83
			points=yes tender=yes 79 ==> class=M79.1 62	0,78
			tender=yes 81 ==> class=M79.1 63	0,78
			points=yes 85 ==> class=M79.1 66	0,78
			cervical=yes vb=yes 81 ==> class=M54.2 62	0,77
			auriculo=yes cervical=yes 81 ==> class=M54.2 59	0,73
0,04	0,6	<i>false</i>	continuotomar=yes 83 ==> interno=yes 82	0,99
			capsula=yes 83 ==> interno=yes 81	0,98
			shen=yes 82 ==> auriculo=yes 80	0,98
			tender=yes 81 ==> points=yes 79	0,98
			formular=yes 74 ==> interno=yes 70	0,95
			melox=yes 85 ==> famo=yes 80	0,94
			ciclo=yes 84 ==> famo=yes 79	0,94
			janta=yes 105 ==> interno=yes 98	0,93
			cml=yes 75 ==> colo=yes 70	0,93
			points=yes 85 ==> tender=yes 79	0,93
			acido=yes 82 ==> urico=yes 71	0,87
			fibromialgia=yes 92 ==> class=M79.1 76	0,83
			glic=yes urico=yes 89 ==> vldl=yes 73	0,82
			hba=yes vldl=yes 92 ==> glic=yes 75	0,82
			urico=yes vldl=yes 91 ==> glic=yes 73	0,80
			interno=yes 132 ==> janta=yes 98	0,74
			hba=yes 150 ==> glic=yes 111	0,74
			colo=yes 98 ==> cml=yes 70	0,71
glic=yes 189 ==> vldl=yes 132	0,70			

Fonte: Elaborado pelo autor

Como as 19 regras extraídas nesta última configuração contemplavam as 11 regras extraídas no experimento anterior, optou-se em apresentar somente as regras extraídas para os valores de 0,04 e 0,6.

Com a utilização destes valores no conjunto não balanceado, retornaram 2 regras abaixo do limiar de 0,70 de confiança estabelecido, mas que não apareceram nos experimentos com o conjunto balanceado. Por estas diferirem das demais e estarem pouco abaixo do limiar estabelecido não entraram na listagem acima, são elas:

- miosan=yes 209 ==> class=M79.1 141 - confiança de 0,67;
- fluoxetina=yes 180 ==> class=M79.1 115 – confiança de 0,64.

5.5.2.2 Experimentos sobre o conjunto dos 3 *Ms* mais Frequentes

O conjunto dos 3 *Ms* mais Frequentes, é composto por 1.982 textos (1.452 balanceado) associados as 3 classes que formam este conjunto de interesse, com 3.129 termos selecionados.

Seguindo os valores mínimos iniciais utilizados nos experimentos do conjunto anterior para os limiares de suporte e confiança, iniciaram-se os experimentos no conjunto não balanceado, aplicando o algoritmo *Apriori* com CAR.

Os valores iniciais foram reduzidos de forma gradativa, pois não retornaram nenhuma regra, e para que fosse estabelecido um patamar em que pudessem ser extraídas as regras. A extração foi obtida somente com valores 0,05 para suporte e 0,70 de confiança, onde foram obtidas 8 regras e 13 regras quando estes valores foram alterados para 0,04 e 0,70, respectivamente.

A execução do experimento com o conjunto balanceado com valores para suporte e confiança de 0,05 e 0,70, conduziu a extração de 13 regras, porém regras diversificadas, pois estão associadas às 3 classes do conjunto e não somente uma, motivo pelo qual foram selecionadas.

Os valores das métricas utilizadas bem como as quantidades de regras extraídas para cada configuração estão listados na Tabela 27.

Tabela 27 - Quantidade de regras extraídas conjunto 3 *Ms* mais Frequentes

SUPORTE	CONFIANÇA	CAR	QNT. DE REGRAS
0,10	0,70	<i>true</i>	0
0,05	0,70	<i>true</i>	8
0,05	0,70	<i>true</i>	13
0,10	0,70	<i>false</i>	02
0,08	0,70	<i>false</i>	03
0,05	0,70	<i>false</i>	10

Fonte: Elaborado pelo autor

A execução do algoritmo *Apriori*, sem a utilização do CAR, no conjunto não balanceado, utilizando os valores de 0,04 e 0,70 como limiares de suporte e confiança, extraíram 47 regras. Muitas dessas regras, apesar de terem um valor de confiança acima de 0,70, além de serem irrelevantes, pouco ou nada contribuem para geração de conhecimento. Por exemplo, a seguinte regra:

corpo=yes 110 ==> dores=yes 83 conf (0.75)

A regra possui uma confiança de 0,75, mas seu conteúdo pouco contribui, pois a presença de dor sempre será manifestada no corpo do paciente. Regras desse tipo dificultam a análise de outras regras.

Visando aumentar o limiar de suporte e verificar a existência de outras regras que não as já extraídas, iniciou-se o experimento, porém com o conjunto de dados com as classes balanceadas. Utilizando os valores de 0,05 e 0,70 para os limiares de suporte e confiança foram extraídas 10 regras. Estas regras foram selecionadas por possuírem valores de confiança maiores que as encontradas no experimento anterior. A Tabela 28 lista as regras selecionadas nos experimentos com e sem a utilização do CAR.

Tabela 28 - Lista de regras extraídas para o conjunto dos 3 CID mais frequentes

Parâmetros Apriori			Regra	Confiança
Supporte	Confiança	CAR		
0,04	0,7	true	fibromialgia=yes 111 ==> class=M79.1 99	0,89
			dores=yes tender=yes 86 ==> class=M79.1 76	0,88
			dores=yes points=yes tender=yes 85 ==> class=M79.1 75	0,88
			dores=yes points=yes 90 ==> class=M79.1 79	0,88
			tender=yes 111 ==> class=M79.1 95	0,86
			points=yes tender=yes 107 ==> class=M79.1 91	0,85
			points=yes 114 ==> class=M79.1 96	0,84
			idem=yes 121 ==> class=M54.2 93	0,77
			diacereina=yes 124 ==> class=M15.0 95	0,77
			acup=yes idem=yes 109 ==> class=M54.2 83	0,76
			joelhos=yes 135 ==> class=M15.0 97	0,72
			vb=yes 154 ==> class=M54.2 109	0,71
fluoxetina=yes 118 ==> class=M79.1 83	0,70			
0,05	0,7	false	dores=yes tender=yes 86 ==> points=yes 85	0,99
			dores=yes tender=yes class=M79.1 76 ==> points=yes 75	0,99
			tender=yes 111 ==> points=yes 107	0,96
			tender=yes class=M79.1 95 ==> points=yes 91	0,96
			shen=yes 92 ==> auriculo=yes 88	0,96
			dores=yes points=yes class=M79.1 79 ==> tender=yes 75	0,95
			points=yes class=M79.1 96 ==> tender=yes 91	0,95
			dores=yes points=yes 90 ==> tender=yes 85	0,94
			points=yes 114 ==> tender=yes 107	0,94
			janta=yes 78 ==> interno=yes 73	0,94

Fonte: Elaborado pelo autor

Analisando as 47 regras extraídas com o conjunto não balanceado, é possível verificar que a maioria destas estão entre as regras extraídas com o conjunto balanceado. Isso porque regras do conjunto com classes não balanceadas, fazem as mesmas associações ou referências. Algumas dessas 47 regras diferem, são elas:

- fibromialgia=yes 129 ==> class=M79.1 117 - confiança de 0,91;
- melox=yes 96 ==> famo=yes 87 - confiança de 0,91;

- paciente=yes class=M79.1 115 ==> acup=yes 96 - confiança de 0,83;
- paciente=yes 261 ==> acup=yes 210 - confiança de 0,80;
- fluoxetina=yes 151 ==> class=M79.1 115 - confiança de 0,76;
- diacereina=yes 136 ==> class=M15.0 102 - confiança de 0,75;
- acup=yes fibro=yes 131 ==> class=M79.1 95 - confiança de 0,73;
- joelhos=yes 154 ==> class=M15.0 109 - confiança de 0,71.

5.5.2.3 Discussão dos Resultados de Regras de Associação

Tendo em foco uma visão qualificada e um melhor entendimento sobre os resultados obtidos na execução dos experimentos, buscou-se auxílio com dois especialistas da área da Reumatologia para realizar a interpretação das regras extraídas.

O CAR, por levar em consideração as classes do conjunto na extração das regras, foi aplicado visando verificar a coerência do conteúdo textual com o código CID associado. Também foi objetivo de sua utilização a busca de associações de sintomas desconhecidos no domínio da Reumatologia, que pudessem conduzir a produção de conhecimento.

Os experimentos realizados com o conjunto dos 5 CIDs mais Frequentes, utilizando o CAR retornou 8 regras com confiança acima de 0,73, e sem utilizar o CAR retornou 12 regras com confiança acima de 0,83.

Segundo a análise realizada pelos especialistas em Reumatologia, as regras geradas são condizentes com o conteúdo textual e muitas associadas à área médica em geral como a regra *capsula=yes 83 ==> interno=yes 81* (conjunto dos 5 CIDs mais Frequentes), que se refere à prescrição de um medicamento descrevendo seu tipo (capsula) e forma de ser administrado (interno, geralmente via oral), sendo comum a qualquer ramo da medicina, e portanto sem relevância alguma para o contexto deste trabalho.

Entre as regras extraídas do conjunto dos 5 CIDs mais Frequentes, várias são formadas pelos termos *tender*, *points* e *dores*, como:

- dores=yes points=yes 64 ==> class=M79.1 54;
- dores=yes tender=yes 62 ==> class=M79.1 52;

- points=yes tender=yes 79 ==> class=M79.1 62.

Na opinião dos especialistas, a expressão *tender points* significa ponto doloroso ou ponto de dor. É uma série de pontos dolorosos em lugares específicos do corpo que o reumatologista examina para o diagnóstico de Fibromialgia (código CID M79.7). Essas regras se repetem nos resultados dos experimentos do conjunto dos 3 Ms mais Frequentes, e como ocorreu nesse experimento, estão sempre associadas ao código CID M79.1 que codifica Mialgia. Os especialistas consultados afirmam que o CID M79.1 codifica um sintoma, e não uma doença, e a presença de *tender points*, também caracteriza que o diagnóstico correto deveria ser o de Fibromialgia (M79.7). Algumas regras sobre medicações corroboraram essas opiniões, como:

- melox=yes 85 ==> famo=yes 80;
- fluoxetina=yes 118 ==> class=M79.1 83.

Essas regras fazem referência a medicações utilizadas para o tratamento de Fibromialgia, como a Fluoxetina, que é um antidepressivo empregado no controle da dor em doenças crônicas, mais comumente em Fibromialgia (M79.7).

Um dos especialistas consultados afirmou que o CID M79.1 pode ser usado para um diagnóstico preliminar, onde o médico em uma primeira consulta, sem recursos para emitir um diagnóstico mais preciso, lança o CID M79.1, mas que em um segundo momento deveria ter lançado o CID M79.7. A partir desse questionamento, foi feita uma busca pelo CID M79.7 na base de dados completa (antes de ser pré-processada), o resultado não retornou nenhum CID M79.7 (Fibromialgia), caracterizando a utilização de Mialgia de forma genérica para caracterizar quadros de Fibromialgia. Já o outro especialista consultado, após analisar as regras extraídas, afirmou ter ficado com a nítida impressão, que a instituição de saúde geradora dos dados, utiliza o CID que codifica Mialgia como sinônimo de Fibromialgia, conforme declaração no anexo II.

Questionados sobre o impacto da utilização equivocada do CID M79.1, foi respondido que não existem impactos negativos na área médica, o que contribui para a recorrência do erro.

Algumas regras referentes a exames não fizeram sentido na opinião dos especialistas, como:

- *glic=yes urico=yes 89 ==> vldl=yes 73*;
- *hba=yes vldl=yes 92 ==> glic=yes 75*;
- *colo=yes 98 ==> cml=yes 70*.

Segundo um dos especialistas, a regra *glic=yes urico=yes 89 ==> vldl=yes 73* pode ser uma associação da abreviação do termo glicose com VLDL que é um exame realizado para dislipidemia, aparentemente sem relevância.

Em algumas regras, apresentam termos relativos à Acupuntura, como:

- *auriculo=yes cervical=yes 81 ==> class=M54.2 59*;
- *shen=yes 82 ==> auriculo=yes 80*.

Essas regras indicam o uso de Auriculoterapia/Auriculopuntura como terapia complementar em dor cervical, por isso associado ao CID M54.2 Cervicalgia, ou seja dor na cervical, já a segunda regra faz referência ao ponto Shen de Auriculoterapia. A regra *cervical=yes vb=yes 81 ==> class=M54.2 62*, onde VB é uma referência a pontos de acupuntura diversos, onde vários são para dor na região da Cervical, então associados ao CID M54.2 (Cervicalgia). Segundo um dos especialistas consultados, pela presença dessas regras, e pelo teor dos textos da base de dados, a Acupuntura é uma prática comum na instituição de saúde de onde esses dados são oriundos, pois essas expressões são muito específicas para quem não adota tal prática. A ocorrência da regra *paciente=yes 261 ==> acup=yes 210* com uma confiança de 0,80 reforça essa opinião.

As regras apresentadas extraídas no conjunto dos 3 Ms mais Frequentes, em pouco variaram o seu conteúdo ou significado. Tal fato faz sentido, pois o conjunto dos 3 Ms mais Frequentes é um subconjunto do conjunto dos 5 CIDs mais Frequentes. Entretanto era esperado que o conjunto dos 3 Ms mais Frequentes, por ser formado por códigos CIDs específicos da área da Reumatologia revelassem regras de teor mais específico e pontual, o que não ocorreu. Isso pode ter acontecido em função de o conjunto “pai” do conjunto dos

3 *Ms* mais Frequentes ter apresentado pouca diversidade na distribuição das regras extraídas, predominando as regras mais comuns.

6 CONCLUSÃO

Este capítulo apresenta uma análise crítica do trabalho desenvolvido, desde a aquisição da base de dados passando pela sua preparação até a discussão dos resultados obtidos, destacando dificuldades encontradas e fatos positivos.

O desenvolvimento deste trabalho visou estabelecer e descrever um processo de exploração de uma base de dados, com vistas à descoberta de padrões com potencial para produção de conhecimento. A saúde foi escolhida como área de pesquisa por possuir uma produção crescente de trabalhos realizados no campo da computação, mais especificamente em DCT, aliado ao expressivo volume de informação gerado nesse meio. Ao buscar trabalhos relacionados a esse tema na literatura, constatou-se que diversos trabalhos utilizavam bases de dados de PME, porém não foram localizados trabalhos que abordassem a exploração de dados de PME através da técnica de Regras de Associação.

A primeira dificuldade para a realização deste trabalho foi a aquisição da base de dados. Algumas instituições de saúde consultadas tinham a necessidade de processos legais e administrativos que impediram a cedência dos dados em tempo hábil. Outras bases com considerável volume de informações não foram utilizadas por questões técnicas, tal como a falta de conteúdo descritivo associado a campos estruturados, de forma que pudesse caracterizar classes. Alguns setores específicos de algumas instituições de saúde, por não possuírem um sistema adequado para o gerenciamento das informações, faziam uso de meio físico para o armazenamento, inviabilizando o uso desses dados com provável potencial de pesquisa. Para solucionar esse entrave, buscou-se uma base de dados com um dos autores de um dos trabalhos correlatos, conforme relatado na subseção 5.1.1.

A base de dados cedida apresentava quesitos necessários como quantidade de registros e conteúdo descritivo associado ao conteúdo pré-estruturado. A análise da base de dados, visando traçar uma estratégia de exploração, foi concebida sem o auxílio de especialistas. Através de levantamentos estatísticos elaborados com base nos códigos CIDs cadastrados,

chegou-se a principal característica da base, a predominância de conteúdo descritivo associado a códigos CID da área da Reumatologia.

A oscilação nos resultados dos experimentos, face às técnicas de transformação utilizadas, destacou a importância da preparação dos dados a serem submetidos à etapa de mineração. Os melhores resultados foram obtidos por diferentes técnicas de transformação, não havendo concentração em uma única técnica. Desta forma não foi possível afirmar que uma determinada técnica seja melhor que outra, destacando a importância da escolha de mais de uma técnica de transformação para a execução dos experimentos. A escolha de diferentes frequências mínimas também teve influência sobre os resultados, pois o melhor resultado também não foi obtido na mesma frequência em todos os experimentos.

O processo classificatório obteve resultados interessantes como o percentual máximo de Acurácia alcançado no conjunto *M Não M* balanceado, seguido de resultados próximos a este máximo, demonstrando potencial para exploração deste conjunto com outras abordagens, em busca de melhores resultados. Os resultados com o conjunto dos 5 CIDs mais Frequentes não superaram os 61% de Acurácia, ficando abaixo das expectativas. O conjunto dos 3 *Ms* mais Frequentes, apresentou resultados de Acurácia intermediários entre o conjunto *M Não M* e dos 5 CIDs mais Frequentes, demonstrando potencial de melhoria nos resultados deste conjunto.

As regras extraídas com a aplicação do algoritmo *Apriori*, segundo a opinião dos especialistas, são de conhecimento na área da Reumatologia. Determinadas regras pareciam revelar conhecimento novo, o que foi desmistificado. Entretanto a extração de regras associadas ao CID M79.1, quando deveriam estar associadas ao CID M79.7, reforçado pela inexistência de ocorrências deste último CID, denota um fato não comum da utilização do mesmo, revelado pela técnica empregada, podendo ser investigado na sua origem. Algumas destas regras cumpriram um dos objetivos deste trabalho, que era a validação do conteúdo descritivo com o CID associado, fato verificado pelos especialistas consultados.

O algoritmo *Apriori* cumpriu com a tarefa que lhe foi destinada, o que foi comprovado com a concretização de alguns dos objetivos propostos por este

trabalho. Porém para almejar resultados de maior relevância este processo deve ser apurado.

O tempo escasso para a realização deste trabalho impediu que fossem testadas outras técnicas para as distintas etapas apresentadas no capítulo 5, que poderiam ter alavancado os resultados obtidos.

A aplicação, por exemplo, de técnicas de transformação com diferentes abordagens e mais sofisticadas (como Ganho de Informação ou Informação Mútua) (HAN; KAMBER, 2006), nas frequências relatadas na subseção 5.4.5, poderiam ter promovido um aumento de representatividade dos termos e por consequência melhores resultados para o processo classificatório bem como a extração de regras. O acompanhamento de todo o processo investigativo da base de dados por especialistas da área da Reumatologia, poderia ter influenciado no desfecho deste processo, contribuindo para a obtenção de resultados de maior expressão.

A interação com especialistas da área da Reumatologia para avaliação das regras extraídas mostrou-se uma gratificante experiência, tanto pela permuta de conhecimento, como a possibilidade de realização de trabalhos futuros, uma vez que o interesse foi manifestado.

6.1 TRABALHOS FUTUROS

Os diversos contatos realizados em busca de bases de dados para a realização deste trabalho, bem como especialistas para a análise das regras extraídas, despertaram o interesse de profissionais da área da saúde pelo processo de DCBD em geral e possibilidades de execução de pesquisas diversas em bases de dados de outras áreas da saúde.

Este interesse pode conduzir a uma análise criteriosa da base de dados em conjunto com especialistas, para serem estabelecidos focos de exploração dos dados (desta ou de outras bases de dados) que não foram abordados nesse trabalho.

A base de dados, embora seja caracterizada por conteúdo relativo à área da Reumatologia, possui 6.613 registros relativos a outras áreas da saúde

(classe *Não M*) que não foram explorados, e possuem potencial de extração de conhecimento, através de questões como:

- Quais relações ou associações poderiam ser encontradas neste conteúdo com a área da Reumatologia;
- A possibilidade de determinadas doenças da área da Reumatologia derivarem em outras enfermidades presente na base de dados;
- Que associações podem ser descobertas através da extração de regras aplicadas somente no código CID;
- Quais informações mais podem ser extraídas do conteúdo pré-estruturado e descritivo.

Estas questões podem servir para o aprimoramento deste processo através do incremento de outras técnicas que não foram contempladas pela falta de tempo.

A hipótese de elaboração de uma *Stoplist*, com o auxílio de especialistas, para uso na área da medicina também deve ser examinada. Também deve ser considerada uma abordagem diferente na forma de inferir sobre os dados como o uso de conceitos (LOH et al., 2002).

Para isso deve ser submetida à avaliação de diferentes técnicas de Mineração de Dados tais como Naïve Bayes, Redes Neurais ou Clusterização (FRANK; WITTEN, 2009), dependendo do problema a ser investigado.

Um dos especialistas manifestou interesse na realização de um estudo epidemiológico na mesma base de dados utilizada. Embora tal estudo não seja o escopo deste trabalho, o mesmo pode conduzir a outras questões a serem investigadas, como a exploração do conteúdo não ligado à área da Reumatologia (classe *Não M*), devendo assim ser considerado.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. **Proc. of the 20th Int'l Conference on Very Large Databases**. Santiago, set. 1994. Expanded version available as IBM Research Report RJ9839, jun. 1994.

BAEZA, Y. R.; RIBEIRO, N. B. **Modern information retrieval**. Harlow: Addison-Wesley, 1999.

BIREME. **DeCS**: Descritores em Ciências da Saúde. Disponível em: <<http://decs.bvs.br/P/decswebp.htm>>. Acesso em: 05 mar. 2011.

BRASIL. Ministério da Saúde. Secretaria de Assistência à Saúde. **Manual brasileiro de acreditação hospitalar**. 3. ed. rev. e atual. Brasília: Ministério da Saúde, 2002.

CFM – Conselho Federal de Medicina. **Resolução CFM n. 1.639/2002**, de 14 de julho de 2002. Disponível em: <http://www.portalmedico.org.br/resolucoes/cfm/2002/1638_2002.htm>. Acesso em: 12 abr. 2011.

_____. **Resolução CFM n. 1.639/2002**, de 14 de julho de 2002. Disponível em: <http://www.portalmedico.org.br/resolucoes/cfm/2002/1639_2002.htm>. Acesso em: 12 abr. 2011.

_____. **Resolução CFM n. 1.821/07**, de 23 de novembro de 2007. Disponível em: <http://www.portalmedico.org.br/resolucoes/cfm/2007/1821_2007.htm>. Acesso em: 12 abr. 2011.

COHEN, A. M.; HERSH, W. R. A survey of current work in biomedical text mining. **Briefings in Bioinformatics**, v. 6, n. 1, p. 57-71, mar. 2005.

CRM – Conselho Regional de Medicina do Distrito Federal. **Prontuário médico do paciente**: guia para uso. Brasília: CRM, 2006. Disponível em: <<http://www.sbis.org.br/site/arquivos/prontuario.pdf>>. Acesso em: 05 mar. 2011.

DATASUS – Departamento de Informática do SUS. **Classificação estatística internacional de doenças e problemas relacionados à saúde**. 10. rev., 2008. Disponível em: <<http://www.datasus.gov.br/cid10/v2008/cid10.htm>>. Acesso em: 12 mar. 2011.

ENSP – Escola Nacional de Saúde Pública. **Origem e evolução da classificação internacional de doenças**. 2006. Disponível em: <http://www4.ensp.fiocruz.br/biblioteca/dados/txt_316889162.ppt#2>. Acesso em: 10 maio 2011.

FAYYAD, U. M. **Advances in knowledge discovery and data mining**. Califórnia: American Association for Artificial Intelligence, 1996.

FELDMAN, R. **Knowledge Discovery in Textual Databases (KDT)**. California: AAAI, p. 112-117, 1995.

_____; SANGER, J. **The text mining handbook**: advanced approaches in analyzing unstructured data. Cambridge: Cambridge University Press, 2006.

FRANK, E.; WITTEN, I. H. **Data mining**: know it all. Burlington: Elsevier, 2009.

GUPTA, A.; TENNETI, T.; GUPTA, A. **Sentiment Based Summarization of Restaurant Reviews Final Project Report**, jun. 2009.

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. 2. ed. Amsterdam: Elsevier / San Francisco: Morgan Kaufmann, 2006.

JOHN, M. P. Mining knowledge from text collections using automatically generated metadata. **Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management (PAKM '02)**. Dimitris Karagiannis and Ulrich Reimer (Eds.). London: Springer-Verlag, p. 537-548, 2002.

KARPISCHEK, R. U. **Dicionário br.ispell**: versão 2.4. out. 1999. Disponível em: <<http://www.ime.usp.br/~ueda/br.ispell/>>. Acesso em: 15 jun. 2011.

KONCHADY, M. **Text mining application programming**. Brief Bioinform / Charles River Media, 2006.

LIU, B.; HSU, W.; MA, Y. Integrating classification and association rule mining. Proc. **4th Int. Conf. on Knowledge Discovery and Data Mining**. New York, p. 80-86, 1998.

LOH, S.; GAMEIRO, M. A.; GASTAL, F. L.; OLIVEIRA, J. P. M. de. **Descoberta de conhecimento em prontuários eletrônicos**. 2002. Disponível em: <<http://telemedicina.unifesp.br/pub/SBIS/CBIS2002/dados/arquivos/106.pdf>>. Acesso em: 10 abr. 2011.

LOVELL, Brian C.; WALDER, Christian J. Support vector machines for business applications. In: VOGES, K.; POPE, N. (ed.). **Business applications and computational intelligence**. Hershey, p. 267-290, 2006.

MARTHA, A. S.; CAMPOS, C. J. R. de; SIGULEM, D. Recuperação de informação em campos de texto livres de prontuários eletrônicos do paciente baseada em semelhança semântica e ortográfica. **Journal of Health Informatics**. São Paulo, p. 63-71, set. 2010.

MASSAD, E.; MARIN, H. F.; AZEVEDO NETO, R. S. **O prontuário eletrônico do paciente na assistência, informação e conhecimento médico**. São Paulo, 2003.

MEYSTRE, S. M.; SAVOVA, G. K.; KIPPER-SCHULER, K. C.; HURDLE, J. F. Extracting information from textual documents in the electronic health record: a review of recent research. **IMIA Yearbook of Medical Informatics**, p.138-154, 2008.

ORENGO, V.; HUYCK, C. A stemming algorithm for the portuguese language. **Eighth Symposium on String Processing and Information Retrieval (SPIRE'01)**, p. 186, 2001.

PORTER, M. F. **An Algorithm for Suffix Stripping**, v. 14, n. 3, p. 7, 1980. Disponível em: <http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html>. Acesso em: 12 jun. 2011.

PRINCETON. Disponível em: <<http://wordnet.princeton.edu>>. Acesso em: 16 dez. 2011.

SABBATINI, R. M. E. Informatizando o consultório médico. **Revista de Informática Médica**, v. 1, n. 4, ago. 1998. Disponível em: <<http://www.informaticamedica.org.br/informaticamedica/n0104/sabbatini.htm>>. Acesso em: 05 maio 2011.

SAYED, A. el. **Contributions in knowledge discovery from textual data**. Tese de Doutorado. Lyon: Université Lumière Lyon 2, 12 abr. 2008.

SEMOLINI, R. **Support vector machines, inferência transdutiva e o problema de classificação**. Dissertação de mestrado. Campinas: Universidade Estadual de Campinas, 2002.

SIGULEM, D.; CARDOSO, O. L.; GIMENEZ, S. S. F. X.; CEBUKIN, A.; ANÇÃO, M. S. Clinic manager system. **World Congress on Medical Physics and Biomedical Engineering**. Rio de Janeiro: Physics in Medicine and Biology, v. 1, p. 556-556, 1994.

SMOLA, A. J.; SCHÖLKOPF B. **Learning with Kernels**. Cambridge: The MIT Press, 2002.

SOARES, F. A. **Mineração de texto na coleta inteligente de dados na web**. Dissertação de mestrado. Rio de Janeiro: PUC-Rio, 2008.

TAN, A. Text mining: the state of the art and the challenges. **Proceedings of the Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases**. Beijing, p. 65-70, 1999.

VIEIRA, A. F. G.; VIRGIL, J. Uma revisão dos algoritmos de radicalização em língua portuguesa. **Information Research**, v. 12, n. 3, p. 315, 2007. Disponível em: <<http://InformationR.net/ir/12-3/paper315.html>>. Acesso em: 16 jun. 2011.

WHO – World Health Organization. **History Of ICD**. 2004. Disponível em: <<http://www.who.int/classifications/icd/en/HistoryOfICD.pdf>>. Acesso em: 07 maio 2011.

WITTEN, I. H.; FRANK, E. **Data Mining**: practical machine learning tools and techniques. 2 ed. Elsevier; San Francisco: Morgan Kaufmann, 2005.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Tese de doutorado. Porto Alegre: Instituto de Informática, UFRGS, 2004. Disponível em: <<http://www.lume.ufrgs.br/handle/10183/4576>>. Acesso em: 07 maio 2011.

YAN, Y.; FUNG, G.; DY, J. G.; ROSALES, Romer. Medical coding classification by leveraging inter-code relationships. **Proceedings of KDD**, p. 193-202, 2010.

ANEXOS

ANEXO A – RELAÇÃO DE STOPWORDS

A	com	deviam	estes	la
agora	como	disse	estou	la
ainda	contra	disso	eu	la
alguem	contudo	disto		lhe
algum		dito	F	lhes
alguma	D	diz	fazendo	lo
algumas	da	dizem	fazer	
alguns	daquele	do	feita	M
ampla	daqueles	dos	feitas	mas
amplas	das		feito	me
amplo	de	E	feitos	mesma
amplos	dela	e	foi	mesmas
ante	delas	e'	for	mesmo
antes	dele	ela	foram	mesmos
ao	deles	elas	fosse	meu
aos	depois	ele	fossem	meus
apos	dessa	eles		minha
aquela	dessas	em	G	minhas
aquelas	desse	enquanto	grande	muita
aquele	desses	entre	grandes	muitas
aqueles	desta	era		muito
aquilo	destas	essa	H	muitos
as	deste	essas	ha	
ate	deste	esse		N
atraves	destes	esses	I	na
	deve	esta	isso	nao
B	devem	esta	isto	nas
	devendo	estamos		nem
C	dever	estao	J	nenhum
cada	devera	estas	ja	nessa
coisa	deverao	estava		nessas
coisas	deveria	estavam	K	nesta

no	poder	se	todas	algo
nos	poderia	seja	todavia	alo
nos	poderiam	sejam	todo	ambos
nossa	podia	sem	todos	bis
nossas	podiam	sempre	tu	caso
nosso	pois	sendo	tua	certa
nossos	por	sera	tuas	certas
num	porem	serao	tudo	certo
numa	porque	seu		certos
nunca	posso	seus	U	chi
	pouca	si	última	comigo
O	poucas	sido	últimas	conforme
os	pouco	so	último	conosco
ou	poucos	sob	últimos	consigo
outra	primeiro	sobre	um	contigo
outras	primeiros	sua	uma	convosco
outro	propria	suas	umas	cuja
outros	proprias		uns	cujas
	proprio	T		cujo
P	proprios	talvez	V	cujos
para		tambem	vendo	desde
pela	Q	tampouco	ver	eia
pelas	quais	te	vez	embora
pelo	qual	tem	vindo	eram
pelos	quando	tendo	vir	eramos
pequena	quanto	tenha	vos	estar
pequenas	quantos	ter	vos	estariam
pequeno	que	teu		fui
pequenos	quem	teus	X	haver
per		ti		havera
perante	R	tido	Z	havia
pode		tinha		hem
pode	S	tinham	ah	hum
podendo	sao	toda	ai	ih

ir	quaisquer			
irei	qualquer			
iremos	quanta			
logo	quantas			
mais	ser			
menos	sereis			
mim	seremos			
nada	seria			
naquela	seriam			
naquele	sou			
naqueles	tanta			
naquilo	tantas			
naquilos	tanto			
nela	tantos			
nelas	tem			
nele	tera			
neles	teria			
nenhuma	teriam			
nenhumas	tras			
nenhuns	ue			
nesse	uh			
nesses	ui			
nisso	vai			
o	varia			
oba	varias			
oh	vario			
ola	varios			
onde	voce			
opa	vossa			
ora	vossas			
outrem	vosso			
portanto	vossos			
psit	vou			
psiu	quaisquer			

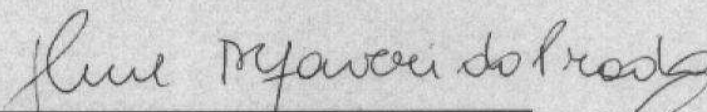
ANEXO B – DECLARAÇÃO DA ESPECIALISTA

Porto Alegre, 09 de Fevereiro de 2012

Prezado Alexandre,

Observando e contextualizando as regras extraídas, a sua grande maioria está relacionada ao CID M79.1 (Mialgia), passando a nítida impressão de que, na clínica de onde foram extraídos estas informações, o CID de Mialgia estava sendo usado como sinônimo para Fibromialgia, reforçado pela tua informação de que nos dados da clínica não existe nenhum lançamento do CID de Fibromialgia. Fibromialgia é um diagnóstico, uma síndrome clínica, enquanto Mialgia é apenas a descrição de um sintoma.

Atenciosamente,



Aline Defaveri do Prado
Médica Reumatologista
CRMRS 27405

Dra. Aline Defaveri do Prado
Médica Reumatologista
CREMERS 27405

Serviço de Reumatologia Hospital São Lucas PUCRS
Avenida Ipiranga 6690 conjunto 220 - Porto Alegre, RS - Brasil