repec

Revista de Educação e Pesquisa em Contabilidade

www.repec.org.br

v. 4, n. 2, art. 2, p. 23-42
ISSN 1981-8610
mai/ago. 2010.

CLASSIFICAÇÃO DO CONTEÚDO DE DOCUMENTOS CONTÁBEIS USANDO APRENDIZAGEM DE MÁQUINA: O CASO DOS FATOS RELEVANTES

CONTENT CLASSIFICATION OF ACCOUNTING DOCUMENTS USING MACHINE LEARNING: THE RELEVANT FACTS CASE

CLASIFICACIÓN DEL CONTENIDO DE DOCUMENTOS CONTABLES USANDO APRENDIZAJE DE MÁQUINA: EL CASO DE LOS HECHOS RELEVANTES

BRUNNA HISLA DA SILVA SENA
Bacharel em Ciências Contábeis pela UnB, Mestre em Ciências Contábeis pela UnB
brunnahisla@gmail.com

CÉSAR AUGUSTO TIBÚRCIO SILVA
Contador pela Unieuro, Mestre em Administração pela UnB, Doutor em Contabilidade pela USP, Professor Titular da UnB, Professor do Programa Multiinstitucional e Inter-Regional de Pós-Graduação em Ciências Contábeis da Universidade de Brasília (UnB), Universidade Federal da Paraíba (UFPB), Universidade Federal de Pernambuco (UFPE) e Universidade Federal do Rio Grande do Norte (UFRN)
cesartiburcio@unb.br

ROBERTO TERNES ARRIAL
Bacharel em Ciências Contábeis pela UnB, Bacharel em Ciências Biológicas
rtarrial@gmail.com

RESUMO

A análise de conteúdo de textos narrativos tem sido estudada, nos últimos anos, com mais frequência. Em diversos trabalhos, verifica-se a pesquisa com relação a sua legibilidade, compreensibilidade e o nível de otimismo, neutralidade e pessimismo. Porém, a análise de classificação quanto a tendências otimistas, pessimistas e neutras tem

sido feita de forma muito trabalhosa, pois demanda uma análise humana dos textos, justificando a criação de uma análise de textos de forma mais rápida e objetiva, além da tentativa de eliminação da subjetividade. Diante disso, o objetivo deste trabalho é propor uma classificação automática de fatos relevantes contábeis, fazendo-se uma análise do conteúdo de textos narrativos, com a utilização de ferramentas computacionais de leitura e classificação de textos. A ideia é procurar contribuir com um exemplo de aplicação de aprendizado de máquina à Ciência Contábil. Esta análise utilizou-se de fatos relevantes já analisados anteriormente no trabalho de Pereira e Silva (2008). Os fatos já classificados foram utilizados como conjunto de treinamento para o programa, para que assim ele pudesse classificar outros dados desconhecidos, não classificados.

PALAVRAS-CHAVE: *Aprendizado de máquina. Aprendizado bayesiano. Fatos relevantes. Análise de conteúdo.*

ABSTRACT

The analysis of narrative texts content has been more often studied in recent years. In several works research is noticed in relation to readability, comprehensiveness and level of optimism, pessimism or neutrality. However, the classification analysis regarding their optimistic, pessimistic or neutral trends has been proven burdensome, because it demands human analysis of texts, justifying the creation of more rapid and objective text analysis procedures, besides the attempt to reduce subjectivity. Therefore, the objective of this work is to propose an automatic classification of the accounting relevant facts , by making an analysis of narrative texts content using computational tools for text reading and classification. The idea is to try to contribute with an example of machine learning application to Accounting Science. The analysis in this work used relevant facts previously analyzed in the study by Pereira and Silva (2008). The already classified facts were used as training set for the program, so that it could classify other unknown and not-classified data.

Keywords: *Machine learning. Bayesian learning. Relevant facts. Content analysis.*

RESUMEN

El análisis de contenido de textos narrativos ha sido estudiado, en los últimos años, con más frecuencia. En diversos trabajos, se verifica la pesquisa con relación a su legibilidad, comprensibilidad y el nivel de optimismo, neutralidad y pesimismo. Pero, el análisis de clasificación en lo referente a tendencias optimistas, pesimistas y neutras ha sido efectuado de forma muy laboriosa, pues demanda un análisis humano de los textos, justificando la

creación de un análisis de textos de forma más rápida y objetiva, además de la tentativa de eliminación de la subjetividad. Delante de eso, el objetivo de este trabajo es proponer una clasificación automática de hechos relevantes contables, haciéndose un análisis del contenido de textos narrativos, con la utilización de herramientas computacionales de lectura y clasificación de textos. La idea es procurar contribuir con un ejemplo de aplicación de aprendizaje de máquina a la Ciencia Contable. Este análisis utilizó hechos relevantes ya analizados anteriormente en el trabajo de Pereira e Silva (2008). Los hechos ya clasificados fueron utilizados como conjunto de entrenamiento para el programa, para que así éste pudiese clasificar otros datos desconocidos, no clasificados.

Palabras clave: *Aprendizaje de máquina. Aprendizaje bayesiano. Hechos relevantes. Análisis de contenido.*

1. INTRODUÇÃO

O método de aprendizagem de máquina é amplamente referenciado na literatura internacional, muitas vezes na área de Finanças, para classificar, por exemplo, se uma empresa está propícia à falência, ou não, além de inúmeras aplicações em diversas áreas do conhecimento. Seu uso estende-se a áreas, como, por exemplo, Biologia, Filosofia, Medicina, Finanças, Telecomunicações, Marketing e Análise de conteúdo na internet (BOSE e MAHAPATRA, 2001).

Uma aplicação possível na área contábil é a classificação de textos narrativos, os quais já vêm sendo cada vez mais estudados, como pode ser visto em trabalhos que classificam quanto à sua legibilidade, compreensibilidade e nível de otimismo, neutralidade e pessimismo, conforme estudos de Rodrigues (2005), Fernandes e Silva (2007) e Pereira e Silva (2008). O posicionamento de um texto possui uma relevância tal que ele, por si só, pode influenciar decisões e julgamentos de leitores. Por exemplo, Devitt e Ahmad (2007) relatam que a polaridade de sentimentos em textos contábeis pode vir a influenciar a decisão de investidores no mercado de ações. Por isso, mais do que uma análise quantitativa e informacional, é importante identificar nesses textos, também, aspectos emocionais implícitos. Além disso, a quantidade de textos disponíveis para análise vem aumentando diante da facilidade de acesso trazida pela disseminação da internet. Apesar da grande utilidade que a classificação desses textos possui, verifica-se uma grande desvantagem na utilização de um método não automatizado para análise, já que é necessária a disponibilidade de pessoas para classificar em cada nível, que, além de demandar tempo, pode ser subjetivo.

O objetivo deste trabalho é propor uma classificação automática dos textos e fazer uma análise do conteúdo de textos narrativos, utilizando-se ferramentas computacionais e

textos já classificados por pesquisadores, com o intuito de contribuir com um exemplo de aplicação de aprendizado de máquina à Ciência Contábil.

Este artigo está dividido em seis seções. A seção dois trata do referencial teórico, com a abordagem de alguns conceitos introdutórios sobre aprendizagem de máquina, aprendizado bayesiano, textos narrativos e fatos relevantes; a terceira seção trata da metodologia adotada neste trabalho; a quarta seção, por sua vez, mostra os resultados e a sua análise; na quinta seção, estão as conclusões do trabalho; e, finalmente, a sexta traz as referências bibliográficas.

2. REFERENCIAL TEÓRICO

2.1. Classificação automática de textos

A classificação automática de textos começou a ser amplamente aplicada a partir dos anos 90 devido, principalmente, ao desenvolvimento de máquinas mais potentes e à maior facilidade de publicação de textos em forma eletrônica. A classificação de textos é considerada uma associação de textos em linguagem natural a rótulos predefinidos. É uma área que engloba conceitos de extração de informação e de aprendizado de máquina, podendo ser aplicada em uma grande variedade de contextos, como, por exemplo, indexação automática de textos, identificação de autores de textos, filtragem de e-mails, classificação hierárquica de páginas da internet e geração automática de métodos (MARON, 1961; MOSTELLER e WALLACE, 1964; GRAHAM, 2002; MCCALLUM et al., 1998; GILES et al., 2003, apud STEINBRUCH, 2006).

2.2. Aprendizagem de máquina

Aprendizagem de máquina é uma área de Inteligência Artificial que tem por objetivo desenvolver técnicas computacionais sobre o aprendizado e construir sistemas capazes de adquirir conhecimento de forma automática. Programas de aprendizagem de máquina são entendidos como programas computacionais que se utilizam de dados existentes para melhorar seu desempenho em uma tarefa de classificação (MITCHELL, 1997).

A aprendizagem de máquina é utilizada em áreas como Finanças, Telecomunicações, Marketing, Análise de documentos de internet. Nessas áreas, podem-se encontrar aplicações, como, por exemplo, predição de falência, de preço de ações e de taxa de juros; aprovação de empréstimo; administração de risco; detecção de fraudes; análise de segmentação de mercado; análise de desempenho de produtos; estimativa de similaridade de padrões de navegação do usuário; estimativas de litígio e de custo de software; reconhecimento de imagens; reconhecimento de genes em uma sequência de DNA; e previsão do movimento da bolsa de valores (BOSE e MAHAPATRA, 2001; STEINBRUTCH, 2006; SOUTO et al., 2003).

As técnicas de aprendizagem de máquina criam um classificador por meio de um processo indutivo de aprendizado. Na tarefa de análise de documentos, este classificador é

criado baseado em um conjunto de relações entre documentos e rótulos associados. Assim, depois de criado o classificador, o algoritmo classifica um documento ainda não conhecido em uma das categorias aprendidas na fase de treinamento, sendo capaz de tomar decisões baseado em experiências acumuladas por meio da solução bem sucedida de problemas anteriores (STEINBRUCH, 2006; MITCHELL, 1997).

Estudos, como o de Steinbruch (2006), propõem classificação automática de textos baseados no algoritmo multinomial *naïve* Bayes, no qual o autor faz uma aplicação em um ambiente *on-line* de classificação automática de notícias, combinando técnicas de aprendizagem de máquina e mineração de textos.

2.3. Aprendizado bayesiano

Segundo Witten e Frank (2005), existem diversos tipos de classificadores em aprendizagem de máquina, entre eles o classificador *naïve* Bayes. Este classificador presume que exista independência entre as palavras de um texto. Embora essa independência seja dificilmente encontrada empiricamente, o que poderia até invalidar aplicações empíricas do programa, mostrou-se teoricamente que essa suposição de independência de palavras, na maioria dos casos, não prejudica a eficiência do classificador (DOMINGOS e PAZZANI, 1997, apud STEINBRUCH, 2006).

Basicamente, existem dois tipos de modelos estatísticos para os classificadores *naïve* Bayes: o modelo binário e o modelo multinomial.

O modelo binário representa um documento por meio de um vetor binário, no qual um valor 0 na posição *k* significa que o documento não possui nenhuma ocorrência de determinado termo, e um valor 1 significa que o documento possui pelo menos uma ocorrência do termo (ou seja, um operador booleano). Por exemplo, se usarmos o conjunto de termos: {“balanço”, “lucros”, “empresa”} na análise da frase: “A adoção de práticas de governança corporativa aumentou os lucros da empresa em comparação aos lucros do mês anterior”, essa frase pelo modelo binário seria representada pelo seguinte vetor binário: {0; 1; 1}. (STEINBRUCH, 2006)

Já o modelo multinomial representa um documento através de um vetor de frequências, no qual um peso representa a frequência de um termo nos documentos analisados. Usando os dados do exemplo acima, e sendo feita sobre eles normalização da frequência de ocorrências dos termos pelo total de palavras da frase, o vetor seria expresso da seguinte forma: {0; 0,11; 0,05} (STEINBRUCH, 2006).

McCallum e Nigam (1998) apud Steinbruch (2006) compararam o modelo binário com o multinomial por meio de experimentos, concluindo que o modelo multinomial apresenta melhores resultados.

2.4. Conjuntos de treinamento e teste

A avaliação do desempenho de um classificador induzido por aprendizagem de máquina é etapa essencial na determinação do desempenho e, portanto, da utilidade do mo-

delo que foi induzido, além de permitir sua comparação a outros modelos e classificadores (WITTEN e FRANK, 2005).

Como procedimento mais comum de cálculo de desempenho, nas tarefas de aprendizagem de máquina, comumente, o conjunto de treinamento é seccionado em dois conjuntos disjuntos: o de treinamento e o de teste. O subconjunto de treinamento deve ser usado com o exclusivo propósito de treinar o classificador. A seguir, o modelo induzido deve ser testado pelo subconjunto de teste, que não foi usado durante a indução do modelo. Os subconjuntos são disjuntos para assegurar que os resultados experimentais obtidos, por meio do conjunto de validação, sejam de um conjunto diferente do usado para realizar o aprendizado, tornando os resultados estatisticamente válidos (WITTEN e FRANK, 2005).

O conjunto de teste é composto por exemplos que tanto o algoritmo de aprendizado de máquina quanto o pesquisador desconhecem os rótulos de saída. Ou seja, a classificação do conjunto de teste é o objetivo principal de se induzir um modelo de aprendizado (STEINBRUCH, 2006; SOUTO et al., 2003).

Nem sempre é possível seccionar o conjunto de treinamento para gerar e testar um modelo de aprendizado. Em alguns casos, a quantidade de exemplos é tão pequena que, ao excluir do subconjunto de treinamento os exemplos que vão compor o subconjunto de validação,

os dados ficam muito escassos e o modelo induzido possui uma capacidade de generalização muito baixa, ou seja, ele é incapaz de classificar os dados do subconjunto de validação e do conjunto de teste, caracterizando o fenômeno de *underfitting* (WITTEN e FRANK, 2005).

Para esses casos em que se dispõe de um conjunto de treinamento muito pequeno, diversas abordagens alternativas podem ser usadas, que possuem boa equivalência ao método tradicional, como, por exemplo: resubstituição, *holdout*, amostragem aleatória, validação cruzada de *k* vezes, leave-one-out e *bootstrap*. Um dos métodos mais populares é a validação cruzada de *k* vezes, que é o método usado nesse trabalho e é descrito a seguir.

Segundo Arrial (2008, p. 42), a validação cruzada funciona como a seguir:

> inicialmente, o conjunto de treinamento é fracionado em k vezes (ou subpartes) com quantidades de exemplos os mais similares possíveis e com uma distribuição homogênea de classes. O parâmetro k pode ser escolhido pelo usuário, no entanto k=10 é um valor tido como ideal tanto por aproximações teóricas como empíricas (WITTEN e FRANK, 2005). O algoritmo consiste em uma quantidade de repetições (iterações) igual ao número k selecionado pelo usuário. A cada iteração, um subconjunto diferente é tomado como único conjunto de treinamento, o qual gera um modelo que é usado para realizar predições sobre o conjunto de teste (que é composto pelos k-1 subconjuntos restantes). Ao final do processo, cada subconjunto terá sido

usado como conjunto de treinamento uma vez, e terá sido testado k-1 vezes pelos demais subconjuntos (Figura 1). Como a cada iteração o modelo criado fornece uma predição para instâncias cujos rótulos (classes) já são previamente conhecidos, pode-se fazer ao final do processo uma avaliação da acurácia desse modelo.

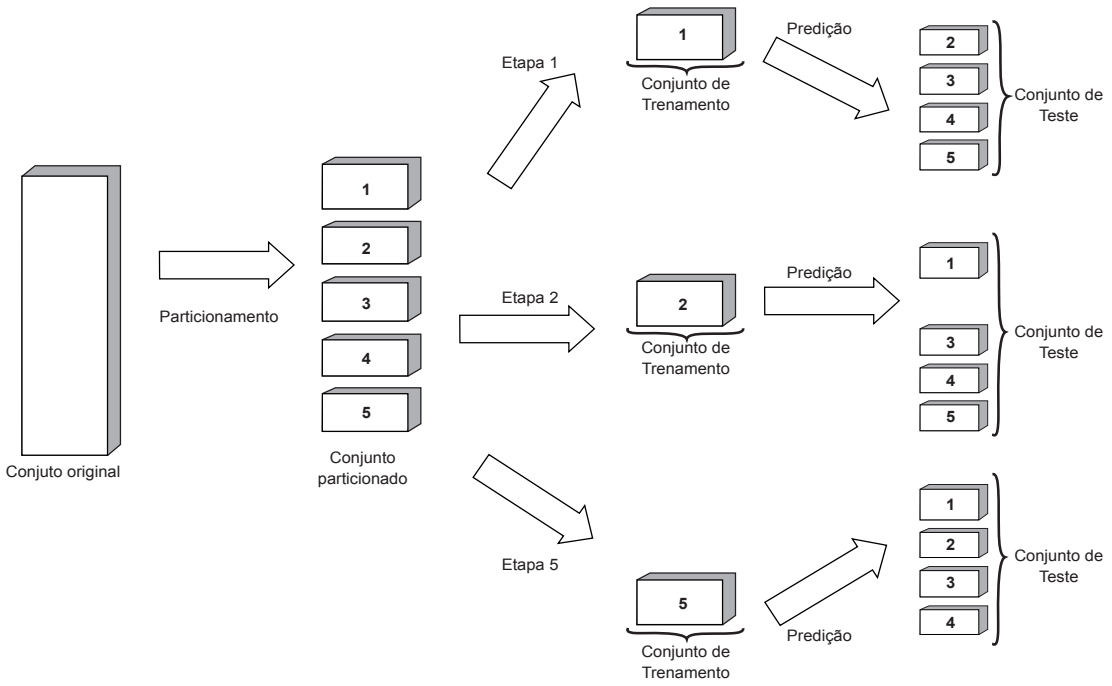

Figura 1: Esquema ilustrativo do processo de validação cruzada para k=5 (5 vezes).
Fonte: Arrial (2008, p. 43)

Conforme ilustrado na Figura 1, após o particionamento do conjunto original, são feitas cinco iterações e a cada iteração um subconjunto diferente é tomado como conjunto de treinamento (em cinza) e os demais k-1 subconjuntos são tomados como teste (em branco). A cada iteração, a acurácia do modelo é estimada de acordo com a contabilização dos acertos e erros da predição. Ao final de todas as iterações, a acurácia global do modelo pode ser estimada fazendo a média aritmética das diversas acurácias obtidas (ARRIAL, 2008).

2.5. Representação de documentos

Uma forma simples de representar documentos é associar a cada documento analisado um conjunto de termos que se julga serem significativos no contexto a que um documento pertence, partindo do pressuposto que esses termos ocorrem nos documentos.

A representação mais abordada na literatura de classificação de textos é conhecida como *bag of words*. Nessa abordagem, cada termo corresponde a uma única palavra no conjunto de palavras do conjunto de treinamento. Lewis (1992) mostrou que representa-

ções de documentos mais sofisticadas, como frases, resultaram em um pior desempenho em experimentos rodados na base de notícias da Reuters. Além disso, Scott e Matwin (1999) acrescentaram informação semântica à tarefa de classificação de textos e não obtiveram resultados satisfatórios (STEINBRUCH, 2006).

Em contraste, outros trabalhos apresentaram melhor desempenho na utilização de frases (MLADENIC e GROBELNIK, 1998) e reconhecimento de nomes próprios (BASILI, 2000), comparados à representação tradicional *bag of word*s (STEINBRUCH, 2006).

2.6. Divulgação de Informações

Segundo a Norma Brasileira de Contabilidade NBCT 1, as demonstrações contábeis objetivam fornecer informações que sejam úteis na tomada de decisões e avaliações por parte dos usuários em geral, com a finalidade de satisfazer às necessidades comuns da maioria dos seus usuários, como: compra e venda de ações, desempenho e prestação de contas, capacidade de pagamento a empregados, segurança de credores, distribuição de lucros e dividendos (CFC, 2008).

Dias Filho e Nakagawa (2001) entendem que o usuário da Contabilidade, em qualquer circunstância, precisa de informações claras para fazer julgamentos adequados e adotar decisões racionais.

A Comissão de Valores Mobiliários dos Estados Unidos (*Securities and Exchange Commission-SEC*) exige o arquivamento do Formulário 8-K para companhias de capital aberto, devendo estas relatar qualquer fato relevante que possa afetar sua situação financeira ou o valor de suas ações, portanto, desde atividades de fusão até alterações nos documentos constitutivos ou estatutos. A SEC considera como relevantes quaisquer questões sobre as quais um investidor comum e prudente deveria ser informado antes de decidir comprar, vender ou deter um valor mobiliário registrado. Normalmente o Formulário 8-K deve ser arquivado dentro de um mês da ocorrência do fato relevante, porém as regras de divulgação oportuna podem exigir que uma companhia divulgue imediatamente por meio da imprensa um comunicado relacionado com um evento que se deverá subsequentemente relatar no Formulário 8-K (DOWNE e GOODMAN, 1993).

2.7. Fatos Relevantes

Os fatos relevantes são comunicados divulgados por sociedades cujo capital é dividido em ações, chamadas de sociedade abertas, que para negociar em Bolsa ou mercado de balcão, seus valores mobiliários, necessitam estar registradas na CVM. Estas, ao se registrarem, precisam manter atualizadas as informações econômicas e financeiras, para que se possa avaliar as condições da empresa e suas perspectivas futuras (BRASIL, 2004 apud Rodrigues, 2005). Estes comunicados geralmente são expressos em textos narrativos e podem ser divulgados em qualquer momento. Diversos estudos mostram a importância da

informação narrativa por ser frequentemente utilizada pelos investidores na tomada de decisão, pela sua tempestividade e por não serem auditados (FERNANDES e SILVA, 2007).

Segundo Hendriksen e Van Breda (1999, p. 511 apud Rodrigues, 2005), a política de divulgação ao público investidor de uma companhia, na medida em que todos os investidores têm necessidade de informação para avaliar os riscos relativos de cada empresa, é necessária para o funcionamento ótimo do mercado de capitais.

Diante disso, os fatos relevantes exercem um papel importante na política de divulgação da empresa, já que a estes estão relacionados as decisões que possam influir na cotação de valores mobiliários, na decisão de investidores com relação a comprar e vender, ou exercer direitos, como descrito no artigo 2º da instrução 358/02 da CVM (CVM, 2008), sendo definido como fato ou ato relevante

> qualquer decisão de acionista controlador, deliberação da assembleia geral ou dos órgãos de administração da companhia aberta, ou qualquer outro ato ou fato de caráter político-administrativo, técnico, negocial ou econômico-financeiro ocorrido ou relacionado aos seus negócios que possa influir de modo ponderável: i) na cotação dos valores mobiliários de emissão da companhia aberta ou a eles referenciados; ii) na decisão dos investidores de comprar, vender ou manter aqueles valores mobiliários; iii) na decisão dos investidores de exercer quaisquer direitos inerentes à condição de titular de valores mobiliários emitidos pela companhia ou a eles referenciados.

A mesma Instrução 358/02 da CVM, em seu artigo 3º, define que ao Diretor de Relações com investidores cumpre divulgar o ato ou o fato relevante ocorrido ou relacionado ao seu negócio, além de disseminar a informação imediatamente, em todos os mercados em que haja negociação dos seus valores mobiliários, por meio de publicação nos jornais de grande circulação. Já com relação à divulgação, o artigo citado diz que esta deve ser feita de modo claro e preciso, em linguagem clara e acessível ao público investidor. Em seu artigo 6º, ressalva-se que os atos ou fatos relevantes podem deixar de ser divulgados se os acionistas controladores ou os administradores entenderem que a revelação do fato relevante poderá pôr em risco interesse legítimo da companhia.

Estudo de Fernandes e Silva (2007), o qual analisou a facilidade de leitura de fatos relevantes divulgados nos anos de 2002 a 2006 pelas companhias brasileiras de capital aberto, mostrou que apenas 10% dos textos são de fácil leitura, além da constatação, ao longo dos anos, da divulgação dos fatos relevantes, sendo que ficaram mais extensos.

2.8. Pesquisas em textos narrativos

Pesquisas relacionadas a textos narrativos têm sido recentemente objeto de estudo, como de Rodrigues (2005), que estuda os relatórios da administração, classificando-os em

otimista, pessimista e neutro. Outra pesquisa é de Fernandes e Silva (2007), na qual se analisou a legibilidade dos fatos relevantes, além de concluir que as empresas estão divulgando menos fatos relevantes.

Um exemplo de trabalho estrangeiro similar é o de Davis et al. (2006), o qual analisa uma amostra de 24.000 *earnings press releases*, que são semelhantes aos fatos relevantes brasileiros; chegou-se à conclusão que os administradores utilizam essas informações para assegurar credibilidade sobre o desempenho futuro da empresa. Esse trabalho, portanto, corrobora a importância desses breves comunicados formais.

A análise de conteúdo tem sido uma ferramenta utilizada em textos narrativos, como apontada em pesquisas como de Jones e Shoemaker (1994 apud Fernandes e Silva, 2007), que examinaram pesquisas sobre análise de conteúdo e legibilidade em relatórios anuais.

2.9. Análise de conteúdo

De acordo com Martins e Theóphilo (2007, p. 96 apud Pereira e Silva, 2008), a análise de conteúdo é uma técnica que tem como principais objetivos identificar intenções, características e apelos dos comunicadores; medir a clareza das mensagens; descobrir estilos de comunicação e desvendar as ideologias dos textos legais, para que se possa fazer inferências (CARNEY, 1972 apud RODRIGUES, 2005).

Bligh e Hess (2005 apud Pereira e Silva, 2008) utilizaram-se desta técnica para verificar os níveis de otimismo, pessimismo, linguagem técnica e urgência dos comunicados do presidente do *Federal Reserve*, tendo os resultados apontado que essas variáveis podem ajudar a prever as oscilações do mercado financeiro.

As etapas da análise de conteúdo podem ser listadas como a definição e a categorização do universo, e a escolha das unidades de análise (FREITAS e JANISSEK, 2000 apud RODRIGUES, 2005).

Diversas pesquisas têm sido realizadas com relação à análise de conteúdo. Uma delas é a de Bryan (1997, apud RODRIGUES, 2005), na qual o autor, para avaliar a eficácia de Relatórios da Administração, utilizou-se da análise de conteúdo para classificar em relatório favorável e desfavorável, e em neutro ou omisso, de acordo com a opinião do administrador e do julgamento do pesquisador, respectivamente. Já Jones e Shoemaker (1994 apud FERNANDES e SILVA, 2007) estudaram questões relativas à legibilidade, verificando a dificuldade de leitura de um relatório anual; além de constatar que alguns relatórios ou algumas partes dos relatórios são mais difíceis de serem lidas do que outras; houve um aumento progressivo na dificuldade de leitura dos relatórios anuais e que há alguma associação entre a legibilidade dos relatórios e outras variáveis.

Já Fernandes e Silva (2007), em seu estudo, analisam fatos relevantes divulgados pelas companhias brasileiras de capital aberto, do período de 2002 a 2006. O trabalho utilizou-se da fórmula de legibilidade Flesh, para verificar quão difícil é ler um fato relevante,

se existem alguns tipos de fatos relevantes mais difíceis de serem lidos do que outros e se os fatos relevantes estão se tornando mais difíceis de serem lidos.

Pesquisas como a de Pereira e Silva (2008) tiveram por objetivo determinar se o grau de otimismo dos fatos relevantes divulgados pelas companhias abertas, no período de 2006/2007, afetou o comportamento do preço das ações dessas entidades. Para tanto, realizaram-se leituras e classificações de todos os 2.350 fatos do período, por duas pessoas, de forma independente e sem comunicação entre si. Além disso, para cada empresa, foram coletadas as cotações de sua ação mais líquida e também o índice Ibovespa correspondente ao período, para que assim fosse feita uma relação entre a variação do preço da ação, a variação do Ibovespa e o fato relevante. Seu resultado conclui que, para 95% dos fatos relevantes neutros, não apresentaram nenhum efeito sobre o preço das ações. Para os fatos relevantes classificados como otimistas, verificou-se que o mercado não reagiu à linguagem utilizada nesses fatos. Com relação aos fatos classificados como pessimistas, os autores concluem que nesta classificação se encaixaram poucos, dificultando fazer generalizações. Na pesquisa citada, sugeriu-se para as próximas pesquisas a utilização de novas metodologias, como tentativa de utilização de nova metodologia neste trabalho.

3. METODOLOGIA

Esta pesquisa foi, inicialmente, realizada por meio das etapas descritas a seguir:

a) escolha do objeto de análise e do tipo de conhecimento que será aprendido;
b) coleta dos dados: fatos relevantes e planilha com as classificações de cada um;
c) separação dos documentos de acordo com a classificação de conteúdo;
d) codificação dos documentos em forma inteligível para o programa de aprendizagem de máquina;
e) escolha dos termos que melhor expressam o conteúdo dos textos, para definição dos atributos que serão utilizados;
f) cálculo da quantidade média de palavras de cada conjunto para escolha dos intervalos de análise.

Diante do fato de a pesquisa ser baseada em pesquisa anterior, feita por Pereira e Silva (2008), na qual foi feita leitura e classificação dos fatos relevantes divulgados nos anos de 2006 e 2007, em “Otimista”, “Pessimista” e “Neutro”, o conhecimento a ser assimilado com a aprendizagem de máquina também se refere à classificação de fatos relevantes em Otimista”, “Pessimista” e “Neutro”. Na pesquisa citada, duas pessoas classificaram cada um dos fatos relevantes de maneira independente. Para a presente pesquisa, foram considerados somente os textos no qual a classificação foi consensual entre os dois pesquisadores, totalizando 1.670 fatos relevantes.

A próxima etapa correspondeu à coleta dos fatos relevantes ocorridos nos anos de 2006 e 2007, estes disponibilizados por Pereira e Silva (2008), mas que também estão disponíveis na página da CVM na internet . Além disso, foi disponibilizada planilha com as respectivas classificações de cada fato relevante.

A etapa seguinte corresponde à separação dos documentos de acordo com a planilha citada. Em um primeiro momento, encontravam-se juntos todos os fatos relevantes, independentemente se a classificação feita pelos classificadores de Pereira e Silva (2008) foi consensual ou não, sendo necessário isolar apenas os textos que tiveram uma classificação consensual. Para a utilização apenas dos dados consensuais, foi feita uma conversão dessa planilha para o formato "texto separado por tabulações" e, para que fossem obtidos somente os nomes dos documentos que tiveram seu conteúdo classificado consensualmente por ambos os analisadores, foram utilizados *scripts* de PERL para separá-los. Depois de separados apenas os arquivos que eram consensuais, foi necessário outro *script* para separar os arquivos em pastas diferentes, de acordo com a classificação de conteúdo que lhe foi atribuída ("Otimista", "Pessimista" ou "Neutro"). Essa etapa eliminou 29% dos arquivos de fatos relevantes, proporção que representa os arquivos que foram classificados de forma não consensual entre classificadores.

Após possuir cada fato relevante separado de acordo com a sua classificação, para a próxima etapa, que corresponde à codificação dos documentos em forma inteligível para o programa de aprendizagem de máquina, foi necessário convertê-los em formato arquivo de texto (TXT), pois, apesar de serem versáteis e largamente utilizados, os formatos de arquivos do programa Word, do pacote Microsoft (DOC) e Adobe Systems (PDF) são muito inconvenientes para análise computacional por *scripts*, devido a seus padrões de codificação. Para esse tipo de análise, comumente utilizam-se os chamados arquivos em formato plano, entre os quais um dos mais conhecidos, e que é usado nesse trabalho, é o formato TXT.

Considerando esse fato, os arquivos em formato DOC e formato PDF precisaram passar por um processo inicial de conversão. Para esse fim, optou-se por converter os arquivos para linguagem oficial de criação de páginas da web, *HyperText Markup Language* (HTML), sendo suas especificações mantidas pelo World Wide Web Consortium (W3C). Esta conversão foi feita pelos softwares *wvware*[2] e *pdftohtml*[3], para DOC e PDF, respectivamente. Depois, os arquivos em formato HTML foram convertidos ao formato TXT por scripts de PERL[4] . Os arquivos de texto resultantes foram verificados por rápida inspeção visual.

Depois da conversão, verificou-se que alguns fatos relevantes não poderiam ser analisados. Os critérios de exclusão foram os seguintes:

a) arquivo com codificação incompatível com processo de conversão (por exemplo, arquivos PDF em formato escaneado);

[1] *http://www.cvm.gov.br*
[2] Implementação GnuWin32 para Windows – disponível em gnuwin32.sourceforge.net/packages/wv.htm
[3] Disponível em *http://pdftohtml.sourceforge.net/*
[4] Disponível em *http://www.omanurkka.net/files/html2txt.pl.txt*

b) arquivos protegidos por senhas, impedindo conversão;
c) conversão bem-sucedida, mas arquivo resultante defeituoso;
d) arquivo com excesso de caracteres inválidos devido a problemas de conversão;
e) documentos em inglês;
f) documentos repetidos.

Assim, a Tabela 1 mostra a quantidade de fatos relevantes antes e após o processo de exclusão.

Tabela 1 – Quantidade remanescente de fatos relevantes em cada classificação, antes e após o processo de exclusão.

Anos - 2006 e 2007	Otimista	Pessimista	Neutro	Total
Antes do processo de exclusão	364	19	1.287	1.670
Depois do processo de exclusão	351	19	1.215	1.585

Fonte: elaboração própria

Após a coleta de dados, a separação, a conversão e a exclusão, seguiu-se a etapa de escolha dos termos que melhor expressam o conteúdo dos textos, ou seja, os termos com maior potencial discriminativo para classificação dos fatos relevantes em Otimista", "Pessimista" e "Neutro".

A seguir, o *script* permite que o usuário forneça quais palavras serão utilizadas na análise. A acentuação gráfica deve ser ignorada pelo usuário, e são aceitos termos compostos por mais de uma palavra. Nesse trabalho, as palavras usadas foram obtidas a partir de prospecção de trabalhos similares na literatura. Pereira e Silva (2008) citam em seu trabalho algumas palavras que aparecem com maior frequência em documentos classificados como otimistas, que foram as seguintes: "maximização", "sinergia", "ganho de escala", "crescimento" e "redução de custos". No trabalho de Davis et al. (2006), no qual documentos contábeis também foram classificados em função de seu conteúdo, as palavras com melhor poder discriminatório foram as seguintes (em termos obtidos por tradução livre seguidos do vocábulo original em inglês):

- palavras que aumentam a probabilidade de o conteúdo ser otimista: "honra" (*praise*), "satisfação/existência/recompensa" (*satisfaction*), "inpiração" (*inspiration*);
- palavras que diminuem a probabilidade de o conteúdo ser otimista: "culpa/responsabilidade/repreensão" (*blame*), "dificuldade" (*hardship*), "negação/recusa" (*denial*).

Além dessas palavras obtidas da literatura contábil, outros termos foram escolhidos a partir de um glossário contábil.

Um *script* de PERL foi criado para contar a ocorrência de cada uma dessas palavras em cada um dos conjuntos de dados, de forma a estimar quais termos possuem um maior poder de discriminar as classes. A Tabela 3 mostra os resultados obtidos a partir desse cálculo de ocorrências para alguns dos termos testados.

Tabela 2 – Quantidade de ocorrência de termos nos documentos que compõem o conjunto de treinamento.

Termos obtidos de trabalhos anteriores				**Termos obtidos de jargões da Contabilidade**			
	Otimista	**Pessimista**	**Neutro**		**Otimista**	**Pessimista**	**Neutro**
Honra	0,57	0,00	1,32	falen-	0,28	0,00	0,82
satisf-	17,95	0,00	0,00	fusao	0,00	0,00	0,00
Existência	0,00	0,00	0,00	**incorpora-**	708,55	0,00	58,44
Culpa	0,57	0,00	0,16	cisao	0,00	0,00	0,00
responsa-	16,24	0,00	4,12	**fluxo de caixa**	14,53	0,00	1,32
dificul-	1,71	5,26	0,82	concordata	0,00	0,00	0,74
Recusa	0,57	0,00	0,16	**divida**	15,67	0,00	4,69
Sinergia	27,35	0,00	0,90	fusao	9,12	0,00	1,81
ganho de escala	0,85	0,00	0,00	**juros**	16,24	0,00	31,44
maximiz-	4,56	0,00	0,08	compra	49,86	0,00	63,46
Crescimento	60,40	5,26	1,15	**lucr-**	33,33	5,26	7,49
custo	93,73	15,79	9,71	venda	67,24	5,26	51,77

Fonte: elaboração própria.

A frequência de cada termo foi normalizada pela quantidade de documentos presentes em cada conjunto de classe (“Otimista”, “Pessimista” e “Neutro”) e multiplicada por 100. Um hífen no final de um termo indica que foi utilizado um radical em vez de palavra inteira. Como algumas palavras são usadas com grande variação de sufixos e prefixos, para esses casos, optou-se por usar como termo apenas o radical de uma palavra, por exemplo: ao buscar o radical “lucr” nos textos, o programa considera como coincidentes os termos derivados “lucro”, “lucrar”, “lucratividade”, “lucramos”, etc.

Termos em negrito, na Tabela 3, são os que possuem maior diferença de ocorrência interclasses e, por isso, foram escolhidos para integrar o vetor de características. A Tabela 3 foi dividida de acordo com a origem dos termos.

A seguir, foi criado outro *script* para que os dados fossem codificados em forma de conjunto de treinamento inteligível para programas de aprendizagem de máquina. O programa resultante possui uma interface por um script de PERL, implementado em Sistema Operacional Windows XP com acesso pelo *prompt* de comando de DOS que exige que o usuário insira a localização do diretório contendo os arquivos de treinamento, os termos que deseja utilizar e, opcionalmente, um diretório com arquivos a serem testados pelo modelo induzido.

4.3. Limitações da pesquisa

Com relação ao conjunto de treinamento, reitera-se que os dados usados nesse trabalho não foram obtidos da forma ideal. Em vez disso, configura-se como documentos claramente destinados ao usuário humano final, sem preparação para um processamento automático computacional subsequente, dada a sua falta de padronização e níveis de desestruturação (ausência de delimitação de campos) e de subjetividade do texto redigido. Deve-se considerar também que os dados usados no trabalho contêm elementos textuais que não são interessantes e nem relevantes, como, por exemplo, cabeçalhos, datas, etc., e que idealmente deveriam estar ausentes antes mesmo do processamento.

Com relação à limitação do método, o uso do algoritmo bayesiano como classificador é considerado simplista para alguns pesquisadores já que pode ser ineficiente em tarefas de classificação que envolvam atributos correlacionados e dados de treinamento redundantes.

O desempenho do classificador depende da qualidade dos termos selecionados e da representatividade e acurácia dos bancos de dados, por isso o classificador deve ser aplicado em áreas que apresentam conceitos consolidados na literatura, fundamentos teóricos estabelecidos e grupos de pesquisa atuantes.

4. RESULTADOS E ANÁLISE

4.1 Resultados

O programa utilizado foi o *Naïve* Bayes Classifier – NBC (Borgelt, 2007), que consiste em um algoritmo de aprendizagem de máquina do tipo supervisionado, com aprendizagem do tipo indutiva estatística, multiclasses (três classes analisadas: “Otimista”, “Pessimista” e “Neutro”), com tarefa de classificação.

A coleta de todos os valores das variáveis é feita pelo método multinomial *bag of words*, exceto para o atributo “quantidade de palavras”, no qual são utilizadas quatro variáveis booleanas de acordo com o método binário.

A coleta dos dados, formatação e fornecimento dos dados ao programa NBC é executada por *scripts* de PERL, sendo que todos os algoritmos usados neste trabalho, ou seja, *scripts* de PERL, *wvware*, *pdftohtml* e NBC, foram desenvolvidos, instalados ou implementados em sistema operacional Microsoft Windows. Com exceção do sistema operacional, todos os softwares citados são disponibilizados gratuitamente.

A escolha dos atributos está baseada nas premissas: alguns termos são mais frequentes em algumas classes do que em outras (como mostrado na Tabela 2); fatos relevantes com classificações diferentes possuem diferentes quantidades de palavras (como mostrado na Tabela 4); e fatos relevantes com classificações diferentes possuem quantidades diferentes de números de algarismos numéricos (como mostrado na Tabela 6).

Para a análise, foram escolhidos 12 atributos, de maneira a se delinearem padrões dentro de cada classificação, que estavam divididos em 17 variáveis, conforme a seguir:

- Quantidade de palavras: i) menor que 200 palavras, ii) entre 200 e 400, iii) entre 400, e iv) 800 e maior do que 800 palavras, portanto 1 atributo e 4 variáveis.
- Quantidade de números: quantidade de números que aparecem no texto dividido pela quantidade de palavras totais de cada texto, portanto 1 atributo e 1 variável.
- Quantidade de termos: termos selecionados para análise: “satisf”, “sinergia”, “maximiz”, “crescimento”, “custo”, “fluxo de caixa”, “incorpora”, “divida”, “juros”, “lucr”, portanto 10 atributos e 10 variáveis.

A quantidade de palavras é verificada de forma categorizada, codificada de forma binária, representada pelas classes: i) menor que 200 palavras, ii) entre 200 e 400, iii) entre 400 e 800 e iv) maior do que 800 palavras. Ou seja: um documento com 160 palavras possui a codificação “1 0 0 0”, outro com 350 é codificado como “0 1 0 0”, um com 1320 é representado por “0 0 0 1”, e assim por diante. Os valores dos limites dessas classes 200, 400 e 800 foram escolhidos baseados no cálculo da quantidade média de palavras descrita na metodologia.

A quantidade de palavras foi, então, normalizada, somando-se todas as palavras de cada fato relevante e dividido pela quantidade de fatos relevantes, descritos na Tabela 5.

Classificação	<200	>200 – 400	>400 - 800	>800
Otimistas	10,26	29,63	21,08	39,03
Pessimistas	47,37	31,58	21,05	0,00
Neutros	29,79	42,80	22,47	4,94

Fonte: elaboração própria

A quantidade de números foi normalizada dividindo-se a quantidade de números de cada documento pela quantidade de palavras em cada fato relevante, resultados mostrados na Tabela 6.

Classificação	Quantidade normalizada de números
Otimistas	0,24
Pessimistas	0,36
Neutros	0,36

Fonte: elaboração própria

O programa automaticamente forma um arquivo único de saída com o conjunto de treinamento contendo instâncias, cada qual classificada com suas respectivas três classes (“Otimista”, “Pessimista” e “Neutro”).

Esse arquivo de saída é preparado e fornecido de forma automática ao classificador nB, que induz um modelo de classificação. Após induzido o modelo, este é usado para clas-

sificar o próprio conjunto de treinamento que o gerou, o que dá uma ideia para o usuário sobre o desempenho do classificador.

A seguir, é facultado ao usuário fornecer uma pasta com os arquivos que ele porventura deseje classificar. Esses arquivos de teste devem estar em formato TXT. O *script* se encarrega de montar um arquivo único de teste, que é então classificado pelo modelo induzido durante a etapa de treinamento.

Ao se utilizar o próprio conjunto de treinamento como conjunto de teste do modelo induzido, o índice de acerto deste foi 99,43%. Esse índice, no entanto, não deve ser interpretado como o desempenho esperado do programa em um conjunto de teste desconhecido, mas, sim, uma estimativa de adaptação do modelo ao conjunto de treinamento pelo qual ele foi induzido.

4.2. Desempenho do programa

Como estimativa do desempenho do programa, obteve-se uma acurácia de 80,89% no procedimento de validação cruzada. Esse índice sugere que o classificador foi bem-sucedido em sua tarefa. Ressalta-se que o índice de acertos melhorou muito após análise e escolha como atributos apenas de termos que apresentaram diferença de frequência entre os dados das três classes analisadas.

Como a quantidade de fatos relevantes classificados como pessimistas foi baixa relativamente às outras classes, procedeu-se à indução e ao teste de outro classificador, treinado e testado apenas em fatos relevantes otimistas e neutros. Esse novo classificador teve uma acurácia de 82,37% no procedimento de validação cruzada. Portanto, pode-se inferir que a baixa quantidade de dados na classe pessimista “confunde” o classificador, degradando seu desempenho.

5. CONCLUSÕES

O objetivo do presente trabalho foi propor uma classificação automática dos textos, fazendo uma análise do conteúdo de textos narrativos, utilizando ferramentas computacionais e textos já classificados por pesquisadores, além de buscar uma maior inserção da Ciência Contábil na abordagem de aprendizado de máquina.

Os resultados obtidos demonstraram que é possível classificar textos narrativos contábeis de forma automática e que a Ciência Contábil pode utilizar-se do conhecimento de aprendizado de máquina para fazer análises em outros tipos de textos. O procedimento de validação cruzada permite inferir que o programa desenvolvido nesse trabalho identifica corretamente 80,89% dos fatos relevantes quanto a conteúdo “Otimista “ou “Neutro” e 82,37% quando as classificações de conteúdo são somente Otimista”, “Pessimista” e “Neutro”.

Apesar de a quantidade de exemplos no conjunto de treinamento nesse trabalho ter sido baixa, uma das vantagens do programa aqui descrito é que esse conjunto pode ser ali-

mentado por dados de trabalhos futuros, de forma simples e imediata, conforme forem sendo disponibilizados outros fatos relevantes classificados.

Referentemente às limitações do trabalho, reitera-se que os dados usados nesse trabalho configuram-se como documentos destinados ao usuário humano final e sem preparação para um processamento automático computacional subsequente, a julgar por sua falta de padronização, níveis de desestruturação (ausência de delimitação de campos) e de subjetividade do texto redigido. Ademais, os documentos contêm elementos textuais que não são interessantes e nem relevantes, como, por exemplo, cabeçalhos, datas, etc., e que idealmente deveriam estar ausentes antes mesmo do processamento. Com relação à limitação do método, o uso do algoritmo bayesiano pode ser ineficiente em tarefas de classificação que envolvam atributos correlacionados e dados de treinamento redundantes. Além disso, o desempenho do classificador depende do poder discriminatório dos termos selecionados e da representatividade e qualidade dos bancos de dados, por isso o classificador deve ser aplicado em áreas que apresentam conceitos consolidados na literatura, fundamentos teóricos estabelecidos e grupos de pesquisa atuantes.

Para pesquisas futuras, sugere-se uma classificação mais generalista e precisa com um conjunto de treinamento maior; refinamento na preparação dos arquivos que compõem os conjuntos de treinamento e teste, como, por exemplo, eliminação de cabeçalhos, assinaturas, datas, etc.; integração de análises sintática e de contexto; determinação das palavras com melhor poder discriminativo; adaptação dos programas de leitura automatizada a caracteres com acentuação; uso na classificação de conteúdo de outros tipos de documentos contábeis; uso de técnicas de aprendizagem de máquina mais aprimoradas, como máquinas de vetores de suporte e redes neurais (WITTEN e FRANK, 2005).

O programa desenvolvido pode ser usado durante a elaboração de um documento de fato relevante, para se ter uma estimativa prévia quanto a posicionamento de seu conteúdo. Outro uso interessante seria comparar suas predições às classificações atribuídas por pesquisadores, como um trabalho de determinação de concordância entre classificações emitidas por humanos e por máquinas.

Segundo o levantamento bibliográfico, esse trabalho é pioneiro na utilização de aprendizagem de máquina na classificação de textos contábeis em português quanto ao seu conteúdo. Trabalhos similares foram encontrados apenas com aplicação específica em língua inglesa (DEVITT e AHMAD, 2007; DAVIS et al., 2006).

Os programas e bancos de dados desenvolvidos e utilizados nesse trabalho são disponibilizados gratuitamente pelos autores.

6. AGRADECIMENTOS

Agradecemos a PEREIRA, V.A. e SILVA, C.A.T. pela disponibilização da base de dados de fatos relevantes usada nesse trabalho.

REFERÊNCIAS

ARRIAL, Roberto Ternes. *Predição de RNAs não codificadores no transcriptoma do fungo Paracoccidioides brasiliensis usando aprendizagem de máquina*. Dissertação de mestrado em Biologia Molecular, Universidade de Brasília, 2008. Disponível em: http://bdtd.bce.unb.br/tedesimplificado/tde_busca/arquivo.php?codArquivo=3450

BORGELT, C. Full and *Naïve* Bayes classifiers. Software disponível em http://www.borgelt.net/bayes.html, 2008.

BOSE, Indranil. MAHAPATRA, Radha K. Business data mining – a machine learning perspective. *Information & Management* 39, 2001. p. 211-225

CONSELHO FEDERAL DE CONTABILIDADE (CFC). *Princípios Fundamentais e Normas Brasileiras de Contabilidade*. NBCT – 1. Brasília: CFC, 2000.

COMISSÃO DE VALORES MOBILIÁRIOS – CVM. *Instrução CVM n° 358, de 3 de janeiro de 2002*. Disponível em http://www.cvm.gov.br. Acesso em: 14/05/2008.

DAVIS, Ângela K.; PIGER, Jeremy M.; SEDOR, Lisa M. Beyond the numbers: an analysis of optmistic and pessimistic language in earnings press Releases.

Research Division – Federal Reserve Bank of St. Louis. *Relatório técnico*, 2006. Disponível em: research.stlouisfed.org. Acesso em: 21 de abril de 2006.

DEVITT, Ann. AHMAD, Khurshid. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. P. 984–991, República Tcheca, 2007.

DIAS FILHO, José Maria. NAKAGAWA, Masayuki. *Revista Contabilidade & Finanças FIPECAFI - FEA - USP*, São Paulo, FIPECAFI, v.15, n. 26, p. 42 - 57, maio/agosto 2001.

DOWNE, Johns. GOODMAN, Jordan Elliot. *Dicionário de termos financeiros e de investimento*. Tradução Ana Rocha Tradutores Associados. — São Paulo: Nobel, 1993.

FERNANDES, José Lúcio Tozetti. SILVA, César Augusto Tibúrcio. Análise da Legibilidade dos Textos Narrativos dos Fatos Relevantes Divulgados pelas Empresas Brasileiras de Capital Aberto nos Anos de 2002 a 2006. RAC *Eletrônica*, no prelo.

HENDRIKSEN, Eldon S; BREDA, Michael F. Van. *Teoria da Contabilidade*. Trad. de Antônio Zoratto Sanvicente. 5ª ed. São Paulo: Atlas, 1999.

MITCHELL, T.M. *Machine Learning*. McGraw-Hill, 1997.

PEREIRA, Vinícius Alves dos Santos. SILVA, César Augusto Tibúrcio. Fatos Relevantes e sua Influência no Preço das Ações no Brasil. *Anais do 5º. Congresso USP de Iniciação Científica em Contabilidade*, São Paulo, 2008.

RODRIGUES, Fernanda Fernandes. *Análise das Variáveis que Influenciam as Informações Divulgadas nos Relatórios da Administração das Companhias Abertas Brasileiras: um estudo empírico nos anos de 2001 a 2003*. Dissertação de mestrado do Programa Multiinstitucional e Inter-regional de Pós-Graduação em Ciências Contábeis. 2005.

SOUTO, M. C. P. de; LORENA, A. C. ; DELBEM, A. C. B., CARVALHO A. C. P. L. F. de. *Técnicas de Aprendizado de Máquina para Problemas de Biologia Molecular*. Instituto de Ciências Matemáticas e de Computação Universidade de São Paulo-São Carlos. 2003.

STEINBRUCH, David. *Um estudo de algoritmos para classificação automática de textos utilizando naïve-Bayes*. Pontifícia Universidade Católica do Estado do Rio de Janeiro. Dissertação de Mestrado. Setembro de 2006.

WITTEN, I.A. e FRANK, E. *Data Mining: Practical machine learning tools and techniques*. 2. ed. Elsevier: EUA, 2005.