CEETEPS – PROGRAMA DE PÓS-GRADUAÇÃO
MESTRADO EM TECNOLOGIA: GESTÃO, DESENVOLVIMENTO E FORMAÇÃO

THIAGO FERAUCHE

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTOS PARA CLASSIFICAÇÃO DE EMENTAS DA JURISPRUDÊNCIA DA JUSTIÇA DO TRABALHO DE SÃO PAULO

SÃO PAULO
SETEMBRO DE 2011

THIAGO FERAUCHE

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTOS PARA CLASSIFICAÇÃO DE EMENTAS DA JURISPRUDÊNCIA DA JUSTIÇA DO TRABALHO DE SÃO PAULO

Trabalho de dissertação apresentado como exigência parcial para obtenção do Título de Mestre em Tecnologia no Centro Estadual de Educação Tecnológica Paula Souza, no Programa de Mestrado em Tecnologia: Tecnologia da Informação Aplicada, sob orientação do Prof. Dr. Maurício Amaral de Almeida.

SÃO PAULO
SETEMBRO DE 2011

Ferauche, Thiago

F345a Aplicação de técnicas de mineração de textos para classificação de ementas da jurisprudência da Justiça do Trabalho de São Paulo / Thiago Ferauche. – São Paulo : CEETEPS, 2011.

85 f. : il.

Orientador: Prof. Dr. Maurício Amaral de Almeida.

Dissertação (Mestrado) – Centro Estadual de Educação Tecnológica Paula Souza, 2011.

1. Mineração de textos. 2. Inteligência artificial. 3. Jurisprudência. 4. Informática jurídica. I. Almeida, Mauricio Amaral de. II. Centro Estadual de Educação Tecnológica Paula Souza. III. Título.

THIAGO FERAUCHE

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTOS PARA CLASSIFICAÇÃO DE EMENTAS DA JURISPRUDÊNCIA DA JUSTIÇA DO TRABALHO DE SÃO PAULO

PROF. DR. MAURICIO AMARAL DE ALMEIDA

PROF. DR. EDSON SATOSHI GOMI

PROFª. DRª. MÁRCIA ITO

São Paulo, 23 de setembro de 2011

Dedicatória

Aos meus pais, meus grandes incentivadores pela busca do saber, pelo exemplo de vida, dedicação, amor e carinho em mim depositados.

À minha esposa e filhos, razão de todo o meu esforço e esperança de um futuro onde no final tudo terá valido a pena.

Agradecimentos

À minha família, em especial ao meu pai, que me acompanhou durante toda a trajetória deste trabalho, dando-me suporte nos dias mais difíceis, auxílio nas dificuldades encontradas, e acima de tudo acreditou no esforço depositado nessa empreitada.

Ao meu orientador Maurício Amaral de Almeida, pelo compartilhamento de seu vasto conhecimento e sabedoria, além da confiança a mim atribuída.

Aos professores do programa de pós-graduação do Centro Paula Souza pelo excelente trabalho realizado nas disciplinas do programa. Aos colegas de sala que compartilharam suas experiências e conhecimentos durante as disciplinas do programa.

Aos colegas da Secretaria de Gestão da Informação Institucional, em especial ao setor de Sistematização e Catalogação, pelos esclarecimentos prestados e toda paciência despendida durante o desenvolvimento do trabalho.

Aos colegas da Secretaria de Tecnologia da Informação, em especial ao Serviço de Desenvolvimento de Sistemas, pelo apoio, compreensão e ajuda na caminhada dessa jornada.

“Não existe nenhum caminho lógico para a descoberta das leis elementares do universo – o único caminho é o da intuição”.

Albert Einstein

Resumo

FERAUCHE, T. **Aplicação de Técnicas de Mineração de Textos para Classificação de Ementas da Jurisprudência da Justiça do Trabalho de São Paulo.** Dissertação (Mestrado em Tecnologia), Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2011. 85p.

O objetivo desta dissertação é avaliar a utilização de técnicas de mineração de textos para a classificação das ementas que compõem a jurisprudência do Tribunal Regional do Trabalho da 2ª Região – São Paulo. A ementa da jurisprudência é um resumo da decisão jurídica, relevante o suficiente para ser utilizada como exemplo para outros litígios. O Serviço de Gestão Normativa e Jurisprudencial do Tribunal Regional do Trabalho da 2ª Região – São Paulo realiza a classificação destes documentos por assuntos, com o intuito de auxiliar a pesquisa dos mesmos. Com a aplicação de técnicas de mineração de textos, em conjunto com técnicas de aprendizado supervisionado, utilizando-se de documentos previamente categorizados, foi avaliada eficácia da classificação automática realizada pelo computador de documentos desconhecidos do modelo de aprendizagem, e comparado seus resultados com os de um especialista humano.

Palavras-chave: Mineração de textos, Inteligência Artificial, jurisprudência, informática jurídica.

Abstract

FERAUCHE, T. **Applying Text Mining Techniques for Classification of case law summaries of Labor Court in São Paulo**. Dissertation (Master degree in Technology), Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2011. 85p.

The aim of this dissertation is to evaluate the use of text mining techniques for classification of the summaries that make up the jurisprudence of Labor Court, the Tribunal Regional do Trabalho da 2ª Região – São Paulo. The summary of the jurisprudence summarizes the relevant legal decision enough to be used as an example for other disputes. There is a division of the Labor Court, Serviço de Gestão Normativa e Jurisprudencial of the Tribunal Regional do Trabalho da 2ª Região, that classify these documents by subject, in order to make it easier to search for those in the data base. With the application of text mining techniques, in conjunction with supervised machine learning techniques, using previously classified documents, was evaluated the effectiveness of automatic classification performed by computer, and compared theis results with those of a human expert.

Keywords: Text Mining, Artificial Intelligence, jurisprudence, Information Technology in judiciary power.

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

1.1 Motivação

A Inteligência Artificial fornece um conjunto de técnicas e algoritmos úteis para que sistemas computadorizados consigam resolver problemas que a computação tradicional não consegue resolver, seja por limitações teóricas (o problema não pode ser descrito de maneira prática) ou práticas (a memória ou o tempo de processamento são impraticáveis ou infinitos) (RUSSELL; NORVIG, 2004).

O uso de técnicas de IA pode contribuir para o auxílio a atividades em que se faz necessária a aplicação do conhecimento jurídico, uma vez que as técnicas tradicionais de programação, utilizando a lógica booleana, não são suficientes para tal tarefa (ROVER, 2007).

As ementas das decisões do Tribunal Regional do Trabalho da 2ª. Região São Paulo, forma uma coleção de documentos onde estão armazenados conhecimentos jurídicos de maneira explícita, e o ato de classificá-los requer a identificação de tais conhecimentos.

1.2 Problema de Pesquisa

A ementa é um resumo de uma decisão (acórdão) tomada por um colegiado de desembargadores. As ementas das decisões mais relevantes compõem a jurisprudência de um Tribunal. Com a finalidade de facilitar a pesquisa jurisprudencial do Tribunal Regional do Trabalho da 2ª. Região – São Paulo, um especialista em Direito realiza a tarefa de classificá-las, seguindo uma determinada ontologia, porém de maneira empírica e altamente dependente do nível de conhecimento e experiência do especialista. O grande número de ementas a serem classificadas, sobre os mais variados assuntos, faz com que o procedimento adotado seja bem complexo, sem nenhum auxílio computacional que ajude esta tarefa.

1.3 Objetivo Geral

Realizar a avaliação e analisar os resultados de técnicas de classificação de textos para classificar as ementas da jurisprudência do Tribunal Regional do Trabalho da 2ª. Região – São Paulo, e verificar a sua eficiência através da validação dos resultados junto a um especialista e comparando com ementas previamente classificadas.

1.4 Objetivos Específicos

As técnicas de mineração de textos podem ser divididas em várias tarefas. Este trabalho irá concentrar-se na tarefa de classificação de documentos. Para tanto, é necessário atingir os seguintes objetivos específicos:

- Extração das ementas e análise quantitativa da coleção de documentos e suas categorias;
- Pré-Processamento das ementas e análise quantitativa das informações contidas na coleção de documentos;
- Processamento das ementas, análise do aprendizado supervisionado e da acurácia na predição de novos documentos.

1.5 Hipótese de Pesquisa

Com a utilização de técnicas de Mineração de Textos, aliada ao aprendizado de máquinas supervisionado, é possível que um sistema computacional indique à qual categoria, é mais provável que uma ementa da jurisprudência pertença, desta forma auxiliando o trabalho do especialista classificador.

1.6 Justificativa

A Lei Federal Nº 11.416, de 19 de dezembro de 2006, em seu capítulo III, e regulamentado no Poder Judiciário trabalhista através da Instrução Normativa Nº 30 de 2007 do Tribunal Superior do Trabalho, institui e normatiza o chamado “Processo Eletrônico”, onde exclui a obrigatoriedade dos autos em papel, e habilita a

tramitação das informações oficiais do processo através do meio eletrônico, ou meio digital. O “Processo Eletrônico” está sendo desenvolvido pelo Conselho Nacional de Justiça - CNJ, em conjunto com o Conselho Superior da Justiça do Trabalho – CSJT, e deve ser implantado em 2011 (FRANÇA, 2010).

Uma vez que todos os dados estejam em meio digital, a automatização de atividades puramente manuais tende a diminuir, e a tarefa da área da Tecnologia da Informação em auxiliar atividades que envolvem o intelecto tende a crescer, contribuindo assim para a celeridade processual. Importante ressaltar que a participação humana jamais será substituída pela máquina (ALMEIDA FILHO, 2010).

A utilização de sistemas computacionais possui como foco auxiliar, automatizar e agilizar muitas das tarefas humanas, em seus mais variados campos de atuação. Isto faz com que algumas áreas da Ciência da Computação tornem-se multidisciplinares, ou seja, envolvam o estudo de outras áreas de conhecimento. É possível citar a utilização de sistemas computacionais na área médica, nas engenharias, na administração, no direito, na educação, entre outras diversas áreas onde é necessário o entendimento não só de técnicas computacionais, mas também da área de aplicação destas técnicas para a elaboração de um sistema computacional eficaz e efetivo. O avanço de desempenho, a capacidade de recursos de hardware, e o uso de técnicas da Ciência de Computação, principalmente de técnicas de Inteligência Artificial, fazem com que os sistemas computacionais tornem-se ainda mais específicos para auxiliar em tarefas humanas especializadas em uma determinada área do conhecimento. Tais sistemas computacionais são chamados sistemas especialistas, que se utilizam do poder do desempenho computacional para auxiliar tarefas específicas, geralmente aquelas que exigem o uso do intelecto, e na maioria das vezes de uma área do conhecimento diversa à área computacional.

Todo sistema especialista (SE) é um modelo computacional, dentro de um domínio específico de conhecimento, com poder de especialização na resolução de um problema, poder este comparável ao de um especialista humano. Todo sistema especialista legal (SEL) é basicamente um SE voltado para a manipulação do conhecimento jurídico. Qualquer tentativa em declarar o Direito como um corpo de regras, necessariamente terá muitos predicados complexos que não podem ser definidos facilmente em termos mais fundamentais. É provável que as regras sejam,

em alguns casos, deliberadamente ambíguas, certamente incompletas e provavelmente contraditórias (ROVER, 2007).

Trabalhos semelhantes já foram realizados com documentos da jurisprudência da justiça comum, sendo possível citar os mais recentes: Tribunal de Justiça de Santa Catarina (BEPPLER; FERNANDES, 2005), Tribunal de Justiça de Goiás (MORAIS, 2007) e Tribunal de Justiça do Paraná (MOLINARI; TACLA, 2010). O que demonstra que a jurisprudência é uma fonte de conhecimento jurídico que pode ser trabalhada. Há pouca quantidade de pesquisas semelhantes utilizando jurisprudência nacional dentro da Justiça Comum, e nenhuma pesquisa até o momento no âmbito da Justiça do Trabalho brasileira. A Justiça do Trabalho é um ramo do Direito bem específico, onde o especialista na área, operador do Direito, deve aplicar seus conhecimentos específicos para poder analisar um documento e classificá-lo.

A característica interdisciplinar deste trabalho, aplicando técnicas de Classificação de Textos nas ementas que compõem a jurisprudência trabalhista, apresenta uma nova abordagem do uso da Tecnologia da Informação como apoio às atividades de aplicação do conhecimento jurídico. Jurisprudência é o conjunto uniforme e constante das decisões judiciais sobre casos semelhantes (MONTORO, 2000), é o resultado efetivo da aplicação do conhecido jurídico, o que a torna adequada ao uso de técnicas de Mineração de Textos, para extrair informações relevantes e servir como entrada para uma máquina de aprendizado capaz de formar uma base de conhecimento jurisprudencial a partir de informações textuais não-estruturadas. A classificação das ementas é uma atividade realizada por especialistas do Direito, e o estudo de tal atividade é uma oportunidade para compreender a aplicação do conhecimento jurídico e colher informações para o uso posterior em sistemas especialistas.

2 A Jurisprudência

A jurisprudência é um conjunto de decisões de magistrados, que expressam aplicação da legislação em casos práticos, formando assim o conhecimento jurídico de um Tribunal.

Conforme a obra de De Plácido e Silva (2009) jurisprudência é um derivado da conjugação dos termos, em latim, *jus* (Direito) e *prudentia* (sabedoria), o que entende-se como a Ciência do Direito vista com sabedoria, ou, simplesmente, o Direito aplicado com sabedoria. Já Oliveira (2006) diz que a jurisprudência pode ser encarada em sentido amplo ou restrito. Em sentido amplo, significa a ciência ou o conhecimento do Direito. Já no sentido restrito, jurisprudência significa a interpretação dada pelos tribunais (*rerum perpetuo similiter iudicatorum auctoritas*).

Não são todas as decisões de um Tribunal que formam a jurisprudência, são as decisões mais relevantes e que seguem uma mesma linha de pensamento. Geralmente os magistrados indicam as decisões para compor a jurisprudência.

Oliveira (2006) chama a atenção de que não basta uma sentença[1] isolada ou três acórdãos[2] de um tribunal para serem considerados jurisprudência. Neste caso, são meras decisões isoladas. A jurisprudência em seu sentido restrito significa "revelação do direito que se processa através do exercício da jurisdição, em virtude de uma sucessão harmônica de decisões dos tribunais" (REALE, 1995 *apud* OLIVEIRA, 2006).

A jurisprudência não tem força de lei, porém expressa a aplicabilidade da lei, que pode ser alterada de acordo com o momento sócio-econômico em que a sociedade se encontra. Por isso, pode ser utilizada tanto por magistrados quanto advogados para basear suas interpretações da legislação.

A doutrina diverge quanto a incluir a jurisprudência como fonte formal do Direito. Conforme Orlando Gomes (1995 *apud* OLIVEIRA, 2006) a jurisprudência não pode ser considerada fonte do Direito, por que o juiz é servo da lei, além de o julgado produzir efeitos somente nas partes. Já Miguel Reale (1995 *apud* OLIVEIRA,

---

[1] Decisão individual proferida por Magistrado de 1ª instância
[2] Decisão tomada em colegiado por Desembargadores de 2ª instância

2006) diz que o juiz ao aplicar a norma, não age como um autômato, mas, ao contrário, ao interpretá-la e aplicá-la à realidade social que está julgando, indiscutivelmente está criando Direito. Oliveira (2006) entende que a jurisprudência, ao lado da lei, dos costumes e das manifestações de vontade é fonte formal do Direito. Além disso, a lei é sempre abstrata, contendo, normalmente, normas genéricas que devem, através da jurisprudência, ser concretizadas. Nunes (1999) afirma “os cidadãos necessitam saber como as leis serão aplicadas para poderem planejar suas vidas; todas as pessoas na sociedade têm o direito de saber com certeza o que podem e o que não podem fazer”, e ainda que “a sociedade conta, portanto, com as decisões fixadas na jurisprudência para poder respirar a liberdade assegurada pelo Direito e vivenciada na segurança jurídica”.

Resumidamente, a jurisprudência, possui um importante papel como fonte do Direito, e o seu conteúdo auxilia na interpretação da lei e sua aplicação na solução de um problema jurídico. Jurisprudência é o conjunto uniforme e constante das decisões judiciais sobre casos semelhantes (MONTORO, 2000).

2.1 A Jurisprudência da Justiça do Trabalho de São Paulo

As decisões de 2ª. Instância do Tribunal do Trabalho da 2ª. Região – São Paulo, são proferidas durante as sessões de julgamento. As ementas são citadas dentro do documento que explicita a decisão tomada em colegiado (Acórdão). A secretaria de Turma transcreve as ementas dentro do Sistema Informatizado, para posterior classificação por parte do Serviço de Gestão de Normas e Jurisprudencial.

Após a classificação, o sistema informatizado gera documentos em formato de hipertexto (HTML) a partir de informações do banco de dados, conforme a Figura 1.

O documento segue uma estrutura. Existe uma espécie de cabeçalho, com informações que identificam a origem dos dados processuais, como: Tipo do processo, Data de julgamento, Juiz Relator e Revisor do acórdão, Número do acórdão, Ano do acórdão, Turma do acórdão, Data de publicação, Número do processo e Partes envolvidas.

É possível ainda identificarmos mais duas partes da estrutura: a ementa e o índice. A ementa é onde se encontra a síntese do que foi decidido no acórdão, suas

premissas e justificativas. É na ementa que está concentrado resumidamente todo o conhecimento da jurisprudência.

TIPO: RECURSO ORDINÁRIO
DATA DE JULGAMENTO: 16/11/2004

RELATOR(A): RICARDO ARTUR COSTA E TRIGUEIROS

REVISOR(A): CARLOS ROBERTO HUSEK

ACÓRDÃO Nº: 20040643829

PROCESSO Nº: 01152-1998-445-02-00-5 **ANO:** 2004 **TURMA:** 4ª

DATA DE PUBLICAÇÃO: 26/11/2004

PARTES:

RECORRENTE(S):
INSTITUTO NACIONAL DO SEGURO SOCIAL INSS

RECORRIDO(S):
RODRIMAR S/A TRANSP EQUIPS INDS ARM GER
GENILSON ALMEIDA GOIS

EMENTA:

INSS. RECURSO ORDINÁRIO. NÃO CONHECIMENTO. INADEQUAÇÃO, AUSÊNCIA DE INTERESSE E IRREGULARIDADE DA REPRESENTAÇÃO.Recurso do INSS que não se conhece em razão de: (1) inadequação, vez que é notória a impropriedade do recurso ordinário (art. 895, CLT), cabível apenas na fase cognitiva, para atacar decisão terminativa em sede de execução, para a qual o recurso específico é o agravo de petição (art. 897, CLT), sendo inaplicável à espécie o princípio da fungibilidade; (2) ausência de interesse porquanto o valor previdenciário já foi quitado, configurando sanha arrecadatória a pretensão do Instituto de receber o que já lhe foi pago; (3) irregularidade da representação, em vista da subscrição do apelo por advogado particular e não por procurador autárquico.

ÍNDICE:

PREVIDÊNCIA SOCIAL, Recurso do INSS

Figura 1: Exemplo de uma ementa retirada do site do TRT da 2ª Região – São Paulo.

O índice é a classificação da jurisprudência. O índice utilizado foi elaborado pelo Desembargador Valentin Carrion e aprimorado, ao longo dos anos, pelo Serviço de Jurisprudência e Divulgação, atualmente chamado de Serviço de Gestão

Normativa Jurisprudencial. A partir de dezembro de 2009, o índice passou a trabalhar conjuntamente com a Tabela de Assuntos Processuais da Justiça do Trabalho (CNJ Resolução Nº 46, de 18 de dezembro de 2007). A lista completa com todas as categorias utilizadas como índice está demonstrada no Apêndice 1, como pode ser observado não existe uma regra clara da forma como estão estruturadas as categorias, elas são fruto de anos de trabalho, onde através da tentativa e do erro chegou-se à estrutura atual.

A tarefa de classificação é realizada pelos servidores públicos do Serviço de Gestão Normativa Jurisprudencial. Não é um processo automático e requer conhecimentos específicos no âmbito do Direito. Os servidores públicos deste serviço podem ser identificados como os especialistas do conhecimento jurídico, pois são eles que leem a ementa, identificam relações na área do Direito, e depois classificam a jurisprudência. Esta classificação é utilizada para organizar e facilitar a busca da jurisprudência.

2.2 A Informática Jurídica

A disciplina que trata da utilização otimizada da informática pelos profissionais ou operadores do direito e nas atividades de natureza jurídica é conhecida como Informática Jurídica. Ela pode ser dividida em: (CASTRO, 2005)

1. **Gestão ou Operacional**: relacionada com a mecânica e o funcionamento dos espaços jurídicos e dos trabalhos e fluxos físicos;
2. **Registro ou documental**: relacionada com o acesso rápido e fácil aos vários registros oficiais;
3. **Ajuda à decisão**: relacionada com o tratamento e a recuperação da informação jurídica nos campos da legislação, doutrina e jurisprudência.

Os sistemas amplamente utilizados pelos diversos Tribunais do Poder Judiciário podem ser classificados como sistemas de Gestão ou Operacional e sistemas de Registro ou Documental, são sistemas que utilizam os paradigmas de programação tradicional e podem ser facilmente tratados computacionalmente. Conforme Almeida Filho (2010), a respeito da Informatização Judicial atual, no Brasil não existe processo eletrônico, mas sim procedimentos eletrônicos, ou seja, partes

dos atos processuais são praticados por meio eletrônico, disponibilizando algumas peças em meio digital.

Sistemas de Ajuda à decisão são os menos utilizados, pois necessitam do máximo de informações processuais em meio eletrônico, além de um entendimento mais aprofundado sobre o raciocínio jurídico.

O raciocínio jurídico se distingue em duas partes: o estabelecimento dos fatos relevantes (*quaestio facti*) e a aplicação da norma correspondente (*quaestio iuris*). Esta segunda etapa compreende a qualificação jurídica dos fatos, que pressupõe a interpretação da lei, na tentativa de retirar as consequências previstas pela mesma para aqueles fatos (ROVER, 1994).

Na tentativa de expressar a aplicação das normas de maneira explícita, Rover (2007) utilizou a lógica deôntica cujo objetivo é formalizar conceitos (normas) que têm a ver não só com a prescrição de comportamentos desejados, mas também, e isto é essencial, com a necessidade de admitir que os comportamentos se podem desviar do ideal, e de prescrever o que fazer em tais circunstâncias. Porém, conforme Rover (2007), a representação do conhecimento jurídico através de regras, a princípio parecia ser a solução para a representação do conhecimento jurídico. No entanto, como constatado em suas pesquisas, qualquer tentativa em declarar o Direito como um corpo de regras necessariamente terá muitos predicados complexos que não podem ser definidos facilmente em termos mais fundamentais. É provável que as regras sejam, em alguns casos, deliberadamente ambíguas, certamente incompletas e provavelmente contraditórias.

Desta maneira, para uma representação do conhecimento jurídico, existe a alternativa da utilização de paradigmas computacionais que se utilizam de técnicas de inteligência artificial, através da Mineração de Textos, uma vez que o conhecimento jurídico explícito encontra-se em forma de textos não-estruturados.

A jurisprudência é exatamente a coleção de documentos que expressa o conhecimento jurídico aplicado, portanto o estudo de técnicas de mineração de textos utilizando a jurisprudência contribui para o desenvolvimento ou aprimoramento de sistemas de Ajuda à decisão no âmbito da Informática Jurídica.

3 A Mineração de Textos

A Mineração de Textos (MT) tem como objetivo descobrir informações relevantes através de dados não-estruturados, contidos em formato texto. Uma definição genérica inclui todos os tipos de processamento de texto que tratam de encontrar, organizar e analisar informação (KONCHADY, 2006). A MT é semelhante à Mineração de Dados (MD), porém a diferença encontra-se no tipo de dado a ser minerado. A MD é um conjunto de técnicas estudado na Inteligência Artificial (IA), uma especialidade da Ciência da Computação, que objetiva a descoberta de informações relevantes a partir de dados estruturados, geralmente armazenados em Banco de Dados relacional.

De maneira análoga à MD, a MT procura extrair informações úteis de fontes de dados através da identificação e exploração de padrões, no entanto, no caso da MT as fontes de dados são coleções de documentos, e os padrões são encontrados não em registros de banco de dados, mas sim em dados não estruturados em forma de texto dentro da coleção de documentos (FELDMAN; SENGER, 2007).

A mineração de textos é uma técnica para a descoberta de conhecimento em textos não-estruturados, o que se aplica obviamente aos textos jurídicos. Existem duas maneiras de analisar o texto não-estruturado: a análise semântica, baseada no significado dos termos no texto; e a análise estatística, baseada na frequência com que os termos aparecem no texto. Estes dois modos podem ser aplicados separados ou em conjunto.

A análise estatística de textos demonstra ser a mais interessante para se aplicar a textos jurídicos, pois os textos empregam uma linguagem técnica com muitos termos em latim. Nesse tipo de análise, a importância dos termos é dada basicamente pelo número de vezes que eles aparecem nos textos. É interessante ressaltar que este tipo de estratégia pode ser conduzido independentemente do idioma (EBECKEN; LOPES; COSTA, 2003).

Indiferente ao tipo de análise, o processo de mineração de texto pode ser dividido em quatro etapas, conforme (GONÇALVES; REZENDE, 2002):

- **Coleta de Documentos:** nesta fase, os documentos relacionados com o domínio da aplicação final são coletados.

- **Pré-processamento:** consiste de um conjunto de ações realizadas sobre o conjunto de textos obtido na etapa anterior, com o objetivo de prepará-los para a extração de conhecimento.
- **Extração de Conhecimento:** utilizam-se alguns algoritmos de aprendizado com o objetivo de extrair, a partir de documentos pré-processados, conhecimento na forma de regras de associação, relações, segmentação, classificação de textos, entre outros.
- **Avaliação e Interpretação dos Resultados:** nessa etapa os resultados obtidos são analisados, filtrados e selecionados para que o usuário possa ter um melhor entendimento dos textos coletados. Esse entendimento maior pode auxiliar em algum processo de tomada de decisão.

A mineração de texto possui várias tarefas que podem fazer parte do processo de extração do conhecimento. Cada tarefa extrai um tipo de informação diferente. As tarefas são: clustering, classificação, extração de características, sumarização e indexação.

O processo de clustering, ou agrupamento, torna explícito o relacionamento entre documentos, enquanto a classificação identifica os tópicos-chave de um documento. A extração de características é usada quando é preciso conhecer pessoas, lugares, organizações e objetos mencionados no texto. A sumarização estende o princípio de extração de características, concentrando-se mais em sentenças inteiras do que em nomes ou frases. A indexação temática é útil quando se quer ser capaz de trabalhar preferencialmente com tópicos que com palavras-chave (EBECKEN; LOPES; COSTA, 2003).

3.1 Tarefa de Classificação

A classificação, em Mineração de Textos, visa a identificar os tópicos principais em um documento e associar este documento a uma ou mais categorias predefinidas (Yang; Pedersen, 1997, *apud* EBECKEN; LOPES; COSTA, 2003). Segundo Konchady (2006), o problema da classificação pode ser descrito como a classificação de documentos em múltiplas categorias, onde se tem um conjunto de *n* categorias $\{C_1, C_2, \ldots C_n\}$ para as quais são associados *m* documentos $\{D_1, D_2, \ldots$

D_m}. A Figura 2 demonstra o processo de classificação, onde as *n* categorias são pré-definadas através de palavras-chave que diferenciam qualquer categoria C_i de qualquer outra categoria C_j. O processo de identificar essas palavras-chave é chamado de Extração de Características.

O estudo da automatização da classificação de textos data de meados da década de 60, na época sua aplicação era para indexar literatura científica através de um vocabulário controlado (MARON, 1961 *apud* FELDMAN; SANGER, 2007). Foi apenas na década de 90 que a classificação foi totalmente desenvolvida devido ao progressivo aumento da quantidade de documentos textuais na forma digital e a necessidade de organizá-los (FELDMAN; SANGER, 2007).

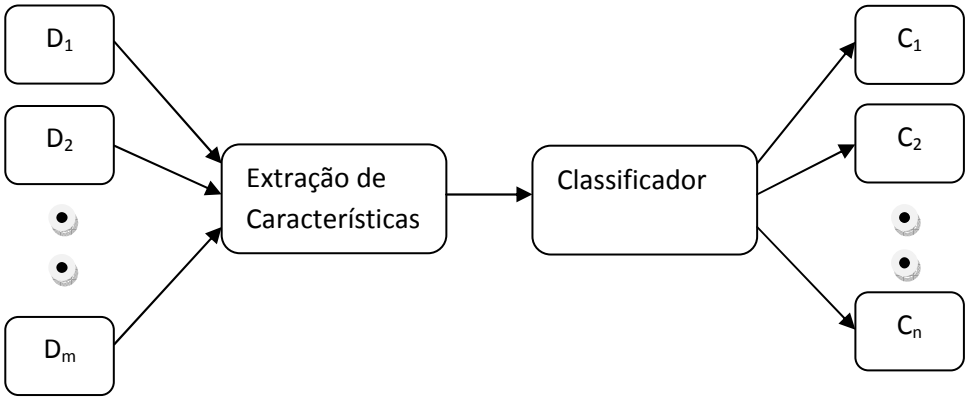


Figura 2: Associação de documentos a categorias (KONCHADY, 2006)

Conforme Feldman e Sanger (2007), assim como em outras tarefas de inteligência artificial, existem duas principais abordagens para a classificação de textos. A primeira é a abordagem da engenharia do conhecimento (*knowledge engineering*) onde o conhecimento de especialistas sobre as categorias está codificado no sistema, seja declarativamente ou na forma de regras procedimentais de classificação. A outra abordagem é o aprendizado de máquina (*machine learning*) onde geralmente através de um processo indutivo é construído um classificador aprendendo a partir de um conjunto de exemplos pré-classificados. A principal desvantagem da abordagem da engenharia do conhecimento é o que pode ser chamado de "gargalo da aquisição de conhecimento" (*knowlege acquisition bottleneck*) – a enorme quantidade de trabalhadores altamente qualificados e especialistas no domínio do conhecimento, necessária para manter as regras do conhecimento implementadas no sistema. Portanto, a maioria dos trabalhos recentes sobre classificação está concentrada na abordagem de aprendizado de máquina,

que requer apenas um conjunto de instâncias de treinamento manualmente classificadas, o que é bem menos custoso para se produzir.

3.2 O Pré-processamento dos documentos

A etapa de pré-processamento diz respeito à limpeza dos dados para facilitar as análises da etapa seguinte. Esta etapa consiste na remoção do que for desnecessário para o entendimento do texto, o documento gerado é utilizado como base para a fase seguinte (MONTEIRO; GOMES; OLIVEIRA, 2006). Segundo Álvarez (2007), a etapa de pré-processamento é responsável por transformar uma coleção de documentos em uma representação estruturada adequada, normalmente no formato de uma tabela atributo-valor, a qual é mais apropriada para processamento do que simples arquivos textos. Dada uma coleção de documentos, é aplicada a abordagem *bag of words,* que consiste em representar cada documento da coleção como um vetor de termos contidos em seu respectivo documento. Cada termo que ocorre no documento pode ser composto por apenas uma palavra (unigrama) ou várias palavras (bigramas, trigramas, ..., n-gramas). Com a finalidade de identificar todos os termos presentes em um documento, um procedimento de marcação (*tokenização)* é realizado, geralmente através do reconhecimento de espaços em branco, tabulações e sinais de pontuação que delimitam termos. Essa representação, no entanto, pode resultar em uma tabela esparsa com alta dimensionalidade, portanto um objetivo da etapa de pré-processamento é reduzir a dimensionalidade dessa representação.

3.2.1 Ferramenta PRETEXT II

O PRETEXT, proposto por Matsubara *et al* (2003) *apud* Soares (2009), é uma ferramenta computacional que realiza o pré-processamento de textos utilizando a abordagem *bag of* words e gera uma tabela atributo-valor. A ferramenta foi desenvolvida utilizando o paradigma de orientação a objetos, na linguagem de programação Perl. O PRETEXT passou por um processo de remodelagem e reimplementação e foi criado o PRETEXT II (SOARES *et al,* 2008 *apud* SOARES, 2009), uma ferramenta com mais funcionalidades e melhor desempenho

computacional. A ferramenta implementa as funcionalidades de *tokenização*, remoção de *stopwords* e *stemming*, utilizadas comumente nas técnicas de mineração de textos com a finalidade de preparar o texto para ser analisado estatisticamente e para ser processado.

- ***Tokenização***

Para realizar a Extração de Características de textos não estruturados (Figura 2) é necessário executar a quebra do fluxo contínuo de caracteres em partes mais significantes, o que pode ser feito em vários níveis diferentes, podendo dividir o texto em capítulos, seções, parágrafos, frases, palavras e até mesmo em sílabas e fonemas. A abordagem mais freqüente encontrada em sistemas de mineração de textos é a divisão do texto em frases e palavras, o que pode ser chamado de "tokenização", em inglês "*tokenization"* (FELDMAN; SANGER, 2007). No caso da ferramenta PRETEXT, como utiliza a técnica de *bag of words*, faz a quebra do texto em palavras, portanto considera um *token como* uma palavra. Para realizar esta tarefa o programa de computador deve remover alguns caracteres indesejados, como sinais de pontuação, separação silábica, marcações especiais e números, os quais, isoladamente, trazem pouca informação (SOARES, 2009).

- ***Remoção de Stopwords***

A tarefa de pré-processamento que remove as palavras irrelevantes é chamada de "seleção de características" (*feature selection*). A maioria dos sistemas de mineração de textos, ao menos remove as *stopwords* (FELDMAN; SANGER, 2007). A remoção de *stopwords* consiste na retirada de palavras que se repetem inúmeras vezes no decorrer do texto ou palavras sem relevância aparente para o processamento do texto, como artigos, conjunções, pronomes, preposições, etc. Este conjunto de palavras recebe o nome de s*topwords* (MONTEIRO; GOMES; OLIVEIRA, 2006). Com essas palavras, é gerada uma lista (*stoplist*), na qual inúmeras *stopwords* podem ser armazendas para que sejam desconsideradas ao processar o texto. Desta forma, a remoção *stopwords* minimiza consideravelmente a quantidade total de *tokens* usados para representar os documentos (SOARES, 2009).

- ***Stemming***

Em inglês, como em muitas outras línguas, palavras ocorrem em textos em mais de uma forma. O processo de *stemming* é responsável por reduzir as diversas formas de um termo a uma forma comum (raiz) denominada *stem*. É possível definir um *stem* de qualquer palavra após retirar o seu prefixo e sufixo (KONCHADY, 2006). Os algoritmos de *stemming* aplicam uma série de normalizações linguisticas para remover prefixos e/ou sufixos de termos, ou inclusive mapear verbos a sua forma no infinitivo (ÁLVAREZ, 2007). O PRETEXT II utiliza o algoritmo de PORTER (1980 *apud* ÁLVAREZ, 2007), um dos algoritmos mais utilizados em sistemas de mineração de textos (KOCHANDY, 2006), aprimorado e adaptado à língua portuguesa por Soares (2009), onde é possível destacar a melhora considerável do tratamento de verbos irregulares.

3.2.2 Problema da dimensionalidade

Segundo Álvarez (2007), em geral, uma das características do processo de mineração de textos é a alta dimensionalidade do conjunto de atributos. Entretanto, em determinadas circunstâncias pode ser desejável aplicar métodos para a redução da representação, pois a alta dimensionalidade pode tornar o custo de processamento e armazenamento, em alguns casos, inviável.

As funcionalidades de *tokenização*, remoção de *stop words* e *stemming*, já contribuem para a redução da alta dimensionalidade dos atributos. Porém, seja pela quantidade elevada de documentos, ou pela quantidade elevada de termos presentes nos documentos, são necessários outros mecanismos para a redução da dimensionalidade. Para tal, a ferramenta PRETEXT II (SOARES, 2009) faz uso de cortes de palavras baseados em frequência, utilizando a Lei de Zipf (ZIPF, 1949 *apud* SOARES, 2009) e os cortes de Luhn (LUHN, 1958 *apud* SOARES, 2009).

A Lei de Zipf é utilizada para encontrar termos considerados pouco representativos em uma determinada coleção de documentos. Luhn usou essa lei como uma hipótese para especificar dois pontos de corte para excluir *tokens* não relevantes em uma coleção de documentos. Os termos que excedem o corte superior são os mais frequentes e são considerados comuns por aparecer em qualquer tipo de documento, como as preposições, conjunções e artigos. Já os

termos abaixo do corte inferior são considerados raros e, portanto, não contribuem significativamente na discriminação dos documentos (SOARES, 2009).

Na Figura 3 é mostrada a curva da Lei de Zipf (I) e os cortes de Luhn aplicados a Lei de Zipf (II), onde o eixo cartesiano *f* representa a frequência das palavras e o eixo cartesiano *r* representa as palavras correspondentes ordenadas segundo essa frequência.

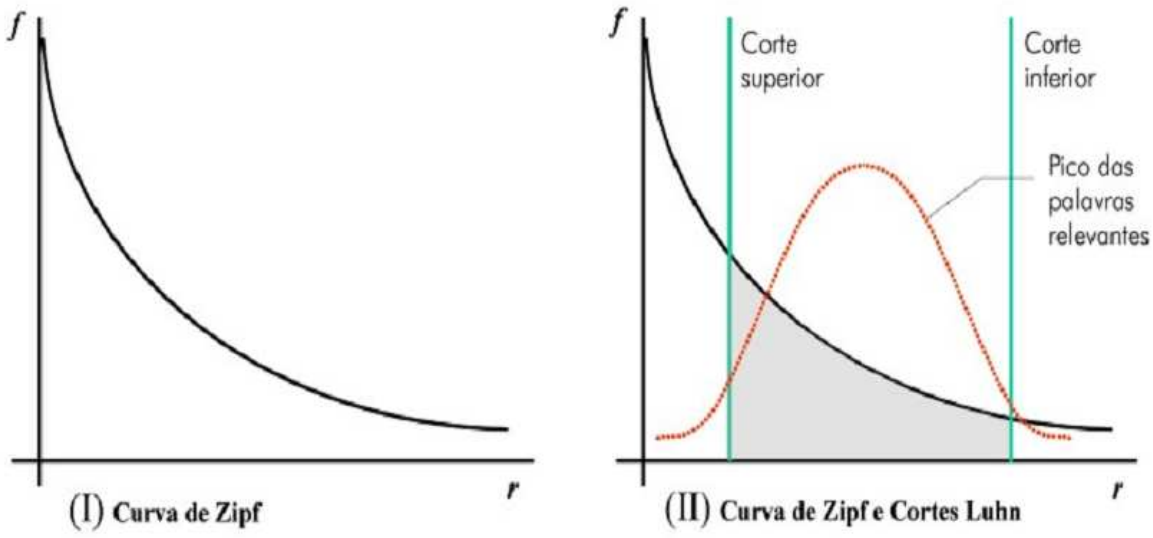


Figura 3: Representação da curva de Zipf e os cortes de Luhn (SOARES, 2009)

Outro mecanismo para redução da alta dimensionalidade é a geração de atributos pela união de duas ou mais palavras consecutivas, podem-se gerar atributos com um maior poder de predição. O *n*-grama é exatamente essa junção de palavras, na qual *n* representa o número de palavras que são geradas por simples acaso, porém, aquelas que apresentam uma frequência maior podem ser úteis para o aprendizado.

Por exemplo, considerar as palavras São e Paulo individualmente pode agregar pouco conhecimento, pois São pode referir-se ao verbo ser e Paulo é um nome próprio relativamente comum no Brasil. Entretanto, o termo composto São Paulo pode agregar muito mais informação se o texto se refere à cidade ou estado de São Paulo (SOARES, 2009).

3.2.3 Valores dos Atributos

Além das funcionalidades para auxiliar na redução do número de atributos visando a melhorar a relevância da informação para a classificação do texto, a ferramenta PRETEXT implementa as medidas mais comuns da literatura para calcular o valor dos atributos na tabela atributo-valor (SEBASTIANI, 2002 *apud* SOARES, 2009).

- ***Boolean*:** atribui o valor um (verdadeiro) ao atributo se ele existe no documento e zero (falso) caso contrário.
- ***Term Frequency*:** conhecida também como *tf*, consiste na contagem de aparições de um determinado atributo (termo) em um documento, atribuindo-se essa contagem ao valor do atributo (frequência absoluta). Pode ser representada pela Equação 1, na qual $freq(t_j,d_i)$ é a frequência do termo t_j no documento d_i

$$a_{ij} = tf(t_j, d_i) = freq(t_j, d_i) \quad (1)$$

- ***Term Frequency Linear*:** indica a frequência com que um termo aparece na coleção de documentos, para tanto um fator de ponderação pode ser utilizado para que os termos que aparecem na maioria dos documentos tenham um peso de representação menor. A chamada *tf-linear* (MATSUBARA *et al,* 2003 *apud* SOARES, 2009) pode ser definida pelas Equações 2 e 3, onde o fator de ponderação é dado por um menos a frequência relativa do número de documentos em que o termo aparece no número total de documentos.

$$a_{ij} = tflinear(t_j, d_i) = freq(t_j, d_i) \times linear(t_j) \quad (2)$$

$$linear(t_j) = 1 - \frac{d(t_j)}{N} \quad (3)$$

- ***Term Frequency – Inverse Document Frequency*:** conhecida como *tf-idf, também* é uma medida ponderada da frequência dos termos na coleção de documentos, de tal maneira que termos que aparecem na

maioria dos documentos tenham um peso de representação menor (JONES, 1972; ROBERTSON, 2004 *apud* SOARES, 2009). O fator de ponderação *idf* é inversamente proporcional ao logaritmo do número de documentos em que o termo aparece no número total *N* de documentos, conforme as Equações 4 e 5.

$$a_{ij} = tfidf(t_j, d_i) = freq(t_j, d_i) \times idf(t_j) \quad (4)$$

$$idf(t_j) = log \frac{N}{d(t_j)} \quad (5)$$

Suavização dos valores: é muito provável que um determinado *token* ocorra em todos os documentos, ocasionando com que o fator de ponderação, linear ou idf, seja nulo, o que faz com que seja atribuído o valor zero ao token. Conforme Monard *et al* (2008) desta maneira perde-se informação, e para que isto não ocorra é possível fazer com que os fatores de ponderação não sejam nulos utilizando um critério de suavização conhecido como *smooth*. Este critério somente é ativado quando o fator de ponderação for nulo. Uma maneira simples de implementar a suavização, é aumentar o valor de N (número de documentos da coleção) em 10% de seu valor, assim o fator de ponderação será diferente de nulo.

- ***Normalização:*** um detalhe importante que deve ser levado em consideração é o tamanho dos documentos na coleção. Quando existe uma diferença grande de tamanho nos documentos de uma coleção, pode ocorrer uma diferença grande na frequência dos termos dos documentos. Uma solução para este problema é a normalização dos valores da tabela atributo-valor, podendo ter seu foco nos atributos (colunas) ou nos documentos (linhas). As equações 6 e 7 demonstram respectivamente a normalização quadrática em linha e em coluna (MONARD, 2008).

$$NormQuadratic(t_j, d_i) = \frac{a_{ij}}{\sqrt{\sum_{k=1}^{N} (a_{kj}{}^2)}} \quad (6)$$

$$NormQuadratic(t_j, d_i) = \frac{a_{ij}}{\sqrt{\sum_{k=1}^{N}(a_{ik}{}^2)}} \quad (7)$$

3.3 Aprendizado de Máquina

Conforme Russell e Norving (2004), o campo do aprendizado de máquina geralmente pode ser distinguido em 3 casos: **aprendizado supervisionado**, **não-supervisionado** e **por reforço**. O problema do aprendizado supervisionado envolve aprender uma função a partir de exemplos de suas entradas e saídas. O problema do aprendizado não-supervisionado envolve aprender padrões de entradas sem que haja o fornecimento de valores de saídas especificadas. Por último, o problema do aprendizado por reforço envolve aprender através de um retorno indicativo que um determinado comportamento não é desejável, o que implica um subproblema de aprender como o ambiente funciona.

O aprendizado supervisionado utiliza a inferência indutiva, ou indução. A tarefa da inferência indutiva pura (ou indução) é essa: “Dada uma coleção de exemplos de *f*, retorne uma função *h* que se aproxima de *f*”. A função *h* é chamada de hipótese. A razão da dificuldade do aprendizado, do ponto de vista conceitual, é que não é fácil dizer quando uma função *h* é uma boa aproximação de *f*. Uma boa hipótese irá generalizar bem, ou seja, irá predizer exemplos não vistos anteriormente (RUSSEL; NORVING, 2004).

Segundo Monard e Baranauskas (2003), a indução é a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos. Na indução, um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados. Portanto, as hipóteses geradas através da inferência indutiva podem ou não preservar a verdade. Ainda Monard e Baranauskas (2003) lembram que mesmo assim, a inferência indutiva é um dos principais métodos utilizados para derivar conhecimento novo e predizer eventos futuros, que foi pelo meio da indução que Arquimedes descobriu a primeira lei da hidrostática e o princípio da alavanca, que Kepler descobriu as leis do movimento planetário, que Darwin descobriu as leis da seleção natural das espécies.

Utilizando o aprendizado supervisionado, através do método da inferência, é possível montar algoritmos classificadores que possam aprender com documentos

previamente classificados, como demonstra Monard e Baranaukas (2003) em sua hierarquia do aprendizado, e pode ser visto demonstrado na Figura 4.

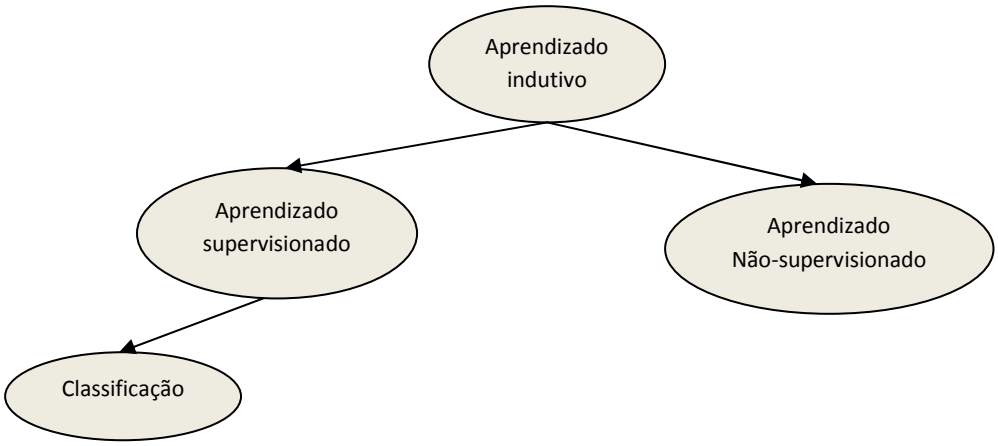


Figura 4: A hierarquia do aprendizado (MONARD; BARANAUSKAS, 2003)

Abaixo seguem alguns conceitos e definições amplamente utilizados, tanto neste trabalho, quanto na literatura de Aprendizado de Máquina (MONARD; BARANAUSKAS, 2003):

- **Indutor**: conhecido como programa de aprendizado, ou algoritmo de indução. Tem como objetivo extrair um bom classificador a partir de um conjunto de exemplos rotulados. A saída do indutor, o classificador, pode ser usada para classificar exemplos novos (ainda não rotulados) com a meta de predizer corretamente o rótulo de cada um.
- **Exemplo**: também denominado caso, registro ou dado. É uma tupla de valores de atributos (ou um vetor de valores de atributos).
- **Atributo**: descreve uma característica ou um aspecto de um exemplo. Normalmente, há pelo menos dois tipos de atributos: nominal, quando não existe uma ordem entre os valores (por exemplo, cor: vermelho, verde, azul) e contínuo, quando existe uma ordem linear nos valores (por exemplo, peso: pertencente ao conjunto dos números reais).
- **Classe**: todo o exemplo possui um atributo especial, denominado rótulo ou classe, que descreve o fenômeno de interesse, isto é, o conceito-meta que se deseja aprender para fazer previsões a respeito.
- **Conjunto de exemplo**: um conjunto de exemplos é composto por exemplos contendo valores de atributos bem como a classe associada.

Na Tabela 1, é mostrado o formato padrão de um conjunto de exemplos *T* com *n* exemplos e *m* atributos. Nessa tabela, a linha *i* refere-se ao *i*-ésimo exemplo (*i* = 1, 2, ... , *n*) e a entrada x_{ij} refere-se ao valor do *j*-ésimo (*j* = 1, 2, ... , *m*) atributo X_j do exemplo *i*.

Como pode ser visto, exemplos são tuplas $T_i = (x_{i1,}\ x_{i2,} \ldots ,\ x_{im,}\ y_i)$, onde fica subentendido o fato que x_i é um vetor, e a última coluna, *Y*, contém o atributo meta, também chamado de classe.

Tabela 1 - Conjunto de exemplos no formato atributo-valor

	X_1	X_2	. . .	X_m	Y
T_1	x_{11}	x_{12}	. . .	x_{1m}	y_1
T_2	x_{21}	x_{22}	. . .	x_{2m}	y_2
:	:	:	:	:	:
T_n	x_{n1}	x_{n2}	. . .	x_{nm}	y_n

Usualmente, um conjunto de exemplos é dividido em dois subconjuntos disjuntos: o conjunto de treinamento, usado para o aprendizado do conceito, e o conjunto de teste usado para medir o grau de efetividade do conceito aprendido. Os subconjuntos são normalmente disjuntos para assegurar que as medida obtidas, utilizando o conjunto de teste, sejam de um conjunto diferente do usado para realizar o aprendizado, tornando a medida estatisticamente válida.

- **Classificador ou Hipótese**: dado um conjunto de exemplos de treinamento, um indutor gera como saída um classificador (também denominado hipótese ou descrição de conceito) de forma que, dado um novo exemplo, ele possa predizer com a maior precisão possível sua classe.
- **Ruído**: são dados imperfeitos, podem ser derivados do próprio processo que gerou os dados, do processo de aquisição dos dados, do processo de transformação dos dados ou mesmo devido a classes rotuladas incorretamente.
- ***Under* e *Overfitting***: como o conjunto de treinamento é apenas uma amostra de todos os exemplos do domínio, é possível induzir hipóteses

que melhorem seu desempenho no conjunto de treinamento, enquanto pioram o desempenho em exemplos diferentes daqueles pertencentes ao conjunto de treinamento. Nesta situação, o erro (ou outra medida) em um conjunto de teste independente evidencia um desempenho ruim da hipótese. Neste caso, diz-se que a hipótese ajusta-se em excesso ao conjunto de treinamento ou que houve um *overfitting.* Também é possível induzir hipóteses que possuam pequena melhora de desempenho no conjunto de treinamento, assim como em um conjunto de teste. Neste caso, diz-se que a hipótese ajusta-se muito pouco ao conjunto de treinamento ou que houve um *underfitting*.

3.3.1 Avaliação do Aprendizado

Existem vários meios para avaliar o aprendizado do algoritmo através do método supervisionado por indução. O primeiro deles é chamado de *Ressubstituição*, onde o conjunto de exemplos para treinamento é o mesmo conjunto para testes. Esse meio fornece uma medida falsa, possuindo uma estimativa muito otimista, ou seja, o desempenho no conjunto de treinamento em geral não se estende a conjuntos independentes de testes.

Para que o algoritmo seja o mais genérico possível, é importante utilizar meios que não utilizem exemplos em comum entre o conjunto de treinamento (ou aprendizado) e o conjunto de teste. Estes meios são conhecidos como métodos e reamostragem, os principais métodos são descritos a seguir: (MONARD; BARANAUSKAS, 2003)

- ***Holdout*:** método de amostragem mais simples, onde divide os exemplos em uma porcentagem fixa de exemplos para treinamento e o restante para teste. É comum a utilização de 2/3 dos exemplos para treinamento e 1/3 para teste. Porém existe um dilema neste método, pois para obter um bom classificador é preciso utilizar o máximo possível de exemplos para treinamento, e ao mesmo tempo para obter uma boa estimativa de erro é necessário utilizar o máximo possível de exemplos para teste (WITTEN, FRANK; 2000).

- ***Cross-validation:*** utilizado para assegurar a representatividade máxima de todas as classes durante o treinamento e teste. Os exemplos são aleatoriamente divididos em *r* partições mutuamente exclusivas (*folds*) de tamanho aproximadamente igual *n* / *r* exemplos. Os exemplos nos *r* – 1 *folds* são usados para treinamento e a hipótese é induzida no *fold* remanescente (MONARD; BARANAUSKAS, 2003).
- ***Stratified cross-validation:*** similar ao *cross-validation*, porém considera a distribuição de classe – proporção de exemplos em cada uma das classes – durante a geração dos *folds* mutuamente exclusivos. Isto significa, por exemplo, que se o conjunto original de exemplos possui duas classes com distribuição de 20% e 80%, cada *fold* também terá esta proporção de classes (MONARD; BARANAUSKAS, 2003).
- ***Leave-One-Out:*** é um caso especial de *cross-validation*. É computacionalmente dispendioso e freqüentemente usado para amostras pequenas. Para uma amostra de tamanho *n* uma hipótese é induzida utilizando (*n* - 1) exemplos, e a hipótese é testada no único exemplo remanescente. Este processo é repetido *n* vezes, cada vez induzindo uma hipótese deixando de considerar um único exemplo (MONARD; BARANAUSKAS, 2003).

Para avaliar os resultados induzidos com o objetivo de extrair a quantidade de acertos, erros e poder analisar melhor as hipóteses geradas, os resultados são colocados em uma matriz de duas dimensões conhecida como “matriz de confusão” (WITTEN; FRANK, 2000). Como mostrado na Tabela 2, os resultados são totalizados em classes verdadeiras e classes preditas, para *k* classes diferentes $\{C_1, C_2, \ldots, C_k\}$. Cada elemento $M\ (C_i, C_j)$ da matriz representa o número de exemplos que realmente pertencem à classe C_i que foram classificados como sendo da classe C_j. O número de acertos, para cada classe, se localiza na diagonal principal $M\ (C_i, C_i)$ da matriz. Os demais elementos representam erros na classificação.

Tabela 2 - Matriz de confusão de um classificador (MONARD; BARANAUSKAS, 2003)

Classe	**Predita** C_1	**Predita** C_2	. . .	**Predita** C_k
verdadeira C_1	$M(C_1,C_1)$	$M(C_1,C_2)$	. . .	$M(C_1,C_k)$
verdadeira C_2	$M(C_2,C_1)$	$M(C_2,C_2)$	. . .	$M(C_2,C_k)$
:	:	:	:	:
verdadeira C_k	$M(C_k,C_1)$	$M(C_k,C_2)$	. . .	$M(C_k,C_k)$

Uma vez os resultados inseridos na matriz, é possível identificar os erros e classificá-los como *falsos positivos* (F_P) e *falsos negativos* (F_N). Por exemplo, na tabela Y é ilustrada uma matriz de confusão de duas classes, rotuladas como "+" (positiva) e "-" (negativa), onde T_P é o número de exemplos positivos classificados corretamente e T_N é o número de exemplos negativos classificados corretamente do total de $n = (T_P + F_P + T_N + F_N)$ exemplos.

Tabela 3 - Matriz de confusão para a classificação com duas classes

Classe	**Predita** C_+	**Predita** C_-	***Taxa de Erro da Classe***	***Taxa de Erro Total***
verdadeira C_+	Verdadeiros Positivos T_P	Falsos Negativos F_N	$\frac{F_N}{T_P + F_N}$	$\frac{F_P + F_N}{n}$
verdadeira C_-	Falsos Positivos F_P	*Verdadeiros Negativos* T_N	$\frac{F_P}{F_P + T_N}$	

Outras medidas também podem ser extraídas da matriz de confusão, como a precisão total (*total accuracy*), dada pela soma dos Verdadeiros Positivos (T_P) com os Verdadeiros Negativos (T_N), dividido pelo total de exemplos n.

3.3.2 Comitê de Classificadores

Conforme a maioria dos pesquisadores, não há um consenso sobre qual é o melhor algoritmo classificador, uma vez que isso depende do tipo de dado e aplicação utilizados. Portanto, não existe um único algoritmo que apresente melhor desempenho para todos os problemas (MONARD; BRANAUSKAS, 2003). Porém os classificadores que trouxeram melhores resultados em pesquisas já realizadas

foram: SVM, AdaBoost, kNN e métodos de regressão. Naive Bayes apesar de não ter apresentado bons resultados, é muito utilizado em conjunto com outros classificadores. As árvores de decisão foram pouco utilizadas como classificadores, e em alguns resultados foram quase tão bem quanto o SVM (FELDMAN; SANGER, 2007).

Como cada algoritmo classificador possui sua característica própria, sendo seu desempenho dependente das características extraídas dos textos, e dos dados utilizados para treinar o algoritmo, é possível elaborar um método utilizando algoritmos classificadores em conjunto de maneira a formar uma combinação de classificadores.

Segundo Dietterich (2000) e Breiman (2000) *apud* Monard e Baranauskas (2003), técnicas de combinação de classificadores tem sido objeto de pesquisas com o intuito de construir um preditor mais preciso pela combinação de vários outros. O resultado dessa combinação é chamado *ensemble*. E ainda, a utilização de *ensembles* tem obtido melhores resultados que a utilização de um único preditor.

A ideia de usar comitês de classificadores partiu da intuição de que uma equipe de especialistas, combinando seus conhecimentos, pode produzir melhores resultados que um único especialista. Utilizando a técnica chamada de *bagging*, os classificadores individuais devem ser treinados em paralelo com a mesma coleção de documentos de treinamento. Para que o comitê funcione, os classificadores devem ser bem diferentes um dos outros, seja pela forma de representação do documento ou pela maneira que aprendem. É necessário combinar o resultado dos classificadores, sendo que a maneira mais simples é o voto majoritário, onde são precisos no mínimo (k + 1) / 2 classificadores, onde k deve ser obviamente um número ímpar (FELDMAN; SANGER, 2007).

Desta maneira, é possível montar um comitê classificador utilizando os algoritmos SVM, Naive Bayes e Árvore de Decisão, uma vez que os três algoritmos diferem bastante da maneira como aprendem, como podemos ver a seguir.

3.3.3 Algoritmo Naive Bayes

O teorema de Bayes provê a base para o tratamento da imperfeição da informação em diversos sistemas baseados em conhecimento (RICH, 1983 *apud*

BITTENCOURT, 2006). Resumidamente este teorema calcula a probabilidade de um dado evento, a partir de um conjunto de observações. Seja:

- $P(H_i|E)$ a probabilidade que a hipótese H_i seja verdadeira dada a evidência E;
- $P(E|H_i)$ a probabilidade que a evidência E será observada se a hipótese H_i for verdadeira;
- $P(H_i)$ a probabilidade "a priori" que a hipótese H_i é verdadeira na ausência de qualquer evidência específica;
- k o número de hipóteses possíveis.

O teorema de Bayes é formulado conforme a equação 8 (BITTENCOURT, 2007).

$$P(H_i|E) = \frac{P(E|H_i)\cdot P(H_i)}{\sum_{j=1}^{k} P(E|H_j)\cdot P(H_j)} \quad (8)$$

O modelo mais comum de redes Bayesianas utilizado em aprendizado de máquinas é o chamado modelo Naive Bayes. Neste modelo, a classe representada pela variável C (que será prevista) é a hipótese e as variáveis de atributo X_i são as evidências. Assumindo variáveis booleanas, os parâmetros são estabelecidos na equação 9 (RUSSEL; NORVING, 2004).

$$\theta = P(C = true), \theta_{i1} = P(X_i = true \mid C = true), \theta_{i2} = P(X_i = true \mid C = false) \quad (9)$$

Uma vez o modelo treinado, ele pode ser utilizado para classificar novos exemplos para os quais a classe C não é conhecida, porém conhecidos os valores dos atributos $x_1, \dots, x_n$. Desta maneira a probabilidade de cada classe é dada pela equação 10 (RUSSEL; NORVING, 2004).

$$P(C|x_1, \dots, x_n) = P(C) \prod_i P(x_i \mid C) \quad (10)$$

O modelo é dito "ingênuo" (naive) por que assume que os atributos são independentes uns dos outros, dada uma determinada classe. A premissa de assumir que os atributos são independentes, para aplicações práticas do cotidiano, é

muito simplista, porém Naive Bayes funciona muito bem quando testado com dados reais (WITTEN; FRANK, 2000). Naive Bayes também não apresenta dificuldades em aprender com ruídos nos dados e é capaz de realizar previsões probabilísticas quando apropriado (RUSSEL; NORVING, 2004).

3.3.4 Naive Bayes para classificação de textos

Como um classificador probabilístico, a matriz atributo-valor é vista como a probabilidade $P(c \mid d)$ que o documento d pertence a uma classe c e calcula sua probabilidade aplicando o teorema de Bayes, conforme a equação 11.

$$P(c \mid d) = \frac{P(d \mid c) \cdot P(c)}{P(d)} \tag{11}$$

Assumindo como premissa de um classificador Naive Bayes que todos os atributos são independentes, e sabendo que um documento é representando por um vetor de características $d = (w_1, w_2, \ldots)$, é possível representar o classificador através da equação 12 (FELDMAN; SANGER, 2007).

$$P(d \mid c) = \prod_i P(w_i|c) \tag{12}$$

Apesar de saber que premissa da independência entre os atributos, na prática, não é real, as tentativas de tratar o modelo com atributos dependentes até o momento não tem produzido melhora significativa no desempenho de classificadores probabilísticos (FELDMAN; SANGER, 2007).

3.3.5 Algoritmo de Árvores de Decisão

A indução através de árvores de decisão é uma das formas de algoritmo de aprendizagem mais simples, e que traz bons resultados (RUSSEL; NORVING, 2004). Os algoritmos que induzem árvores de decisão pertencem à família de algoritmos *Top Down Induction of Decision Trees* –TDIDT. Uma árvore de decisão é uma estrutura de dados definida recursivamente como (MONARD; BARANAUSKAS, 2003):

- Um *nó folha* que corresponde a uma classe ou
- Um *nó de decisão* que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma sub-árvore. Cada sub-árvore tem a mesma estrutura que a árvore.

Uma árvore de decisão toma como entrada um objeto ou situação descrita por um conjunto de atributos e retorna uma "decisão" – o valor da saída prevista. Os valores de entrada podem ser discretos ou contínuos, assim como as saídas. Aprender uma função de valores discretos é conhecido como classificação; e aprender uma função de valor contínuo e chamado de regressão (RUSSEL, NORVING, 2004).

A árvore de decisão alcança seu resultado executando uma sequência de testes. Cada nó interno da árvore corresponde a um teste de valor de um atributo, e os galhos a partir do nó são etiquetados com os valores possíveis do teste. Cada *nó folha* da árvore especifica um valor a ser retornado se aquela folha for alcançada, como na Figura 5.

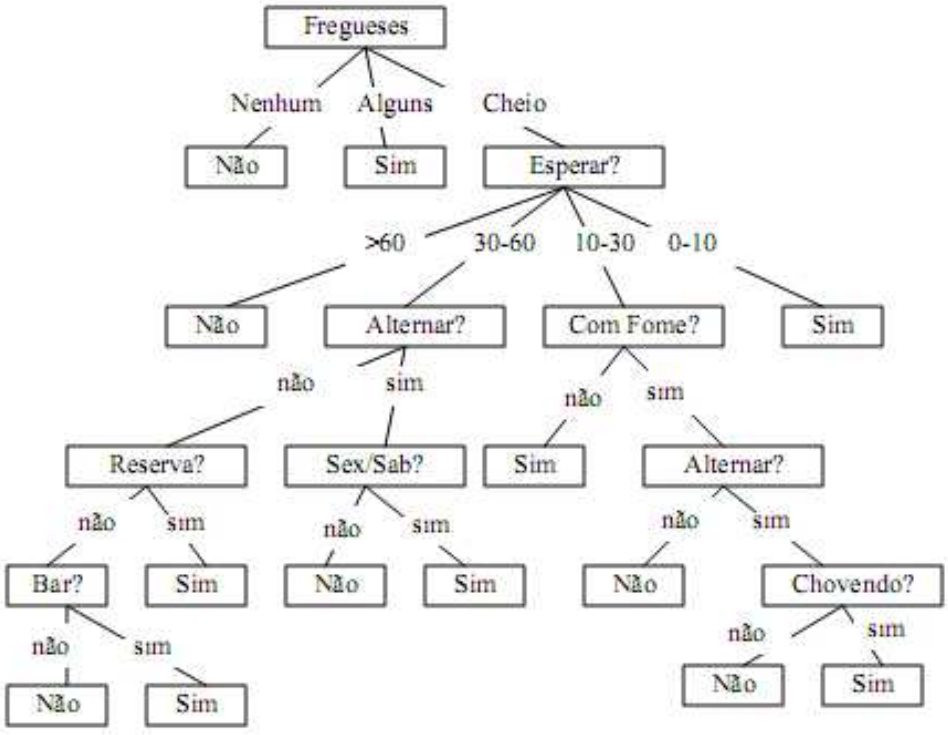


Figura 5: Exemplo de uma árvore de decisão para o problema de espera para jantar em um restaurante (RUSSEL; NORVING, 2004)

Como pode ser visto, as árvores de decisão permitem acompanhar o que ocorre no processo de aprendizagem, facilitando o entendimento do mesmo. Conforme Russel e Norving (2004), a representação da árvore de decisão parece ser natural para humanos, pois se assemelha a forma de um manual (como manuais de equipamentos) que são escritos como uma única árvore de decisão prolongada por centenas de páginas.

3.3.6 Árvores de decisão para classificação de textos

Para Konchady (2006) um classificador baseado em árvores de decisão pode ser visto como um conjunto de regras ordenadas que é utilizado para classificação. Dado qualquer documento, ele é testado pela árvore de decisão para descobrir se o documento pertence ou não a uma determinada categoria. A cada nó da árvore é estabelecida uma probabilidade de pertencer a uma categoria.

Geralmente, a árvore é construída recursivamente tomando uma característica *f* a cada passo do algoritmo e dividindo a coleção de treinamento em dois subconjuntos, um contendo *f* e outro que não contém *f*, até que apenas documentos de uma única categoria restem (FELDMAN; SANGER, 2007). A Figura 6 representa a divisão formada entre os elementos positivos, que pertencem à categoria e os que não pertencem.

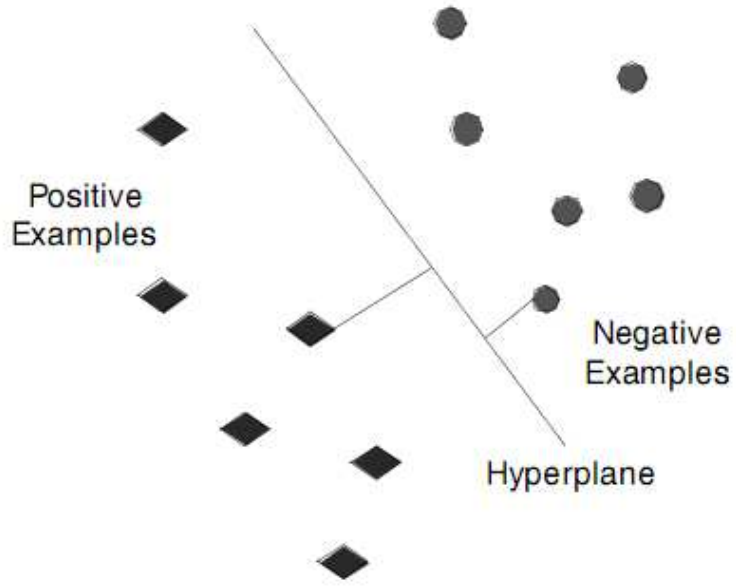


Figura 6: Um classificador baseado em árvore de decisão (FELDMAN; SANGER, 2007)

A escolha da característica em cada passo do algoritmo é feita por alguma medida teórico-informacional como ganho de informação ou entropia. Classificadores baseados em árvore de decisão são muito utilizados como uma linha base de comparação com outros classificadores e também como membro de um comitê de classificadores (FELDMAN; SANGER, 2007).

O problema de *overfitting*, causado por um excesso de atributos presentes em uma coleção de treinamentos, pode ser facilmente evitado realizando a poda (*prunning*) da árvore, onde os *nós folha* com menor probabilidade de classificação são removidos. O desempenho do algoritmo não é tido como um dos melhores, porém sua facilidade de entendimento facilita a análise das características do texto, permitindo identificar termos mais relevantes para a escolha de uma determinada categoria (KONCHADY, 2006).

3.3.7 Algoritmo SVM (*Support Vector Machine*)

O algoritmo *Support Vector Machine* (SVM) (VAPNIK, 1995 *apud* JÚNIOR, 2007) é muito utilizado em problemas de mineração de textos e categorização textual (JOACHIMS, 1998; GONÇALVES, 2002 *apud* JÚNIOR, 2007), principalmente quando os textos estão modelados no formato *bag of words*, pois este algoritmo baseia-se no aprendizado estatístico.

Segundo Witten e Frank (2000) *Support Vector Machine* utiliza modelos lineares para implementar limites de classes não-lineares.

3.3.7.1 Problema dos Modelos Lineares

A Equação 13 mostra um exemplo de um modelo linear para dois atributos, incluindo todos os seus produtos com três fatores.

$$x = w_1 {a_1}^3 + w_2 {a_1}^2 a_2 + w_3 a_1 {a_2}^2 + w_4 {a_2}^3 \quad (13)$$

Onde x é a saída, a_1 e a_2 são os dois valores de atributos, e existem quatro pesos w_i a serem aprendidos. O resultado pode ser utilizado para classificação treinando um sistema linear para cada classe e associando uma instância desconhecida a classe que retornar com o valor mais alto.

Dois problemas envolvendo complexidade computacional surgem ao aplicar essa técnica devido ao grande número de coeficientes introduzidos em uma transformação utilizando a Equação 13 envolvendo problemas com dados reais. O primeiro é o problema da praticalidade, por exemplo, utilizando 10 atributos, incluindo todos os seus produtos com 5 fatores, o algoritmo de aprendizagem terá que determinar mais de 2000 coeficientes. O segundo problema é conhecido como *overfitting,* onde o número de coeficiente é relativamente maior comparado ao número de instâncias treinadas. O resultado do modelo fica é linear devido ao excesso de parâmetros no modelo.

3.3.7.2 Maximum Margin Hyperplane

Support Vector Machine é baseado em um algoritmo que encontra um tipo especial de modelo linear conhecido como *maximum margin hyperplane*. É um hiperplano no espaço das instâncias que classifica todas as instâncias treinadas corretamente; é a maior distância entre instâncias de classes diferentes. As distâncias mais próximas do *maximum margin hyperplane* são chamadas de *support vectors* . Há pelo menos um *support vector* para cada classe.

Um hiperplano separando duas classes pode ser escrito conforme a Equação 14, e em termos de *support vector* conforme a Equação 15 (WITTEN; FRANK, 2000).

$$x = w_0 + w_1 a_1 + w_2 a_2 \tag{14}$$

$$x = b + \sum \alpha_i y_i a(i) \cdot a \tag{15}$$

Na Equação 15, *a* corresponde a uma instância de teste, e *a(i)* corresponde aos s*upport vectors.* Os parâmetro α_i e *b* na Equação 13 são parâmetros a serem descobertos assim como os pesos w_i da Equação 14.

Porém, o problema envolvendo a complexidade computacional ocorre quando são tratados modelos não-lineares, devido à alta dimensionalidade do espaço gerado para poder tratá-los. Isto ocorre tanto no processo de aprendizado do

algoritmo, quanto no processo de classificação, já que tanto para um quanto para outro é necessário calcular o produto escalar dos vetores *a(i)* e *a*.

É possível calcular o produto escalar dos dois vetores antes do mapeamento não-linear, direto no conjunto de atributos originais. Uma versão para tratar a alta dimensionalidade é apresentada na equação 16 (WITTEN; FRANK, 2000).

$$x = b + \sum \alpha_i y_i (a(i) \cdot a)^n \quad (16)$$

Onde *n* é escolhido como o número de fatores na transformação. Devido a esta equivalência matemática, os produtos escalares podem ser calculados diminuindo a dimensionalidade de seu espaço. O recurso de elevar o produto escalar a uma potência *n* é chamado de *polynomial kernel*. Conforme Witten e Frank (2000) uma boa maneira de escolher o valor de *n* é começar com 1 (um modelo linear) e incrementá-lo até que o erro estimado pare de crescer. Podemos citar como outras funções kernel utilizadas para implementar diferentes mapeamentos não-lineares, a *sigmoid kernel* e a *radial basis kernel*.

Utilizando *support vectors* o problema de *overfitting* é raro de ocorrer, pois este problema é ocasionado por muita flexibilidade nos limites de decisão, e os *support vectors* são representações globais de todo o conjunto de pontos de treinamento, o que dá uma certa estabilidade nos limites de decisão.

3.3.8 SVM para categorização de textos

O algoritmo SVM é um classificador binário. Cada categoria tem um classificador separado e os documentos são individualmente comparados com cada categoria. Ele procura por um hiperplano com o máximo de margem entre exemplos de documentos treinados positivos e negativos. A entrada do SVM é um conjunto de *N* pares de documentos e categorias treinadas, {(x_1,c_1), . . (x_n,c_n)}. Cada c_i contém o valor positivo 1 (pertence à categoria) ou negativo -1. O objetivo de se treinar uma algoritmo SVM para uma categoria é criar uma função $f(x) = \pm 1$ onde *x* é um vetor documento. A função *f(x)* deve atribuir corretamente categorias a documentos não vistos a partir da mesma distribuição probabilística. Classificadores SVM pertencem a uma classe de funções do tipo $f(x) = sign(w \cdot x + b)$ onde *w* e *x* são vetores. A função

decisão *f(x)* encontra o hiperplano ideal para que a margem de separação entre duas classes seja maximizada, conforme pode ser visto na Figura 7 (KONCHADY, 2006).

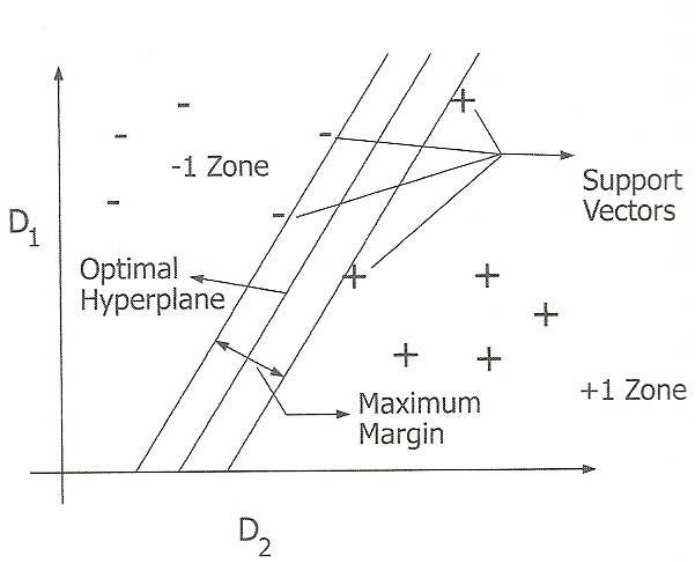


Figura 7: Um classificador SVM com *maximum margin* (KONCHADY, 2006)

Um processo de classificação pode ser descrito conforme a Figura 8, onde o vetor a ser testado, vetor **x**, é transformado do espaço de entrada para o seu espaço de características usando uma função *sigmoid kernel*. Os vetores $\mathbf{x_1}$, .. $\mathbf{x_n}$ também são transformados da mesma maneira. O produto escalar (*dot product)* de **x** com *n support vectors* é calculado, e são aplicados os pesos $\alpha_i y_i$ encontrados aplicando um algoritmo QP (*Quadratic Programming*), soma-se o *bias* (*b*) e se tem a decisão de qual categoria pertence o documento.

Feldman e Sanger (2007) chamam a atenção para o fato de que os hiperplanos do SVM são totalmente determinados por uma quantidade relativamente pequena de instâncias treinadas, que são os chamados vetores de suporte (*support vectors*). O restante dos dados treinados possui pouca influência no classificador treinado, o que é uma característica, aparentemente, presente somente no algoritmo SVM, o que o faz ser único entre os diferentes tipos de algoritmos para categorização.

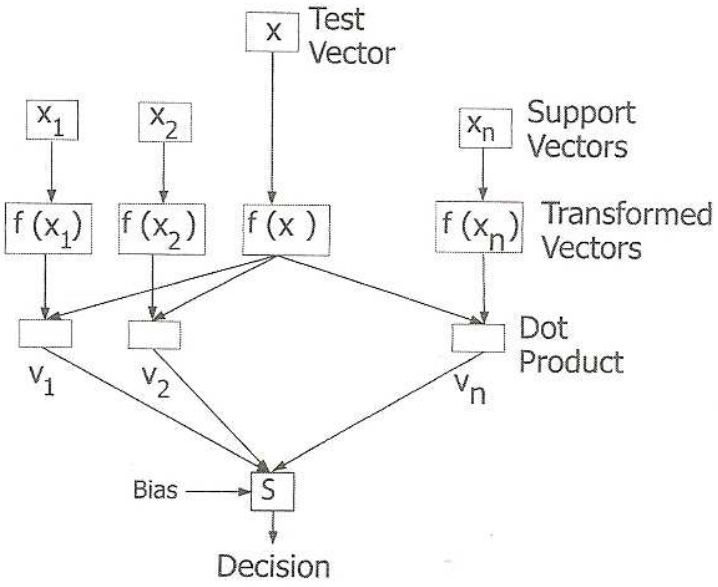


Figura 8: Categorização de um documento desconhecido (KONCHADY, 2006)

3.3.9 Algoritmo SMO (Sequential Minimal Optimization)

Conforme Park (2010) o SMO surgiu da necessidade de implementação de um algoritmo SVM de maneira rápida, simples e capaz de tratar conjuntos de dados mais extensos. Além disso, possui a capacidade de tratar um conjunto de dados esparsos, que possuem um número substancial de elementos com valor zero. Park (2010) afirma que a otimização realizada no SMO encontra-se na programação quadrática analítica, ao invés da abordagem numérica tradicional. O algoritmo SMO escolhe a resolução dos problemas de otimização, optando pelas menores otimizações possíveis em cada passo. Nos problemas de programação quadrática em SVM, a menor otimização possível envolve dois multiplicadores de Lagrange, pois eles devem obedecer a restrição de igualdade linear. Em cada passo, o método SMO escolhe a otimização de dois multiplicadores, buscando valores ótimos para eles e atualizando-os para refletir os novos valores ótimos. A vantagem está em utilizar um otimizador analítico ao invés de toda uma biblioteca de rotinas de programação quadrática. Além disso, não há necessidade de armazenar matrizes externas, o que permite manipular problemas com conjunto de treinamento volumoso.

4 Método e Resultados

A pesquisa foi realizada, dividindo o trabalho em 3 fases conforme os objetivos citados no item 1.4, sendo elas: a Fase de Extração das Ementas, onde são capturados os conteúdos dos textos a serem minerados; o Pré-processamento das Ementas, onde os textos são transformados em valores que demonstram as características dos textos; e o Processamento das Ementas, onde as características são utilizadas para o aprendizado dos classificadores. Estas fases são sequenciais, iniciando pela Fase de Extração de Ementas, passando pelo Pré-processamento das Ementas e finalizando com o Processamento das Ementas. Estas fases formam um método para realizar a classificação das ementas de maneira sistemática apoiada pelo computador.

4.1 Fase de Extração das Ementas

Nesta fase, os conteúdos das Ementas foram extraídos a partir de arquivos textos, originalmente utilizados para o envio às Editoras que compõem a revista jurisprudencial. Estes arquivos possuem a extensão “.JUR” e layout próprio, como pode ser visto no Apêndice 2. Os arquivos utilizados correspondem às ementas dos Acórdãos publicados desde janeiro de 2008 até janeiro de 2011.

Um aplicativo foi desenvolvido para ler os arquivos citados anteriormente, e gerar arquivos textos, sendo gerado um arquivo para cada ementa, somente com as informações contidas nos limitadores: “..EMEN:”,”..DECI:”. O texto entre os limitadores “..EMEN:” e “..DECI:” forma o conteúdo do resultado da decisão. O texto entre o limitador “..INDE:” é a categoria da jurisprudência atribuída à ementa, portanto o mesmo aplicativo cria uma estrutura de diretórios, sendo cada diretório uma categoria, e copia os arquivos correspondentes a cada categoria em seu respectivo diretório, totalizando 187 diretórios.

Esta estrutura de diretório gera coleções de documentos, cada coleção de uma determinada categoria, conforme a Figura 9. A organização é importante para poder identificar a quantidade de documentos de cada categoria, e a quantidade de informação (quantidade de bytes) presente em cada categoria. Isto permite analisar

a coleção de documentos e entender as particularidades de cada categoria. O conhecimento de tais características da coleção forma uma base para a tomada de decisão e interpretação dos resultados das fases subsequentes.

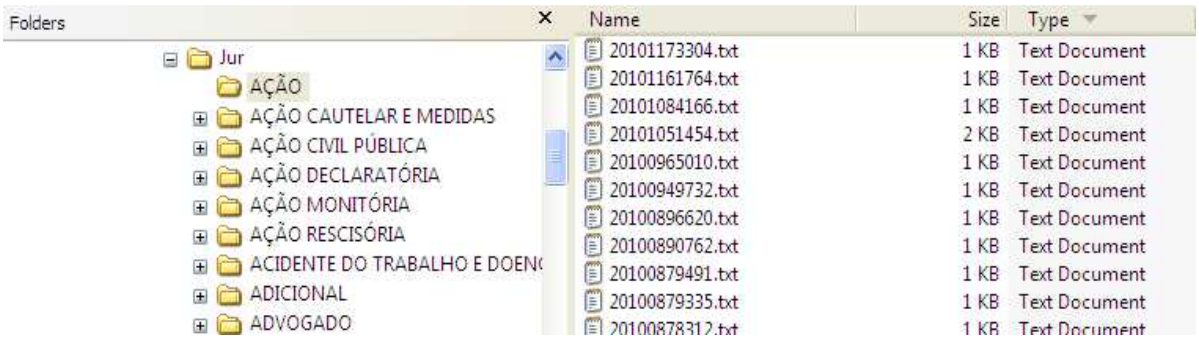


Figura 9: Estrutura de diretórios das categorias e suas ementas

O resultado final desta fase é uma coleção de documentos textuais, sem as informações de cabeçalho, como número do processo, data de julgamento e juiz relator, pois estas informações são irrelevantes. Os documentos possuindo apenas o conteúdo puro do texto da ementa são organizados em diretórios, onde cada diretório é uma categoria da jurisprudência trabalhista da 2ª Região São Paulo. Assim formou-se a organização da coleção de documentos, com cada diretório representando uma categoria, existem documentos representando o conhecimento explícito das decisões.

Pode ser verificado nesta fase que a quantidade de documentos não era distribuída de forma equânime pelas categorias, muito pelo contrário existiam categorias com unidades de documentos, e categorias com milhares de documentos, como pode ser visto no Apêndice 4. Existiam 37 categorias com até 10 documentos, sendo que 9 categorias possuíam apenas 1 documento, 17 categorias possuíam de 2 a 5 documentos e apenas 11 categorias possuíam mais de 5 documentos. Outras 41 categorias possuíam entre 11 e 50 documentos bem distribuídos entre as 41 categorias. Outras 36 categorias possuíam entre 51 e 200 documentos, sendo que a metade destas categorias tinha entre 100 e 200 documentos. Mais 25 categorias possuíam entre 201 a 400 documentos, sendo que 16 categorias possuíam entre 201 e 300 documentos. Outras 20 categorias continham entre 301 e 600 documentos. Apenas 11 categorias possuíam entre 601

e 800 documentos. Somente 5 categorias possuíam entre 801 e 900 documentos. Não havia categorias com quantidade de documentos entre 901 e 1100 documentos. Existiam 11 categorias contendo entre 1100 e 2000 documentos, outras 6 categorias contendo entre 2001 e 3000 documentos. Apenas 1 categoria contendo documentos na faixa de 3000 documentos, 2 categorias na faixa de 4000 documentos, 1 categoria na faixa de 5000 documentos e a uma categoria única na faixa de 1200 documentos.

4.2 Pré-Processamento das Ementas

Nesta etapa é onde ocorre a preparação dos documentos e extração de um conjunto de características dos mesmos, chamado de vetor atributo-valor, onde cada termo é um atributo do vetor, com um valor para cada atributo.

A primeira ação desta etapa foi definir qual a técnica de valorização dos atributos a ser utilizada, e a utilização ou não de critérios de suavização e normalização. Para tal, foi necessário verificar a distribuição da quantidade de documentos dentro de cada categoria e da quantidade de informações[3].

As categorias das ementas apresentaram uma distribuição irregular quanto à relação entre quantidade de documentos e tamanho de bytes de cada categoria, de forma que existem categorias com menos documentos, porém com mais informação, assim como existem categorias com menos informação, porém com mais documentos, conforme pode ser visto na Tabela 4. Utilizando o método *bag of words* para extrair as características dos textos, esta irregularidade pode ocasionar a exclusão de termos menos freqüentes que estariam presente em categorias com menos documentos, uma vez que a frequencia é dada em relação a todos os documentos da coleção.

Portanto, para a geração do vetor atributo-valor, foi utilizado critério de medida *Term Frequency – Inverse Document Frequency* (tf-idf), e critérios de suavização e normalização quadrática por atributo (coluna), com o objetivo de amenizar o problema da irregularidade de distribuição da quantidade de documentos

[3] Quantidade de informação medida em bytes.

e de informação nas categorias, capturando assim o máximo das características relevantes dos documentos.

Tabela 4 - As dez categorias com mais documentos.

CATEGORIA	QUANTIDADE DE DOCUMENTOS	TAMANHO (Bytes)
PREVIDENCIA SOCIAL	12865	10.250.190
EXECUÇÃO	5370	4.191.805
MÃO-DE-OBRA	4308	4.743.027
EMBARGOS DECLARATÓRIOS	4248	2.561.049
PROVA	3689	2.867.747
RELAÇÃO DE EMPREGO	2922	2.583.514
PRESCRIÇÃO	2834	2.728.237
DANO MORAL E MATERIAL	2532	2.529.515
COMPETÊNCIA	2151	2.397.916
SINDICATO OU FEDERAÇÃO	2094	2.079.449

4.2.1 Seleção dos Exemplos de Treinamento

A teoria do aprendizado computacional, conhecida como *PAC-learning,* criada por Leslie Valiant em 1984, mostra a importância da complexidade do relacionamento entre aprendizado computacional e a complexidade dos exemplos utilizados no conjunto de treinamento. Sinteticamente, a teoria leva em consideração a distribuição de exemplos positivos e negativos dentro do conjunto de treinamento, para o caso de uma predição *booleana.* De forma que a quantidade de exemplos deve ser restrita e distribuída de maneira proporcional, sem haver uma diferença grande entre exemplos positivos e negativos, caso contrário a complexidade dos exemplos fará com que o algoritmo não seja capaz de aprender, portanto é necessário restringir o espaço de exemplos (RUSSEL; NORVING, 2004).

Desta maneira, formar uma única tabela atributo-valor contendo todas as 187 categorias causaria a ineficiência de aprendizado de um algoritmo de aprendizado indutivo, e exigiria grandes recursos computacionais de processamento. As 187 categorias podem ser agrupadas em conjuntos levando em consideração a quantidade de documentos, formando conjuntos de até 9 documentos, de dezenas de documentos, de centenas de documentos e que contém milhares de documentos. Cada conjunto poderia ser pré-processado para posterior processamento, havendo

um vetor atributo-valor para cada conjunto de categoria. Porém, para as categorias com milhares de documentos, para esta pesquisa não foi possível obter poder computacional para realizar o pré-processamento, e mesmo que houvesse sua tabela atributo-valor seria muito grande, o que causaria a necessidade de um poder computacional ainda maior para processá-la.

Com o objetivo de restringir o espaço de exemplos, para esta pesquisa foram escolhidas 10 categorias que possuem no mínimo 500 documentos, distribuídos da seguinte maneira: uma categoria que possui até 1000 documentos, duas categorias que possuem entre 1000 e 2000 documentos, duas categorias que possuem entre 2000 e 3000 documentos, uma categoria que possue entre 3000 e 4000 documentos, duas categorias que possuem entre 4000 a 5000 documentos e duas categorias que possuem acima de 5000 documentos. Porém não foram selecionados todos os documentos das categorias escolhidas, pelo mesmo motivo de não haver poder computacional disponível. Foram selecionados aleatoriamente 500 documentos de um categoria, confrontados com mais 500 documentos de 5 das 177 categorias restantes, selecionadas também aleatoriamente respeitando a distribuição proporcional da quantidade de documentos real das 187 categorias, seguindo a teoria *PAC-learning*, compondo um conjunto de exemplos para treinamento, que contenha uma distribuição de exemplos positivos (da categoria que pretende-se aprender) e de exemplos negativos (das outras categorias diversas), conforme pode ser visto uma amostra na Tabela 5. O conjunto total de exemplos está presente no Apêndice 5.

Os documentos foram pré-processados utilizando a ferramenta PRETEXT 2 (SOARES, 2009) que gera os vetores atributo-valor da coleção de documentos. A ferramenta foi configurada para utilizar a métrica tf-idf, com método de suavização e normalização quadrática (por coluna).

O resultado final desta fase é a geração dos vetores atributo-valor de cada categoria a ser aprendida, em conjunto com outras categorias selecionadas de maneira aleatória. Portanto, foram geradas 10 vetores atributo-valor. A ferramenta trabalha com um formato próprio de tabela atributo-valor, e para servir de entrada para a fase seguinte deve ser traduzido para o formato ARFF (Attribute-Relation File Format).

Tabela 5 - Exemplo de 3 Categorias utilizadas e quantidade de exemplos selecionados.

Categoria	Real[4]	Selec[5]	Outras	Real[4]	Selec[5]
EXECUÇÃO	5370	500	EMBARGOS DECLARATÓRIOS	4248	181
			RELAÇÃO DE EMPREGO	2922	125
			SINDICATO OU FEDERAÇÃO	2094	89
			MANDADO DE SEGURANÇA	1612	69
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	39
			Total de Outros	**11771**	**503**
MÃO-DE-OBRA	4308	500	PROVA	3689	212
			SINDICATO OU FEDERAÇÃO	2094	121
			RECURSO	1297	75
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	52
			PROCESSO	726	42
			Total de Outros	**8701**	**502**
EMBARGOS DECLARATÓRIOS	4248	500	PRESCRIÇÃO	2834	171
			SINDICATO OU FEDERAÇÃO	2094	126
			CONCILIAÇÃO	1377	83
			HORAS EXTRAS	1136	69
			NORMA COLETIVA (EM GERAL)	885	54
			Total de Outros	**8326**	**503**

4.3 Processamento das Ementas

As tabelas atributos-valor, após a tradução para o formato ARFF (*Attribute-Relation File Format*) foram inseridas na ferramenta WEKA - *Waikato Environment for Knowledge Analysis* (WITTEN; FRANK, 2000), para que os dados fossem processados por algoritmos de aprendizado de máquina, e assim fossem criados modelos de conhecimento.

Como a seleção das categorias, e seus exemplos, foi realizada de maneira aleatória, e assim sendo impossível prever quais as características dos exemplos a serem aprendidos, optou-se por não apenas um algoritmo de aprendizado, mas um comitê de classificadores, formado por três algoritmos de aprendizado distintos: Árvore de Decisão, Naive Bayes e SVM.

Foi utilizado o algoritmo J4.8 como implementação do algoritmo de árvore de decisão disponível através da ferramenta WEKA. É uma implementação posterior,

[4] Quantidade real de exemplos presentes na categoria
[5] Quantidade de exemplos selecionados aleatoriamente

com poucas melhorias do algoritmo C4.5 *revision* 8. A ferramenta WEKA possui também a implementação do classificador probabilístico Naive Bayes, utilizando a distribuição normal para modelar os atributos (WITTEN; FRANK, 2000).

Uma variante do algoritmo SVM, denonimada SMO (*Sequential Minimal Optimization*), foi utilizado como algoritmo classificador SVM, sendo implementada através da ferramenta WEKA.

Portanto, foram montados 3 modelos de aprendizado, utilizando as 3 implementações de algoritmos de aprendizado (J4.8, Naive Bayes e SMO), para cada uma das 10 categorias selecionadas.

A técnica utilizada para a avaliação do treinamento dos algoritmos foi o *cross-validation*. Essa técnica quebra o conjunto de exemplos em dois, um conjunto usado para treinar o algoritmo e outro utilizado para testá-lo, de forma a poder avaliar a precisão do algoritmo treinado. A escolha dos exemplos para cada conjunto é realizada de forma aleatória. Para que o algoritmo aprenda com uma diversidade maior de exemplos, e possa ir ajustando sua taxa de erro, é recomendado repetir o processo várias vezes, alternando os exemplos dos conjuntos (WITTEN; FRANK, 2000).

É possível fixar o número de *folds*, ou partições dos exemplos a serem utilizados. Foram utilizados 3 *folds* para o treinamento dos algoritmos. Portanto, os exemplos foram divididos em 3 partes aproximadamente iguais, sendo uma parte para testar, enquanto o restante foi utilizado para treinar, ou seja, foram utilizados dois terços para treinar e um terço para testar, sendo repetido o processo por três vezes, para que no final cada parte tenha sido utilizada para teste. Essa maneira de treinar é conhecida como *threefold cross-validation* (WITTEN; FRANK, 2000).

4.3.1 Resultados do Treinamento

A taxa de acertos durante os testes do treinamento foram altas, com poucas variações entre os algoritmos, sendo que na maioria das vezes o algoritmo SMO obteve melhores taxas, porém a diferença dele para os outros algoritmos foi muito pouca, o que dificulta afirmar qual o algoritmo que teve melhor índice de acerto, como pode ser visto na Tabela 6.

A diferença principal entre os algoritmos foi a saída apresentada do modelo de aprendizado de cada algoritmo. O J4.8 como uma implementação de uma árvore de decisão, permitiu identificar facilmente termos (*stems*) relevantes para diferenciar uma categoria das demais, como pode ser notado no desenho da árvore da categoria SINDICATO, conforme a Figura 10.

É possível identificar facilmente a relevância dos *stems* “contribu”, “sindicat”, “sindical” e “fat”, ou seja, palavras derivadas destes *stems* aparecem com muita freqüência e se sobressaem como características principais dos documentos da categoria SINDICATO. Podemos citar como derivados destes *stems* as palavras: “contribuição”, “sindicato”, “sindical”, “fato” e “fator”.

Tabela 6 - Taxa de acertos dos algoritmos durante o treinamento.

Categorias	Acertos durante treinamento (cross-validation)		
	J4.8	Naïve Bayes	SMO
EXECUÇÃO	92,30%	93,90%	95%
PREVIDÊNCIA SOCIAL	97,30%	98,30%	98,20%
MÃO-DE-OBRA	92,91%	91,91%	93,21%
EMBARGOS DECLARATÓRIOS	99,20%	97,70%	98,50%
PROVA	90,90%	87,93%	94,50%
RELAÇÃO DE EMPREGO	93,50%	94,40%	97,60%
SINDICATO OU FEDERAÇÃO	97,10%	97,90%	97,40%
HONORÁRIOS	97,40%	97,30%	97,30%
NULIDADE PROCESSUAL	96,30%	92,10%	95,40%
RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	98%	97,70%	98,40%

Os algoritmos Naive Bayes e SMO não permitiram tal visualização do aprendizado, apresentaram em suas saídas apenas o cálculo do bias e taxas de erros durante a criação do modelo de aprendizado.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

contribu <= 0
|   sindical <= 0
|   |   sindicat <= 0: OUTROS (497.0/6.0)
|   |   sindicat > 0
|   |   |   fat <= 0: SINDICADO_FEDERACAO (23.0/4.0)
|   |   |   fat > 0: OUTROS (3.0)
|   sindical > 0: SINDICADO_FEDERACAO (88.0/1.0)
contribu > 0: SINDICADO_FEDERACAO (392.0/4.0)

Number of Leaves  :       5

Size of the tree :  9
```

Figura 10. Árvore de decisão do classificador J4.8 da categoria SINDICATO.

4.3.2 Resultados dos testes dos classificadores

Os modelos de aprendizado dos algoritmos foram salvos, e depois confrontados com exemplos dessas mesmas categorias, que são desconhecidos pelos modelos de aprendizado.

Foram selecionados aleatoriamente, 5 exemplos, desconhecidos para os modelos aprendidos, de cada uma das categorias usadas em treinamento, totalizando 50 documentos a serem preditos. Os documentos foram pré-processados, porém sem informar a categoria a que pertencem, para que os algoritmos façam a predição de suas categorias.

Assim, os classificadores binários, devidamente treinados, receberam 50 exemplos desconhecidos para realizarem a sua predição individual. O resultado da predição de cada algoritmo por categoria, forma o resultado do comitê de classificadores, que utilizou como critério a maior votação entre os algoritmos.

4.3.2.1 Taxa de erro por categoria e Taxa de erro total

A Tabela 7 demonstra a taxa de erro da categoria e a taxa de erro total do algoritmo. A taxa de erro da categoria demonstra a predição incorreta (falsos negativos) dentro dos exemplos verdadeiros do classificador de uma categoria, e a

taxa de erro total demonstra a predição incorreta dentro de todos os exemplos a serem preditos (soma dos falsos negativos e falsos positivos).

O algoritmo Naive Bayes foi o algoritmo que obteve maior taxa de erro por categoria, ou seja, foi o algoritmo que menos conseguiu predizer verdadeiros positivos, chegando a predizer nenhum verdadeiro positivo na categoria “Execução”, obteve uma taxa de 60% de erros na categoria “Mão-de-obra”, e 20% nas categorias “Previdência social” e “Prova”. O algoritmo SMO apesar de apresentar taxas de erros somente em duas categorias, “Execução” e “Prova”, teve uma taxa de erro de 80% na categoria “Execução”. Já o algoritmo J4.8, apesar de apresentar taxas de erros nas categorias “Execução”, “Mão-de-obra” e “Prova”, todas as taxas foram inferiores a 40%.

Tabela 7 - Taxa de Erro da Categoria e Taxa de erro total.

Categorias	J4.8		Naïve Bayes		SMO		Comitê	
	Taxa de Erro da Categoria	Taxa de Erro Total	Taxa de Erro da Categoria	Taxa de Erro Total	Taxa de Erro da Categoria	Taxa de Erro Total	Taxa de Erro da Categoria	Taxa de Erro Total
EMBARGOS DECLARATÓRIOS	0,00%	4,00%	0,00%	12,00%	0,00%	10,00%	0,00%	4,00%
EXECUÇÃO	40,00%	4,00%	100,00%	14,00%	80,00%	12,00%	80,00%	8,00%
HONORÁRIOS	0,00%	2,00%	0,00%	38,00%	0,00%	4,00%	0,00%	8,00%
MÃO-DE-OBRA	20,00%	10,00%	60,00%	18,00%	0,00%	8,00%	0,00%	2,00%
NULIDADE PROCESSUAL	0,00%	4,00%	0,00%	36,00%	0,00%	14,00%	0,00%	12,00%
PREVIDÊNCIA SOCIAL	0,00%	0,00%	20,00%	8,00%	0,00%	2,00%	0,00%	0,00%
PROVA	20,00%	14,00%	20,00%	28,00%	20,00%	12,00%	20,00%	28,00%
RELAÇÃO DE EMPREGO	0,00%	44,00%	0,00%	36,00%	0,00%	38,00%	0,00%	32,00%
RESPOSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	0,00%	12,00%	0,00%	78,00%	0,00%	10,00%	0,00%	32,00%
SINDICATO OU FEDERAÇÃO	0,00%	16,00%	0,00%	28,00%	0,00%	2,00%	0,00%	22,00%

O comitê classificador, unindo os resultados de predição dos três algoritmos classificadores, conseguiu anular a taxa de erro do algoritmo Naive Bayes na categoria “Previdência Social” e “Mão-de-obra”, registrando taxas de erros somente na categoria “Execução” e “Prova”, acompanhando os resultados do algoritmo SMO, apresentando taxa de 0% de erro nas demais categorias, portanto, 100% de acerto nas predições de 8 categorias, como pode ser visto no gráfico da Figura 11.

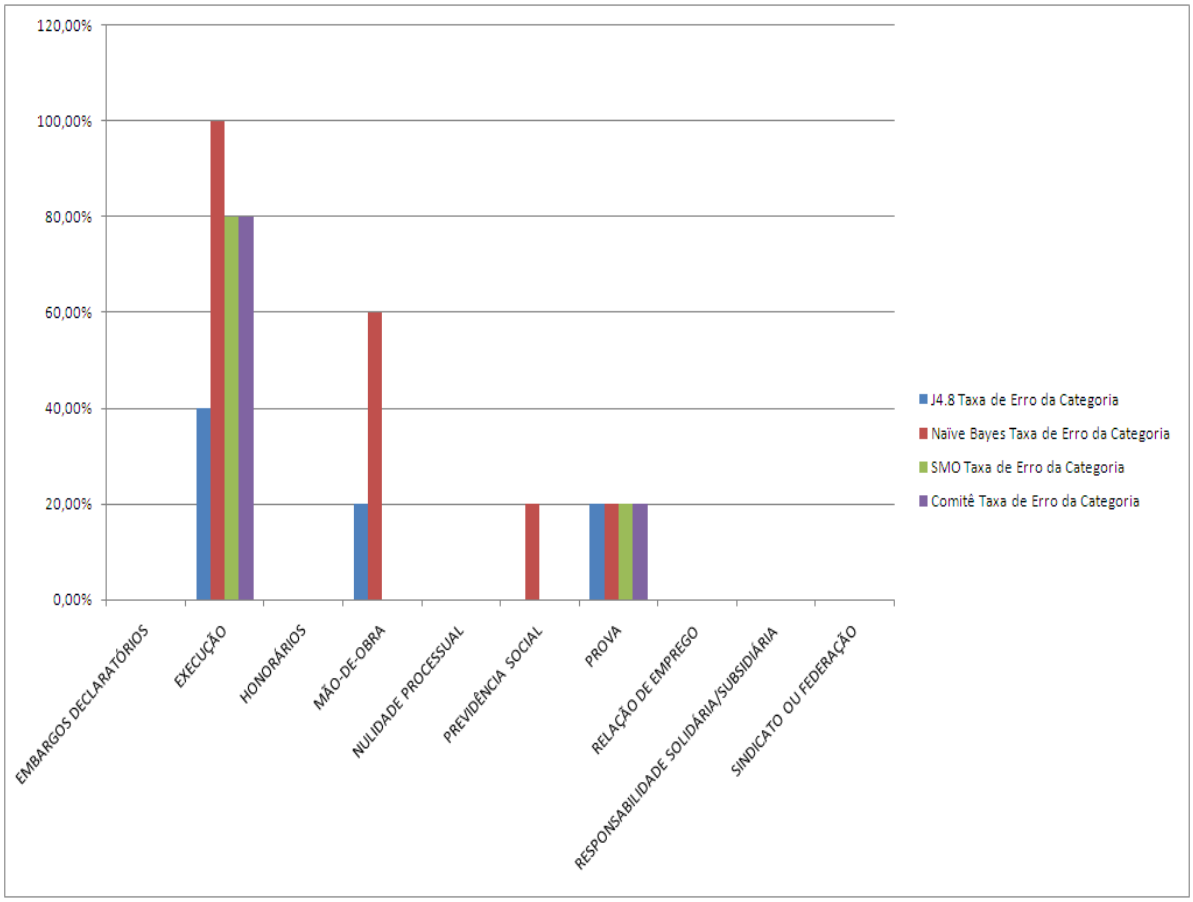


Figura 11: Gráfico indicando as taxas de erro por categoria apresentadas pelos algoritmos

A taxa de erro total esteve presente em todas as categorias para os classificadores Naive Bayes e SMO. O algoritmo J.48 foi o que apresentou menos taxa de erro total. O comitê classificador conseguiu anular a taxa de erro total apresentada pelos algoritmos Naive Bayes e SMO na categoria “Previdência Social”, porém teve taxas de erro total maiores que o algoritmo J4.8 nas categorias “Responsabilidade Solidária/Subsidiária” e “Sindicato ou Federação”. O gráfico da Figura 12 representa distribuição das taxas de erro total entre os algoritmos.

Naive Bayes novamente é o algoritmo que apresentou maior taxa de erro total chegando a 78% de taxa de erro total na categoria “Responsabilidade Solidária/Subsidiária”. A categoria “Execução” foi a terceira categoria a apresentar menor taxa de erro total, sendo que foi a categoria dentre todas que apresentou maior taxa de erro por categoria, ou seja, apesar de não ter conseguido predizer corretamente os documentos da categoria “Execução”, os algoritmos conseguiram relativo sucesso em reconhecer quais os documentos não pertencem a “Execução”.

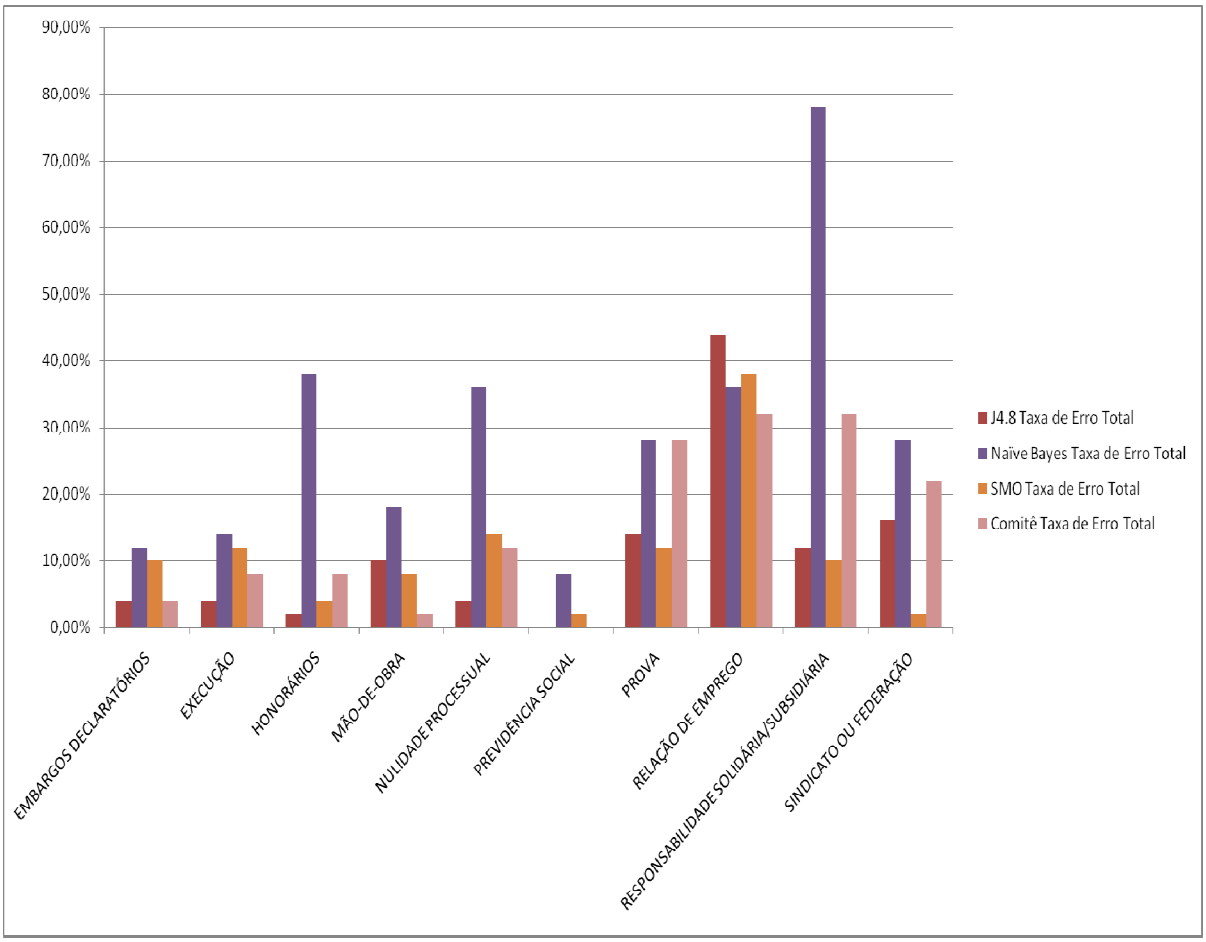


Figura 12: Gráfico indicando o erro total de cada algoritmo.

4.3.2.2 Acuidade Total

A acuidade total individual (verdadeiros positivos, somados aos verdadeiros negativos) de cada um dos algoritmos classificadores e do comitê classificador, após a predição dos 50 exemplos desconhecidos, pode ser analisada através da Tabela 8. É possível notar que não há um classificador que obteve uma acuidade maior em todas as categorias.

Na maioria das categorias o algoritmo J4.8 teve mais acuidade em relação aos algoritmos Naive Bayes e SMO, mas por uma diferença pequena de menos de 10%. O algoritmo SMO foi superior ao algoritmo J4.8 para as categorias “Embargos Declaratórios” e “Honorários”, mas também por um diferença de no máximo 10%. O comitê classificador obteve suas taxas de acuidade por categoria, próximas ao do algoritmo que apresentou melhor acuidade por categoria, quando a diferença de

acuidade entre os algoritmos era pequena. E obteve a acuidade média dos algoritmos classificadores quando estes divergiam muito de suas taxas de acuidade.

Tabela 8 - Acuidade dos algoritmos classificadores e do comitê classificador.

Categorias	Acuidade Total			
	J4.8	Naïve Bayes	SMO	Comitê
EMBARGOS DECLARATÓRIOS	96,00%	88,00%	90,00%	96,00%
EXECUÇÃO	96,00%	86,00%	92,00%	92,00%
HONORÁRIOS	98,00%	62,00%	88,00%	92,00%
MÃO-DE-OBRA	86,00%	78,00%	96,00%	94,00%
NULIDADE PROCESSUAL	96,00%	64,00%	78,00%	88,00%
PREVIDÊNCIA SOCIAL	100,00%	92,00%	100,00%	100,00%
PROVA	86,00%	72,00%	54,00%	72,00%
RELAÇÃO DE EMPREGO	56,00%	64,00%	68,00%	68,00%
RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	88,00%	22,00%	60,00%	68,00%
SINDICATO OU FEDERAÇÃO	84,00%	72,00%	82,00%	78,00%

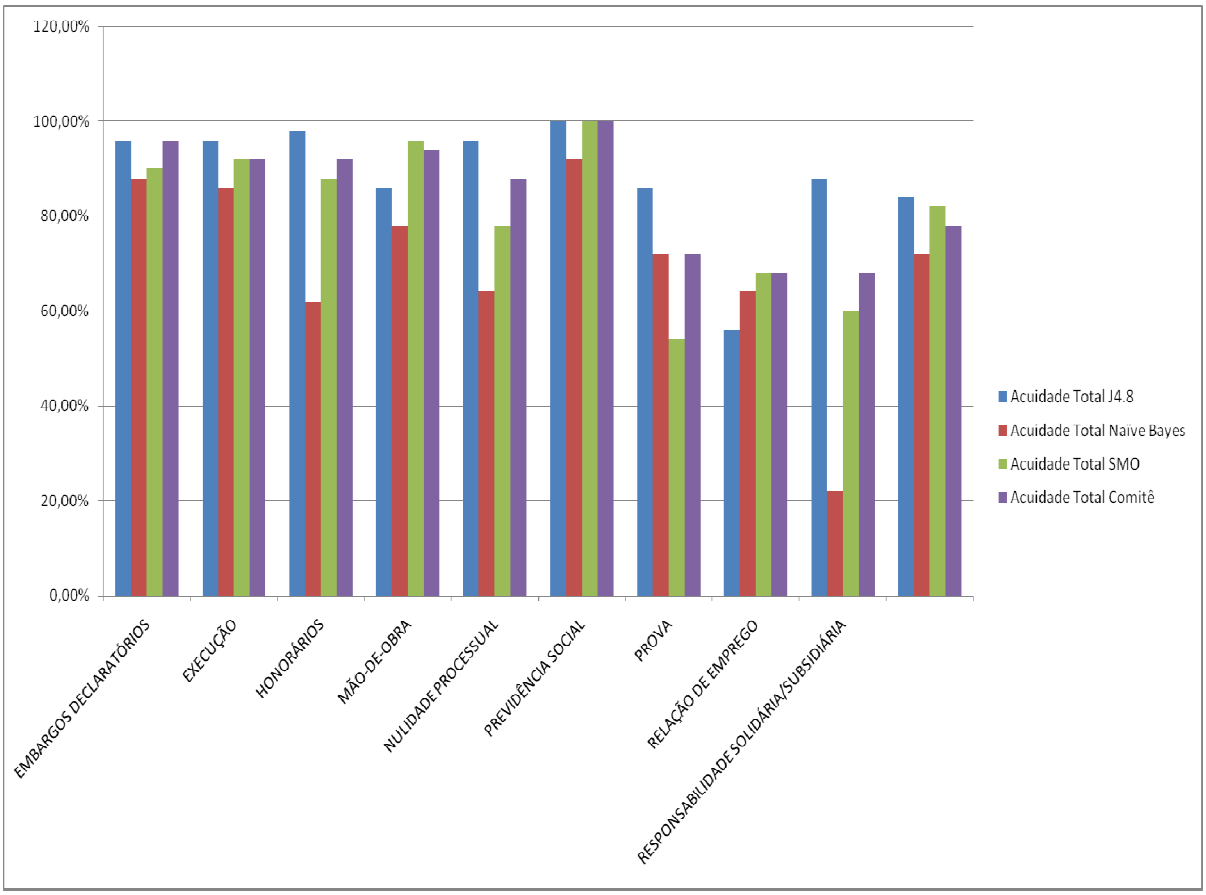


Figura 13: Gráfico indicando a acuidade total de cada algoritmo.

Como é o caso dos resultados das categorias “Responsabilidade Solidária/Subsidiária” e “Prova”, onde o comitê classificador obteve uma acuidade inferior ao algoritmo J4.8, devido à baixa acuidade dos algoritmos SMO e Naive Bayes para essas categorias. Porém o comitê sempre manteve uma taxa de acuidade total superior a 60%, como pode ser observado melhor no gráfico da Figura 13.

4.3.2.3 Verdadeiros Positivos e Verdadeiros Negativos do Comitê Classificador

Realizada a contagem do total de verdadeiros positivos (Tp) e verdadeiros negativos (*Tn*), preditos pelo comitê de classificadores, e levando em consideração que foram testados 50 exemplos, sendo 5 exemplos de cada categoria, ou seja, o máximo de verdadeiros positivos por categoria é 5 e o máximo de verdadeiros negativos é 45. Os valores de *Tp* e *Tn* de cada categoria foram normalizados e colocados no gráfico apresentado na Figura 14, onde é possível analisar a acuidade do comitê em predizer verdadeiros positivos e verdadeiros negativos.

É possível verificar que o comitê de classificadores conseguiu predizer corretamente todos os verdadeiros positivos e negativos da categoria “Previdência Social”. Conseguiu predizer todos os verdadeiros positivos de todas as categorias, exceto das categorias “Execução” e “Prova”, sendo que a categoria “Execução” teve a menor quantidade de verdadeiros positivos preditos. Predisse corretamente todos os verdadeiros negativos somente da categoria “Execução”. As categorias “Embargos Declaratórios”, “Honorários”, “Mão-de-obra” e “Nulidade Processual” obtiveram a quantidade de verdadeiros negativos muito próxima do total por categoria.

As categorias “Prova”, “Relação de Emprego”, “Responsabilidade Solidária/Subsidiária” e “Sindicato ou Federação” obtiveram a quantidade de verdadeiros negativos mais baixas, entre 0,6 e 0,8 pontos normalizados. O que indica que entre 0,2 a 0,4 dos exemplos foram preditos como falsos positivos destas categorias.

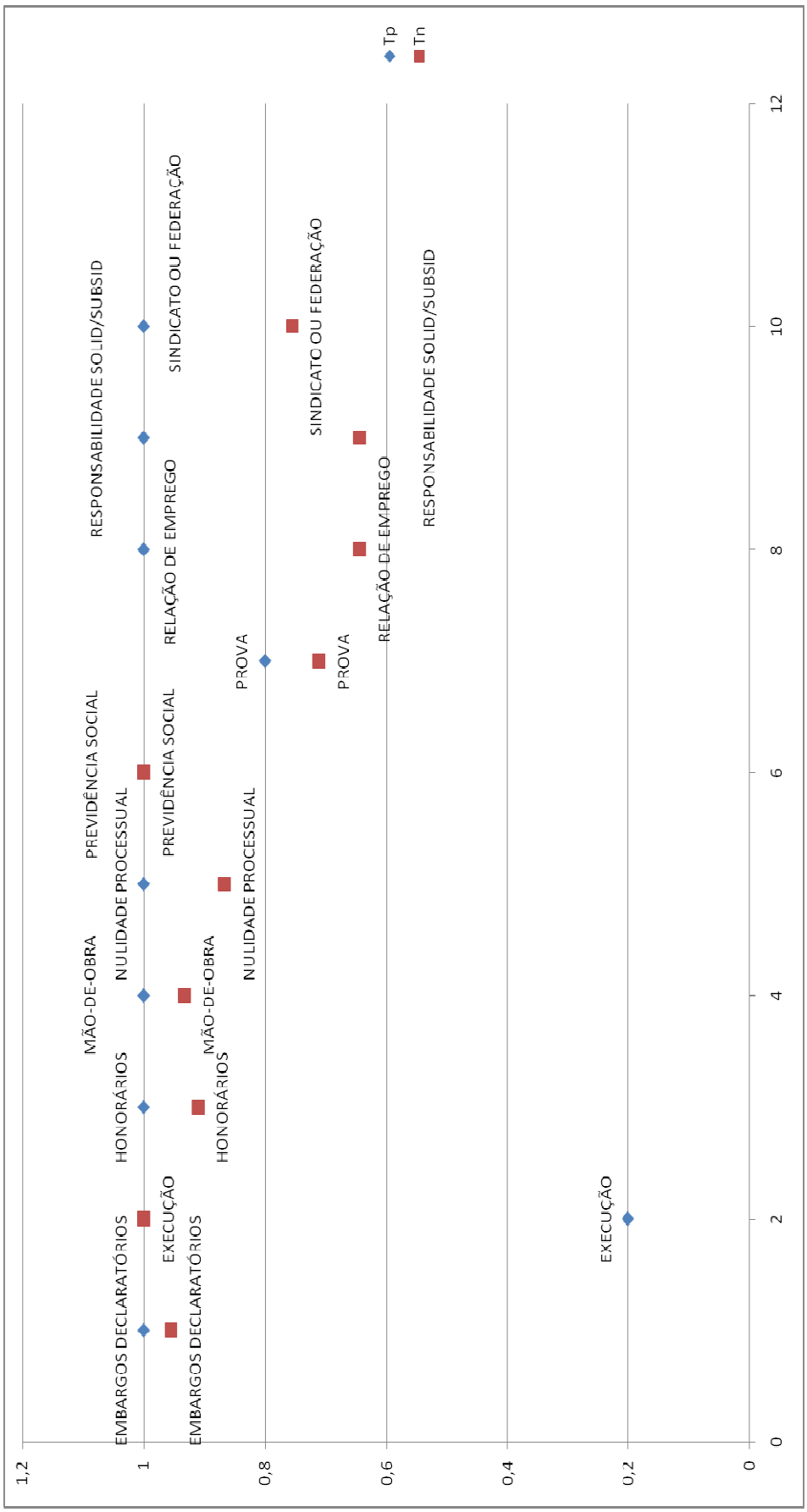


Figura 14: Gráfico normalizado da acuidade total do comitê classificador

Através da tabela de predição do comitê classificador, apresentada no Apêndice 6, foi realizada a análise dos falsos positivos das categorias “Prova”, “Relação de Emprego”, “Responsabilidade Solidária/Subsidiária” e “Sindicato ou Federação” e foi constatado que:

- O comitê classificador da categoria “Prova” classificou todos os exemplos da categoria “Nulidade Processual” e “Responsabilidade Solidária/Subsidiária” como falsos positivos;
- O comitê classificador da categoria “Relação de Emprego” e “Responsabilidade Solidária/Subsidiária”, ambos classificaram todos os exemplos da categoria “Mão-de-obra” como falsos positivos.

4.3.2.4 Avaliação do especialista humano

A alta taxa de erro por categoria apontada na categoria “Execução”, que é ratificada com a quantidade baixa de verdadeiros positivos, apresentada pelo comitê classificador, fez com que os documentos da categoria “Execução” fossem novamente classificados por um especialista humano, para avaliar a classificação atribuída pelo comitê classificador.

A unanimidade do comitê classificador da categoria “Prova” em classificar os documentos das categorias “Nulidade Processual” e “Responsabilidade Solidária/Subsidiária”, fez com que documentos de ambas as categorias fossem submetidas à nova classificação pelo especialista humano.

Assim como, a unanimidade dos comitês classificadores das categorias “Relação de Emprego” e “Responsabilidade Solidária/Subsidiária”, em classificar todos os documentos da categoria “Mão-de-obra” como falsos positivos, fez com que documentos da categoria “Mão-de-Obra” fossem relacionados para nova classificação do especialista humano.

Foi elaborado um formulário, conforme Apêndice 7, exibindo somente o conteúdo do texto da decisão e da ementa dos documentos das referidas categorias, sem identificação do documento original, numerando os documentos seguindo a numeração do mesmo na tabela de predição apresentada Apêndice 6.

Os resultados da reclassificação realizada pelo especialista, no total, divergiram pouco da classificação anterior. Porém, alguns resultados podem ser destacados:

- Um dos documentos da categoria “Execução” (documento de identificador 10), foi reclassificado pelo especialista humano, como pertencente à outra categoria diferente das 10 selecionados para o experimento. Analisando as predições dos comitês classificadores para este documento, é possível notar que apenas o comitê classificador da categoria “Nulidade Processual” induziu este documento como sendo da categoria “Nulidade Processual”. Todos os outros comitês de classificadores induziram como verdadeiro negativo, exceto o comitê da categoria “Execução” que induziu como falso negativo, ou seja, não identificaram o referido documento como sendo de suas respectivas categorias;
- Um dos documentos da categoria “Mão-de-obra” (documento de identificador 20), foi classificado desta vez pelo especialista humano, como pertencente à categoria “Responsabilidade Solidária/Subsidiária”, que foi o comportamento do comitê classificador da categoria “Responsabilidade Solidára/Subsidiária” para todos os documentos da categoria “Mão-de-obra”;
- Um dos documentos da categoria “Responsabilidade Solidária/Subsidiária” (documento de identificador 45), foi reclassificado pelo especialista como pertencente à categoria “Mão-de-obra”. Analisando o comitê classificador da categoria “Mão-de-obra”, o referido documento foi induzido como pertencente a esta categoria;
- Outro documento da categoria “Responsabilidade Solidára/Subsidiária” (documento de identificador 43) foi reclassificado como sendo da categoria “Relação de Emprego”. O comitê classificador da categoria “Relação de Emprego” induziu o documento como sendo pertencente a categoria;

O especialista reportou que os textos inseridos no formulário não são textos de fácil interpretação, e que muitas vezes existe a incerteza em classificar estes textos nas categorias existentes. Informou também que é possível ocorrer a troca de categoria de documentos, principalmente documentos pertencentes à categorias semanticamente mais genéricas como “Mão-de-Obra” e “Relação de Emprego”.

5 Conclusão

A Fase de extração das ementas constitui a fase mais simples e elementar do método apresentado neste trabalho. Nela são formadas as coleções de textos com suas respectivas categorias, sendo possível a análise do quantitativo de documentos e também de informações (bytes) da coleção. Essas informações são úteis para a decisão de qual métrica utilizar para valorar os atributos com o intuito de melhor extrair as características do texto.

O pré-processamento utilizando a técnica *bag of words* contendo todos os termos (*stems*) encontrados, gera vetores atributo-valor muito grandes, causando o hiper-dimensionamento dos vetores, impossibilitando a utilização da coleção de exemplos completa para treinamento. A união de termos (*stems*) formando conceitos não foi utilizada neste trabalho, ela reduziria consideravelmente o tamanho dos vetores, uma vez que os atributos iriam diminuir. Porém exige a necessidade da construção dos conceitos, e para tal seria necessária a participação do especialista, que não esteve disponível nessa fase do trabalho. No entanto, a solução adotada da seleção aleatória por amostragem, respeitando a distribuição da quantidade de documentos da coleção real, formando conjuntos de exemplos para treinamento e formação de classificadores binários, apresentou bons resultados como pode ser visto nos resultados de treinamento.

O processamento aplicando as implementações dos algoritmos J4.8, Naive Bayes e SMO obtiveram excelente desempenho durante o treinamento dos modelos de aprendizagem. Todavia, não é possível afirmar qual o melhor, pois a diferença entre eles foi mínima. Os testes de predição demonstraram que apesar de não apresentarem muita diferença durante o treinamento, durante a predição os resultados obtidos pelos algoritmos foram bem distintos, onde o algoritmo Naive Bayes obteve a pior desempenho e o J4.8 obteve melhor desempenho quanto a acuidade total em todas as categorias, exceto a categoria “Relação de Emprego” onde o SMO obteve o maior desempenho. A formação do comitê classificador por categoria, unindo os resultados dos classificadores binários, não trouxe grande benefício na acuidade total, porém na acuidade por categoria auxiliou na redução da taxa de erro por categoria, fazendo com que um número maior de verdadeiros

positivos fosse predito. Desta forma, é possível afirmar que a combinação dos resultados de algoritmos classificadores, forma um classificador mais preciso para esta aplicação.

A análise dos falsos positivos preditos pelo comitê classificador permitiu identificar documentos que poderiam estar classificados em mais de uma categoria. A identificação destes documentos e a reclassificação realizada pelo especialista demonstram que o comitê classificador pode ser utilizado como ferramenta de auxílio ao especialista humano, sugerindo possíveis categorias.

O relato do especialista humano em afirmar a dificuldade de classificar documentos em categorias mais genéricas, como por exemplo “Mão-de-obra” e “Relação Trabalhista”, coincide com o fato de vários termos iguais estarem presentes em ambas as categorias, com frequencias próximas, o que dificulta também a predição do comitê de classificadores. Trabalhos futuros podem ser realizados com o intuito de agregar valor semântico aos termos, objetivando uma diferenciação maior entre as categorias por parte do comitê classificador. No entanto, o fato do próprio especialista humano em afirmar a existência de categorias muito próximas, indica uma possível falha na ontologia das categorias, que deve ser avaliada no âmbito das ciências jurídicas.

Os objetivos específicos da pesquisa foram alcançados, ou seja, foram estudadas e aplicadas, técnicas de mineração de texto para a Extração das ementas, Pré-processamento das ementas e Processamento das ementas. Um método para trabalhar com a classificação das ementas foi estabelecido e seus resultados foram analisados e submetidos ao crivo dos especialistas.

A hipótese foi confirmada, com a utilização de técnicas de Classificação de Textos, aliadas ao aprendizado de máquina supervisionado é possível que um sistema computacional indique à qual categoria é mais provável que uma ementa de jurisprudência pertença, desta forma auxiliando o trabalho do especialista classificador.

6 Referências

ÁLVAREZ, A. C. **Extração de Informação de Artigos Científicos: uma abordagem baseada em indução de regras de etiquetagem.** 2007. 131 f. Dissertação (Mestrado em Ciência de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2007.//

ALMEIDA FILHO, J. C. A. **Processo Eletrônico e Teoria Geral do Processo Eletrônico: a informatização judicial no Brasil – 3ª. Edição:** Forense, 2010. ISBN 978-85-309-3122-3

BEPPLER, M. D.; FERNANDES, A. M. R. **Aplicação de Text Mining para Extração de Conhecimento Jurisprudencial.** In: I Congresso Sul Catarinense de Computação, out. 2005, Criciúma, SC. ISBN 85-88390-29-9. Disponível na Internet: <http:// www.dcc.unesc.net/sulcomp/05/Art081SulComp2005.pdf>. Acesso em 08/09/2010.

BITTENCOURT, G. **Inteligência Artificial: Ferramentas e Teorias – 3ª. Edição:** UFSC, 2006. ISBN 85-328-0138-2

CASTRO, A. A.; **Manual de Informática Jurídica e Direito da Informática**: Forense, 2005. ISBN 853091919X

Conselho Nacional de Justiça. **Resolução Nº 46, de dezembro de 2007.** Disponível na Internet:< http://www.cnj.jus.br/images/stories/docs_cnj/resolucao/rescnj_46.pdf>. Último acesso em 01/11/2010.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. A. Mineração de Textos. *In*: REZENDE, S. O. **Sistemas Inteligentes**: fundamentos e aplicações. Barueri, SP: Manole, 2003. ISBN 85-204-1683-7.

FELDMAN, R; SANGER, J. **The Text Mining Handbook**: Cambridge University Press, 2007. ISBN 0-521-83657-3

FRANÇA, M. M. **Pronunciamento de Abertura Colégio de Presidentes e Corregedores dos TRTs – Reunião de 28 de Setembro de 2010.** Disponível em: <http://www.csjt.jus.br/noticias/base_dados/abertura_coleprecor.pdf>. Acesso em 31/10/2010.

GONÇALVES, L. S. M.; REZENDE, S. O. **Categorização em Text Mining.** Disponivel em: http://www.icmc.usp.br/~std-cd/Artigos/Computacao/IC/LeaSilviaMG.pdf. Acesso em: 05/09/2007

JÚNIOR, L. C. G. **Avaliação automática da qualidade de escrita de resumos científicos em inglês** 2007. 165 f. Dissertação (Mestrado em Ciência de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2007.

KONCHADY, M. **Text Mining Application Programming**: Charles River Media, 2006. ISBN 1-58450-460-9.

MOLINARI, A. H.; TACLA, C. A. **Titulação Automática de Acórdãos Baseado em Ontologia Jurisprudencial.** In: Revista Democracia Digital e Governo Eletrônico, v.2, n. 3, 2010, ISSN 2175-9391. Disponível na internet: http://www.buscalegis.ufsc.br/revistas/index.php/observatoriodoegov/article/download/34037/33046. Acesso em 31/08/2011.

MONARD, M. C.; PRATI, R. C.; SOARES, M. V. B. **PreText II: Descrição da Reestruturação da Ferramenta Pré-processamento de Textos**: Relatórios Técnicos do Instituo de Ciências Matemática e de Computação, Universidade de São Paulo, São Paulo, 2008. ISSN 0103-2569. Disponível na Internet: http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_333.pdf. Acesso em 01/03/2011.

MONARD, M. C.; BARANAUSKAS, J. A. Indução de Regras e Árvores de Decisão. *In*: REZENDE, S. O. **Sistemas Inteligentes**: fundamentos e aplicações. Barueri, SP: Manole, 2003. ISBN 85-204-1683-7.

MONTEIRO, O. L.; GOMES, R. I; OLIVEIRA, T. **Etapas do Processo de Mineração de Textos – uma abordagem aplicada a textos em Português do Brasil.** In: Anais do XXVI Congresso da SBC, WCOMPA I Workshop de Computação e Aplicações. Jul. 2006, Campo Grande, MS.

MONTORO, A. F. **Introdução a ciência do direito.** 25ª Ed. São Paulo: Revista dos Tribunais, 2000.

MORAIS, E. A. M. **Contextualização de Documentos em Domínios Representados por Ontologias Utilizando Mineração de Textos**. 2007. 130 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal de Goiás, Goiânia, 2007.

NUNES, L. A. R. **Manual de Introdução ao Estudo do Direito**: 2ª ed. São Paulo: Saraiva, 1995. ISBN 850201942.

OLIVEIRA, J. M. L. L. **Introdução ao Direito**: 2ª ed. Rio de Janeiro: Lumen Juris, 2006. ISBN 8573879327.

PARK, A. F. M I. **Aplicação de Técnicas de Mineração de Textos para categorização de eventos de Segurança no CITR Gov.** 2010. 82 f. Dissertação (Mestrado em Informática) – Pós-Graduação da Universidade de Brasília, Brasília, 2010.

Presidência da República, Casa Civil, Sub-chefia para Assuntos Jurídicos. **Lei Nº 11.419, de 19 de dezembro de 2006**. Disponível na Internet: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11419.htm>. Último acesso em 31/10/2010.

ROVER, A. J. **Aplicação de Sistemas Especialistas no Direito – algumas questões de ordem epistemológica.** In: JAIIO 2007, 36º International Conference of the Argentine Computer Science and Operational Research Society, Mar del Plata, Argentina, 2007. Disponível na Internet em: <http://www.infojur.ufsc.br/aires/arquivos/jaiio%20epistemologia%20e%20Sistemas%20Especialistas%20Legais.pdf > Acesso em 09/09/2010.

ROVER, A. J. **Sistemas Especialistas Legais: Pensando o Sistema Jurídico.** In: Revista Eletrônica BuscaLegis dez. 1994, Florianópolis, SC.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**: trad. da 2ª ed. Rio de Janeiro: Elsevier, 2004. ISBN 85-352-1177-2.

SILVA, O. J. P. **Vocabulário Jurídico**: 28ª ed. São Paulo: Forense, 2009. ISBN 9788530927424.

SOARES, M. V. B. **Aprendizado de máquina parcialmente supervisionado multidescrição para realimentação de relevância em recuperação de informação para a WEB 2009.** 95 f. Dissertação (Mestrado em Ciência de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2009.

Tribunal Superior do Trabalho. **Instrução Normativa Nº 30 de 2007**. Disponível na Internet: <http://www.tst.gov.br/DGCJ/instrnorm/30.htm>. Último acesso em 31/10/2010.

WITTEN, I. H.; FRANK E. **Data Mining – Pratical Machine Learning Tools e Tecnhiques with JAVA Implementations**: Morgan Kaufmann, 2000. ISBN 1-55860-552-3.

7 Apêndice 1

Lista de Categorias das Ementas da Jurisprudência do TRT 2ª Região SP

"FACTUM PRINCIPIS"
"HABEAS CORPUS"
"HABEAS DATA"
AÇÃO
AÇÃO CAUTELAR E MEDIDAS
AÇÃO CIVIL PÚBLICA
AÇÃO DE PRESTAÇÃO DE CONTAS
AÇÃO DECLARATÓRIA
AÇÃO MONITÓRIA
AÇÃO PENAL
AÇÃO RESCISÓRIA
ACIDENTE DO TRABALHO E DOENÇA PROFISSIONAL
ADICIONAL
ADVOGADO
AERONAUTA
AEROVIÁRIO
AGRAVO DE INSTRUMENTO
AGRAVO REGIMENTAL
ALIENAÇÃO FIDUCIÁRIA
ALTERAÇÃO CONTRATUAL
APOSENTADORIA
ARQUIVAMENTO
ARTISTA
ASSÉDIO
ASSISTÊNCIA JUDICIÁRIA
ATLETA PROFISSIONAL
AUDIÊNCIA OU SESSÃO DE JULGAMENTO
AUTOS
AUXÍLIO ENFERMIDADE
AVISO PRÉVIO
BANCÁRIO
BOLSISTA
CARGO DE CONFIANÇA
CARTÃO PONTO OU LIVRO
CARTEIRA DE TRABALHO
CARTÓRIO
CHAMAMENTO AO PROCESSO OU DENUNCIAÇÃO À LIDE
COISA JULGADA
COMISSIONAMENTO
COMISSIONISTA
IMPOSTO DE RENDA
INCIDENTE DE FALSIDADE
INDENIZAÇÃO
INQUÉRITO JUDICIAL
INSALUBRIDADE OU PERICULOSIDADE (ADICIONAL)
INSALUBRIDADE OU PERICULOSIDADE (EM GERAL)
ISONOMIA
JORNADA
JORNALISTA
JUIZ CLASSISTA
JUIZ OU TRIBUNAL
JUROS
JUSTA CAUSA
JUSTIFICAÇÃO JUDICIAL
LICENÇA PATERNIDADE
LIQUIDAÇÃO EXTRAJUDICIAL
LITIGÂNCIA DE MÁ-FÉ
LITISCONSÓRCIO
MANDADO DE SEGURANÇA
MÃO-DE-OBRA
MARÍTIMO
MÉDICO E AFINS
MENOR
MINISTÉRIO DO TRABALHO E EMPREGO
MINISTÉRIO PÚBLICO
MULTA
NORMA COLETIVA (AÇÃO DE CUMPRIMENTO)
NORMA COLETIVA (EM GERAL)
NORMA JURÍDICA
NOTIFICAÇÃO E INTIMAÇÃO
NULIDADE MATERIAL
NULIDADE PROCESSUAL
PAGAMENTO
PARTE
PERÍCIA
PETIÇÃO INICIAL
PETROLEIRO
PIS-PASEP
PODER DISCIPLINAR
PORTUÁRIO

COMPENSAÇÃO
COMPETÊNCIA
CONCILIAÇÃO
CONCURSO DE CREDORES
CONFISSÃO FICTA
CONTESTAÇÃO
CONTRATO DE EQUIPE
CONTRATO DE EXPERIÊNCIA
CONTRATO DE TRABALHO (EM GERAL)
CONTRATO DE TRABALHO (PRAZO DETERMINADO OU OBRA CERTA)
CONTRATO DE TRABALHO (SUSPENSÃO E INTERRUPÇÃO)
CONTRIBUIÇÃO SINDICAL (LEGAL OU VOLUNTÁRIA)
COOPERATIVA
CORREÇÃO MONETÁRIA
CULPA RECÍPROCA
CUSTAS
DANO MORAL E MATERIAL
DECADÊNCIA
DÉCIMO TERCEIRO
DEFICIENTE FÍSICO
DEPOSITÁRIO INFIEL
DEPÓSITO RECURSAL
DESERÇÃO
DESPEDIMENTO INDIRETO
DIREITO ADQUIRIDO
DIRETOR DE S/A
DOCUMENTOS
DOMÉSTICO
EMBARGOS DE TERCEIRO
EMBARGOS DECLARATÓRIOS
EMBARGOS INFRINGENTES
EMPREGADOR
EMPRESA (CONSÓRCIO)
EMPRESA (SUCESSÃO)
ENGENHEIRO E AFINS
ENTIDADES ESTATAIS
EQUIPAMENTO
EQUIPARAÇÃO SALARIAL
ESTABILIDADE OU GARANTIA DE EMPREGO
ESTADO MEMBRO
EXCEÇÃO
EXECUÇÃO
FALÊNCIA
FALTA GRAVE
PRAZO
PREPOSTO JUDICIAL DO EMPREGADOR
PRESCRIÇÃO
PRESTAÇÃO DE SERVIÇOS
PREVIDÊNCIA SOCIAL
PROCESSO
PROCURADOR
PROFESSOR
PROFISSÃO
PROMOÇÃO
PROVA
QUADRO DE CARREIRA
QUITAÇÃO
RADIODIFUSÃO
RADIOTELEGRAFISTA
REAJUSTE SALARIAL GENÉRICO
RECLAMAÇÃO CORRECIONAL
RECONVENÇÃO
RECURSO
RECURSO DE EMBARGOS
RECURSO DE REVISTA (CABIMENTO)
RECURSO DE REVISTA (EM GERAL)
RECURSO EXTRAORDINÁRIO
RECURSO ORDINÁRIO
RELAÇÃO DE EMPREGO
REPOUSO SEMANAL REMUNERADO
REPRESENTAÇÃO OU ASSISTÊNCIA
RESCISÃO CONTRATUAL
RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA
REVELIA
RITO SUMARIÍSSIMO
RURAL
SALÁRIO (EM GERAL)
SALÁRIO MÍNIMO
SALÁRIO NORMATIVO E PISO SALARIAL
SALÁRIO PROFISSIONAL
SALÁRIO-FAMÍLIA
SALÁRIO-UTILIDADE
SEGURO DESEMPREGO
SENTENÇA OU ACÓRDÃO
SERVIDOR PÚBLICO (EM GERAL)
SERVIDOR PÚBLICO (RELAÇÃO DE EMPREGO)
SINDICÂNCIA
SINDICATO OU FEDERAÇÃO

8 Apêndice 2

Exemplo de arquivo texto com dados originais.

```
*** BRS DOCUMENT BOUNDARY ***
..TRIB:
2
..NRAC:
2008002851
..DTDE:
2008 11 27
..TPPR:
9
..NRPR:
20378
..ANPR:
2007
..VARA:
000
..NSEQ:
00
..NUUN:
20378-2007-000-02-00-3
..NULK:
20378-2007-000-02-00-3
..TURM:
SDC
ÓRGÃO JULGADOR - Secretaria de Dissídios Coletivos
..TPEX:
Dissídio Coletivo
..TOCP:
PROCESSO - TIPO: 9 NUM: 20378 ANO: 2007
..FONT:
DOE SP, PJ, TRT 2ª    Data: 09/12/2008    PG:
..PART:
SUSCITANTE(S):
SINDICATO DOS EMPREGADOS EM FISCALIZAÇÃO, INSPEÇÃO E CONTROLE
OPERACIONAL NAS EMPRESAS DE TRANSPORTE DE PASSAGEIROS E TRABALHADORES
NO SISTEMA DE VEÍCULOS LEVES SOBRE CANALETAS E PNEUS
NO ESTADO DE SÃO PAULO
SUSCITADO(S):
ETC - EMPRESA DE TRANSPORTE COLETIVO DE SÃO BERNARDO DO CAMPO
..TRAL:
ODETTE SILVEIRA MORAES
..TROL:
..TRAD:
..TROD:
..TREV:
RILMA APARECIDA HEMETÉRIO
..EMEN:
EMBARGOS DE DECLARAÇÃO - AUSÊNCIA DE RACIOCÍNIO LÓGICO -
LITIGÂNCIA DE MÁ-FÉ - O embargante lança, a esmo, meras
alegações, sem qualquer relação com o processado. Provoca
esta Corte através de petição manifestamente infundada, sem
qualquer critério em suas articulações, destoando, por
completo, da realidade dos autos. Litigância de má-fé
declarada de ofício.
..DECI:
por unanimidade de votos, rejeitar os presentes embargos de
declaração,opostos pelo Sindicato dos Empregados em Fiscalização,
Inspeção e Controle Operacional nas Empresas de Transportes
de Passageiros e Trabalhadores no Sistema de Veículos
Leves sobre Canaletas e Pneus no Estado de São Paulo - SINDFICOT
- VLP, e condenar o suscitante a pagar ao suscitado multa
por litigância de má-fé e multa por interposição de embargos
protelatórios, consoante fundamentação do voto. (V. ACÓRDÃO
EMBARGADO SDC Nº 0201/2008-5)
..INDE:
LITIGÂNCIA DE MÁ-FÉ, Geral
```

9 Apêndice 3

WEKA é um sistema desenvolvido pela Universidade de Waikato na Nova Zelândia. WEKA é um acrônimo de *Waikato Environment for Knowledge Analysis*. O sistema é escrito em JAVA, uma linguagem de programação orientada a objeto que é largamente disponibilizada para a maioria das plataformas, sendo o WEKA testado nos sistemas operacionais Linux, Windows e Macintosh. Existem vários meios de se utilizar o WEKA. Primeiramente ele provê implementações de algoritmos de aprendizado de máquina em seu estado-da-arte em que se pode aplicar em um conjunto de dados a partir da linha de comando. Uma maneira de utilizar o WEKA é aplicar um método de aprendizado a um conjunto de dados e analisar sua saída para extrair informações sobre os dados (WITTEN; FRANK, 2000).

Os métodos de aprendizado no WEKA são chamados de *Classifiers*, e as implementações para o pré-processamento dos dados são chamados de *Filters*. A ferramenta WEKA implementa os *Classifiers* SMO (SVM), NaiveBayes (Naive Bayes), J48 (Árvore de Decisão) que foram utilizados neste trabalho para realizar o processamento dos dados extraídos a partir do pré-processamento dos textos realizado através da ferramenta PRE-TEXT II.

Os algoritmos podem ser executados tanto por linhas de código em Java quanto por opções indicadas por meio de interfaces com o usuário. Essa é uma das principais vantagens do WEKA em relação a outros pacotes e bibliotecas para aprendizado de máquina e mineração de dados. Existem duas escolhas de interface: *GUI* ou *Simple SLI*, em que os usuários entram com linhas de comandos.

10 Apêndice 4

CATEGORIA	QUANTIDADE	CATEGORIA	QUANTIDADE
ASSÉDIO SEXUAL	1	SOLIDARIEDADE	11
IDOSO	1	AUDIÊNCIA OU SESSÃO DE JULGAMENTO	12
ISONOMIA	1	FACTUM PRINCIPIS	13
JUIZ CLASSISTA	1	FÉRIAS PROPORCIONAIS	13
LICENÇA PATERNIDADE	1	SALÁRIO FAMÍLIA	13
LITISCONSÓRCIO	1	TELEFONISTA	14
RURAL	1	CARTÓRIO	15
TAREFEIRO	1	CONTESTAÇÃO	15
TRABALHADOR AVULSO	1	RADIODIFUSÃO	15
ALIENAÇÃO FIDUCIARIA	2	EQUIPAMENTO	17
ATLETA PROFISSIONAL	2	VALOR DA CAUSA	19
EMBARGOS INFRIGENTES	2	AÇÃO DECLARATÓRIA	20
RECURSO DE REVISTA (CABIMENTO)	2	MENOR	20
RECURSO DE REVISTA (EM GERAL)	2	RECONVENÇÃO	20
RECURSOS DE EMBARGOS	2	COMISSIONAMENTO	21
CULPA RECÍPROCA	3	JORNALISTA	22
FALTAS AO SERVIÇO	3	PODER DISCIPLINAR	23
MINISTÉRIO PÚBLICO DO TABALHO	3	CORREIÇÃO PARCIAL	24
PETROLEIRO	3	MÉDICO E AFINS	24
ARTISTA	4	REAJUSTE SALARIAL GENÉRICO	24
CONCURSO DE CREDORES	4	SALÁRIO MÍNIMO	24
DÉCIMO TERCEIRO	4	AUTOS	26
NULIDADE MATERIAL	4	SÚMULAS DA JURISPRUDÊNCIA	27
INQUÉRITO JUDICIAL	5	ADVOGADO	30
MARÍTIMO	5	SUBSTITUIÇÃO	30
SINDICÂNCIA	5	EMPREGADOR	31
INCIDENTE DE FALSIDADE	6	SUCESSÃO "CAUSA MORTIS"	32
PROMOÇÃO	7	AÇÃO CIVIL PÚBLICA	34
REPRESENTAÇÃO OU ASSISTÊNCIA	7	AEROVIÁRIO	34
MINISTÉRIO PÚBLICO	8	FORÇA MAIOR	34
AUXÍLIO ENFERMIDADE	9	PAGAMENTO	34
BOLSISTA	9	DIRETOR DE S/A	35
ENGENHEIROS E AFINS	9	VIGIA E VIGILANTE	35
PIS/PASEP	9	SALÁRIO NORMATIVO E PISO SALARIAL	38
AÇÃO MONITÓRIA	10	QUADRO DE CARREIRA	40
ADICIONAL	10	SOCIEDADE DE ECONOMIA MISTA	43
DEFICIENTE FÍSICO	10	CHAMAMENTO AO PROESSO OU DENUNCIAÇÃO A LIDE	44

JUSTA CAUSA	678
SENTEÇA OU ACÓRDÃO	678
SALÁRIO (EM GERAL)	721
PORTUÁRIO	723
PROCESSO	726
EQUIPARAÇÃO SALARIAL	806
AÇÃO RECISÓRIA	818
AGRAVO DE INSTRUMENTO	821
NORMA COLETIVA (EM GERAL)	885
RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895
HORAS EXTRAS	1136
NULIDADE PROCESSUAL	1160
ESTABILIDADE OU GARANTIA DE EMPREGO	1184
APOSENTADORIA	1248
INSALUBRIDADE OU PERICULOSIDADE (EM GERAL)	1260
RECURSO	1297
MULTA	1321
CONCILIAÇÃO	1377
HONORÁRIOS	1559
ASSITÊNCIA JUDICIÁRIA	1564
MANDADO DE SEGURANÇA	1612
JORNADA	1979
SERVIDOR PÚBLICO (EM GERAL)	2020
SINDICATO OU FEDERAÇÃO	2094
COMPETÊNCIA	2151
DANO MORAL E MATERIAL	2532
PRESCRIÇÃO	2834
RELAÇÃO DE EMPREGO	2922
PROVA	3689
EMBARGOS DECLARATÓRIOS	4248
MÃO-DE-OBRA	4308
EXECUÇÃO	5370
PREVIDENCIA SOCIAL	12865
187 categorias	**91.616 documentos**

11 Apêndice 5

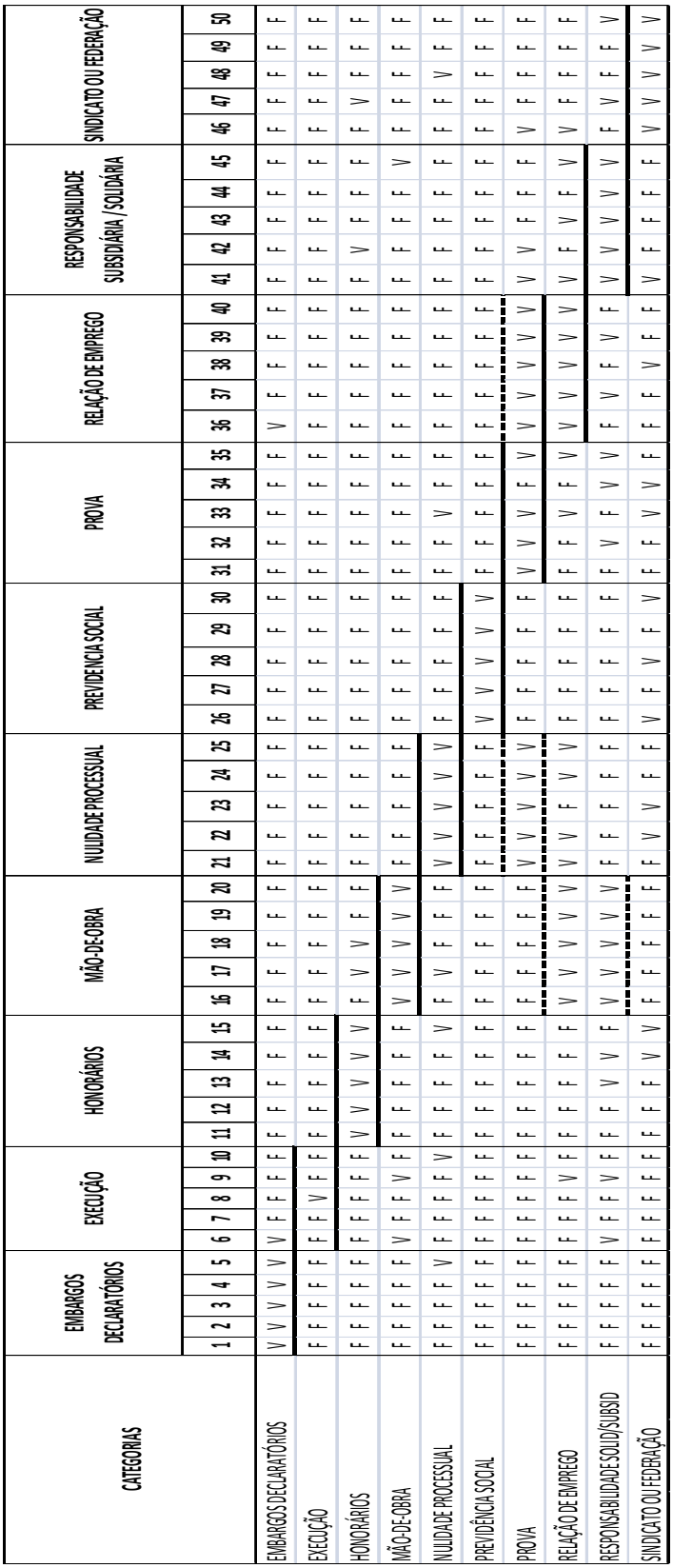
Tabela 9 - Categorias selecionadas para a pesquisa e quantidade de exemplos selecionados.

Categoria	Real[6]	Selec[7]	Outras	Real[4]	Selec[5]
EXECUÇÃO	5370	500	EMBARGOS DECLARATÓRIOS	4248	181
			RELAÇÃO DE EMPREGO	2922	125
			SINDICATO OU FEDERAÇÃO	2094	89
			MANDADO DE SEGURANÇA	1612	69
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	39
			Total de Outros	**11771**	**503**
PREVIDENCIA SOCIAL	12865	500	EMBARGOS DECLARATÓRIOS	4248	164
			DANO MORAL E MATERIAL	2532	98
			PROVA	3689	143
			MANDADO DE SEGURANÇA	1612	63
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	35
			Total de Outros	**12976**	**503**
MÃO-DE-OBRA	4308	500	PROVA	3689	212
			SINDICATO OU FEDERAÇÃO	2094	121
			RECURSO	1297	75
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	52
			PROCESSO	726	42
			Total de Outros	**8701**	**502**
EMBARGOS DECLARATÓRIOS	4248	500	PRESCRIÇÃO	2834	171
			SINDICATO OU FEDERAÇÃO	2094	126
			CONCILIAÇÃO	1377	83
			HORAS EXTRAS	1136	69
			NORMA COLETIVA (EM GERAL)	885	54
			Total de Outros	**8326**	**503**
PROVA	3689	500	COMPETÊNCIA	2151	183
			ASSITÊNCIA JUDICIÁRIA	1564	133
			HORAS EXTRAS	1136	97
			TEMPO DE SERVIÇO	592	51
			RESPONSABILIDADE	452	39
			Total de Outros	**5895**	**503**
RELAÇÃO DE EMPREGO	2922	500	PREVIDENCIA SOCIAL	12865	278
			EMBARGOS DECLARATÓRIOS	4248	92
			HONORÁRIOS	1559	34
			DANO MORAL E MATERIAL	2532	55
			JORNADA	1979	43

[6] Quantidade real de exemplos presentes dentro da categoria
[7] Quantidade de exemplos selecionados aleatoriamente

			Total de Outros	**23183**	**502**
SINDICATO OU FEDERAÇÃO	2094	500	RELAÇÃO DE EMPREGO	2922	133
			MÃO-DE-OBRA	4308	196
			MULTA	1321	61
			NORMA COLETIVA (EM GERAL)	885	41
			ASSITÊNCIA JUDICIÁRIA	1564	72
			Total de Outros	**11000**	**503**
HONORÁRIOS	1559	500	PROVA	3689	203
			COMPETÊNCIA	2151	119
			RECURSO	1297	72
			HORAS EXTRAS	1136	63
			AGRAVO DE INSTRUMENTO	821	46
			Total de Outros	**9094**	**503**
NULIDADE PROCESSUAL	1160	500	EXECUÇÃO	5370	196
			PROVA	3689	135
			COMPETÊNCIA	2151	79
			CONCILIAÇÃO	1377	51
			HORAS EXTRAS	1136	42
			Total de Outros	**13723**	**503**
RESPONSABILIDADE SOLIDÁRIA/ SUBSIDIÁRIA	895	500	EXECUÇÃO	5370	203
			PROVA	3689	139
			SERVIDOR PÚBLICO (EM GERAL)	2020	77
			MULTA	1321	50
			NORMA COLETIVA (EM GERAL)	885	34
			Total de Outros	**13285**	**503**

12 Apêndice 6



CATEGORIAS	EMBARGOS DECLARATÓRIOS					EXECUÇÃO					HONORÁRIOS					MÃO-DE-OBRA					NULIDADE PROCESSUAL					PREVIDENCIA SOCIAL					PROVA					RELAÇÃO DE EMPREGO					RESPONSABILIDADE SUBSIDIÁRIA /SOLIDÁRIA					SINDICATO OU FEDERAÇÃO				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
EMBARGOS DECLARATÓRIOS	V	V	V	V	V	V	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	V	F	F	F	F	F	F	F	F	F	F	F	F	F	F
EXECUÇÃO	F	F	F	F	F	F	F	V	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
HONORÁRIOS	F	F	F	F	F	F	F	F	F	F	V	V	V	V	V	F	V	V	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	V	F	F	F	F	F	F	F	F	F
MÃO-DE-OBRA	F	F	F	F	F	V	F	F	V	F	F	F	F	F	F	V	V	V	V	V	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	V	F	F	F	F	F
NULIDADE PROCESSUAL	F	F	F	F	V	F	F	F	F	V	F	F	F	F	V	F	V	F	F	F	V	V	V	V	V	F	F	F	F	F	F	F	V	F	F	F	F	F	F	F	F	F	F	F	F	F	V	F	F	F
PREVIDÊNCIA SOCIAL	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	V	V	V	V	V	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
PROVA	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	V	V	V	V	V	F	F	F	F	F	V	V	V	F	V	V	V	V	V	F	F	F	F	F	F	F	F	F	V	F
RELAÇÃO DE EMPREGO	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	V	V	V	V	V	F	V	V	F	V	F	F	F	F	F	F	V	F	F	V	V	V	V	V	V	V	F	V	V	F	F	F	F	F	F
RESPONSABILIDADE SOLID/SUBSID	F	F	F	F	F	V	F	F	V	F	F	F	F	V	F	V	V	V	V	V	F	F	F	F	F	F	F	F	F	F	F	V	F	F	F	V	F	V	V	V	V	V	V	V	F	V	F	F	V	F
SINDICATO OU FEDERAÇÃO	F	F	F	F	F	F	F	F	F	F	F	F	F	V	V	F	F	F	F	F	F	V	V	F	F	V	F	V	F	V	F	V	V	F	F	F	V	F	F	F	V	F	F	F	V	V	V	V	V	V

Figura 15: Tabela de Predição do Comitê Classificador

13 Apêndice 7

Pesquisa sobre Classificação das Ementas que compõem jurisprudência do TRT-SP 2ª. Região

Trabalho de Pesquisa para Dissertação de Mestrado Profissional em Gestão de TI Aplicada
Pós Graduação do Centro Paula Souza – Governo do Estado de São Paulo

Caro avaliador, por favor, leia os textos abaixo, e relacione-os às classes (títulos) da jurisprudência:

DOC 6 – 20110342946.txt

Procurador de sócio da empresa. Ausência de responsabilidade na execução. Aquele que possui procuração de sócio da empresa executada não responde pelas dívidas desta, figurando como mero representante do sujeito passivo da ação. Agravo de petição a que se dá provimento para acolhimento dos embargos de terceiro. Decisão por unanimidade de votos, acolher em parte as preliminares argüidas na contraminuta para determinar o desentranhamento das cópias de documentos de fls. 232/263 e, no mérito, por igual votação, DAR PROVIMENTO ao agravo de petição para excluir o agravante da execução e liberar a penhora que recaiu sobre suas contas bancárias, nos termos da fundamentação do voto da Relatora.

( ) EMBARGOS DECLARATÓRIOS ( ) EXECUÇÃO ( ) RESPONSABILIDADE

SUBSIDIÁRIA/SOLIDÁRIA ( ) OUTROS_____

DOC 7 - 20110343233.txt

Diferença entre juros bancários e juros trabalhistas. É devida a diferença pois o depósito foi feito sem a finalidade de quitar a execução. Decisão por unanimidade de votos, DAR PROVIMENTO ao recurso interposto pela agravante, a fim de dar prosseguimento à execução para pagamento das diferenças de juros existentes entre a data do depósito e a data do respectivo levantamento, conforme fundamentação constante do voto da Relatora.

( ) EXECUÇÃO ( ) HONORÁRIOS ( ) OUTROS

Doc 8 – 20110354790.txt

FRAUDE À EXECUÇÃO. CONFIGURAÇÃO. Resta configurada, no caso em epígrafe, a ocorrência de fraude à execução, nos exatos termos do artigo 593, II do CPC, aplicado subsidiariamente ao processo do trabalho (art. 769/CLT), traduzida no reconhecimento de firma no Instrumento Particular de Compromisso de Compra e Venda somente dois anos após a suposta transação imobiliária e 1 mês depois de distribuída a reclamação trabalhista, revelando que a venda somente se operou após o ajuizamento da ação, com o intuito de afastar o imóvel em questão da constrição que insofismavelmente lhe seria imposta. Decisão por unanimidade de votos, REJEITAR a preliminar arguida em contraminuta pelo agravado; no mérito, por igual votação, NEGAR PROVIMENTO ao agravo de petição interposto, tudo nos termos da fundamentação do voto da relatora.

( ) EXECUÇÃO ( ) NULI DADE PROCESSUAL ( ) OUTROS:

DOC 9 - 20110371750.txt

Sócio. Fase de conhecimento. Pólo passivo. Legitimação. Não mais se exige a inclusão dos sócios na fase de conhecimento. Entendimento adotado no Tribunal Superior do Trabalho, com o cancelamento da Súmula 205. A responsabilização do sócio não depende de declaração prévia, decorre da lei. Assim, caberá ao juiz, em caso de insuficiência de bens das empresas responsáveis, determinar a execução dos sócios, nos termos do art. 592, II, e 596 do Código de Processo Civil. Recurso do autor a que se nega provimento. Decisão por unanimidade de votos, NEGAR PROVIMENTO a ambos os recursos.

( ) EXECUÇÃO ( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) OUTROS:

DOC 10 - 20110374716.txt

AÇÃO ANULATÓRIA DE ARREMATAÇÃO. DECADÊNCIA. O prazo para a interposição da ação de anulação de arrematação é de dois anos contados do ato judicial que, no caso da arrematação, ocorreu com a assinatura do auto pelo Juiz. Transcorrido tal prazo, configura-se a decadência da ação. Recurso a que se nega provimento. Decisão por unanimidade de votos, negar provimento ao apelo.

( ) EXECUÇÃO ( ) NULI DADE PROCESSUAL ( ) OUTROS:

DOC 14 - 20110357595.txt

Honorários advocatícios. Justiça do Trabalho. Cabimento. Os princípios do acesso á justiça da ampla defesa e do contraditório (artigo 5º, incisos XXXV e LV da Constituição Federal) pressupõem a defesa técnica do trabalhador, por profissional qualificado, não sendo possível restringir o direito do hipossuficiente, em optar pela nomeação de advogado particular, nos termos do art. 133 da Carta Magna. Em que pese a inaplicabilidade do princípio da sucumbência e a possibilidade do "jus postulandi" no Processo do Trabalho, a condenação em honorários advocatícios tem amparo no princípio da restituição integral, expresso nos artigos 389, 404 e 944 do Código Civil. Além disso, a Lei 10.288/2001 revogou o art.14 da Lei 5.584/70, não havendo óbice legal para a condenação em honorários advocatícios, nos casos em que o reclamante não estiver assistido pelo sindicato, nos termos da Lei 10.537/2002, que acrescentou o parágrafo 3º ao art. 790 da CLT. Decisão por maioria de votos, vencida parcialmente a Exmª Srª Desembargadora Ivani Contini Bramante, DAR PROVIMENTO PARCIAL ao recurso ordinário oposto pela Reclamada para expungir da condenação a devolução das contribuições assistenciais (sindicais) e, por igual votação, vencido parcialmente o Exmº Sr. Juiz Paulo Sérgio Jakútis, DAR PROVIMENTO PARCIAL ao apelo adesivo do Reclamante para condenar a ré ao pagamento de honorários advocatícios, no importe de 15% sobre o valor da condenação. Cumpre ressaltar que os honorários ora deferidos serãodirecionados ao reclamante, e não aos seus patronos, pois visam ressarcir as despesas ocorridas com o advogado particular. Custas inalteradas.

( ) HONORÁRIOS ( ) SINDICATO OU FEDERAÇÃO ( ) OUTROS:

DOC 16 – 20110332576.txt

TERCEIRIZAÇÃO. INADIMPLEMENTO DE OBRIGAÇÕES LEGAIS. RESPONSABILIDADE SUBSIDIÁRIA DA TOMADORA. O provimento de mão-de-obra através de empresa terceirizada que vem a revelar-se inidônea, torna a tomadora subsidiariamente responsável pelas obrigações legais inadimplidas pela agenciadora de pessoal. Incidência da Súmula 331, do C. TST. Decisão por unanimidade de votos,rejeitar a preliminar de ilegitimidade de parte; no mérito, por igual votação, DAR PARCIAL PROVIMENTO ao recurso ordinário interposto pela segunda reclamada., para restringir a condenação em horas extras, considerando a jornada das 08h00 às 20h00, de segunda a sextafeira,com exclusão dos domingos e feriados, consoante fundamentação do voto do Relator, mantendo, no mais, na íntegra a respeitável decisão de origem, inclusive quanto ao valor da condenação e das custas processuais.

( ) MÃO-DE-OBRA ( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO

( ) OUTROS: _____

DOC 17 - 20110332584.txt

TERCEIRIZAÇÃO. INADIMPLEMENTO DE OBRIGAÇÕES LEGAIS. RESPONSABILIDADE SUBSIDIÁRIA DA TOMADORA. O provimento de mão-de-obra através de empresa terceirizada que vem a revelar-se inidônea, torna a tomadora subsidiariamente responsável pelas obrigações legais inadimplidas pela agenciadora de pessoal. Incidência da Súmula 331, do C. TST. Decisão por unanimidade de votos, rejeitar a preliminar de nulidade por negativa de prestação jurisdicional e, no mérito, por igual votação, DAR parcial PROVIMENTO ao recurso ordinário interposto pela segunda reclamada, para excluir da condenação a multa estabelecida nos embargos declaratórios, tudo nos termos da fundamentação do voto do Relator, mantendo, no mais,íntegra a r. Decisão de origem, inclusive quanto ao valo da condenação e das custas processuais.

( ) MÃO-DE-OBRA ( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO

( ) OUTROS: _____

DOC 18 – 20110338019.txt

TOMADOR DE SERVIÇOS. RESPONSABILIDADE SUBSIDIÁRIA. O reconhecimento da responsabilidade subsidiária faz com que o tomador de serviços se torne responsável pelo adimplemento de todas as verbas da condenação, inclusive quanto às multas dos arts. 467 e/ou 477 da CLT, bem como pelo pagamento da multa fundiária e recolhimentos previdenciários e fiscais. AVISO PRÉVIO. NÃO COMPROVAÇÃO PELA RECLAMADA QUE CONCEDEU A REDUÇÃO DA JORNADA PREVISTA NA LEI. A inobservância da redução de que trata o artigo 488 da CLT desvirtua a finalidade de propiciar ao empregado a busca de nova colocação no mercado de trabalho e autoriza a condenação do empregador ao pagamento de novo período de aviso prévio. Decisão por unanimidade de votos, REJEITAR a preliminar arguida e, no mérito, por maioria de votos, vencido parcialmente o Exmº Sr. Desembargador Sérgio Winnik, DAR PARCIAL PROVIMENTO AO Recurso Ordinário interposto, tudo nos termos da fundamentação do voto da Relatora.

( ) MÃO-DE-OBRA ( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO

( ) OUTROS: _____

DOC 19 – 20110343004.txt

Responsabilidade subsidiária. Contrato de prestação de serviços. A empresa tomadora de serviços, ao contratar empresa prestadora, tem obrigação de diligenciar se esta cumpre a legislação trabalhista, eis que se beneficia diretamente da força de trabalho do empregado que lhe prestaserviços. Portanto, havendo inadimplemento do empregador, a tomadora de serviços responde de forma subsidiária perante o trabalhador, com fundamento jurídico nos artigos 927 e 186 do Código Civil. Decisão por unanimidade de votos, NEGAR PROVIMENTO ao apelo para manter na íntegra a r. sentença de primeiro grau, conforme fundamentação constante do voto da Relatora.

( ) MÃO-DE-OBRA ( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO

( ) OUTROS: _____

DOC 20 - 20110356840.txt

TERCEIRIZAÇÃO. PODER PÚBLICO COMO TOMADOR DOS SERVIÇOS. DESRESPEITO A DIREITOS TRABALHISTAS PELA CONTRATADA. RESPONSABILIZAÇÃO DEVIDA. Quando o Poder Público celebra contratos administrativos tem o dever-poder de fiscalização (Lei n.º 8.666, artigos 58, III e 67). (De modo que, f)Ficando inerte frente ao desrespeito aos direitos trabalhistas, com prejuízo aos trabalhadores, há de responder subsidiariamente pelas conseqüências da ilegalidade perpetrada por culpa in eligendo e in vigilando. Recurso ordinário a que se dá parcial provimento. Decisão Por unanimidade de votos, DAR PROVIMENTO PARCIAL ao recurso ordinário interposto para: 1. autorizar a dedução de valores comprovadamente pagos sob igual título; 2. que a correção monetária se dê nos termos da Súmula nº 381 do C. TST; 3. isentar a União do pagamento das custas processuais, mantendo,no mais, a r. sentença nos termos da fundamentação.

( ) MÃO-DE-OBRA ( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO

( ) OUTROS: _____

DOC 21 – 20110332819.txt

DO CERCEAMENTO DE DEFESA. O juiz, a quem incumbe a direção do processo, pode indeferir provas desnecessárias para o deslinde da causa, não caracterizando tal medida cerceamento de defesa, nos termos do art. 765 da CLT c/c art. 130 do CPC. Decisão por unanimidade de votos, NEGAR PROVIMENTO AO RECURSO ORDINÁRIO da reclamada, TRUFER COMERCIO DE SUCATAS LTDA, para manter incólume a r. sentença de origem.

( ) NULI DADE PROCESSUAL ( ) PROVA ( ) RELAÇÃO DE EMPREGO ( ) OUTRO:

DOC 25 - 20110360464.txt

CERCEAMENTO DE DEFESA. Formado o convencimento do magistrado com base nas provas já carreadas, o indeferimento de novas provas não constitui cerceamento de defesa. Preliminar rejeitada. Recurso do Reclamante a que se nega provimento. Decisão Por unanimidade de votos, I - REJEITAR a preliminar arguida; II - NEGAR PROVIMENTO ao recurso do reclamante, nos termos do fundamentado.

( ) NULIDADE PROCESSUAL ( ) PROVA ( ) OUTRO:

DOC 30 - 20110332070.txt

Contribuições Previdenciárias. Fato gerador. Antes da sentença, o direito discutido na ação se traduz em res dubia. É a sentença que constitui o fato gerador das contribuições previdenciárias, pois é ela que reconhece e certifica o direito. Indevida, neste caso, a atualização das contribuições previdenciárias a partir da prestação dos serviços. O art. 276 do Decreto nº 3.048/99 determina que no caso de pagamento de verbas trabalhistas, de natureza salarial, reconhecidas em sentença, o recolhimento da contribuição previdenciária deve ser feito "no dia dois do mês seguinte ao da liquidação de sentença". Decisão por unanimidade de votos, NEGAR PROVIMENTO AO AGRAVO DE PETIÇÃO, mantendo incólume a decisão de fls. 463/464.

( ) PREVIDÊN CIA SOCIAL ( ) SIN DICATO OU FEDERAÇÃO ( ) OUTROS:

DOC 41 - 20110268541.txt

QUARTEIRIZAÇÃO - O fenômeno não teve outro objeto senão fraudar direitos trabalhistas, sendo nulo de pleno direito nos termos do artigo 9º da CLT. Mormente, diante da manifesta intenção de se estabelecer uma cadeia de diversas pessoas intermediárias, de molde a distanciar-se da real responsabilidade da reclamada, o reconhecimento da fraude é de rigor. Decisão Conhecer dos recursos e, no mérito, por maioria de votos, dar provimento parcial ao recurso da reclamada SÃO PAULO TRANSPORTE S/A, extinguir feito sem resolução do mérito em relação a esta, vencido o voto da Exma. Des. Wilma Gomes da Silva Hernandes, que entendia ser matéria de mérito, julgando a ação improcedente. Por unanimidade de votos, negar provimento aos demais recursos das reclamadas e do reclamante, mantendo os demais tópicos da r. sentença de primeiro grau pelos seus próprios e jurídicos fundamentos. Fica mantido o valor da condenação para efeitos de alçada.

( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO ( ) OUTRO:

DOC 43 - 20110342504.txt

RESPONSABILIDADE SUBSIDIÁRIA. DONO DA OBRA. ORIENTAÇÃO JURISPRUDENCIAL Nº 191 DA SDI-I DO C. TST. Diante da comprovação de que a terceira reclamada contratou a segunda para execução de obra certa, desvinculada de sua atividade fim, não se pode falar em responsabilidade subsidiária pelos créditos trabalhistas, eis que não se trata da hipótese de terceirização de serviços. Por não ostentar a condição de tomadora de serviços, mas sim de dona da obra, aplica-se a hipótese o disposto na Orientação Jurisprudencial nº 191 da SDI-I do C. TST, o que impede a aplicação da Súmula nº 331 da mesma Corte. Responsabilidade Subsidiária que resta afastada. Recurso a que se dá provimento. Decisão por unanimidade de votos, DAR PROVIMENTO ao recurso da terceira reclamada, para declarar sua condição de dona da obra e afastar a responsabilidade subsidiária pelos créditos trabalhistas, nos termos da fundamentação do voto da Relatora. Custas inalteradas.

( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO ( ) OUTRO:

DOC 45 - 20110358915.txt

ADMINISTRAÇÃO PÚBLICA DIRETA. RESPONSABILIDADE SUBSIDIÁRIA CONFIGURADA. APLICAÇÃO DA SÚMULA 331, IV, TST. Tendo em vista que a ré, Fazenda Pública, beneficiou-se dos serviços

prestados pelo autor, deve responder pelos riscos da terceirização da mão-de-obra, nos termos da súmula 331, IV, TST. Decisão por unanimidade de votos, DAR PROVIMENTO ao recurso, a fim de que a Fazenda Pública do Estado de São Paulo seja responsabilizada subsidiariamentena hipótese de inadimplemento das obrigações trabalhistas por parte do empregador, nos termos do voto desta Relatora.

( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) MÃO-DE-OBRA ( ) OUTRO:

DOC 46 - 20110268045.txt

Contribuições assistencial. Devolução. Ausência de prova da filiação ou autorização para o desconto. Das contribuições sindicais elencadas no nosso sistema, o trabalhador só está mesmo obrigado àquela de que tratam os artigos 580 e 582 da CLT. Para as demais, como a assistencial e a confederativa, é imperiosa a expressa concordância do empregado. Hipótese em que a concordância não foi demonstrada. Recurso da corré a que se nega provimento, nesse ponto. Decisão por maioria de votos, DAR PROVIMENTO EM PARTE a ambos os recursos, para excluir da condenação as horas extras decorrentes do trabalho em feriados e a multa do art. 477 da CLT, vencido o voto do Exmo. Juiz Antero Arantes Martins, que afastava acondenação na devolução de descontos assistenciais e deferia à Fazenda juros especiais de que trata o art. 1º-F daLei 9.494/97 a partir de 30/06/2009. Custas inLeieradas.

( ) RESPONSABILI DADE SUBSIDIÁRIA/SOLIDÁRIA ( ) RELAÇÃO DE EMPREGO ( ) OUTRO:

DOC 47 - 20110342407.txt

ENQUADRAMENTO SINDICAL. ARTIGO 511 DA CLT. Nos termos do artigo 511 da CLT o enquadramento sindical dá-se pela atividade preponderante da empresa. Assim, comprovado nos autos que a empresa recorrida recolhe as contribuições assistenciais e sindicais a sindicato diverso do recorrente e legalmente constituído, não há se falar na procedência do pedido formulado pelo sindicato que não figura como legítimo representante dos empregados da recorrida. Recurso ordinário a que se nega provimento. Decisão por unanimidade de votos, NEGAR PROVIMENTO ao recurso do reclamante e, por igual votação,DAR PROVIMENTO ao recurso da reclamada, para condenar o autor a pagar os honorários advocatícios, fixados em 10% sobre o valor da causa, nos termos da fundamentação do voto da Relatora.

( ) SINDICATO OU FEDERAÇÃO ( ) HONORÁRIOS ( ) OUTROS:

IDENTIFCAÇÃO

Nome do(a) Especialista Avaliador(a): _____

Assinatura do(a) Especialista Avaliador(a): _____