UNISINOS

Programa Interdisciplinar de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Adriano Cagol

SUMARIZAÇÃO E EXTRAÇÃO DE CONCEITOS DE NOTAS EXPLICATIVAS EM RELATÓRIOS FINANCEIROS:

Ênfase nas notas das principais práticas contábeis

São Leopoldo, 2017

Adriano Cagol

SUMARIZAÇÃO E EXTRAÇÃO DE CONCEITOS DE NOTAS EXPLICATIVAS EM RELATÓRIOS FINANCEIROS:

Ênfase nas notas das principais práticas contábeis

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

Orientador: Dr. João Francisco Valiati

São Leopoldo

2017

C131s Cagol, Adriano.
Sumarização e extração de conceitos de notas explicativas em relatórios financeiros: ênfase nas notas das principais práticas contábeis/Adriano Cagol. – 2018.
90f. : il.; 30 cm.

Dissertação(mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2018.
"Orientador: Dr. João Francisco Valiati."

1. Mineração de textos. 2.Sumarização automática. 3. Extração de conceitos. 4.Notas explicativas. 5. Demonstrações financeiras.I. Título.

CDU 657:004

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Bruna Sant'Anna – CRB 10/2360)

Adriano Cagol

SUMARIZAÇÃO E EXTRAÇÃO DE CONCEITOS DE NOTAS EXPLICATIVAS EM RELATÓRIOS FINANCEIROS:
Ênfase nas notas das principais práticas contábeis

Dissertação apresentada à Universidade do Vale do Rio dos Sinos – Unisinos, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Aprovado em 27 de abril de 2017.

BANCA EXAMINADORA

Prof. Dr. Leonardo Dagnino Chiwiacowsky - UCS

Prof. Dr. Sandro José Rigo - UNISINOS

Prof. Dr. João Francisco Valiati (Orientador)

Visto e permitida a impressão
São Leopoldo,

Prof. Dr. Sandro José Rigo
Coordenador PPG em Computação Aplicada

AGRADECIMENTOS

Para realização deste trabalho, alguns aspectos foram fundamentais. Dentre eles, pessoas com quem convivi e que de diferentes maneiras, foram importantes para a realização deste trabalho. Desta forma, presto meus sinceros agradecimentos:

Ao professor Dr. João Francisco Valiati, orientador deste trabalho, pelos seus conhecimentos, atenção, dedicação e orientações, que foram fundamentais para o meu crescimento pessoal e acadêmico;

Aos demais professores do corpo docente do Programa Interdisciplinar de Pós-Graduação em Computação Aplicada da UNISINOS que através de aulas ou conversas informais contribuíram para a aquisição de conhecimento, desenvolvimento de senso crítico e capacidade de abstração;

Ao professor Dr. Clóvis Antônio Kronbauer, que através de indagações e esclarecimentos da área contábil ajudou no desenvolvimento deste trabalho;

Aos meus colegas de curso, que diretamente ou indiretamente, contribuíram para a realização deste trabalho, com apoio, companheirismo e compartilhamento de conhecimento.

A minha família que, desde sempre, foi essencial na minha formação em todos os sentidos.

A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro.

RESUMO

As demonstrações financeiras apresentam o desempenho financeiro das empresas e são uma importante ferramenta para análise da situação patrimonial e financeira, bem como para tomada de decisões de investidores, credores, fornecedores, clientes, entre outros. Nelas constam as notas explicativas que descrevem em detalhes as práticas e políticas de comunicação dos métodos de contabilidade da empresa, além de informações adicionais. Dependendo dos objetivos, não é possível uma correta análise da situação de uma entidade através das demonstrações financeiras, sem a interpretação e análise das notas explicativas que as acompanham. Porém, apesar da importância, a análise automática das notas explicativas das demonstrações financeiras ainda é um obstáculo. Em vista desta deficiência, este trabalho propõe um modelo que aplica técnicas de mineração textual para efetivar a extração de conceitos e a sumarização das notas explicativas, relativas à seção de principais práticas contábeis adotadas pela empresa, no sentido de identificar e estruturar os principais métodos de apuração de contas contábeis e a geração de resumos. Um algoritmo de extração de conceitos e seis algoritmos de sumarização foram aplicados sobre as notas explicativas das demonstrações financeiras de empresas da Comissão de Valores Mobiliários do Brasil. O trabalho mostra que a extração de conceitos gera resultados promissores para identificação do método de apuração da conta contábil, visto que apresenta acurácia de 100% na nota explicativa do estoque e do imobilizado e acurácia de 96,97% na nota explicativa do reconhecimento da receita. Além disso, avalia os algoritmos de sumarização com a medida ROUGE, apontando os mais promissores, com destaque para o LexRank, que no geral conseguiu as melhores avaliações.

Palavras-chave: Mineração de textos. Sumarização automática. Extração de conceitos. Notas explicativas. Demonstrações financeiras. Demonstrações contábeis.

ABSTRACT

Financial statements present the financial performance of companies and are an important tool for analyzing the financial and equity situation, as well as for making decisions of investors, creditors, suppliers, customers, among others. These are listed explanatory notes that describe in detail how practices and policies of accounting methods of the company. Depending on the objectives, a correct analysis of the situation of a company on the financial statements is not possible without an interpretation and analysis of the footnotes. However, despite the importance, an automatic analysis of the footnotes to the financial statements is still an obstacle. In view of this deficiency, this work proposes a model that applies text mining techniques without the sense of identifying the main methods of calculating the accounting accounts, the reports in the footnotes, with concept extraction, as well as generating a summary that contemplates the main idea of these, through summarization. A concept extraction algorithm and six summarization algorithms are applied in financial statements of companies of Brazilian Securities and Exchange Commission. The work shows that concept extraction generates promising results for the identification of the method of calculating the accounting account, since it presents a 100% accuracy in the footnote of inventory and property, plant and equipment, and accuracy of 96.97% in the footnote on revenue recognition. In addition, it evaluates the algorithms for summarization with the ROUGE measure, pointing out the most promising ones, especially LexRank, which in general obtained the best evaluations.

Keywords: Text mining. Automatic summarization. Concepts extraction. Footnotes. Financial statements. Accounting statements.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

CPC	Comitê de Pronunciamentos Contábeis
CVM	Comissão de Valores Mobiliários
DFC	Demonstração dos Fluxos de Caixa
DMPL	Demonstração das Mutações do Patrimônio Líquido
DRA	Demonstração do Resultado Abrangente
DRE	Demonstração de Resultado do Exercício
DUC	*Document Understanding Conference*
DVA	Demonstração do Valor Adicionado
EDGAR	*Electronic Data Gathering, Analysis, and Retrieval system*
EUA	Estados Unidos da América
FN	Falsas Negativas
FP	Falsas Positivas
IASB	*International Accounting Standards Board*
IR	*Information Retrieval*
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery from Texts*
K-NN	*K Nearest Neighbors*
LCS	*Longest Common Subsequence*
LSA	*Latent Semantic Analysis*
NIST	*National Institute of Standards and Technology*
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
RST	*Rhetorical Structure Theory*
RU	*Relative Utility*
SCUs	*Summarization Content Units*
SGBDs	Sistemas de Gerenciamentos de Bancos de Dados
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency – Inverse Document Frequency*
VN	Verdadeiras Negativas
VP	Verdadeiras Positivas
XBRL	*eXtensible Business Reporting Language*
XML	*eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

As demonstrações contábeis, também usualmente chamadas de demonstrações financeiras, apresentam o desempenho financeiro das empresas e são uma importante ferramenta para análise da situação patrimonial e financeira das mesmas, bem como para tomada de decisões de investidores, credores, fornecedores, clientes, entre outros.

Segundo o Comitê de Pronunciamentos Contábeis do Brasil[1] (CPC), entidade que é responsável pelo estudo, preparo e emissão de pronunciamentos técnicos de contabilidade no Brasil, as demonstrações contábeis são uma representação estruturada da posição patrimonial e financeira e do desempenho da entidade, e tem o objetivo de proporcionar informações a um grande número de usuários em suas avaliações e tomada de decisões econômicas.

As notas explicativas, também chamadas de notas de rodapé, são uma parte inseparável das demonstrações financeiras das empresas e desempenham um papel fundamental no processo de análise destas demonstrações. Elas descrevem, em detalhes, as práticas e políticas de comunicação de métodos de contabilidade da empresa e divulgam informações adicionais que não podem ser mostradas nas próprias declarações. Em outras palavras, as notas de rodapé expandem os dados sobre as demonstrações financeiras quantitativas, fornecendo informações qualitativas, o que permite uma maior compreensão do verdadeiro desempenho financeiro de uma empresa durante um período de tempo especificado (SCHWAB, 2016). Elas são extremamente valiosas para o analista financeiro, que pode discernir a partir das notas de rodapé como as várias políticas contábeis utilizadas por uma empresa estão impactando seus resultados relatados e posição financeira (BRAGG, 2016).

Apesar de tamanha importância, a análise de forma automática das notas explicativas é um obstáculo, pois são redigidas em texto livre e possuem um formato não estruturado. Sendo necessário, dessa forma, um trabalho manual árduo para poder analisá-las (HEIDARI; FELDEN, 2015a).

Estudos realizados por Heidari e Felden (2014) demonstraram a relevância das notas de rodapé e a demanda pelas informações presentes nelas. Além disso, seus estudos afirmam que baseado em diferentes casos de uso, o principal método para análise das notas de rodapé ainda está sendo a leitura manual e não há nenhum método automático identificado para integrar os métodos de análise financeira de valores estruturados com notas de rodapé não estruturadas. Apesar da utilização do padrão XBRL[2] para intercâmbio de dados unificados de relatórios financeiros (JANVRIN; MASCHA, 2010) e mesmo com a marcação detalhada de notas, a análise automática de informações entre a narrativa e as notas de rodapé não é possível.

Existem várias bases com dados estruturados das demonstrações financeiras, entre elas a EDGAR[3] da Comissão de Valores Mobiliários dos Estados Unidos, a Economatica[4], e a Comissão de Valores Mobiliários do Brasil[5] (CVM), com dados das empresas que poss'uem ações na bolsa de valores do Brasil. Porém nenhuma dessas bases estruturadas contempla as informações constantes nas notas explicativas das demonstrações financeiras.

---

[1] Comitê de Pronunciamentos Contábeis do Brasil, disponível em <http://www.cpc.org.br>.

[2] Padrão baseado em XML para definir a informação financeira.

[3] Sistema de coleta eletrônica, análise e recuperação, em inglês *Electronic Data Gathering, Analysis, and Retrieval system* (EDGAR), da Comissão de Valores Mobiliários dos EUA, disponível em <http://www.sec.gov>.

[4] Banco de dados do mercado de capitais de empresas de vários países das Américas pertencente à Economatica que é uma empresa especializada em sistemas de análise financeira, disponível em <http://economatica.com>.

[5] Comissão de Valores Mobiliários do Brasil, disponível em <http://www.cvm.gov.br>.

A maioria das empresas pode escolher métodos para apuração de determinadas contas contábeis, como: receita, despesa, estoque, entre outras. O método utilizado pela empresa para apuração destas contas deve ser claramente identificado em notas explicativas na seção de principais práticas de contabilidade utilizadas por ela e brevemente explicado (PUTRA, 2008). O método de apuração utilizado pela empresa é de extrema importância para análise, pois interfere diretamente nos resultados financeiros apresentados, porém para identificação deste método é necessária a leitura das notas explicativas.

1.1 Objetivo

Tendo em vista a deficiência relativa à manipulação automática das notas explicativas das demonstrações financeiras das empresas, este trabalho tem como objetivo realizar uma investigação teórica e prática sobre formas de identificar e estruturar, de maneira automática, os métodos de apuração das principais contas contábeis utilizados pela empresa, bem como gerar um resumo das notas explicativas da seção de principais práticas de contabilidade utilizadas, com o uso de técnicas de mineração textual.

Para isso, definem-se como objetivos específicos do trabalho as seguintes tarefas:

- Consolidação de uma base de dados com as notas explicativas das principais práticas contábeis e sua respectiva classificação a partir das demonstrações financeiras de empresas disponíveis na CVM do Brasil;
- Realização dos experimentos de extração de conceitos e sumarização sobre a base de dados consolidada;
- Avaliação dos resultados obtidos na extração de conceitos para detecção do método de contabilidade utilizado pela empresa, por meio da validação com os rótulos previamente definidos;
- Avaliação dos resumos utilizando a medida ROUGE (LIN, 2004) que avalia a qualidade dos mesmos;
- Produção de uma análise crítica dos resultados obtidos, apontando situações em que a metodologia se mostrou eficiente e situações em que não foram obtidos bons resultados.

1.2 Organização do Trabalho

Este trabalho apresenta no capítulo 2 o estudo sobre as demonstrações financeiras das empresas, além de conceitos de mineração textual, extração de conceitos e sumarização automática de textos. O capítulo 3 relata alguns trabalhos relacionados com a extração de dados de notas explicativas de demonstrações financeiras. No capítulo 4 é apresentada a abordagem experimental que envolve a metodologia empregada, bem como os experimentos realizados e análise dos resultados. As considerações finais e conclusões são apresentadas no capítulo 5.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são abordados os principais conceitos tratados nesse trabalho. A seção 2.1 apresenta o processo de descoberta de conhecimento em modo geral. Na sequência, a seção 2.2 detalha as particularidade do processo de descoberta de conhecimento em textos.

Na seção 2.3, são abordadas as etapas do processo de mineração textual, e a extração de conceitos e sumarização automática de textos são abordados nas seções 2.4 e 2.5, respectivamente. A seção 2.6 apresenta e define as métricas de avaliação para a extração de conceitos e sumarização. Já as demonstrações financeiras das empresas são detalhadas na seção 2.7, com destaque para as notas explicativas e, por fim, na seção 2.8 são feitas as considerações.

2.1 Processo de Descoberta de Conhecimento

A capacidade de conseguir adquirir, interpretar as informações e poder tomar as decisões adequadas, sempre foi um fator de diferencial para o sucesso. Embora o conhecimento sempre tenha sido necessário, sua importância aumentou gradativamente com o desenvolvimento da ciência e da tecnologia, que nos conduziu à Era do Conhecimento em substituição a Era Industrial.

O conhecimento é um conceito bastante abstrato e muitas vezes é utilizado como sinônimo de informação. Porém existem algumas diferenças entre esses dois conceitos. Informação é a interpretação dos dados, com significado, o qual pode ser diferente para cada pessoa. Já o conhecimento é o que permite e o que é necessário para esta interpretação, ou seja, o conhecimento tem a função de transformar dados em informação (AAMODT; NYGÅRD, 1995). Em outras palavras, pode-se dizer que a informação é a fonte para o conhecimento, ao mesmo tempo que o conhecimento é necessário para extrair novas informações.

Na computação, a descoberta de conhecimento de um modo geral não é uma área nova. Segundo Rich e Knight (1993), ela surgiu na inteligência artificial a partir de projetos voltados ao Aprendizado de Máquina, se preocupando não somente em descobrir conhecimento, mas também, em descobrir formas de aquisição e armazenamento deste conhecimento.

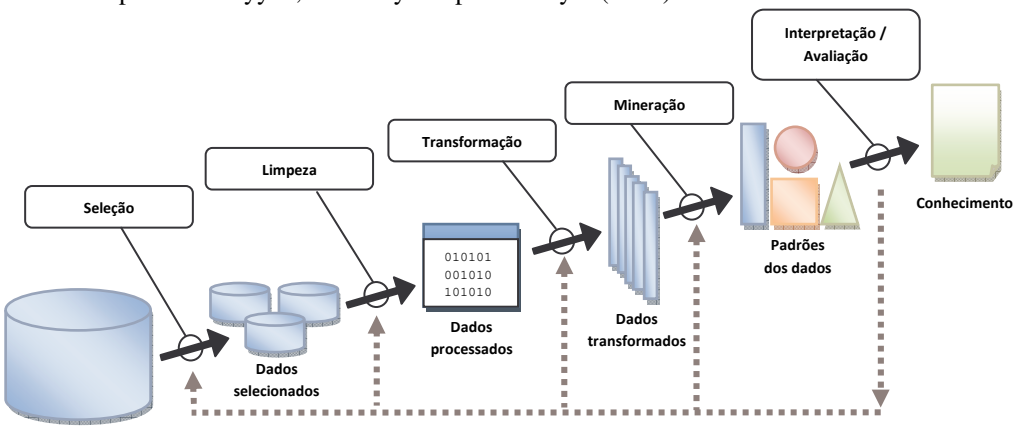
Como uma aplicação prática, os pesquisadores da área de Sistemas de Gerenciamento de Banco de Dados (SGBDs) viram que além das informações tradicionais armazenadas nesses sistemas, poderiam descobrir informações implícitas (que não estavam disponíveis de forma clara) que pudessem ser úteis e assim passaram a pesquisar novas aplicações para estas informações armazenadas.

Com isso, começaram a surgir os primeiros sistemas de análise de dados e de mineração de dados relacionais, dando início às pesquisas na área de Descoberta de Conhecimento em Banco de Dados, em inglês *Knowledge Discovery in Databases* (KDD).

O KDD, segundo Han, Pei e Kamber (2011), é a extração automatizada ou conveniente de padrões que representam o conhecimento implicitamente armazenado ou capturados em grandes bases de dados, *data warehouses*, da Web, entre outros repositórios de informação em massa ou fluxos de dados. Já Fayyad, Piatetsky-Shapiro e Smyth (1996) o definem como um processo não trivial de identificar padrões válidos, não conhecidos, potencialmente úteis e interpretáveis.

Figura 1: Fluxo básico das etapas do KDD.

Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).



Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), a descoberta de conhecimento é um processo que envolve diversas etapas que vão desde a limpeza dos dados, passando pela aplicação de um algoritmo de extração de dados e finalizando com a análise e interpretação dos resultados. Segundo eles, o processo de KDD é interativo e iterativo, envolvendo vários passos com muitas decisões feitas pelo utilizador. Eles amplamente delimitam o processo em nove passos:

1. O entendimento do domínio da aplicação e do conhecimento relevante já existente e identificação do processo de KDD do ponto de vista do cliente;
2. A criação de um conjunto de dados de destino, que se baseia em selecionar um conjunto de dados ou simplesmente um conjunto de variáveis ou amostras de dados onde o processo de descoberta será executado;
3. A limpeza dos dados e pré-processamento. As operações básicas desta etapa incluem a remoção de ruídos, se for o caso, a coleta de informações para definir ou representar algum ruído e decidir estratégias para lidar com alguma possível falta de dados;
4. A redução de dados e projeção. Através de métodos de redução de dimensionalidade ou de transformação, o número efetivo de variáveis utilizado pode ser reduzido, ou representações invariáveis para os dados podem ser encontradas;
5. O objetivo do KDD, onde é selecionado um método de mineração de dados. Conforme o objetivo, pode-se selecionar um algoritmo de sumarização, classificação, agrupamento, entre outros;
6. A análise exploratória e a seleção e parametrização do algoritmo de mineração de dados que será utilizado em busca de padrões de dados;
7. A execução do processo de mineração de dados. Todas as etapas anteriores auxiliam significativamente no desempenho do processo;
8. Os resultados obtidos da mineração devem ser interpretados e, se necessário, retornar a qualquer uma das etapas anteriores para outra iteração;
9. O conhecimento é descoberto e pode ser usando diretamente ou incorporado em outro sistema para novas ações, ou simplesmente documentando e apresentando às partes interessadas.

O processo de KDD envolve significativa interação e pode conter laços entre os passos. É um processo cíclico e, ao final de cada etapa, os resultados devem ser analisados e caso não

sejam satisfatórios ou não estejam devidamente refinados, deve-se realizar ajustes no processo para a realização de um novo ciclo. A Figura 1 apresenta o fluxo básico das etapas do KDD.

2.2 Processo de Descoberta de Conhecimento em Textos

A maior parte dos dados disponível no mundo não está armazenada de forma estruturada em tabelas de bancos de dados relacionais. Ao invés disso, a maior parte dos dados digitalmente disponíveis está armazenada em arquivos texto não estruturados. O processo de descoberta de conhecimento em textos, do inglês *Knowledge Discovery from Texts* (KDT), sinônimo de mineração de textos e garimpagem de textos, envolve uma série de tecnologias para análise de textos não estruturados com o objetivo de organizar, estruturar, extrair informações e até mesmo descobrir conhecimento oculto nesses textos.

A mineração de textos é uma excitante área de pesquisa computacional que tenta resolver a crise da sobrecarga de informação, combinando técnicas de mineração de dados, aprendizado de máquina, processamento de linguagem natural, recuperação de informação e gestão do conhecimento (FELDMAN; SANGER, 2007). Ela pode ser entendida como a aplicação de técnicas de KDD sobre dados extraídos de textos (não necessariamente valores numéricos, mas podendo ser também valores nominais, como palavras do texto) (FELDMAN; HIRSH,1997).

De acordo com Miner et al. (2012), mineração de textos é um termo amplo que descreve uma gama de tecnologias para a análise e processamento de dados de texto semi-estruturados e não estruturados. O foco dessas tecnologias está na necessidade de transformar o texto em dados que o computador consiga entender.

Como na mineração de dados, a mineração de textos visa a extrair informações úteis buscando por padrões em fontes de dados. Porém na mineração de textos, as fontes de dados são documentos textuais e os padrões interessantes não são encontrados nos registros dos bancos de dados e sim nos dados textuais não estruturados em linguagem natural destes documentos. Sendo assim, a mineração textual também se baseia nos avanços feitos em outras áreas da ciência da computação ligados à manipulação da linguagem natural (FELDMAN; SANGER, 2007).

O processo de descoberta de conhecimento em textos não se detém somente em encontrar os textos que contenham informações e deixar que o usuário procure o que lhe interessa, e sim, se preocupa em encontrar informações dentro dos textos e tratá-las de forma a apresentar ao usuário algum tipo de conhecimento útil e novo. Mesmo que tal conhecimento novo não seja a resposta direta às indagações do usuário, ele deve contribuir para satisfazer as necessidades de informação do usuário.

A mineração textual tem como objetivos descobrir informações (padrões e anomalias) de maneira automatizada em textos, facilitar o manuseio da informação, bem como facilitar o processo de descoberta de conhecimento humano. Para isso, o processo pode extrair informações relevantes em um ou mais textos a partir de termos chaves definidos com o objetivo de estruturar a informação. Além disso, pode agrupar diferentes documentos textuais, por similaridade, com o objetivo de categorizá-los e organizá-los, e pode ainda criar resumos de um ou mais textos, tendo ou não por base um assunto definido, com o objetivo de agilizar o processo de aquisição de conhecimento do usuário. Resumindo, pode-se dizer que o processo de mineração textual cria facilidades ao usuário e pode ser uma importante ferramenta para

organizar, estruturar e gerar conhecimento, tanto para as pessoas como para as organizações de um modo geral.

A mineração textual pode ser aplicada em um único arquivo textual ou a um conjunto de arquivos distintos. Um arquivo textual é uma unidade de dados textuais discretos dentro de uma coleção que normalmente, mas não necessariamente, se correlaciona com algum documento do mundo real como um relatório de negócios, um memorando legal, um e-mail, um relatório de pesquisa, um manuscrito, um artigo, uma notícia, etc.

Apesar dos arquivos em formato texto serem rotulados como não estruturados, a partir de muitas perspectivas ele pode ser visto como estruturado, pois um texto apresenta uma rica quantidade de estruturas semântica e sintática, além de sinais de pontuação, capitalizações, espaços em branco, tabelas, colunas e outros componentes que podem fornecer pistas para ajudar a identificar sub-componentes e informações do texto (FELDMAN; SANGER, 2007).

A mineração textual opera baseada nos recursos dos documentos textuais, sendo os mais utilizados:

- Caracteres: são as letras, símbolos e espaços em branco, individuais, no nível de componente;
- Palavras: podem ser descritas como o nível base da riqueza semântica;
- Termos: são formados por palavras específicas e expressões encontradas no documento que geralmente possuem grande representatividade;
- Conceitos: são recursos gerados manualmente com base em regras, estatísticas ou metodologias de categorização para representar algo significativo e ao contrário dos demais, pode consistir de palavras não encontradas no documento.

2.3 Etapas da Mineração Textual

As etapas do processo de descoberta de conhecimento em textos são as mesmas que as da descoberta de conhecimento em banco de dados, porém com algoritmos específicos para tratar dados textuais não estruturados. A seguir, são descritas as etapas de maior impacto na mineração de textos.

2.3.1 Pré-processamento

Levando-se em conta que o problema abordado nesse trabalho se refere a dados textuais, esta seção explana brevemente algumas técnicas que podem ser aplicadas na etapa de pré-processamento sobre dados textuais, com o objetivo de refinar a informação e melhorar a representação do conhecimento.

Um dos desafios da mineração de textos é a conversão do texto não estruturado e semi-estruturado para um modelo estruturado de espaço vetorial, como representação do conhecimento. Este é o primeiro passo e as possíveis etapas de pré-processamento de texto são as mesmas para todas as tarefas de mineração de textos, no entanto, são escolhidas dependendo da tarefa. Conforme Miner et al. (2012), os passos básicos da etapa de pré-processamento são os seguintes:

- Escolha do âmbito do texto a ser processado: determina se o processo de mineração vai ser aplicado em todo o arquivo textual de uma só vez ou se este vai ser divido em partes;
- *Tokenização*: consiste em quebrar as unidades de texto em palavras individuais ou *tokens*;
- Remoção de *stopwords*: consiste em remover as palavras comuns que não têm muita representatividade, com base em um dicionário de palavras, a fim de economizar espaço de armazenamento e acelerar o processamento;
- *Stemming*: também chamado de *lematização*, consiste em remover prefixos e sufixos para normalizar as palavras em um único radical;
- Normalização ortográfica: consiste em corrigir erros ortográficos automaticamente a fim de unificar erros ortográficos e outras variações de ortografia em um único *token*;
- Detectar o limite das frases: marcar o fim das frases;
- Normalizar maiúsculas e minúsculas: deixar o texto todo em maiúsculo ou todo em minúsculo, com exceção dos nomes próprios.

2.3.2 Transformação dos dados

Depois do pré-processamento, os *tokens* que representam as palavras devem ser transformados em uma representação vetorial apropriada, que servirá de entrada para o algoritmo de mineração textual. Esta representação vetorial pode ser de três formas: binária, com contador inteiro de frequência ou como um vetor com pesos representativos de valores decimais. A representação binária indica a presença ou ausência de um determinado termo no texto; a inteira por frequência indica o número de vezes que o termo aparece no texto; e a baseada em pesos se utiliza de um esquema para calcular os pesos. Para o cálculo de pesos, as duas formas mais empregadas são:

- TF-IDF (*Term Frequency – Inverse Document Frequency*): É o esquema mais utilizado e atribui para cada termo da base textual um peso que representa a sua importância considerando todo o *corpus* (MANNING et al., 2008). Este peso aumenta proporcionalmente ao número de vezes que o termo aparece no documento e é compensado pela frequência do termo no *corpus* (SALTON; ALLAN, 1994). A contagem de termos *tf(i,j)* indica o número de vezes que um termo t_i aparece em um documento d_j, porém é normalizada para evitar polarização em documentos longos, sendo definida por:

$$tf(i,j) = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{2.1}$$

onde $n_{i,j}$ é o número de ocorrências do termo t_i no documento d_j, e o denominador é o número de ocorrências de todos os termos k pertencentes ao documento d_j.

A frequência inversa do documento dada por *idf(i)*, na Equação (2.2), representa a importância geral do t_i, obtida pelo logaritmo do quociente entre o número total de documentos D e o número d de documentos que contém o termo t_i.

$$idf(i) = log\frac{|D|}{|\{d: t_i \epsilon\, d\}|} \tag{2.2}$$

Com a relação das medidas das Equações (2.1) e (2.2) é calculado o valor *TF-IDF* para o termo t_i, dado pela Equação (2.3), com a suposição de que palavras que aparecem com maior frequência devam receber peso maior, a menos que sua frequência no conjunto de documentos também seja elevada.

$$tf - idf(i,j) = tf_{i,j} \times idf_i \tag{2.3}$$

- Modelo Bayesiano: Outro esquema de cálculo é a transformação baseada no teorema de Bayes. O teorema de Bayes afirma que a probabilidade de ocorrência de um evento é igual à probabilidade intrínseca (calculada a partir dos dados presentes) vezes a probabilidade de que isso vai acontecer novamente no futuro (com base no conhecimento de sua ocorrência no passado). O modelo bayesiano faz uma suposição muito simplista (*naïve*) de que todos os objetos a serem classificados são completamente independentes uns dos outros. Apesar de sua simplicidade e sua suposição ingênua (que quase nunca ocorre no mundo real), esses classificadores podem ser extremamente eficientes e precisos, particularmente quando o número de variáveis é elevado (MINER; ELDER IV; HILL, 2012).

2.4 Extração de Conceitos

A extração de conceitos tem como objetivo identificar palavras que partilham do mesmo significado semântico, usando o contexto em que as palavras aparecem (MINER; ELDER IV; HILL, 2012). Por exemplo, um conceito poderia ser uma coleção de palavras que definem animais domésticos: cães, gatos, peixes, aves, hamsters, e assim por diante.

Segundo Faber et al. (1994), a premissa subjacente de extração de conceito é que, a fim de interpretar dados, os seres humanos rápida e naturalmente extraem e identificam os conceitos significativos. Uma pessoa pode contemplar uma paisagem, receber dados através dos órgãos dos sentidos, e pode então fazer um julgamento razoável sobre se vai chover, por exemplo, ou se vai nevar. Uma pessoa pode examinar a tabela de uma revista de conteúdos e facilmente selecionar os artigos que dizem respeito a um assunto de interesse. Humanos estão constantemente assimilando, categorizando, sintetizando e analisando dados na maioria das vezes sem qualquer consciência de que estão a fazê-lo. A grande complexidade, subjetividade e ambiguidade dos conceitos humanos os fazem extremamente difíceis, se não impossíveis de serem definidos de forma quantitativa adequada para a utilização por computadores.

De acordo com Parameswaran, Garcia-Molina e Rajaraman (2010), conceitos são sequências de palavras que representam entidades reais ou imaginárias ou ideias que os usuários estão interessados.

Tradicionalmente, a conversão de palavras em conceitos é baseada em um dicionário de sinônimos. Os dicionários utilizados são especificamente criados para a tarefa ou são um modelo de linguagem pré-existente. Isso faz o processo de extração de conceitos ser totalmente dependente de domínio, pois os conceitos foram criados para aquele objetivo em especial.

Os conceitos podem ser representados utilizando-se vários modelos. Dentre os mais utilizados está o modelo binário, modelo de espaço de vetores e o modelo contextual. Porém para todos os casos, parte-se do pressuposto de que conceitos são expressos por palavras, mas que as palavras sozinhas não são adequadas para representar um conceito (CHEN et al., 1994). Dessa forma, um conjunto suficiente de termos ou palavras deve ser utilizado para representar cada conceito. Em todos os modelos, os termos descritores de um conceito podem incluir

sinônimos, quase sinônimos (palavras semanticamente relacionadas), variações léxicas (conjugações verbais, verbos e substantivos correlatos), entre outros.

O modelo binário é o mais simples de todos e nele cada conceito é representado por um vetor de termos simples, sem conexão e com um indicador da presença ou ausência de determinado termo, para definir um conceito.

Outro modelo simples é o modelo espaço de vetores que facilita as tarefas de definição e identificação dos conceitos. Nesse modelo, cada conceito é representado por um vetor de termos simples, sem conexão entre eles e com um peso associado a cada termo, que indica seu grau de importância para descrever ou identificar o conceito.

Já no modelo contextual, a relação entre os termos influencia na representação do conceito e minimiza o problema de interpretações erradas. Nesse modelo, a análise do contexto em que os termos aparecem é que descreve ou identifica o conceito.

2.5 Sumarização Automática de Textos

Os primeiros trabalhos de sumarização automática de texto remontam à década de 1950 (STEINBERGER; JEŽEK, 2009). A sumarização automática de texto tem como objetivo a criação de um resumo, que segundo Radev, Hovy e Mckeown (2002), é um texto produzido a partir de um ou mais textos, transmitindo as informações importantes desses textos e que não possui mais de metade do tamanho do(s) texto(s) original(is) ou significativamente menos do que isso. O objetivo principal de um resumo é apresentar as principais ideias em um documento em menos espaço. Conforme Nenkova e Mckeown (2012), a sumarização automática de textos deve produzir um resumo conciso e transmitir com fluência o tema principal do texto.

Como uma ferramenta de mineração textual, a sumarização automática de textos teve seu interesse renovado recentemente, pois o resumo pode ajudar a lidar com a sobrecarga de informações existente. Além disso, o aumento no uso de dispositivos móveis, que muitas vezes são utilizados em movimento e com déficit de atenção, seria importante apresentar ao usuário apenas as informações mais relevantes, isto é, um resumo, para que se interessado, solicite em seguida o texto inteiro (SANKARASUBRAMANIAM; RAMANATHAN; GHOSH, 2014).

2.5.1 Principais abordagens da sumarização automática

A literatura apresenta diversas abordagens de sumarização automática de textos. As principais abordagens de acordo com Steinberger e Ježek (2009) são relacionadas a seguir:

- Abordagem ao nível de superfície: Esta abordagem de sumarização se baseia na análise da superfície do documento de texto, através da extração de partes deste por critérios de relevância, como em Edmundson (1969). Um dos critérios mais utilizados é a frequência de um determinado termo no documento, onde a relevância do termo é considerada proporcional à sua frequência no documento, ou seja, o termo que aparece com maior frequência indica o tema principal do documento. Outros indicadores de relevância utilizados são a posição de uma frase no documento, a presença de palavras do título ou certas palavras-chave como, por exemplo, “importante”, “relevante”;
- Abordagem baseada no *corpus*: Essa abordagem parte da ideia de que os termos muito comuns no documento não carregam informações importantes e, portanto sua

relevância deve ser reduzida. Dessa forma, para calcular a relevância dos termos, essa abordagem utiliza na etapa de transformação dos dados o esquema *TF-IDF* (SALTON; 1989) ou modelo bayesiano (KUPIEC; PEDERSEN; CHEN, 1995) vistos na seção 2.3.2;

- Abordagem baseada em coesão: Os métodos de extração podem deixar de capturar as relações entre conceitos em um texto, relações essas expressas através de expressões anafóricas[6] que remetem a eventos e entidades no texto que precisam de seus antecedentes para serem compreendidas. Um resumo pode se tornar difícil de entender se uma frase que contém uma ligação anafórica for extraída sem o seu contexto anterior. Dessa forma, essa abordagem é baseada na coesão textual, que compreende as relações entre expressões que determinam a conectividade do texto, como em Barzilay e Elhadad (1999);
- Abordagem baseada na teoria da estrutura retórica: Teoria da estrutura retórica, em inglês *Rhetorical Structure Theory* (RST), é uma teoria descritiva que tem por objetivo o estudo da organização dos textos, caracterizando as relações que se estabelecem entre as partes do mesmo. A estrutura é constituída a partir das relações retóricas que unem as unidades de texto. Nesta teoria, os textos são formados por grupos organizados de orações que se relacionam hierarquicamente, as quais podem ser descritas com base na intenção comunicativa do enunciador e na avaliação que o enunciador faz do enunciatário, e refletem as escolhas do enunciador para organizar e apresentar os conceitos, como em Ono, Sumita e Miike (1994). A maioria das relações retóricas que se estabelecem é do tipo núcelo-satélite, onde uma parte do texto serve de subsídio para outra. A representação destas relações compõe uma árvore, de onde as unidades de texto são extraídas para o resumo;
- Abordagem baseada em grafo: Nessa abordagem, um grafo é construído através da adição de um vértice para cada frase no texto, e arestas entre vértices são estabelecidas usando a interconexão das frases. Estas conexões são definidas utilizando uma relação de similaridade, que é medida através de uma função de sobreposição. A sobreposição de duas frases pode ser definida como o número de *tokens* que existem em comum entre as representações lexicais das duas frases. Com o grafo construído, um algoritmo de classificação é executado e as sentenças são ordenadas em ordem inversa de pontuação, como em Mihalcea e Tarau (2005). Ao final as frases melhores classificadas são incluídas no resumo;
- Abordagem além da extração de sentenças: Os resumos produzidos por sumarizadores automáticos a partir da extração de frases se distinguem muito dos escritos por profissionais humanos. Isso porque nem sempre os sistemas conseguem identificar corretamente os temas importantes do texto. Ainda, a maioria dos sumarizadores automáticos simplesmente extrai as frases identificadas como principais. Porém, se as frases extraídas estiverem desconectadas do texto original e forem anexadas ao resumo, o resultado pode ser incoerente ou até enganoso.

  Os resumos produzidos pela abordagem além da extração de sentenças são os que mais se assemelham aos produzidos pelos humanos e Jing e McKeown (2000) introduzem um método com essa abordagem.

[6] Expressão anafórica é uma palavra ou frase que remete para alguma palavra expressada anteriormente (tipicamente, os pronomes, como ele, ela).

2.6 Métricas de Avaliação dos Resultados

Os resultados de cada uma das técnicas que são utilizadas, extração de conceitos e sumarização, serão avaliados com métricas diferentes. Como para a extração de conceitos é possível identificar se houve acerto ou erro, pode-se usar o esquema de Matriz de Confusão. Isso já não é possível para a sumarização, onde só é possível avaliar o quão bom foi o resultado. As métricas de avaliação para ambas as técnicas são descritas nas seções a seguir.

2.6.1 Métricas de Avaliação para a Extração de Conceitos

Os resultados do processo de Extração de Conceitos podem ser comparados com os valores desejados e sendo possível dessa forma, para cada classe, a obtenção do número de amostras classificadas de forma correta e incorreta. Para isso, existem métricas que podem ser calculadas pelos valores de VP, VN, FP e FN, abaixo descritos, que são utilizados para montar a Matriz de Confusão, conforme Tabela 1.

Tabela 1: Matriz de Confusão para duas classes

Resultado desejado	Resultado predito	
	Positivo	Negativo
Positivo	VP	FN
Negativo	FP	VN

Fonte: Elaborado pelo autor.

Em casos de polaridade, amostras positivas classificadas corretamente compõem o valor de Verdadeiras Positivas (VP) e as negativas incorretamente preditas como positivas o de Falsas Positivas (FP). De forma análoga, as amostras classificadas corretamente como negativas fazem parte do valor de Verdadeiras Negativas (VN) e as positivas preditas incorretamente como negativas compõem o de Falsas Negativas (FN) (HAN; PEI; KAMBER, 2011).

Através dos componentes da Matriz de Confusão, podem ser calculadas diferentes métricas (LIU, 2011), das quais destaca-se as que são utilizadas neste trabalho:

- acurácia: é o resultado da relação entre amostras classificadas corretamente e o total de amostras;
- especificidade (*precision*): determina o quanto o classificador é capaz de rejeitar as outras classes (em relação a classe a qual a *precision* foi calculada);
- sensibilidade (*recall*): apresenta a capacidade de classificação de amostras pertencentes à classe analisada;
- *f-measure*: também chamada de *f-score*, é a média harmônica entre *precision* e *recall*.

As fórmulas das métricas apresentadas são expostas nas equações abaixo listadas.

$$acuracia = \frac{VP+TN}{VP+FP+VN+FN} \tag{2.4}$$

$$precision = \frac{VP}{VP+FP} \tag{2.5}$$

$$recall = \frac{VP}{VP+FN} \tag{2.6}$$

$$f\text{-}measure = \frac{2 \times recall \times precision}{recall+precision} \tag{2.7}$$

2.6.2 Métricas de Avaliação para Sumarização

Avaliar a qualidade de um resumo não é uma tarefa trivial, principalmente porque não há nenhuma indicação para o que é um resumo ideal (RADEV; HOVY; MCKEOWN, 2002). Tradicionalmente, a avaliação de resumos envolve julgamentos humanos com diferentes métricas de qualidade, por exemplo, a coerência, concisão, gramaticalidade, legibilidade e conteúdo (MANI, 2001).

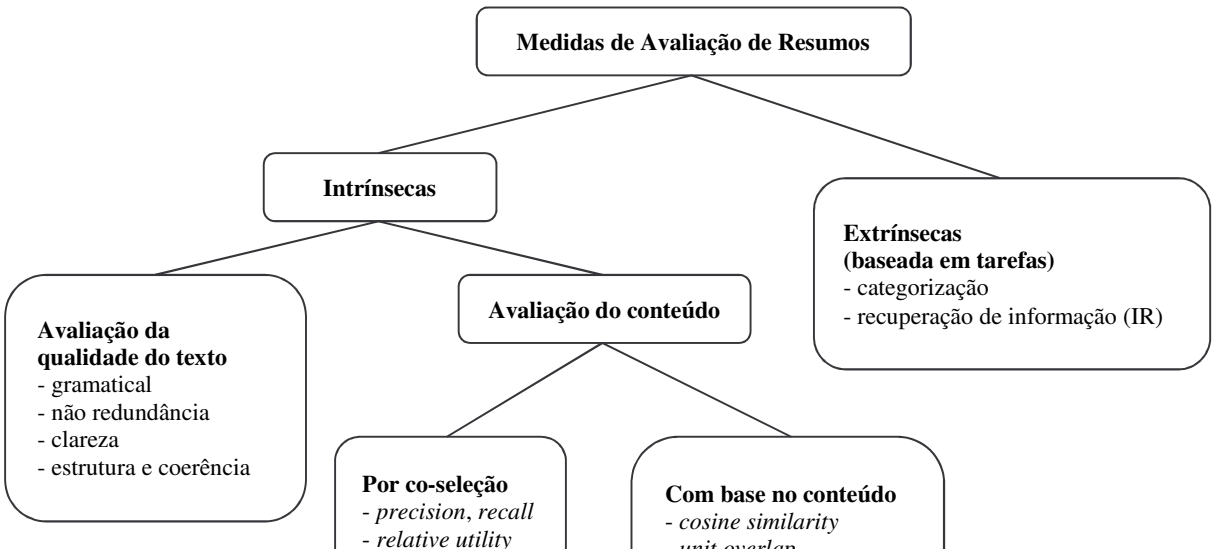
A taxonomia das medidas de avaliação de resumo é apresentada na Figura 2, a qual mostra que os resumos podem ser avaliados pela qualidade do texto, pelo seu conteúdo ou ainda com base em tarefas. Quanto à qualidade, a avaliação do texto do resumo é frequentemente feita por humanos e não pode ser feita automaticamente.

Já quanto ao conteúdo, a principal abordagem para determinação da qualidade do resumo é a avaliação intrínseca do conteúdo, que muitas vezes é feita por comparação com um resumo ideal (STEINBERGER; JEŽEK, 2009).

Os resumos por extração de frases podem ser medidos por co-seleção, que indica quantas frases que representariam um resumo ideal, definido por juízes humanos, foram selecionadas pelo resumo automático. Para esse fim, as principais medidas são *precision*, *recall* e *f-score*. O principal problema dessas métricas está no fato de que juízes humanos muitas vezes discordam sobre quais frases são as mais importantes em um documento, fazendo com que dois extratos de resumos igualmente bons sejam julgados de forma diferente.

Para resolver o problema com *precision* e *recall*, Radev, Jing e Budzikowska (2000) introduziram *Relative Utility* (RU), onde o resumo modelo apresenta todas as sentenças do documento original com valores de importância para sua inclusão no resumo. Como exemplo, imagina-se um documento com cinco frases [1 2 3 4 5] que fica representado no resumo modelo como [1/5 2/4 3/4 4/1 5/2], onde o segundo número, chamado de número de utilidade, representa a importância da frase para inclusão no resumo, de acordo com um juiz humano, sendo que os valores maiores indicam as frases de maior importância. Nesse exemplo, o sistema que selecionar as frases [1 2] não vai obter pontuação maior que outro que selecionar as frases [1 3], pois ambos somam a mesma pontuação considerando o número de utilidade (5 + 4) = 9. Com isso, dois resumos diferentes, mas igualmente bons, são julgados da mesma forma.

Figura 2: Taxonomia das medidas de avaliação de resumos.



Fonte: Adaptado de Steinberger e Ježek (2009).

O método de avaliação por medidas de co-seleção pode avaliar situações em que se usam exatamente as mesmas frases. Porém, isso ignora o fato de que duas frases podem conter a mesma informação, mesmo que escritas de maneira diferente. As medidas de co-seleção não conseguem fazer essa avaliação, mas medidas de similaridade baseadas em conteúdo podem. As medidas de similaridade baseadas no conteúdo são: *Cosine Similarity*, cosseno de similaridade, introduzido por Salton (1989); *Unit Overlap*, por Saggion et al. (2002); *Longest Common Subsequence* (LCS), que se baseia na maior subsequência comum introduzido por Radev et al. (2003); *Pyramid*, introduzido por Nenkova e Passonneau (2004), cuja ideia básica é identificar unidades de conteúdo de sumarização, do inglês *summarization content units* (SCUs) que são usadas para comparação de informações nos resumos; *Latent Semantic Analysis* (LSA), que é um método de avaliação automática de resumos, proposto por Steinberger e Ježek (2009), baseado na análise semântica latente onde os resumos são classificados de acordo com a similaridade dos temas entre o resumo e o documento original; e ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), por Lin e Hovy (2003), é melhor descrito a seguir por ser o método de avaliação a ser utilizado neste trabalho.

ROUGE é uma família de medidas de avaliação baseada na similaridade de N-gramas[7] entre um ou mais resumo(s) modelo(s) e o resumo automático gerado. É o método de avaliação de resumos mais popular, já consolidado e está sendo utilizado nas edições do DUC[8] para avaliação automática dos resumos e é calculado seguindo a expressão:

$$ROUGE - N = \frac{\sum_{S \in \{resumo\ de\ referência\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{resumo\ de\ referência\}} \sum_{gram_n \in S} Count(gram_n)}, \quad (2.8)$$

onde *n* representa o comprimento do N-grama, e $Count_{match}(gram_n)$ é o número máximo de N-gramas que co-ocorrem no resumo candidato e no resumo de referência e $Count\ (gram_n)$ é o número de N-gramas no resumo de referência. É claro que ROUGE-N é uma medida relacionada com o *Recall*, pois o denominador da equação é a soma total do número de N-gramas que ocorrem no lado do resumo de referência (LIN; HOVY, 2003). O resultado obtido com a aplicação do ROUGE vai de zero a um e a proximidade com o um, indica maior similaridade com o resumo de referência e consequentemente o melhor resumo.

---

[7] Um N-grama é uma subsequência de n palavras a partir de um determinado texto (FIGUEROA; ATKINSON, 2012)

[8] O Instituto Nacional de Padrões e Tecnologia em inglês *National Institute of Standards and Technology* (NIST) criou uma Conferência para compreensão de documentos, em inglês *Document Understanding Conference* (DUC) para avaliar a sumarização automática de texto, onde o objetivo é promover o progresso da sumarização e permitir aos pesquisadores a troca de experiências em larga escala.

Já os métodos baseados em tarefas não analisam as sentenças do resumo e sim sua perspectiva na utilização para uma determinada tarefa. Nesse sentido, um resumo pode ser avaliado de acordo com sua aptidão para categorização de documentos, onde é verificado se a correta classificação do documento pode ser obtida com base no resumo. Outra tarefa para avaliação dos resumos é a recuperação de informação, do inglês *Information Retrieval* (IR), em que se usa a correlação relevante, que é uma medida de IR, para avaliar a diminuição relativa do desempenho de recuperação de informação, quando se desloca de documentos completos para o resumo.

2.7 Demonstrações Financeiras das Empresas

As demonstrações contábeis, também chamadas de demonstrações financeiras, apresentam o desempenho financeiro das empresas e são uma importante ferramenta para análise da situação patrimonial e financeira, bem como para tomada de decisões de investidores, credores, fornecedores, clientes, entre outros.

Conforme o Comitê de Pronunciamentos Contábeis (CPC), as demonstrações contábeis representam de forma estruturada a posição patrimonial e financeira de uma entidade, bem como seu resultado e fluxo financeiro, que são úteis para uma ampla variedade de usuários na tomada de decisões.

Todas as sociedades por ações são obrigadas a elaborar e publicar suas demonstrações financeiras (MARTINS; ASSAF NETO, 1996). No Brasil, com a Lei nº 11.638/2007, que produziu efeito a partir do exercício de 2008, todas as demonstrações contábeis devem seguir os padrões contábeis internacionais, além de estabelecer para a CVM o poder/dever de emitir normas para as companhias abertas em consonância com esses padrões internacionais. Em função do disposto no § 5º do artigo 177 adicionado pela Lei nº 11.638/2007, as normas contábeis emitidas pela CVM deverão estar obrigatoriamente em consonância com os padrões contábeis internacionais adotados nos principais mercados de valores mobiliários, ou seja, de acordo com as normas emitidas pelo *International Accounting Standards Board* (IASB), que hoje é considerado como a referência internacional dos padrões de contabilidade.

Conforme o CPC, para satisfazer seu objetivo, as demonstrações contábeis proporcionam informação da entidade acerca do seguinte:

- ativos;
- passivos;
- patrimônio líquido;
- receitas e despesas, incluindo ganhos e perdas;
- alterações no capital próprio mediante integralizações dos proprietários e distribuições a elas;
- fluxos de caixa.

Essas informações, juntamente com outras informações constantes das notas explicativas, ajudam os usuários das demonstrações contábeis a prever os futuros fluxos de caixa da entidade e, em particular, a época e o grau de certeza de sua geração.

Ainda, de acordo com o CPC, o conjunto completo de demonstrações contábeis é composto por um conjunto de demonstrativos, os quais são definidos de acordo com De

Iudícibus et al. (2010), por ser um livro atual, consolidado com as normas internacionais de contabilidade e escrito por autores renomados da área. Esses demonstrativos são descritos a seguir:

- Relatório da Administração: Apesar de não fazer parte das demonstrações contábeis propriamente ditas, a lei exige a apresentação do Relatório da Administração. Ele deve evidenciar os negócios sociais e principais fatos ocorridos no exercício, os investimentos em outras empresas, a política de distribuição de dividendos e de reinvestimento de lucros etc.;
- Balanço Patrimonial: Tem por finalidade apresentar a posição financeira e patrimonial da empresa em determinada data, representando, portanto, uma posição estática. Conforme o art. 178 da Lei nº 6.404176, "no balanço, as contas serão classificadas segundo os elementos do patrimônio que registrem, e agrupadas de modo a facilitar o conhecimento e a analise da situação financeira da companhia" e é composto por três elementos básicos: o Ativo, que representa os bens e os direitos; o Passivo, que compreende as exigibilidades e obrigações; e o Patrimônio Líquido, que representa o valor líquido da empresa, o valor que os sócios e acionistas tem na entidade;
- Demonstração de Resultado do Exercício (DRE): Tem por finalidade apresentar as receitas e despesas da entidade, bem como o lucro ou prejuízo resultantes. A Lei nº 6.404/76 define o conteúdo do DRE, que deve ser apresentado na forma dedutiva, com os detalhes necessários das receitas, despesas, ganhos e perdas e definindo claramente o lucro ou prejuízo líquido do exercício, e por ação;
- Demonstração do Resultado Abrangente (DRA): Apresenta a soma do resultado do período com os demais lucros ou prejuízos que não foram oriundos da receita e despesa relacionados no DRE. Nela aparecem as demais variações do patrimônio líquido (reservas de reavaliação, certos ajustes de instrumentos financeiros, variações cambiais de investimentos no exterior e outros), que poderão transitar no futuro pelo resultado do período ou irem direto para Lucros ou Prejuízos Acumulados;
- Demonstração das Mutações do Patrimônio Líquido (DMPL): Evidencia a mutação do patrimônio líquido em termos globais (novas integralizações de capital, resultado do exercício, ajustes de exercícios anteriores, dividendos, ajuste de avaliação patrimonial etc.) e em termos de mutações internas (incorporações de reservas ao capital, transferências de lucros acumulados para reservas e vice-versa, etc.);
- Demonstração dos Fluxos de Caixa (DFC): Visa mostrar como ocorreram as movimentações de disponibilidades em um dado período de tempo. Essa demonstração divide todos os fluxos de entrada e saída de caixa em três grupos: os derivados das atividades operacionais, das atividades de investimento e das atividades de financiamento;
- Demonstração do Valor Adicionado (DVA): Tem como objetivo principal informar o valor da riqueza criada pela empresa e a forma de sua distribuição. Não deve ser confundida com a DRE, pois esta tem suas informações voltadas quase que exclusivamente para os sócios e acionistas, principalmente na apresentação do lucro líquido, enquanto a DVA esta dirigida para a geração de riquezas e sua respectiva distribuição pelos fatores de produção (capital e trabalho) e ao governo;

- Demonstrações comparativas: A Lei das Sociedades por Ações obriga a comparação das demonstrações contábeis de dois exercícios. O grande objetivo da comparação é que a análise de uma empresa é feita sempre com vista no futuro. Por isso, é fundamental verificar a evolução passada, e não apenas a situação de um momento;
- Notas explicativas: Compreende as políticas contábeis significativas e outras informações elucidativas. As notas explicativas serão melhor descritas na seção a seguir por serem o objeto de estudo deste trabalho.

2.7.1 Notas Explicativas

As demonstrações contábeis devem ser complementadas por notas explicativas, quadros analíticos ou outras demonstrações contábeis necessárias a plena avaliação da situação e da evolução patrimonial da empresa.

As notas explicativas podem estar expressas tanto na forma descritiva como na forma de quadros analíticos, ou mesmo englobar outras demonstrações contábeis que forem necessárias ao melhor e mais completo esclarecimento dos resultados e da situação financeira da empresa, tais como: demonstração das origens e aplicações de recursos, balanço social e demonstrações contábeis em moeda constante. As notas podem ser usadas para descrever práticas contábeis utilizadas pela companhia, para explicações adicionais sobre determinadas contas ou operações específicas e ainda para composição e detalhes de certas contas. A utilização de notas para dar composição de contas auxilia também a estética do Balanço, pois se pode fazer constar dele determinada conta por seu total, com os detalhes necessários expostos por meio de uma nota explicativa, como no caso de Estoques, Ativo Imobilizado, Investimentos, Empréstimos e Financiamentos e outras contas (DE IUDÍCIBUS et al., 2010). No Apêndice A apresenta-se um exemplo da seção de principais práticas contábeis das notas explicativas, extraídas de uma demonstração financeira.

A legislação contábil enumera o mínimo dessas notas e induz a sua ampliação quando for necessário para o devido esclarecimento da situação patrimonial e dos resultados do exercício. Segundo o CPC, “as demonstrações serão complementadas por notas explicativas e outros quadros analíticos ou demonstrações contábeis necessários para esclarecimento da situação patrimonial e dos resultados do exercício”.

Dentro das notas explicativas, é obrigatória a apresentação da seção das principais práticas contábeis. As principais notas dessa seção são:

- Aplicações financeiras: diz respeito à forma com que os valores que compõem essa conta são contabilizados;
- Direitos e obrigações: trata da forma com que os valores que compões os direitos e obrigações são contabilizados;
- Estoque: indica a forma com que os valores que compõe o estoque são apurados e contabilizados, e pode também conter um detalhamento do estoque;
- Imobilizado: diz respeito à forma com que os valores que compõem o imobilizado são contabilizados e pode também conter um detalhamento do imobilizado;
- Ajuste de avaliação patrimonial: descreve ajustes de avaliação patrimonial ou indica que a empresa nunca efetuou ajuste de avaliação patrimonial;
- Investimentos em empresas coligadas e controladas: descreve participações em outras sociedades ou indica que a empresa não participa de outras sociedades;

- Impostos federais: indica o regime e forma de tributação dos impostos federais.

A seguir, apresentam-se três exemplos de notas explicativas da seção de principais práticas contábeis, com indicação de algumas informações que podem ser extraídas:

- Exemplo 1:

> "Estoques: Os estoques são apresentados pelo menor valor entre o valor de custo e o valor líquido realizável. Os custos dos estoques são determinados pelo método do custo médio. O valor líquido realizável corresponde ao preço de venda estimado dos estoques, deduzido de todos os custos estimados para conclusão e custos necessários para realizar a venda."

Qual o método de apuração do estoque (PEPS, UEPS ou custo médio)? Nesse exemplo, a nota explicativa descreve que o método de apuração do estoque utilizado é o custo médio.

- Exemplo 2:

> "Imobilizado: Terrenos, edificações, imobilizações em andamento, móveis e utensílios e equipamentos estão demonstrados ao valor de custo, deduzidos de depreciação e perdas por redução ao valor recuperável acumuladas. São registrados como parte dos custos das imobilizações em andamento os honorários profissionais e, no caso de ativos qualificáveis, os custos de empréstimos capitalizados de acordo com a política contábil do Grupo. Tais imobilizações são classificadas nas categorias adequadas do imobilizado quando concluídas e prontas para o uso pretendido. A depreciação desses ativos inicia-se quando eles estão prontos para o uso pretendido na mesma base dos outros ativos imobilizados.
> Os terrenos não sofrem depreciação.
> A depreciação é reconhecida com base na vida útil estimada de cada ativo pelo método linear, de modo que o valor do custo menos o seu valor residual após sua vida útil seja integralmente baixado (exceto para terrenos e construções em andamento). A vida útil estimada, os valores residuais e os métodos de depreciação são revisados no fim da data do balanço patrimonial e o efeito de quaisquer mudanças nas estimativas é contabilizado prospectivamente.
> Ativos mantidos por meio de arrendamento financeiro são depreciados pela vida útil esperada, da mesma forma que os ativos próprios, ou por um período inferior, se aplicável, conforme termos do contrato de arrendamento em questão.
> Um item do imobilizado é baixado após alienação ou quando não há benefícios econômicos futuros resultantes do uso contínuo do ativo. Quaisquer ganhos ou perdas na venda ou baixa de um item do imobilizado são determinados pela diferença entre os valores recebidos na venda e o valor contábil do ativo e são reconhecidos no resultado."

Qual o método de depreciação do imobilizado (linear, soma dos dígitos ou unidades produzidas)? Nesse exemplo, a nota explicativa descreve que o método de depreciação do imobilizado utilizado é o linear.

- Exemplo 3:

> "Reconhecimento da Receita: A receita é mensurada pelo valor justo da contrapartida recebida ou a receber, deduzida de quaisquer estimativas de devoluções, descontos comerciais e/ou bonificações concedidos ao comprador e outras deduções similares.
> Vendas de produtos
> A receita de vendas de produtos é reconhecida quando os produtos são entregues e a posse foi passada nesse prazo de tal forma que todas as seguintes condições forem satisfeitas:
> o Grupo transferiu para o comprador os riscos e benefícios significativos relacionados à propriedade dos produtos;

> o Grupo não mantém envolvimento continuado na gestão dos produtos vendidos em grau normalmente associado à propriedade nem controle efetivo sobre tais produtos;
> o valor da receita pode ser mensurado com confiabilidade;
> é provável que os benefícios econômicos associados à transação fluirão para o Grupo;
> e os custos incorridos ou a serem incorridos relacionados à transação podem ser mensurados com confiabilidade.
> Mais especificamente, a receita de venda de produtos é reconhecida quando os produtos são entregues e a titularidade legal é transferida.
> As vendas de produtos que resultam na emissão de créditos de prêmios para clientes, na forma de pontos ou milhagens de acordo com o programa de fidelidade a clientes do Grupo, são contabilizadas como transações com receitas de elementos múltiplos, e o valor justo da contrapartida recebida ou a receber é alocado entre as mercadorias entregues e os créditos de prêmio concedidos. A contrapartida destinada aos créditos de prêmios é mensurada pelo valor justo na data da venda. Essa contrapartida não é reconhecida como receita na data da venda inicial, mas é diferida e reconhecida como receita quando os créditos de prêmio são resgatados e as obrigações do Grupo são cumpridas."

Qual o regime de apuração da receita (competência ou caixa)? Nesse exemplo, a nota explicativa descreve que o regime de apuração da receita é o de competência.

2.8 Considerações

Neste capítulo foram tratados conceitos que são necessários para o entendimento do trabalho proposto. Inicialmente foi apresentada a grande área do processo de descoberta de conhecimento e na sequência foi feito um aprofundamento no processo de descoberta de conhecimento em textos e mineração textual, visto que o trabalho proposto consiste em sumarizar e extrair conceitos de forma automática das notas explicativas das demonstrações financeiras das empresas, as quais são em formato texto não estruturado.

Foram também apresentadas e definidas as etapas pelas quais o processo de mineração de textos deve passar para então poder aplicar a extração de conceitos e a sumarização automática de textos. A fim de permitir um melhor entendimento e aprofundamento, as técnicas de mineração textual, extração de conceitos e sumarização, também foram contextualizados.

Seguindo, para entender a estrutura e o mecanismo de funcionamento das demonstrações financeiras das empresas, onde constam as notas explicativas, foi feito um relato demonstrando sua estrutura, seguida por uma breve explicação, destacando-se as notas explicativas.

Os conceitos de mineração textual em notas explicativas de demonstrações financeiras das empresas são também tratados nos trabalhos relacionados.

3 TRABALHOS RELACIONADOS

Nesse capítulo, são apresentados os trabalhos encontrados na literatura pela revisão bibliográfica realizada, que tratam de mineração textual em notas explicativas de demonstrações financeiras. Existem poucos estudos nessa área e fortalecendo esta afirmação, o primeiro trabalho discutido apresenta uma pesquisa do estado da arte, que afirma que as principais análises envolvendo as notas explicativas ainda continuam sendo feitas manualmente.

3.1 O Estado da Arte com o Objetivo de Apoiar Tarefas Analíticas em Notas Explicativas de Demonstrações Financeiras

Heidari e Felden (2014) realizaram um estudo do estado da arte no domínio de análise financeira para avaliar a importância da análise das notas explicativas de demonstrações financeiras e comprovar que nelas residem informações relevantes para uma completa análise financeira. Também verificaram os métodos existentes de análise financeira para extrair informações da parte não estruturada das demonstrações financeiras.

Os autores realizaram uma revisão da literatura para identificar os métodos de análise financeira existentes em três diferentes áreas: de negócios, científica e em casos de uso. Com isso concluíram que ambos os tipos de dados, estruturados e não estruturados, desempenham papel fundamental na análise das demonstrações financeiras. Ainda, mais especificamente com relação às notas explicativas, os autores concluem que para várias finalidades de análise financeira é necessária a análise dos dados estruturados em conjunto com as notas explicativas e sem essa análise em conjunto o processo de análise é incompleto e de alguma forma enganoso.

Além disso, o estudo baseado em 44 pesquisas de análise financeira, entre os anos de 1990 e 2012, mostrou que no caso de métodos de análise financeira, as principais análises envolvendo as notas explicativas foram feitas manualmente e que até então não haviam identificado nenhum método automático para integrar a métodos de análise financeira valores estruturados com notas explicativas em demonstrações financeiras.

3.2 Desenvolvimento de uma Abordagem de Mineração de Textos em Análise Financeira de Notas Explicativas

A abordagem baseada em mineração de textos desenvolvida por Heidari e Felden (2015a) aplica procedimentos de classificação baseados em classes pré-definidas a fim de extrair informações das notas explicativas das demonstrações financeiras das empresas, automaticamente.

O estudo utiliza como base de dados os relatórios financeiros (10K e 10Q) de 120 diferentes empresas da Comissão de Valores Mobiliários dos EUA, presentes na base de dados denominada EDGAR. Para toda a implementação do processo de mineração de textos, os autores utilizaram a ferramenta RapidMiner, que é uma ferramenta de código aberto para mineração de dados e análise preditiva.

Os autores focaram o estudo na classificação de uma nota explicativa específica das demonstrações financeiras, a nota do imposto de renda. Focaram nela por estar presente na

maioria das demonstrações financeiras. Foram pré-definidas seis classes para classificação da nota explicativa: o imposto diferido; taxa efetiva de imposto; perda operacional líquida; benefício fiscal não reconhecido; autoridade fiscal e provisão de avaliação.

Após o processo de mineração, com os dados prontos, os autores aplicaram algoritmos de classificação com o objetivo de obter os melhores resultados e identificar o mais apropriado para o objetivo da pesquisa.

Os algoritmos de classificação utilizados no estudo foram K-NN, Naïve Bayes, SVM e Árvore de Decisão. Esses algoritmos são supervisionados, onde os dados são separados em dados de treino e dados de teste, sendo que os dados de treino já são previamente rotulados com suas respectivas classes. Esses algoritmos passam por um processo de aprendizagem sobre a base de treinamento e após são aplicados sobre a base de teste com o objetivo de identificar a classificação correta de cada classe. Os resultados desses algoritmos no estudo são exibidos na Tabela 2, onde para o objetivo dos autores, que além da melhor acurácia, buscavam também menor tempo de execução, o Naïve Bayes acabou obtendo os melhores resultados com uma acurácia de 82,86% e menor tempo de execução.

Tabela 2: Resultados obtidos no estudo de Heidari e Felden (2015a).

Algoritmo	Tempo de execução	Acurácia
K-NN	7s	81,82%
Naïve Bayes	4s	82,86%
SVM	28s	79,22%
Árvore de Decisão	1m45s	90,65%

Fonte: Adaptado de Heidari e Felden (2015a).

Para efeito de análise, os autores também testaram um algoritmo de classificação não supervisionado, o K-means, configurando-o para separar os dados em seis classes (K=6). Porém conforme os autores concluem, para fins de análise financeira onde se procura termos específicos em notas explicativas, a classificação através de algoritmos supervisionados, onde as classes são previamente definidas, gera resultados mais confiáveis.

3.3 O Impacto da Mineração Textual na Análise de Notas Explicativas de Demonstrações Financeiras

Heidari e Felden (2015b), mesmos autores do artigo anterior, apresentam outro artigo que no geral se assemelha muito ao anterior, sendo que aplicam os mesmos testes e obtém os mesmos resultados. Porém adicionalmente apresentam uma tabela comparativa entre o método proposto por eles e os métodos de análise de notas explicativas existentes, que é reproduzida na Tabela 3.

Neste estudo, os autores apresentam um comparativo entre as diferentes abordagens para tratamento da parte textual dos relatórios financeiros e ao tipo de relatório financeiro que estas se aplicam. No comparativo, os autores descrevem o processo de análise que é aplicado na abordagem, inclusive na proposta por eles, destacando os benefícios e desafios das mesmas.

Tabela 3: Comparativo entre o método proposto por Heidari e Felden (2015b) e outros métodos de análise existentes

Abordagem para parte textual dos relatórios financeiros	Tipo de relatório financeiro	Processo de análise das notas explicativas	Benefícios e desafios
Abordagem tradicional	Relatórios tradicionais	Processo manual	Processo demorado e propenso a erros.
Em blocos com identificação detalhada	Relatórios XBRL	Tabelas e números são identificados, porém a informação descritiva das notas explicativas precisa ser analisada manualmente	Procedimentos de identificação ajudam para a unificação da estrutura dos relatórios financeiros. As partes textuais das notas explicativas precisam ser lidas manualmente, o que torna o processo demorado e propenso a erros.
Análise de conteúdo baseada em sistemas de codificação	Relatórios tradicionais	Atribuição de código manual para o texto	Os códigos devem ser atribuídos manualmente aos textos para cada relatório. Possíveis falhas na codificação por humanos. Mais preciso com relação ao processo manual.
Método de classificação do texto	Aplicado aos relatórios tradicionais e XBRL	Classificação automática de notas explicativas baseada em classes pré-definidas.	As notas explicativas são classificadas automaticamente. Ganho de tempo. O aumento do número de classes reduz a precisão do modelo.

Fonte: Adaptado de Heidari e Felden (2015b).

Com o estudo, os autores afirmam que com relação a sua pesquisa proposta e a lacuna existente na literatura, a utilização da solução de mineração por eles proposta oferece um caminho apropriado para superar as dificuldades da análise manual das notas explicativas das demonstrações financeiras.

3.4 Detecção de Fraudes em Demonstrações Financeiras Utilizando Mineração Textual

Gupta e Gill (2012) partem de pressuposto de que técnicas de mineração de dados são enormemente utilizadas pela comunidade de pesquisadores na detecção de fraudes em demonstrações financeiras, porém a maioria dessas pesquisas se baseia apenas nos números (informação quantitativa) e há pouca ou quase nenhuma investigação sobre a análise do texto dessas demonstrações. Com isso eles propõem uma abordagem de mineração de textos para a detecção de fraudes em demonstrações financeiras, analisando a informação qualitativa (o texto) presente nesses demonstrativos.

Os autores afirmam que para detectar a fraude é necessário examinar as informações qualitativas nas notas explicativas das demonstrações financeiras, bem como os números (informação quantitativa) associados nas demonstrações financeiras. Porém destacam que as informações textuais presentes nas demonstrações financeiras não são estruturadas e devem ser estruturadas antes da aplicação de qualquer técnica preditiva de mineração de dados.

O modelo conceitual apresentado pelos autores parte do pré-processamento da mineração textual e gera um vetor de palavras com a frequência em que as mesmas aparecem no documento para cada demonstração financeira, gerando o chamado “saco de palavras”. Este vetor de palavras serve de entrada para uma SVM que irá classificar o relatório financeiro como fraudulento ou não fraudulento. Os autores apenas sugeriram a utilização de uma SVM, mas não a aplicaram.

Por ser um modelo conceitual, não apresenta a aplicação e consequentemente não apresenta os resultados, que poderiam permitir uma avaliação.

3.5 Considerações

Nesse capítulo, foram apresentados quatro trabalhos da literatura que tratam de mineração textual em notas explicativas de demonstrações financeiras. Uma área ainda pouco explorada, sendo que a pesquisa de estado da arte feita por Heidari e Felden (2014) afirma que as principais análises envolvendo as notas explicativas são feitas manualmente, e que não identificaram nenhum método que integre os valores estruturados com as notas explicativas das demonstrações financeiras, automaticamente. Além disso, cita-se a pesquisa realizada por Fisher, Garnsey e Hughes (2016), na qual os autores fazem uma síntese da literatura de processamento de linguagem natural em contabilidade, auditoria e finanças, na qual relacionam apenas um único trabalho envolvendo notas explicativas, que é uma análise manual da década de 70.

Os três últimos trabalhos relacionados aplicam técnicas de classificação para permitir uma estruturação das notas explicativas das demonstrações financeiras. Os trabalhos de Heidari e Felden (2015a, 2015b) têm exatamente o mesmo foco deste trabalho. Porém, nesse trabalho sugere-se a aplicação de extração de conceitos e sumarização ao invés de classificação, por entender que os métodos de apuração das contas contábeis utilizados pela empresa e descritos nas notas explicativas na seção de principais práticas contábeis da empresa, que são o foco deste trabalho, são bem objetivos e claros. Com isso, pretende-se, através da extração de conceitos, a obtenção de resultados melhores e mais rápidos.

Os trabalhos de Heidari e Felden (2015a, 2015b) ainda estimulam a utilização da base de dados da Comissão de Valores Mobiliários. Assim como nos EUA que possuem o EDGAR, a Comissão de Valores Mobiliários do Brasil também disponibiliza as demonstrações financeiras das empresas com ações na Bolsa de Valores.

Ainda analisando a aplicação de Heidari e Felden (2015a, 2015b), verifica-se que os autores rotularam as demonstrações financeiras para teste e para avaliação dos resultados em seis classes distintas, porém não especificaram como e com base em que isso foi feito. Além disso, não apresentaram quaisquer especificações com relação às configurações dos classificadores supervisionados.

4 ABORDAGEM EXPERIMENTAL

Tendo em vista que a manipulação automática das notas explicativas das demonstrações financeiras das empresas é deficiente e que sua manipulação ocorre praticamente de forma manual, é proposta uma forma para identificar e estruturar de maneira automática os métodos de apuração das principais contas contábeis utilizados pelas empresas, juntamente com um resumo dessas notas, da seção de principais práticas de contabilidade utilizadas, com o uso de técnicas de mineração textual.

Este capítulo descreve a metodologia e o processo de descoberta de conhecimento realizado pelo modelo proposto, desde a etapa de consolidação da base de dados, seguida pelos experimentos de mineração textual, com a aplicação de algoritmos de sumarização e extração de conceitos, finalizando com a análise dos resultados.

4.1 Metodologia Aplicada

Para a aplicação do processo de descoberta de conhecimento neste estudo, é necessária a consolidação de uma base de dados para realização dos experimentos. Com a base de dados consolidada, o processo de mineração textual é aplicado e os resultados são obtidos, para uma posterior avaliação e análise. Essas etapas são descritas nas seções a seguir.

4.1.1 Consolidação da Base de Dados

A consolidação de uma base de dados se torna necessária, pois no trabalho de pesquisa realizado não foi encontrada na literatura nenhuma base já consolidada contendo as notas explicativas das demonstrações financeiras para o português. As bases de dados encontradas se limitavam às informações quantitativas e não continham as informações qualitativas que se referem às notas explicativas.

A maioria dos trabalhos relacionados com relatórios financeiros, presentes na literatura, utiliza os dados da Comissão de Valores Mobiliários (CVM), em especial a dos EUA. Seguindo esse princípio, para compor esta base de dados, foram utilizadas as demonstrações financeiras das empresas constantes na CVM do Brasil. Estas declarações são de domínio público e foram utilizadas com o objetivo de promover estudos da aplicação da mineração textual para a língua portuguesa. Além disso, os padrões das demonstrações financeiras adotados pelo Brasil seguem os padrões internacionais, o que faz acreditar que o processo de mineração adotado nesse trabalho pode se estender às demonstrações financeiras de outros países.

Porém, o conjunto de empresas que pertencem a CVM do Brasil é grande para o experimento proposto e por isso adotou-se como critério de seleção as empresas que integram o índice Bovespa. Este critério se baseou no fato de que as empresas que integram o índice Bovespa são as de maior representatividade no mercado financeiro nacional.

Dessa forma, para a criação da base de dados, foram obtidas da página da CVM do Brasil as demonstrações financeiras anuais completas das empresas que compõem o IBOVESPA, com data de referência de 31 de dezembro de 2015. Ao final, obtiveram-se as demonstrações financeiras das 55 empresas que compõem o índice. Todas as demonstrações encontram-se em arquivo formato PDF, sendo que cada demonstração é composta por um único arquivo.

A partir destas demonstrações financeiras, foi criado um arquivo texto contendo somente as notas explicativas, que são o foco desse estudo, de todas as empresas relacionadas. Como a intenção deste trabalho não é a extração automática de textos dos arquivos PDF, criou-se esse arquivo texto com o objetivo de centralizar as notas explicativas das demonstrações em um único local. Nesse arquivo se inseriu marcações indicando as empresas e os títulos das notas explicativas, pois por ser puramente textual não possui formatação.

Durante o processo de criação do arquivo texto, com dados das demonstrações em formato PDF, houve dificuldade com as demonstrações de duas empresas, onde não foi possível extrair as informações textuais. Em um dos casos, o texto encontrava-se em forma de imagem e em outro o arquivo estava protegido e não permitia a extração dos dados.

Dessa forma, ao final do processo, produziu-se um arquivo em formato texto contendo as notas explicativas das demonstrações financeiras de 31 de dezembro de 2015, de 53 empresas do IBOVESPA, listadas no Apêndice B.

4.1.2 Definição das Classes

Este trabalho tem como foco as notas explicativas das principais práticas contábeis, onde estão descritos os métodos de apuração das respectivas contas contábeis. No presente estudo é utilizada a extração de conceitos, com o objetivo de detectar esse método de apuração descrito na respectiva nota explicativa. Para isso é necessária a definição de classes que representam os diferentes métodos de apuração de cada conta contábil. O trabalho foi focado na identificação dos seguintes métodos de apuração utilizados pelas empresas:

- Método de apuração do estoque;
- Método de depreciação do imobilizado;
- Regime de apuração das receitas.

Focou-se na detecção desses três métodos por serem de contas contábeis presentes na maioria das demonstrações financeiras e por possuírem grande impacto no resultado financeiro da empresa. O método de apuração dessas contas está diretamente relacionado ao resultado financeiro, como por exemplo:

- Método de apuração do estoque: se for utilizado um método em que o valor do estoque tende a ficar maior, o lucro final da empresa também será maior;
- Método de depreciação do imobilizado: se for utilizado um método que acelere a depreciação do imobilizado, resultará em menor lucratividade para a empresa;
- Regime de apuração das receitas: se utilizado um regime que antecede as receitas, o resultado da empresa tende a ser melhor naquele período.

Com a aferição de um contador[9], foram definidas as classes para representar os métodos de apuração das contas contábeis selecionadas. Para desenvolver essa tarefa, o contador leu as notas explicativas e indicou os respectivos métodos de apuração das contas de cada empresa.

Dessa forma, foram definidas três classes para identificar o método de apuração do estoque:

- Custo médio: agrupa todas as notas explicativas da conta estoque que descrevem como método de apuração do estoque o custo médio, que é obtido através da média

[9] Maria Soares de Lima, com registro no CRC/RS 092498, é contadora com pós-graduação em controladoria e finanças e tem 9 anos de experiência em contabilidade empresarial.

ponderada dos custos de aquisição dos produtos, conseguindo refletir com exatidão os dados sobre os custos por período e dos estoques remanescentes;

- PEPS: agrupa todas as notas explicativas do estoque que descrevem como método de apuração do estoque o sistema PEPS (Primeiro a Entrar, Primeiro a Sair), onde o produto que chega antes ao estoque deve ir embora primeiro e o que chega por último vai embora por último, fazendo com que toda a operação realizada em estoques resulte em custo e lucro real;
- UEPS: agrupa todas as notas explicativas do estoque que descrevem como método de apuração do estoque o sistema UEPS (Último a Entrar, Primeiro a Sair), onde o cálculo do custo do estoque parte dos últimos itens que chegaram ao depósito. Como o custo dos últimos itens adquiridos normalmente é mais alto, este método tende a elevar o custo do produto e automaticamente diminuir o lucro tributável ao final do exercício, e por essa razão não é permitido pela legislação brasileira.

Para identificar o método de depreciação do imobilizado também foram definidas três classes:

- Linear: agrupa todas as notas explicativas do imobilizado que descrevem como método de depreciação do imobilizado o sistema linear, onde o valor a depreciar do bem é obtido pelo quociente entre o custo total de aquisição e o número de anos de sua vida útil, resultando em um mesmo valor de depreciação por período;
- Soma dos dígitos: agrupa todas as notas explicativas do imobilizado que descrevem como método de depreciação do imobilizado o sistema de soma dos dígitos dos anos, onde o valor da depreciação é calculado com base na fração da vida útil do bem, resultando em um valor de depreciação maior no início e menor ao final de sua vida útil;
- Unidades produzidas: agrupa todas as notas explicativas do imobilizado que descrevem como método de depreciação do imobilizado o sistema de unidades produzidas, onde o valor da depreciação do bem é calculado com base na capacidade produtiva do equipamento.

Já para identificar o regime de apuração das receitas foram definidas duas classes:

- Caixa: agrupa todas as notas explicativas de reconhecimento de receitas, que descrevem como método de apuração das receitas o regime de caixa, onde o registro do evento se dá apenas na data de recebimento ou de pagamento;
- Competência: agrupa todas as notas explicativas de reconhecimento de receitas, que descrevem como método de apuração das receitas o regime de competência, onde o registro do evento se dá na data em que aconteceu, ou seja, na data do documento, não importando quando vai ser pago ou recebido.

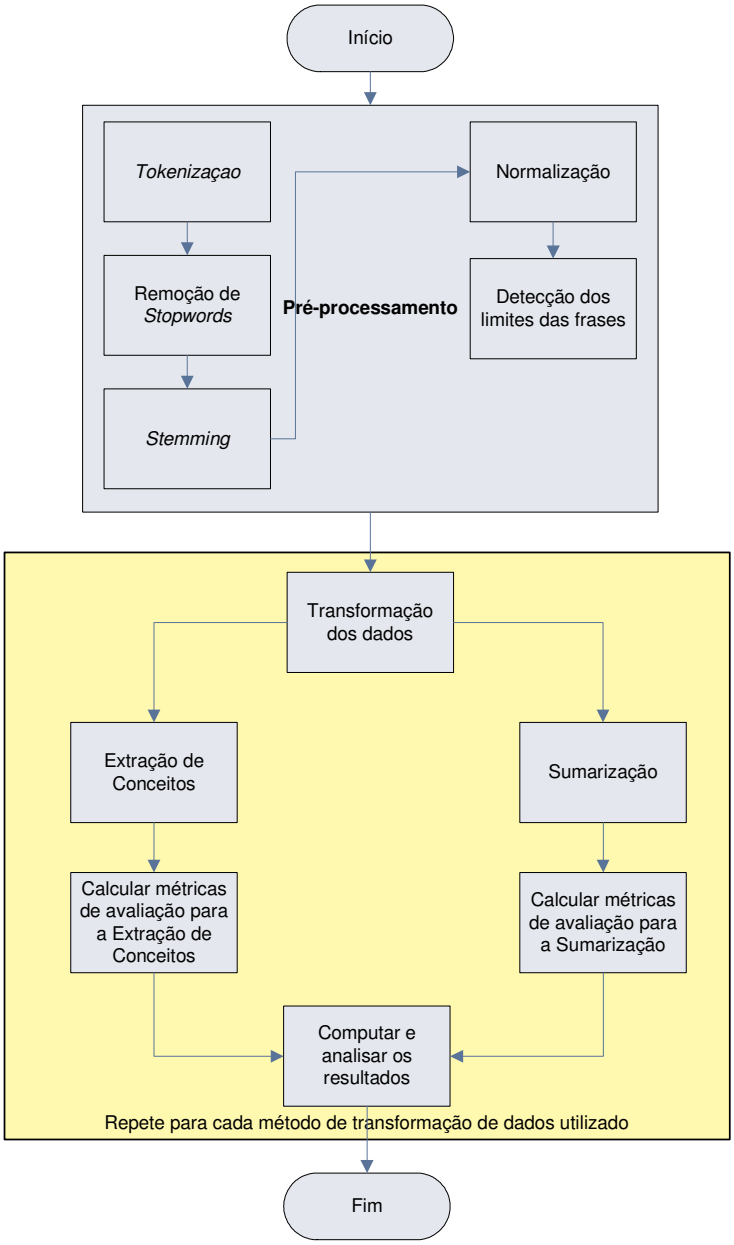
Essas indicações de classe pelo contador permitem um confrontamento do resultado obtido pela extração de conceitos com a correta classificação, sendo possível dessa forma, para cada classe, a obtenção do número de amostras classificadas de forma correta e incorreta. Isso permite a avaliação da extração de conceitos com as métricas mencionadas na seção 2.6.1.

4.1.3 Mineração Textual

Para a aplicação dos algoritmos de mineração textual de extração de conceitos e sumarização, os dados das notas explicativas primeiramente devem passar pela etapa de pré-

processamento, que envolve processos descritos no capítulo 2, como a *tokenização*, remoção de *stopwords*, *stemming*, normalização de escrita e de ortografia e detecção dos limites das frases.

Figura 3: Fluxograma do processo da metodologia proposta



Fonte: Elaborado pelo autor.

Após o pré-processamento, os *tokens* que representam as palavras são transformados em uma representação vetorial apropriada, que serve de entrada para o algoritmo de extração de conceitos e sumarização. Os principais esquemas para esta transformação vetorial estão descritos no capítulo 2 e estão diretamente relacionados ao resultado da extração de conceitos e principalmente ao resultado da sumarização. Esta etapa é considerada a mais importante do processo de sumarização e, dessa forma, todos os principais esquemas de transformação de dados são testados em uma parte representativa da base de dados e depois avaliados através de um quadro comparativo que identifica qual esquema traz os melhores resultados, quando aplicado a notas explicativas de demonstrações financeiras.

Dessa forma, para cada um dos principais métodos de transformação de dados, é executado o algoritmo de sumarização e extração de conceitos adequado. A Figura 3 apresenta o fluxograma do processo, que parte da etapa de pré-processamento que é única para cada nota explicativa e segue para a etapa de transformação de dados que é particular para cada algoritmo de extração de conceitos e sumarização. O algoritmo de extração de conceitos e os de sumarização são executados, as métricas de avaliação são calculadas e ao final os resultados são computados e avaliados conforme descrito na seção a seguir.

Porém, para realização do processo de extração de conceitos, é necessária a criação e definição dos conceitos que representam cada uma das classes definidas que fazem referência ao método de contabilidade utilizado pela empresa para apuração de determinada conta. Dessa forma, os conceitos são criados através de um agrupamento de palavras e expressões de forma que, estas estando presentes na nota explicativa, determinem a classe a que ela pertence.

4.1.4 Análise dos Resultados

Os resultados são avaliados com métricas diferentes para a extração de conceitos e para a sumarização. Isto porque para a extração de conceitos é possível identificar se houve acerto ou erro na classificação da nota explicativa. Já para a sumarização, não é possível avaliar se o resumo está certo ou errado e sim avaliar o quão bom ele está.

Dessa forma, para a extração de conceitos é aplicado o esquema de Matriz de Confusão com base nos acertos e erros com relação aos rótulos de classes pré-estabelecidos. Com estes valores, são calculadas as medidas de acurácia, *precision*, *recall* e *f-measure*, possibilitando assim a comparação dos resultados para avaliação da eficiência do processo de extração de conceitos em cada uma das iterações.

Já para a sumarização é utilizada a medida ROUGE, por ser uma métrica já consolidada e que vem sendo utilizada para avaliação dos resumos no DUC. Esta métrica permite uma avaliação da qualidade dos resumos, possibilitando assim a comparação dos resultados para identificar a eficiência do processo de sumarização em cada iteração. Apesar das métricas *precision* e *recall* permitirem uma avaliação do conteúdo dos resumos, estas não são utilizadas, pois se aplicam somente aos resumos por extração de frases e necessitam para o cálculo de um resumo pré-estabelecido, com as frases definidas como ideais para cada nota, o que envolveria uma minuciosa análise de uma equipe de contadores, algo fora do escopo desse trabalho.

Ao final, é feita uma análise crítica dos resultados obtidos e apontadas as situações do processo em que foram obtidos bons resultados, bem como situações em que os resultados ficaram a desejar.

4.2 Experimentos

Para a realização dos experimentos no estudo de caso foi utilizada a linguagem de programação Python 3.5, por ser muito utilizada no estudo de mineração textual e por possuir um vasto ferramental para manipulação de texto.

Primeiramente, sobre a base de dados consolidada, buscou-se pela nota explicativa que descreve o método de apuração da conta desejada, para cada empresa, através da busca por um conjunto de palavras-chave que compõem o título da nota explicativa, transformado para minúsculo. As palavras-chave utilizadas para cada nota explicativa foram as seguintes:

- Estoque: {estoque};

- Imobilizado: {imobilizado};
- Reconhecimento da receita: {reconhecimento da receita, reconhecimento de receita, receita de venda, receita de serviços, receitas de serviço, receitas de vendas}.

Esse processo resultou na obtenção de 31 notas explicativas referentes ao estoque, 42 do imobilizado e 33 relativas ao reconhecimento da receita.

Em seguida, foi feito o pré-processamento do texto da nota explicativa. Aplicou-se primeiramente a *tokenização*, onde identificou-se as palavras e frases do texto da nota explicativa. Em seguida, utilizou-se o processo de remoção de *stopwords*, com o uso do dicionário de *stopwords* da biblioteca NLTK 3.0[10] para o português. Após, as palavras do texto foram submetidas ao processo de *steamming*, com a utilização do RSLPstemmer (*stemmer* para o português), proposto por Orengo e Huyck (2001). Por fim, o texto passou pelo processo de normalização, onde as palavras foram transformadas para minúsculo.

Após essas tarefas relacionadas, o texto da nota explicativa está preparado para aplicação dos métodos de sumarização e extração de conceitos com seus respectivos processos de transformação de dados.

4.2.1 Extração de Conceitos

A extração de conceitos é utilizada com o objetivo de classificar as notas explicativas, através da presença ou não do conceito definido. Com a aferição de um contador, foram criados os conceitos para identificar cada uma das classes das notas explicativas selecionadas. Os conceitos foram criados como um vetor de palavras, que pudessem identificar a classe no texto das notas explicativas. O vetor foi criado com base no modelo espaço de vetores, sendo formado por palavras e/ou expressões, juntamente com o valor 1 ou -1, onde 1 indica que a palavra e/ou expressão deve aparecer e -1 penaliza, indicando que a palavra e/ou expressão não deve aparecer. Ao final, os valores das expressões encontradas são somados e indicam o valor de representatividade do conceito, onde os resultados maiores ou igual a 1 indicam a presença do conceito para classificação.

Assim, foram criados os seguintes conceitos para identificar as três classes do método de apuração do estoque, onde a primeira posição do vetor possui a palavra e/ou expressão que deve estar presente no texto e a segunda posição possui o valor que indica se a palavra deve aparecer ou não no texto:

- Custo médio: {[custo médio,1], [preço médio,1], [média ponderada,1], [custo histórico,-1]};
- PEPS: {[peps, 1], [custo histórico, 1], [média, -1]};
- UEPS: {[ueps, 1], [último custo, 1], [custo recente, 1], [média, -1]}.

Já os conceitos para identificar as três classes do método de depreciação do imobilizado são:

- Linear: {[método linear, 1], [linear,1]};
- Soma dos dígitos: {[soma dos dígitos,1]};
- Unidades produzidas: {[unidades produzidas,1], [base na quantidade,1]}.

---

[10] Conjunto de ferramentas de linguagem natural, Natural Language ToolKit em inglês, disponível em < http://www.nltk.org/>.

Por fim, os conceitos para identificar as duas classes do regime de apuração das receitas são:

- Caixa: {[caixa,1], [momento do recebimento,1], [momento do pagamento,1], [competência, -1]};
- Competência: {[competência,1], [realizado,1], [venda for faturada,1], [entregue,1], [transferida,1], [efetiva prestação,1], [contrapartida receber,1]}.

Para a aplicação do algoritmo de extração de conceitos, que se baseia em uma comparação simples e pura, as palavras e expressões que compõem os conceitos também foram submetidas ao pré-processamento textual, onde passaram pelo processo de remoção de *stopwords* e *steaming*. Com isso, os conceitos foram pré-processados da mesma forma que o texto da nota explicativa e o algoritmo de extração de conceitos pôde ser aplicado e os resultados computados.

A aplicação da extração de conceitos para identificar o método de apuração do estoque nas 31 notas explicativas correspondentes obteve uma acurácia de 100% e os resultados obtidos são apresentados na Matriz de Confusão exibida na Tabela 4 e avaliação exposta na Tabela 5.

Tabela 4: Matriz de confusão com os resultados da extração de conceitos na nota explicativa do estoque

	Resultado predito	
Resultado desejado	Custo Médio	PEPS
Custo Médio	29	0
PEPS	0	2

Fonte: Elaborado pelo autor.

Tabela 5: Avaliações da extração de conceitos na nota explicativa do estoque

Classe	***Recall***	***Precision***	***f-measure***
Custo Médio	100%	100%	1
PEPS	100%	100%	1

Fonte: Elaborado pelo autor.

Os resultados obtidos na extração de conceitos em busca do método de depreciação do imobilizado, nas 42 notas explicativas correspondentes são apresentados na Matriz de Confusão exibida na Tabela 6. No imobilizado existe a particularidade de que a empresa pode utilizar mais que um método de depreciação, dependendo do bem do ativo imobilizado e dessa forma ocorre a identificação de mais que um método na mesma nota explicativa. Nesse estudo de caso, quatro empresas adotam dois métodos concomitantes para depreciação do imobilizado, elevando o resultado para 46 classificações. A acurácia obtida no processo também foi de 100% e sua avaliação é apresentada na Tabela 7.

Tabela 6: Matriz de Confusão com os resultados da extração de conceitos na nota explicativa do imobilizado

	Resultado predito	
Resultado desejado	Linear	Unidade produzida

Linear	42	0
Unidade produzida	0	4

Fonte: Elaborado pelo autor.

Já os resultados da extração de conceitos para identificar o método de apuração de receitas, nas 33 notas explicativas correspondentes, são exibidos na Tabela 8, que apresenta a Matriz de Confusão com os resultados obtidos. Nesse processo obteve-se 96,97% de acurácia e os resultados da avaliação são apresentados na Tabela 9. Diferentemente das outras notas explicativas, nessa de apuração da receita, a extração de conceitos classificou erroneamente uma nota explicativa

Tabela 7: Avaliações da extração de conceitos na nota explicativa do imobilizado

Classe	***Recall***	***Precision***	***f-measure***
Linear	100%	100%	1
Unidade produzida	100%	100%	1

Fonte: Elaborado pelo autor.

Tabela 8: Matriz de Confusão com os resultados da extração de conceitos na nota explicativa do reconhecimento da receita

	Resultado predito	
Resultado desejado	Caixa	Competência
Caixa	1	0
Competência	1	31

Fonte: Elaborado pelo autor.

Tabela 9: Avaliações da extração de conceitos na nota explicativa do reconhecimento da receita

Classe	***Recall***	***Precision***	***f-measure***
Caixa	100%	50%	0,6667
Competência	96,88%	100%	0,9841

Fonte: Elaborado pelo autor.

A extração de conceitos apresentou resultados promissores nas notas explicativas, visto que identificou corretamente o método de apuração do estoque para todas as empresas, bem como do método de depreciação do imobilizado, no qual também obteve 100% de acurácia. No experimento para identificar o método de apuração das receitas obteve-se uma boa acurácia, porém a metodologia proposta apurou erroneamente a identificação de uma classe, o que gerou *precision* e *f-measure* bem baixo para identificação da classe caixa. Isto ocorreu, pois um número muito pequeno de empresas da base utiliza o método de caixa para apuração das receitas e fez com que um pequeno erro gerasse uma avaliação ruim.

O resultados alcançados demonstram que a extração de conceitos se mostrou uma ferramenta promissora para identificar os métodos de apuração das contas contábeis. Grande parte disso se deve ao fato de que os métodos de apuração das contas contábeis normalmente são bem explícitos e descritos com clareza na maioria das vezes nas notas explicativas das demonstrações financeiras das empresas.

4.2.2 Sumarização

A sumarização tem o objetivo de apresentar o que a nota explicativa diz de mais importante. Para isso, buscou-se na literatura vários algoritmos, desde os mais simples, baseados simplesmente na frequência das palavras, até os mais complexos, baseados em grafo, para serem aplicados nas notas explicativas e poder avaliar o seu desempenho. Os algoritmos selecionados foram:

- LSA: algoritmo de sumarização baseado na análise semântica latente, desenvolvido por Steinberger e Ježek (2004);
- LexRank: algoritmo de sumarização baseado em grafo, desenvolvido por Erkan e Radev (2004);
- KLSum: algoritmo de sumarização que usa, como critério para selecionar as principais sentenças para formação do resumo, a frequência das palavras, introduzido por Haghighi e Vanderwende (2009);
- SumBasic: um algoritmo de sumarização simples e eficaz, introduzido por Vanderwende et al. (2007), que se baseia na observação da frequência relativa das palavras para selecionar as melhores frases para compor o resumo;
- SimpleSummarizer: baseado na simples extração das sentenças que contêm a palavra que aparece com maior frequência no texto;
- NaiveSummarizer: algoritmo de sumarização com abordagem de classificação das sentenças que compõem o resumo, baseada no modelo bayesiano, com base no trabalho de Luhn (1958).

Para avaliação dos resumos obtidos utilizou-se a medida ROUGE, a qual vai de 0 a 1, e quanto mais próxima de 1, melhor o resumo. Porém esta medida depende de uma comparação do resumo obtido com outro resumo de referência. Por ser inviável a criação e validação de um resumo para cada nota explicativa de cada empresa, os algoritmos foram avaliados usando a base CSTNews[11]. Esta base possui vários textos de notícias em português com seu respectivo resumo ideal, criado por um humano, e é amplamente utilizada pelo meio acadêmico para experimentos na língua portuguesa, como em (CARDOSO et al., 2015; RIBALDO et al., 2012; CAMARGO et al., 2012; CASTRO JORGE; PARDO, 2010).

Os algoritmos foram aplicados sobre os textos da base que possuíam uma maior relação com o meio tributário e financeiro, por serem mais próximos dos textos da base deste trabalho. Foram selecionados os textos que tivessem alguma relação com assuntos tributários, financeiros, auditoria e governança, que melhor se encaixam ao perfil dos textos das demonstrações contábeis. A Tabela 10 apresenta os textos utilizados da base e a informação do tamanho resultante do respectivo resumo humano atrelado.

[11] Base de documentos textuais em língua portuguesa com o respectivo resumo criado por humanos, criada pelo grupo de linguística computacional do NILC, disponível em < http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html>.

Os algoritmos de sumarização foram configurados para obter um resumo com 30% do tamanho do texto original, que corresponde aproximadamente ao tamanho do resumo humano disponível, que é apresentado na Tabela 10. Os resumos obtidos pelos algoritmos foram avaliados com o ROUGE, tendo como resumo base os seus respectivos resumos ideais. Como comparativo, foi criado um resumo para cada um dos textos, utilizando a ferramenta de resumo automático do Microsoft Word, que já é uma ferramenta consolidada e amplamente utilizada, como em (LLORET; ROMÁ-FEFRI; PALOMAR, 2013; CHUANG; YANG, 2000; SUANMALI; SALIM; BINWAHLAN, 2009; REEVE; HAN; BROOKS, 2007). Os resumos obtidos pelo Microsoft Word também foram avaliados pelo ROUGE (R) com base no respectivo resumo humano. Os resultados obtidos são apresentados na Tabela 11, bem como o tamanho do resumo resultante, com base no texto original, destacando-se os melhores resultados.

Tabela 10: Relação de textos utilizados da base CSTNews 5.0

Texto	Nome do arquivo	Tamanho do resumo humano
1	CSTNews 5.0\C20_Politica_CPMF\Textos-fonte\D1_C20_Folha_19-09-2007_13h39.txt	30%
2	CSTNews 5.0\C20_Politica_CPMF\Textos-fonte\D4_C20_JB.txt	31%
3	CSTNews 5.0\C20_Politica_CPMF\Textos-fonte\D5_C20_GPovo_19-09-2007_12h44.txt	30%
4	CSTNews 5.0\C30_Dinheiro_LucroItau - concordanciaRST\Textos-fonte\D2_C30_Estadao_07-08-2007_07h59.txt	31%
5	CSTNews 5.0\C30_Dinheiro_LucroItau - concordanciaRST\Textos-fonte\D3_C30_OGlobo_07-08-2007_09h10.txt	32%
6	CSTNews 5.0\C34_Cotidiano_MalhaFina\Textos-fonte\D1_C34_Folha_20-08-2007_12h22.txt	31%
7	CSTNews 5.0\C34_Cotidiano_MalhaFina\Textos-fonte\D2_C34_Estadao_20-08-2007_13h22.txt	30%
8	CSTNews 5.0\C34_Cotidiano_MalhaFina\Textos-fonte\D3_C34_JB_20-08-2007_16h22.txt	31%
9	CSTNews 5.0\C9_Politica_Desvio\Textos-fonte\D1_C9_Folha_04-08-2006_13h20.txt	28%
10	CSTNews 5.0\C9_Politica_Desvio\Textos-fonte\D2_C9_Estadao_04-08-2006_09h53.txt	32%
11	CSTNews 5.0\C9_Politica_Desvio\Textos-fonte\D3_C9_OGlobo_04-08-2006_14h09.txt	32%

Fonte: Elaborado pelo autor.

Nessa avaliação, pode-se concluir que o algoritmo que mais se destacou foi o LexRank que obteve resultado maior ou igual aos demais em 64% dos textos. Em segundo lugar ficaram o naiveSum e o Microsoft Word, com resultado maior ou igual aos demais em 36% dos textos. Empatados na terceira colocação ficaram os algoritmos LSA e simpleSum com resultados melhores ou iguais aos demais em 27% dos textos. O algoritmo que apresentou o pior desempenho foi o KLSum que não obteve resultados superiores aos demais em nenhum dos textos deste experimento.

Tabela 11: Resultados da avaliação dos algoritmos de sumarização sobre os textos da base CSTNews

Texto	LSA		LexRank		KLSum		sumBasic		simpleSum		naiveSum		MS Word	
	R	Tam.	R	Tam.	R	Tam.	R	Tam.	R	Tam.	R	Tam.	R	Tam.
1	0,41	33%	**0,57**	32%	0,28	35%	0,28	28%	**0,57**	34%	0,22	31%	0,22	25%
2	**0,93**	37%	**0,93**	37%	0,00	31%	0,07	27%	**0,93**	37%	**0,93**	37%	**0,93**	29%
3	0,26	33%	**0,29**	33%	0,21	28%	0,24	24%	0,22	34%	**0,29**	34%	**0,29**	25%
4	**0,80**	35%	0,37	22%	0,37	28%	0,29	32%	0,47	30%	0,55	31%	0,55	27%
5	**0,39**	35%	0,21	34%	0,26	22%	0,21	22%	0,12	34%	0,34	34%	0,34	26%

6	0,14	29%	**0,35**	25%	0,28	26%	0,30	25%	0,28	35%	0,34	35%	0,34	26%
7	0,48	32%	**0,53**	30%	0,16	22%	0,24	20%	**0,53**	31%	0,45	30%	0,45	26%
8	0,37	22%	0,25	28%	0,25	34%	**0,56**	30%	0,37	28%	0,25	28%	0,25	27%
9	0,42	27%	**0,54**	30%	0,16	22%	0,07	29%	0,39	29%	0,39	29%	0,39	24%
10	0,19	24%	0,35	32%	0,22	31%	0,19	26%	0,30	34%	**0,36**	34%	**0,36**	28%
11	0,18	27%	**0,56**	35%	0,40	35%	0,38	30%	0,48	35%	**0,56**	35%	**0,56**	26%
$\bar{x}$	0,42		**0,45**		0,24		0,26		0,42		0,43		0,43	
σ	0,25		0,21		0,11		0,14		0,22		0,20		0,20	

Fonte: Elaborado pelo autor.

Nas demonstrações financeiras, os algoritmos de sumarização foram aplicados para cada uma das notas explicativas selecionadas, configurando-os para resultar em um resumo com 30% do seu tamanho original, onde espera-se expressar a ideia principal da nota explicativa de uma forma sucinta. Exemplos de resumos obtidos são apresentados no Apêndice C.

Porém, em busca de uma alternativa de avaliação dos algoritmos de sumarização dentro do contexto contábil, sem a disponibilidade de resumos humanos das notas explicativas, avaliou-se com a medida ROUGE os resumos gerados por um algoritmo, tendo como base os resumos gerados pelos outros cinco algoritmos. Dessa forma, gerou-se o resumo da nota explicativa com cada um dos algoritmos, resultando em seis resumos, onde cada um é avaliado pela medida ROUGE contra cada um dos outros cinco. Por exemplo, o resumo obtido pelo algoritmo LSA é avaliado pelo ROUGE com base nos resumos obtidos da mesma nota explicativa, pelos algoritmos LexRank, KLSum, sumBasic, simpleSum e naiveSum.

Optou-se por avaliar com todos os algoritmos e não apenas com um em especial, pois os algoritmos com técnicas similares geram resumos também similares. Caso fosse avaliar os resumos obtidos com um único resumo modelo, resultante de um algoritmo em especial, beneficiaria as técnicas similares as do resumo modelo, que obteriam resultados superiores devido à similaridade dos dois resumos.

Ao final do processo, obtêm-se cinco avaliações para cada algoritmo, em cada nota explicativa. Pelo fato de que um número elevado de avaliações gera confusão no momento de análise, optou-se por fazer a média entre as avaliações. Assim, obteve-se uma avaliação média de cada algoritmo, para cada nota explicativa, de cada empresa, com base na média dos resultados ROUGE, obtendo-se ao final 53 avaliações para cada nota explicativa. Os resultados obtidos nesta avaliação, para a nota explicativa do estoque, imobilizado e reconhecimento da receita, são apresentados no Apêndice D, Apêndice E e Apêndice F, respectivamente.

Porém ainda assim têm-se vários resultados, pois para cada nota explicativa tem-se uma média de avaliação para cada empresa. Em busca por uma única avaliação para cada algoritmo, em cada nota explicativa, calculou-se a média ($\bar{x}$) e o desvio padrão (σ) entre as avaliações de cada nota explicativa, apresentados na Tabela 12.

Tabela 12: Resultado médio e desvio padrão do ROUGE sobre os algoritmos de sumarização empregados para sumarização das notas explicativas das 53 empresas analisadas

	LSA		LexRank		KLSum		sumBasic		simpleSum		naiveSum	
Nota explicativa	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Estoque	0,29	±0,28	**0,47**	±0,23	0,39	±0,26	0,43	±0,25	**0,47**	±0,23	0,29	±0,27
Imobilizado	0,27	±0,17	0,34	±0,17	0,36	±0,14	0,29	±0,17	0,35	±0,17	**0,38**	±0,14
Reconhecimento da receita	0,38	±0,20	0,43	±0,19	0,39	±0,20	0,35	±0,19	**0,43**	**±0,18**	0,37	±0,21

Fonte: Elaborado pelo autor.

O fato de não se ter resumos das notas explicativas criados por humanos para poder avaliar os algoritmos dificultou muito a análise. A alternativa encontrada e aplicada nesse estudo não é a melhor, porém foi a forma adotada para poder obter resultados que fornecessem uma estimativa de qual algoritmo possui o melhor desempenho para resumir notas explicativas de demonstrações financeiras. Algo semelhante é feito pelo método automático de avaliação de resumos *Pyramid*, introduzido por Nenkova e Passonneau (2005), o qual gera automaticamente um resumo para servir de modelo para comparar com o resumo a ser avaliado.

Na avaliação dos algoritmos de sumarização sobre as notas explicativas, com base na média das medidas ROUGE, nas notas explicativas do estoque, se destacaram o LexRank e o simpleSum, com maior ROUGE médio e menor desvio padrão. Já para as notas explicativas do imobilizado, destaque para o naiveSum que obteve o maior ROUGE médio, e mesmo descontando-se o desvio padrão, ainda permanece melhor classificado. Para as notas explicativas do reconhecimento da receita, destaca-se o simpleSum que também obteve o melhor ROUGE médio, e descontando-se o desvio padrão, ainda permanece melhor classificado, seguido muito de perto pelo LexRank, com o mesmo ROUGE médio, porém com um maior desvio padrão. De um modo geral, destacaram-se o simpleSum que obteve melhor desempenho em duas das três notas explicativas e o LexRank, que praticamente se igualou ao simpleSum, perdendo apenas no desvio padrão em uma das duas notas explicativas em que obteve melhor avaliação.

Confrontando-se os resultados médios obtidos sobre os textos da base CSTNews, que possui resumos humanos para avaliação, apresentados na Tabela 11, com os resultados médios obtidos com a aplicação dos algoritmos de sumarização sobre as notas explicativas, apresentados na Tabela 12, é possível identificar uma similaridade nos resultados, mesmo não dispondo de resumos humanos para avaliação das notas explicativas.

5 CONCLUSÕES

Foi realizada uma ampla busca por pesquisas que utilizaram mineração textual nas demonstrações financeiras das empresas com algum objetivo específico, em especial trabalhos de pesquisa com mineração textual em notas explicativas das demonstrações financeiras. Buscou-se por trabalhos de sumarização e extração de conceitos com o uso de mineração textual em notas explicativas, mas não se obteve sucesso. Os trabalhos de pesquisa encontrados na literatura que mais se aproximaram deste utilizavam mineração textual nas notas explicativas e posteriormente algoritmos de classificação das mesmas. Foram poucas as pesquisas encontradas com mineração textual em notas explicativas nas demonstrações financeiras, tendo sido identificada a real necessidade de sua interpretação de maneira automática.

Em busca de formas para ajudar na análise das notas explicativas de forma automática, foram realizados experimentos com mineração textual. Foram selecionados seis algoritmos de sumarização automática de textos, com técnicas distintas, envolvendo algoritmos simples e complexos de sumarização. Quando avaliados com a medida ROUGE em textos com conteúdo financeiro da CSTNews, destacou-se o LexRank que se mostrou igual ou superior aos demais em sete dos onze textos avaliados. Além disso, apresentou um desempenho superior quando comparado aos resultados obtidos pela ferramenta de sumarização do Microsoft Word, que foi igual ou superior aos demais em apenas quatro dos onze textos.

Esses mesmos algoritmos foram aplicados para sumarizar as notas explicativas selecionadas da base consolidada das demonstrações financeiras. Devido ao fato de não se ter resumos ideais para fins de comparação para obtenção da medida ROUGE, o resultado de um algoritmo foi avaliado com base nos resumos obtidos pelos outros cinco algoritmos. No resultado médio das avaliações, se destacaram os algoritmos simpleSum e LexRank, os quais obtiveram melhor desempenho, similarmente em duas das três notas explicativas, ao mesmo tempo que seu desempenho nas demais notas explicativas ficou próximo da melhor avaliação. E mesmo não dispondo de resumos humanos para avaliação, o método de avaliação automática aplicado, com base nos outros cinco algoritmos, apresentou resultados médios bem próximos das médias dos resultados obtidos sobre os textos da base CSTNews, com resumos humanos.

Na união das duas avaliações, a primeira em textos financeiros e a segunda nas notas explicativas das demonstrações financeiras, pode-se concluir que entre os algoritmos avaliados, o LexRank é o mais promissor para sumarizar textos do meio financeiro.

Seguindo com o objetivo de identificar de forma automática os métodos de apuração das contas contábeis, a extração de conceitos apresentou resultados promissores, visto que no experimento obteve acurácia de 100% nas notas explicativas do estoque e imobilizado e acurácia de 96,97% na nota explicativa do reconhecimento da receita. Os trabalhos relacionados à classificação de notas explicativas encontrados na literatura utilizaram K-NN, Naïve Bayes, SVM e Árvore de Decisão. Essas técnicas são mais complexas e exigem um maior processamento que a extração de conceitos. A extração de conceitos apresentou resultados promissores pelo fato de que, normalmente, os métodos de apuração das contas contábeis estão bem explícitos e claramente especificados no texto da nota explicativa. A extração de conceitos é um método simples, porém exige o auxílio de um especialista, enquanto que os métodos dos trabalhos relacionados encontrados na literatura, apesar de mais complexos não têm essa necessidade.

A análise de algumas classes que identificam o método de apuração das contas contábeis, bem como dos conceitos que às representam, não foi a esperada. Isso ocorreu pelo fato de poucas notas explicativas da base utilizar aquele método, não permitindo uma avaliação

plena. Com mais dados os conceitos poderiam ser aprimorados e melhor analisados. Porém, no geral, os resultados indicam que o uso de sumarização e extração de conceitos pode ajudar na interpretação e análise das demonstrações financeiras das empresas sem a necessidade de uma leitura de todo o texto das mesmas.

No entanto, há muito que fazer para a análise automática das notas explicativas das demonstrações financeiras. Também, essa mesma abordagem pode ser avaliada com outros estudos de caso, podendo demonstrar a sua utilidade em outras situações de análise.

Outro ponto que pode ser estudado é a conexão dos resultados obtidos na análise da nota explicativa com os valores financeiros do corpo principal da demonstração financeira, buscando atingir a análise ideal, que integra os dados quantitativos e qualitativos das demonstrações financeiras.

REFERÊNCIAS

AAMODT, A.; NYGÅRD, M. Different roles and mutual dependencies of data, information, and knowledge - An AI perspective on their integration. **Data & Knowledge Engineering**, [S.l.], v. 16, n. 3, p. 191–222, 1995.

BARZILAY, R.; ELHADAD, M. Using lexical chains for text summarization. **Advances in automatic text summarization**, [S.l.], p. 111–121, 1999.

BRAGG, S. **What are financial statement footnotes?** Disponível em: < http://www.accountingtools.com/questions-and-answers/what-are-financial-statement-footnotes.html>. Acesso em 14 abr. 2016.

CAMARGO, R.T.; MAZIERO, E.G.; PARDO T.A.S. Corpus analysis of aspects in multi-document summaries - the case of news texts from 'world' section. In: ONLINEPROCEEDINGS OF THE 11 TH CORPUS LINGUISTICS SYMPOSIUM, 2012. **Proceedings. . .** [S.l.: s.n.], 2012.

CARDOSO, P.C.F.; JORGE, M.L.R.C.; PARDO, T.A.S. Exploring the Rhetorical Structure Theory for multi-document summarization. In: CONGRESO DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL, XXXI, 2015. **Proceedings. . .** Alicante, 2015.

CASTRO JORGE, M.L.R.; PARDO, T.A.S. Experiments with CST-based multidocument summarization. In: WORKSHOP ON GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 2010. **Proceedings. . .** [S.l.: s.n.], 2010. p. 74–82.

CHEN, H.; HSU, P.; ORWIG, R.; HOOPES, L.; NUNAMAKER, J. F. Automatic concept classification of text from electronic meetings. **Communications of the ACM**, New York, NY, USA, v. 37, n. 10, p. 56–73, Oct. 1994.

CHUANG, W. T.; YANG, J. Extracting Sentence Segments for Text Summarization: a machine learning approach. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., 2000, New York, NY, USA. **Proceedings. . .** ACM, 2000. p. 152–159.

DE IUDÍCIBUS, S.; MARTINS, E.; GELBCKE, E. R.; DOS SANTOS, A. **Manual de Contabilidade Societária**: aplicável a todas as sociedades. São Paulo: Atlas, 2010.

EDMUNDSON, H. P. New Methods in Automatic Extracting. **J. ACM**, New York, NY, USA, v. 16, n. 2, p. 264–285, Apr. 1969.

ERKAN, G.; RADEV, D. R. LexRank: graph-based lexical centrality as salience in text summarization. **Journal of Artificial Intelligence Research**, USA, v. 22, n. 1, p. 457–479, Dec. 2004.

FABER, V.; HOCHBERG, J. G.; KELLY, P.M.; THOMAS, T. R.; WHITE, J. M. Concept extraction: a data-mining technique**. Los Alamos Science**, [S.l.], v. 22, p. 145–149, 1994.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, [S.l.], v. 17, n. 3, p. 37, 1996.

FELDMAN, R.; HIRSH, H. Exploiting background information in knowledge discovery from text. **Journal of Intelligent Information Systems**, [S.l.], v. 9, n. 1, p. 83–97, 1997.

FELDMAN, R.; SANGER, J. **The text mining handbook**: advanced approaches in analyzing unstructured data. [S.l.]: Cambridge university press, 2007.

FIGUEROA, A.; ATKINSON, J. Contextual Language Models for Ranking Answers to Natural Language Definition Questions. Computational Intelligence, v. 28, n. 4, 2012.

FISHER, I. E.; GARNSEY, M. R.; HUGHES, M. E. Natural language processing in accounting, auditing and finance: a synthesis of the literature with a roadmap for future research. **Intelligent Systems in Accounting, Finance and Management**, [S.l.], 2016.

GUPTA, R.; GILL, N. S. Financial Statement Fraud Detection using Text Mining. **International Journal of Advanced Computer Science and Applications**, v. 3, n. 12, 2012.

HAGHIGHI, A.; VANDERWENDE, L. Exploring content models for multi-document summarization. In: HUMAN LANGUAGE TECHNOLOGIES: THE 2009 ANNUAL CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2009. **Proceedings. . .** Boulder, USA, 2009. p. 362–370.

HAN, J.; PEI, J.; KAMBER, M. **Data mining**: concepts and techniques. [S.l.]: Elsevier, 2011.

HEIDARI, M.; FELDEN, C. Toward Supporting Analytical Tasks in Financial Footnotes Analysis-A State of the Art. In: MULTIKONFERENZ WIRTSCHAFTSINFORMATIK (MKWI), 2014. **Proceedings. . .** Paderborn, Deutschland, 2014, p. 26–28.

HEIDARI, M.; FELDEN, C. Financial Footnote Analysis: developing a text mining approach. In: INTERNATIONAL CONFERENCE ON DATA MINING (DMIN), 2015. **Proceedings...** Las Vegas, USA, 2015a. p. 10.

HEIDARI, M.; FELDEN, C. Impact of Text Mining Application on Financial Footnotes Analysis. In: INTERNATIONAL CONFERENCE ON DESIGN SCIENCE RESEARCH IN INFORMATION SYSTEMS, 2015. **Proceedings. . .** Springer, Dublin, 2015b, p. 463–470.

JANVRIN, D.; MASCHA, M. F. The process of creating XBRL instance documents: a research framework. **Review of Business Information Systems (RBIS)**, [S.l.], v. 14, n. 2, p. 11–34, 2010.

JING, H.; MCKEOWN, K.R. Cut and Paste Based Text Summarization. In: PROCEEDINGS OF THE 1ST NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS CONFERENCE, 2000. **Proceedings. . .** Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, p. 178–185.

KUPIEC, J.; PEDERSEN, J.; CHEN, F. A Trainable Document Summarizer. In: PROCEEDINGS OF THE 18TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1995. **Proceedings. . .** New York, NY, USA: ACM, 1995, p. 68–73.

LIN, C.-Y. ROUGE: a package for automatic evaluation of summaries. In: TEXT SUMMARIZATION BRANCHES OUT: PROCEEDINGS OF THE ACL-04 WORKSHOP,

2004. **Proceedings. . .** Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81.

LIN, C.-Y.; HOVY, E. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOGY - VOLUME 1, 2003. **Proceedings. . .** Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. p. 71–78.

LIU, B. **Web data mining**: exploring hyperlinks, contents, and usage data. 2th Ed. New York: Springer, 2011.

LLORET, E.; ROMÁ-FERRI, M. T.; PALOMAR, M. COMPENDIUM: a text summarization system for generating abstracts of research papers. **Data & Knowledge Engineering**, [S.l.], v. 88, p. 164–175, 2013.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of research and development**, [S.l.], v. 2, n. 2, p. 159–165, 1958.

MANI, I. **Automatic Summarization**. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001.

MANNING, C.; RAGHAVAN, P.; SCHUTZE, H. **An introduction to information retrieval**. Cambridge: Cambridge university press, 2008.

MARTINS, E.; ASSAF NETO, A. **Administração financeira**: as finanças das empresas sob condições inflacionárias. São Paulo: Atlas, 1996.

MIHALCEA, R.; TARAU, P. A language independent algorithm for single and multiple document summarization. In: PROCEEDINGS OF THE INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2005. **Proceedings. . .** Korea, 2005.

MINER, G.; DELEN, D.; ELDER, J.; FAST, A.; HILL, T.; NISBET, R. A. **Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications**. Waltham: Academic Press, 2012.

NENKOVA, A.; MCKEOWN, K. A survey of text summarization techniques. In: MINING TEXT DATA, 2012. **Proceedings. . .** Boston: Springer, 2012, p. 43–76.

NENKOVA, A.; PASSONNEAU, R. J. Evaluating Content Selection in Summarization: the pyramid method. In: NAACL-HLT, 2004. **Proceedings. . .** Boston, 2004, p. 145–152.

ONO, K.; SUMITA, K.; MIIKE, S. Abstract Generation Based on Rhetorical Structure Extraction. In: CONFERENCE ON COMPUTATIONAL LINGUISTICS - VOLUME 1, 15,1994. **Proceedings. . .** Stroudsburg, PA, USA: Association for Computational Linguistics, 1994. p. 344–348.

ORENGO, V. M.; HUYCK, C. R. A Stemming Algorithmm for the Portuguese Language. In: 8th International Symposium on String Processing and Information Retrieval (SPIRE), 2001. **Proceedings. . .** Laguna de San Raphael, Chile, 2001, p. 186–193.

PARAMESWARAN, A.; GARCIA-MOLINA, H.; RAJARAMAN, A. Towards the Web of Concepts: extracting concepts from large datasets. **Proceedings of the VLDB Endowment**, [S.l.], v. 3, n. 1-2, p. 566–577, Sept. 2010.

PUTRA, L. D. **Understanding footnotes to financial statements**. Disponível em: <http://accounting-financial-tax.com/2008/08/understanding-footnotes-to-financial-statement>. Acesso em 12 mai. 2016.

RADEV, D. R.; HOVY, E.; MCKEOWN, K. Introduction to the Special Issue on Summarization. **Computational linguistics**, Cambridge, MA, USA, v. 28, n. 4, p. 399–408, Dec. 2002.

RADEV, D. R.; JING, H.; BUDZIKOWSKA, M. Centroid-based Summarization of Multiple Documents: sentence extraction, utility-based evaluation, and user studies. In: NAACL-ANLPWORKSHOP ON AUTOMATIC SUMMARIZATION - VOLUME 4, 2000. **Proceedings. . .** Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. p. 21–30.

RADEV, D. R.; TEUFEL, S.; SAGGION, H.; LAM, W.; BLITZER, J.; QI, H.; CELEBI, A.; LIU, D.; DRABEK, E. Evaluation Challenges in Large-scale Document Summarization. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS - VOLUME 1, 41, 2003. **Proceedings. . .** Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, p. 375–382.

REEVE, L. H.; HAN, H.; BROOKS, A. D. The Use of Domain-specific Concepts in Biomedical Text Summarization. **Information Processing & Management**, Tarrytown, NY, USA, v. 43, n. 6, p. 1765–1776, Nov. 2007.

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. Graph-based methods for multi-document summarization: exploring relationship maps, complex networks and discourse information. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 2012. **Anais. . .** [S.l.: s.n.], 2012. p. 260–271.

RICH, E.; KNIGHT, K. **Inteligência Artificial**. São Paulo: Makron Books, 1993.

SAGGION, H.; RADEV, D.; TEUFEL, S.; LAM, W.; STRASSEL, S. M. Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In: LREC 2002. **Proceedings. . .** Las Palmas, Gran Canaria, Spain, 2002, p. 747–754.

SALTON, G. **Automatic Text Processing**: the transformation, analysis, and retrieval of information by computer. Boston: Addison-Wesley Longman Publishing Co., Inc., 1989.

SALTON, G.; ALLAN, J. **Text retrieval using the vector processing model**. [S.l.]: Nevada Univ., Las Vegas, NV (United States), 1994.

SANKARASUBRAMANIAM, Y.; RAMANATHAN, K.; GHOSH, S. Text summarization using Wikipedia. **Information Processing & Management**, [S.l.], v. 50, n. 3, p. 443–461, 2014.

SCHWAB; C. **Financial Footnotes: Start Reading The Fine Print**. Disponível em: <http://www.investopedia.com>. Acesso em: 14 abr. 2016.

STEINBERGER, J.; JEŽEK, K. Using latent semantic analysis in text summarization and summary evaluation. In: ISIM'04, 2004. **Proceedings. . .** Amsterdam, 2004, p. 93–100.

STEINBERGER, J.; JEŽEK, K. Evaluation measures for text summarization. **Computing and Informatics**, [S.l.], v. 28, n. 2, p. 251–275, 2009.

SUANMALI, L.; SALIM, N.; BINWAHLAN, M. S. Fuzzy logic based method for improving text summarization. **International Journal of Computer Science and Information Security**, [S.l.], v. 2, n. 1, p. 4–10, 2009.

VANDERWENDE, L.; SUZUKI, H.; BROCKETT, C.; NENKOVA, A. Beyond SumBasic: task-focused summarization with sentence simplification and lexical expansion. **Information Processing & Management**, [S.l.], v. 43, n. 6, p. 1606–1618, 2007.

APÊNDICE A – EXEMPLO DE NOTAS EXPLICATIVAS DA SEÇÃO DE PRINCIPAIS PRÁTICAS CONTÁBEIS

Abaixo se apresenta um exemplo de uma parte da seção de principais práticas contábeis das notas explicativas da demonstração financeira de encerramento anual de 2015 da empresa Gerdau S.A. Nessa demonstração, antes das notas explicativas estavam o Relatório da Administração e todos os demonstrativos contábeis como o Balanço Patrimonial, DRE, DRA, DMPL, DFC e DVA. Porém para identificação do início das notas explicativas, sempre existirá um título constando a expressão "Notas Explicativas". Dentro das notas explicativas existem várias notas, porém aqui estão exemplificadas algumas notas referentes às principais práticas contábeis. Nesse exemplo elas constam na "Nota 2", mas não necessariamente tem que ser assim, e para identificar o início das principais práticas contábeis pode-se utilizar a expressão "Principais Práticas Contábeis" após a expressão "Notas Explicativas". Aqui são apresentados apenas alguns exemplos de notas explicativas das principais práticas contábeis, mas existem outras, totalizando mais de vinte notas em uma demonstração financeira.

GERDAU S.A.

NOTAS EXPLICATIVAS DA ADMINISTRAÇÃO ÀS DEMONSTRAÇÕES FINANCEIRAS INDIVIDUAIS DA CONTROLADORA E CONSOLIDADAS EM 31 DE DEZEMBRO DE 2015

....

NOTA 2 -RESUMO DAS PRINCIPAIS PRÁTICAS CONTÁBEIS

...

2.3 – Ativos financeiros

A Companhia valoriza os instrumentos financeiros derivativos pelo seu valor justo na data das Demonstrações Financeiras, sendo a principal evidência do valor justo a consideração das cotações obtidas junto aos participantes do mercado. O valor de mercado reconhecido em suas Demonstrações Financeiras da Controladora e Consolidadas pode não necessariamente representar o montante de caixa que a Companhia receberia ou pagaria, conforme apropriado, se a Companhia liquidasse as transações na data das Demonstrações Financeiras da Controladora e Consolidadas.

A Companhia classifica seus ativos financeiros, no reconhecimento inicial, sob as seguintes categorias: mensurados ao valor justo reconhecido no resultado, empréstimos e recebíveis e disponíveis para venda (quando aplicável). A classificação depende da finalidade para a qual os ativos financeiros foram adquiridos, como detalhado na nota 15.

a) Ativos financeiros ao valor justo reconhecido no resultado

Os ativos financeiros ao valor justo reconhecido no resultado são ativos financeiros mantidos para negociação e incluem Certificados de Depósitos Bancários - CDB e investimentos em títulos e valores mobiliários. Os ativos financeiros ao valor justo reconhecido no resultado são, inicialmente, reconhecidos pelo valor justo, e os custos da transação são debitados à demonstração do resultado.

b) Empréstimos e recebíveis

Os empréstimos e recebíveis são ativos financeiros não derivativos, com pagamentos fixos ou determináveis, que não são cotados em um mercado ativo. Os empréstimos e recebíveis da Companhia compreendem "Contas a receber de clientes e demais contas a receber", "Caixa e equivalentes de caixa" e "Depósitos judiciais". São apresentados como ativo circulante, exceto aqueles com prazo de vencimento superior a 12 meses após a data de emissão do balanço, os quais são classificados

como ativos não circulantes.

c) Instrumentos financeiros derivativos e atividades de hedge

Inicialmente, os derivativos são reconhecidos pelo valor justo na data em que um contrato de derivativos é celebrado e são, subsequentemente, remensurados ao seu valor justo. O método para reconhecer o ganho ou a perda resultante depende do fato do derivativo ser designado ou não como um instrumento de hedge nos casos de adoção da contabilidade de hedge (hedge accounting). Sendo este o caso, o método depende da natureza do item que está sendo protegido por hedge. Como descrito na nota 15, a Companhia adota a contabilidade de hedge (hedge accounting).

d) Derivativos mensurados ao valor justo reconhecido no resultado

Certos instrumentos derivativos não se qualificam para a contabilização de hedge. As variações no valor justo de qualquer um desses instrumentos derivativos são reconhecidas imediatamente na demonstração do resultado em "(Perdas) Ganhos com instrumentos financeiros, líquido".

e) Caixa e equivalentes de caixa

Caixa e equivalentes de caixa incluem caixa, contas bancárias e investimentos de curto prazo com liquidez imediata e vencimento original de 90 dias ou menos e com baixo risco de variação no valor de mercado, sendo demonstrados pelo custo e acrescido de juros auferidos, quando aplicável.

f) Aplicações financeiras

As aplicações financeiras estão classificadas como títulos para negociação são mensurados pelo seu valor justo reconhecido com contrapartida no resultado (títulos para negociação), em virtude do propósito do investimento ser a aplicação de recursos para obter ganhos de curto prazo. Os juros, correção monetária e variação cambial, quando aplicável, assim como as variações decorrentes da avaliação ao valor justo, são reconhecidos no resultado quando incorridos.

g) Contas a receber de clientes

Estão apresentadas a valores de custo amortizado, sendo que as contas a receber de clientes no mercado externo estão atualizadas com base nas taxas de câmbio vigentes na data das Demonstrações Financeiras. A provisão para riscos de crédito foi calculada com base na análise de riscos dos créditos, que contempla o histórico de perdas, a situação individual dos clientes, a situação do grupo econômico ao qual pertencem, as garantias reais para os débitos e a avaliação dos consultores jurídicos, e é considerada suficiente para cobrir eventuais perdas sobre os valores a receber. Informações referentes à abertura de contas a receber em valores a vencer e vencidos, além da provisão para risco de crédito estão demonstradas na nota 5.
A exposição máxima ao risco de crédito da Companhia, líquida da provisão para risco de crédito, é o valor das contas a receber. A qualidade do crédito do contas a receber a vencer é considerada adequada, sendo que o valor do risco efetivo de eventuais perdas no contas a receber de clientes encontra-se apresentado como provisão para risco de crédito.

h) Avaliação da recuperabilidade de ativos financeiros

Ativos financeiros são avaliados a cada data de balanço para identificação da recuperabilidade de ativos (impairment). Estes ativos financeiros são considerados ativos parcialmente ou totalmente não recuperáveis quando existem evidências de que um ou mais eventos tenham ocorrido após o reconhecimento inicial do ativo financeiro e que tenham impactado negativamente o fluxo estimado de caixa futuro do investimento. Os critérios utilizados para determinar se há evidência objetiva de uma perda por impairment incluem, entre outros fatores: (i) dificuldade financeira relevante do emissor ou devedor; e (ii) condições econômicas nacionais ou locais que se correlacionam com as inadimplências sobre os ativos na carteira.

2.4 – Estoques

Os estoques são avaliados com base no menor valor entre o custo histórico de aquisição e produção e o valor líquido realizável. O custo de aquisição e produção é acrescido de gastos relativos a transportes, armazenagem e impostos não recuperáveis.

O valor líquido realizável é o preço estimado de venda no curso normal dos negócios, deduzido dos custos estimados para conclusão e despesas de vendas diretamente relacionadas. Informações referentes à abertura do valor líquido realizável estão demonstradas na nota 6.

2.5 – Imobilizado

A Companhia utilizou o custo histórico, acrescido de correção monetária, quando aplicável nos termos da IAS 29, deduzido das respectivas depreciações, à exceção dos terrenos, que não são depreciados. A Companhia agrega mensalmente ao custo de aquisição do imobilizado em formação os custos de empréstimos e financiamentos considerando os seguintes critérios para capitalização: (a) o período de capitalização ocorre quando o imobilizado encontra-se em fase de construção, sendo encerrada a capitalização dos custos de empréstimos quando o item do imobilizado encontra-se disponível para utilização; (b) os custos de empréstimos são capitalizados considerando a taxa média ponderada dos empréstimos vigentes da data da capitalização ou a taxa específica, no caso de empréstimos para a aquisição de imobilizado; (c) os custos de empréstimos capitalizados mensalmente não excedem o valor das despesas de juros apuradas no período de capitalização; e (d) os custos de empréstimos capitalizados são depreciados considerando os mesmos critérios e vida útil determinados para o item do imobilizado ao qual foram incorporados.

A depreciação é calculada pelo método linear ajustado pelo nível de utilização de certos ativos, a taxas que levam em consideração a vida útil estimada dos bens e o valor residual estimado dos ativos no final de sua vida útil. O valor residual ao final da vida útil e a vida útil estimada dos bens são revisados e ajustados, se necessário, na data de encerramento do exercício.

Custos subsequentes são incorporados ao valor residual do imobilizado ou reconhecidos como item específico, conforme apropriado, somente se os benefícios econômicos associados a estes itens forem prováveis e os valores mensurados de forma confiável. O saldo residual do item substituído é baixado. Demais reparos e manutenções são reconhecidas diretamente no resultado quando incorridas.

Direitos de exploração mineral são classificados como Terrenos, Prédios e Construções no grupo de imobilizado. Gastos com exploração são reconhecidos como despesas até se estabelecer a viabilidade da atividade de mineração e após esse período os custos subsequentes são capitalizados. Custos para o desenvolvimento de novas jazidas de minério, ou para a expansão da capacidade das minas em operação são capitalizados e amortizados com base na quantidade de minério extraída. Os gastos de remoção de estéril (custos associados com remoção de estéril e outros materiais residuais), incorridos durante a fase de desenvolvimento de uma mina, antes da fase de produção, são contabilizados como parte dos custos depreciáveis de desenvolvimento. Subsequentemente, estes custos são depreciados durante o período de vida útil da mina. Os gastos com remoção de estéril, após o início da fase produtiva da mina, são tratados como custo de produção. A exaustão das minas é calculada com base na quantidade de minério extraída.

O valor residual dos itens do imobilizado é reduzido imediatamente ao seu valor recuperável quando o saldo residual exceder o valor recuperável.

...

APÊNDICE B – LISTA DAS EMPRESAS DO IBOVESPA

Empresa	Notas Explicativas extraída
AMBEV	Sim
Banco do Brasil	Sim
BB Seguridade	Sim
BM&F Bovespa	Sim
BR Malls	Sim
Bradesco	Sim
Bradespar	Sim
Braskem	Sim
BRF	Sim
CCR	Sim
CEMIG	Sim
CESP	Sim
CETIP	Sim
CIELO	Sim
COPEL	Sim
Cosan	Sim
CPFL	Sim
CYRELA	Sim
Ecorodovias	Sim
Embraer	Sim
Energias BR	Sim
Equatorial	Não
Estacio	Sim
Fibria	Sim
Gerdau	Sim
Gerdau Met	Sim
Hypermarcas	Sim
Itau	Sim
ItauUnibanco	Sim
JBS	Sim
Klabin	Sim
Kroton	Sim
Localiza	Sim
Lojas Americanas	Sim
Lojas Renner	Sim
Marfrig	Sim
MRV	Sim
Multiplan	Não
Natura	Sim
Pão de Açúcar	Sim
Petrobras	Sim
Qualicorp	Sim
Raiadrogasil	Sim
Rumo Log	Sim
SABESP	Sim
Santander	Sim
Siderúrgica Nacional	Sim
Smiles	Sim
Suzano Papel	Sim
Telefonica	Sim
TIM Participações	Sim
Ultrapar	Sim
Usiminas	Sim
Vale	Sim
WEG	Sim

APÊNDICE C – EXEMPLOS DE RESUMOS OBTIDOS COM A APLICAÇÃO DOS ALGORITMOS DE SUMARIZAÇÃO SOBRE AS NOTAS EXPLICATIVAS

Exemplo de nota explicativa do Estoque

Estoques: Os estoques são apresentados pelo menor valor entre o valor de custo e o valor líquido realizável. Os custos dos estoques são determinados pelo método do custo médio. O valor líquido realizável corresponde ao preço de venda estimado dos estoques, deduzido de todos os custos estimados para conclusão e custos necessários para realizar a venda.

- Resumo obtido pelo algoritmo de sumarização LSA

O valor líquido realizável corresponde ao preço de venda estimado dos estoques, deduzido de todos os custos estimados para conclusão e dos custos necessários para realizar a venda.

- Resumo obtido pelo algoritmo de sumarização LexRank

Os estoques são apresentados pelo menor valor entre o valor de custo e o valor líquido realizável.

- Resumo obtido pelo algoritmo de sumarização KLSum

Os estoques são apresentados pelo menor valor entre o valor de custo e o valor líquido realizável.

- Resumo obtido pelo algoritmo de sumarização SumBasic

Os estoques são apresentados pelo menor valor entre o valor de custo e o valor líquido realizável.

- Resumo obtido pelo algoritmo de sumarização SimpleSummarizer

Os estoques são apresentados pelo menor valor entre o valor de custo e o valor líquido realizável.

- Resumo obtido pelo algoritmo de sumarização NaiveSummarizer

O valor líquido realizável corresponde ao preço de venda estimado dos estoques, deduzido de todos os custos estimados para conclusão e dos custos necessários para realizar a venda.

Exemplo de nota explicativa do Imobilizado

Imobilizado: Terrenos, edificações, imobilizações em andamento, móveis e utensílios e equipamentos estão demonstrados ao valor de custo, deduzidos de depreciação e perdas por redução ao valor recuperável acumuladas. São registrados como parte dos custos das imobilizações em andamento os honorários profissionais e, no caso de ativos qualificáveis, os custos de empréstimos capitalizados de acordo com a política contábil do Grupo. Tais imobilizações são classificadas nas categorias adequadas do imobilizado quando concluídas e prontas para o uso pretendido. A depreciação desses ativos inicia-se quando eles estão prontos para o uso pretendido na mesma base dos outros ativos imobilizados.
Os terrenos não sofrem depreciação.
A depreciação é reconhecida com base na vida útil estimada de cada ativo pelo método linear, de modo que o valor do custo menos o seu valor residual após sua vida útil seja integralmente baixado (exceto para terrenos e construções em andamento). A vida útil estimada, os valores residuais e os métodos de depreciação são revisados no fim da data do balanço patrimonial e o efeito de quaisquer mudanças nas estimativas é contabilizado prospectivamente.
Ativos mantidos por meio de arrendamento financeiro são depreciados pela vida útil esperada, da mesma forma que os ativos próprios, ou por um período inferior, se aplicável, conforme termos do contrato de arrendamento em questão.
Um item do imobilizado é baixado após alienação ou quando não há benefícios econômicos futuros resultantes do uso contínuo do ativo. Quaisquer ganhos ou perdas na venda ou baixa de um item do imobilizado são determinados pela diferença entre os valores recebidos na venda e o valor contábil do ativo e são reconhecidos no resultado.

- Resumo obtido pelo algoritmo de sumarização LSA

A depreciação é reconhecida com base na vida útil estimada de cada ativo pelo método linear, de modo que o valor do custo menos o seu valor residual após sua vida útil seja integralmente baixado (exceto para terrenos e construções em andamento).
Um item do imobilizado é baixado após alienação ou quando não há benefícios econômicos futuros resultantes do uso contínuo do ativo.

- Resumo obtido pelo algoritmo de sumarização LexRank

A depreciação é reconhecida com base na vida útil estimada de cada ativo pelo método linear, de modo que o valor do custo menos o seu valor residual após sua vida útil seja integralmente baixado (exceto para terrenos e construções em andamento).
Um item do imobilizado é baixado após alienação ou quando não há benefícios econômicos futuros resultantes do uso contínuo do ativo.

- Resumo obtido pelo algoritmo de sumarização KLSum

Terrenos, edificações, imobilizações em andamento, móveis e utensílios e equipamentos estão demonstrados ao valor de custo, deduzidos de depreciação e perdas por redução ao valor recuperável acumuladas.
A depreciação é reconhecida com base na vida útil estimada de cada ativo pelo método linear, de modo que o valor do custo menos o seu valor residual após sua vida útil seja integralmente baixado (exceto para terrenos e construções em andamento).

- Resumo obtido pelo algoritmo de sumarização SumBasic

A depreciação desses ativos inicia-se quando eles estão prontos para o uso pretendido na mesma base dos outros ativos imobilizados.
Os terrenos não sofrem depreciação.
A depreciação é reconhecida com base na vida útil estimada de cada ativo pelo método linear, de modo que o valor do custo menos o seu valor residual após sua vida útil seja integralmente baixado (exceto para terrenos e construções em andamento).

- Resumo obtido pelo algoritmo de sumarização SimpleSummarizer

São registrados como parte dos custos das imobilizações em andamento os honorários profissionais e, no caso de ativos qualificáveis, os custos de empréstimos capitalizados de acordo com a política contábil do Grupo.
A depreciação desses ativos inicia-se quando eles estão prontos para o uso pretendido na mesma base dos outros ativos imobilizados.
Ativos mantidos por meio de arrendamento financeiro são depreciados pela vida útil esperada, da mesma forma que os ativos próprios, ou por um período inferior, se aplicável, conforme termos do contrato de arrendamento em questão.

- Resumo obtido pelo algoritmo de sumarização NaiveSummarizer

A depreciação é reconhecida com base na vida útil estimada de cada ativo pelo método linear, de modo que o valor do custo menos o seu valor residual após sua vida útil seja integralmente baixado (exceto para terrenos e construções em andamento).
Ativos mantidos por meio de arrendamento financeiro são depreciados pela vida útil esperada, da mesma forma que os ativos próprios, ou por um período inferior, se aplicável, conforme termos do contrato de arrendamento em questão.

Exemplo de nota explicativa do Reconhecimento da Receita

Reconhecimento da Receita: A receita é mensurada pelo valor justo da contrapartida recebida ou a receber, deduzida de quaisquer estimativas de devoluções, descontos comerciais e/ou bonificações concedidos ao comprador e outras deduções similares.
Vendas de produtos
A receita de vendas de produtos é reconhecida quando os produtos são entregues e a posse foi passada nesse prazo de tal forma que todas as seguintes condições forem satisfeitas:
o Grupo transferiu para o comprador os riscos e benefícios significativos relacionados à propriedade dos produtos;
o Grupo não mantém envolvimento continuado na gestão dos produtos vendidos em grau normalmente associado à propriedade nem controle efetivo sobre tais produtos;
o valor da receita pode ser mensurado com confiabilidade;

é provável que os benefícios econômicos associados à transação fluirão para o Grupo; e os custos incorridos ou a serem incorridos relacionados à transação podem ser mensurados com confiabilidade.
Mais especificamente, a receita de venda de produtos é reconhecida quando os produtos são entregues e a titularidade legal é transferida.

- Resumo obtido pelo algoritmo de sumarização LSA

Vendas de produtos A receita de vendas de produtos é reconhecida quando os produtos são entregues e a posse foi passada nesse prazo de tal forma que todas as seguintes condições forem satisfeitas: o Grupo transferiu para o comprador os riscos e benefícios significativos relacionados à propriedade dos produtos; o Grupo não mantém envolvimento continuado na gestão dos produtos vendidos em grau normalmente associado à propriedade nem controle efetivo sobre tais produtos; o valor da receita pode ser mensurado com confiabilidade; é provável que os benefícios econômicos associados à transação fluirão para o Grupo; e os custos incorridos ou a serem incorridos relacionados à transação podem ser mensurados com confiabilidade.

- Resumo obtido pelo algoritmo de sumarização LexRank

A receita é mensurada pelo valor justo da contrapartida recebida ou a receber, deduzida de quaisquer estimativas de devoluções, descontos comerciais e/ou bonificações concedidos ao comprador e outras deduções similares.

- Resumo obtido pelo algoritmo de sumarização KLSum

A receita é mensurada pelo valor justo da contrapartida recebida ou a receber, deduzida de quaisquer estimativas de devoluções, descontos comerciais e/ou bonificações concedidos ao comprador e outras deduções similares.
Mais especificamente, a receita de venda de produtos é reconhecida quando os produtos são entregues e a titularidade legal é transferida.

- Resumo obtido pelo algoritmo de sumarização SumBasic

Mais especificamente, a receita de venda de produtos é reconhecida quando os produtos são entregues e a titularidade legal é transferida.

- Resumo obtido pelo algoritmo de sumarização SimpleSummarizer

Vendas de produtos A receita de vendas de produtos é reconhecida quando os produtos são entregues e a posse foi passada nesse prazo de tal forma que todas as seguintes condições forem satisfeitas: o Grupo transferiu para o comprador os riscos e benefícios significativos relacionados à propriedade dos produtos; o Grupo não mantém envolvimento continuado na gestão dos produtos vendidos em grau normalmente associado à propriedade nem controle efetivo sobre tais produtos; o valor da receita pode ser mensurado com confiabilidade; é provável que os benefícios econômicos associados à transação fluirão para o Grupo; e os custos incorridos ou a serem incorridos relacionados à transação podem ser mensurados com confiabilidade.

- Resumo obtido pelo algoritmo de sumarização NaiveSummarizer

Vendas de produtos A receita de vendas de produtos é reconhecida quando os produtos são entregues e a posse foi passada nesse prazo de tal forma que todas as seguintes condições forem satisfeitas: o Grupo transferiu para o comprador os riscos e benefícios significativos relacionados à propriedade dos produtos; o Grupo não mantém envolvimento continuado na gestão dos produtos vendidos em grau normalmente associado à propriedade nem controle efetivo sobre tais produtos; o valor da receita pode ser mensurado com confiabilidade; é provável que os benefícios econômicos associados à transação fluirão para o Grupo; e os custos incorridos ou a serem incorridos relacionados à transação podem ser mensurados com confiabilidade.

APÊNDICE D – RESULTADO DAS AVALIAÇÕES NA NOTA EXPLICATIVA DO ESTOQUE

LSA

Empresa	ROUGE com base no resumo de					Média
	LexRank	KLSum	sumBasic	simpleSum	naiveSum	
AMBEV	0,49	0,02	0,13	0,13	0,02	0,16
BRF	0,00	1,00	0,00	0,00	0,00	0,20
CEMIG	0,14	0,00	0,14	0,14	0,00	0,09
COPEL	1,00	0,00	0,00	0,00	1,00	0,40
Cosan	1,00	0,10	1,00	1,00	0,00	0,62
Embraer	0,04	0,29	0,02	0,00	0,02	0,07
Energias BR	0,00	0,00	0,00	0,00	1,00	0,20
Natura	0,00	0,00	0,00	0,00	0,00	0,00
Sid Nacional	0,64	0,00	0,10	0,07	0,00	0,16
Fibria	0,05	0,03	0,05	0,05	0,00	0,04
Gerdau Met	0,14	1,00	0,14	0,14	0,00	0,29
Gerdau	0,14	1,00	0,14	0,14	0,00	0,29
Hypermarcas	1,00	0,00	0,14	0,00	0,00	0,23
Itau	0,54	0,09	0,00	0,54	0,00	0,23
JBS	0,00	1,00	0,00	0,00	0,00	0,20
Klabin	1,00	0,00	0,04	1,00	1,00	0,61
Kroton	0,13	0,13	0,13	0,13	1,00	0,31
Lojas Americanas	1,00	1,00	1,00	1,00	0,11	0,82
Lojas Renner	0,00	0,00	0,00	0,00	0,00	0,00
Marfrig	1,00	1,00	1,00	1,00	1,00	1,00
Pão de Açúcar	0,00	0,00	0,00	0,03	0,03	0,01
Petrobras	0,71	0,09	0,09	0,02	0,09	0,20
RaiaDrogasil	0,07	0,03	0,05	0,07	0,21	0,09
SABESP	1,00	1,00	1,00	1,00	1,00	1,00
Suzano Papel	0,00	0,00	0,00	0,00	0,00	0,00
Telefonica	0,04	0,04	0,04	0,06	1,00	0,23
TIM Participações	0,00	0,00	0,07	0,07	1,00	0,23
ULTRAPAR	0,00	0,73	0,42	0,00	1,00	0,43
USIMINAS	1,00	0,09	0,09	1,00	1,00	0,64
Vale	0,00	0,00	0,00	0,00	0,00	0,00
WEG	1,00	0,00	0,00	1,00	0,00	0,40
Média geral						0,29
Desvio padrão						0,28

LexRank

Empresa	ROUGE com base no resumo de					Média
	LSA	**KLSum**	**sumBasic**	**simpleSum**	**naiveSum**	
AMBEV	0,46	0,24	0,43	0,43	0,07	0,32
BRF	0,00	0,00	0,00	1,00	0,00	0,20
CEMIG	0,08	0,00	1,00	1,00	0,00	0,42
COPEL	1,00	0,00	0,00	0,00	1,00	0,40
Cosan	1,00	0,10	1,00	1,00	0,00	0,62
Embraer	0,04	0,69	0,06	1,00	0,06	0,37
Energias BR	0,00	1,00	1,00	1,00	0,00	0,60
Natura	0,00	1,00	0,00	0,00	0,00	0,20
Sid Nacional	0,62	0,02	0,55	0,56	0,04	0,36
Fibria	0,05	0,63	1,00	1,00	0,00	0,54
Gerdau Met	0,18	0,18	1,00	1,00	0,08	0,49
Gerdau	0,18	0,18	1,00	1,00	0,08	0,49
Hypermarcas	1,00	0,00	0,14	0,00	0,00	0,23
Itau	0,47	0,44	1,00	1,00	0,03	0,59
JBS	0,00	0,00	0,00	1,00	0,00	0,20
Klabin	1,00	0,00	0,04	1,00	1,00	0,61
Kroton	0,07	1,00	1,00	1,00	0,07	0,63
Lojas Americanas	1,00	1,00	1,00	1,00	0,11	0,82
Lojas Renner	0,00	1,00	1,00	1,00	1,00	0,80
Marfrig	1,00	1,00	1,00	1,00	1,00	1,00
Pão de Açúcar	0,00	0,13	0,13	0,03	0,03	0,06
Petrobras	1,00	0,27	0,09	0,02	0,09	0,30
RaiaDrogasil	0,03	0,00	0,34	1,00	0,05	0,28
SABESP	1,00	1,00	1,00	1,00	1,00	1,00
Suzano Papel	0,00	0,06	1,00	1,00	0,06	0,42
Telefonica	0,03	0,08	1,00	1,00	0,03	0,43
TIM Participações	0,00	1,00	0,11	0,11	0,00	0,24
ULTRAPAR	0,00	0,02	0,06	1,00	0,00	0,22
USIMINAS	1,00	0,09	0,09	1,00	1,00	0,64
Vale	0,00	1,00	1,00	1,00	0,00	0,60
WEG	1,00	0,00	0,00	1,00	0,00	0,40
Média geral						0,47
Desvio padrão						0,23

KLSum

Empresa	ROUGE com base no resumo de					Média
	LSA	**LexRank**	**sumBasic**	**simpleSum**	**naiveSum**	
AMBEV	0,02	0,28	0,08	0,08	1,00	0,29
BRF	1,00	0,00	0,00	0,00	0,00	0,20
CEMIG	0,00	0,00	0,00	0,00	0,00	0,00
COPEL	0,00	0,00	1,00	1,00	0,00	0,40
Cosan	0,13	0,13	0,13	0,13	0,00	0,11
Embraer	0,28	0,67	0,06	1,00	0,06	0,42
Energias BR	0,00	1,00	1,00	1,00	0,00	0,60
Natura	0,00	1,00	0,00	0,00	0,00	0,20
Sid Nacional	0,00	0,03	0,10	0,04	1,00	0,23
Fibria	0,05	1,00	1,00	1,00	0,00	0,61
Gerdau Met	1,00	0,14	0,14	0,14	0,00	0,29
Gerdau	1,00	0,14	0,14	0,14	0,00	0,29
Hypermarcas	0,00	0,00	0,04	1,00	1,00	0,41
Itau	0,09	0,54	1,00	0,54	0,03	0,44
JBS	1,00	0,00	0,00	0,00	0,00	0,20
Klabin	0,00	0,00	0,17	0,00	0,00	0,03
Kroton	0,07	1,00	1,00	1,00	0,07	0,63
Lojas Americanas	1,00	1,00	1,00	1,00	0,11	0,82
Lojas Renner	0,00	1,00	1,00	1,00	1,00	0,80
Marfrig	1,00	1,00	1,00	1,00	1,00	1,00
Pão de Açúcar	0,00	0,25	1,00	0,08	0,08	0,28
Petrobras	0,19	0,39	1,00	0,10	1,00	0,54
RaiaDrogasil	0,03	0,00	0,63	0,00	0,05	0,14
SABESP	1,00	1,00	1,00	1,00	1,00	1,00
Suzano Papel	0,00	0,05	0,05	0,05	1,00	0,23
Telefonica	0,03	0,07	0,07	0,12	0,03	0,06
TIM Participações	0,00	1,00	0,11	0,11	0,00	0,24
ULTRAPAR	0,73	0,02	0,00	0,02	1,00	0,35
USIMINAS	0,04	0,04	1,00	0,04	0,04	0,23
Vale	0,00	1,00	1,00	1,00	0,00	0,60
WEG	0,00	0,00	1,00	0,00	1,00	0,40
Média geral						0,39
Desvio padrão						0,26

sumBasic

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	simpleSum	naiveSum	
AMBEV	0,10	0,36	0,06	1,00	0,07	0,32
BRF	0,00	0,00	0,00	0,00	0,08	0,02
CEMIG	0,08	1,00	0,00	1,00	0,00	0,42
COPEL	0,00	0,00	1,00	1,00	0,00	0,40
Cosan	1,00	1,00	0,10	1,00	0,00	0,62
Embraer	0,02	0,07	0,07	0,10	1,00	0,25
Energias BR	0,00	1,00	1,00	1,00	0,00	0,60
Natura	0,00	0,00	0,00	1,00	1,00	0,40
Sid Nacional	0,08	0,44	0,07	0,52	0,11	0,24
Fibria	0,05	1,00	0,63	1,00	0,00	0,54
Gerdau Met	0,18	1,00	0,18	1,00	0,08	0,49
Gerdau	0,18	1,00	0,18	1,00	0,08	0,49
Hypermarcas	0,15	0,15	0,03	0,03	0,03	0,08
Itau	0,00	0,43	0,35	0,43	0,03	0,25
JBS	0,00	0,00	0,00	0,00	1,00	0,20
Klabin	0,03	0,03	0,17	0,03	0,03	0,06
Kroton	0,07	1,00	1,00	1,00	0,07	0,63
Lojas Americanas	1,00	1,00	1,00	1,00	0,11	0,82
Lojas Renner	0,00	1,00	1,00	1,00	1,00	0,80
Marfrig	1,00	1,00	1,00	1,00	1,00	1,00
Pão de Açúcar	0,00	0,25	1,00	0,08	0,08	0,28
Petrobras	0,15	0,11	0,78	0,10	1,00	0,43
RaiaDrogasil	0,06	1,00	0,72	1,00	0,11	0,58
SABESP	1,00	1,00	1,00	1,00	1,00	1,00
Suzano Papel	0,00	1,00	0,06	1,00	0,06	0,42
Telefonica	0,03	1,00	0,08	1,00	0,03	0,43
TIM Participações	0,13	0,43	0,43	1,00	0,13	0,42
ULTRAPAR	0,25	0,04	0,00	0,04	0,00	0,07
USIMINAS	0,04	0,04	1,00	0,04	0,04	0,23
Vale	0,00	1,00	1,00	1,00	0,00	0,60
WEG	0,00	0,00	1,00	0,00	1,00	0,40
Média geral						0,43
Desvio padrão						0,25

simpleSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	sumBasic	naiveSum	
AMBEV	0,10	0,36	0,06	1,00	0,07	0,32
BRF	0,00	1,00	0,00	0,00	0,00	0,20
CEMIG	0,08	1,00	0,00	1,00	0,00	0,42
COPEL	0,00	0,00	1,00	1,00	0,00	0,40
Cosan	1,00	1,00	0,10	1,00	0,00	0,62
Embraer	0,00	0,67	0,69	0,06	0,06	0,30
Energias BR	0,00	1,00	1,00	1,00	0,00	0,60
Natura	0,00	0,00	0,00	1,00	1,00	0,40
Sid Nacional	0,05	0,42	0,02	0,48	0,04	0,20
Fibria	0,05	1,00	0,63	1,00	0,00	0,54
Gerdau Met	0,18	1,00	0,18	1,00	0,08	0,49
Gerdau	0,18	1,00	0,18	1,00	0,08	0,49
Hypermarcas	0,00	0,00	1,00	0,04	1,00	0,41
Itau	0,47	1,00	0,44	1,00	0,03	0,59
JBS	0,00	1,00	0,00	0,00	0,00	0,20
Klabin	1,00	1,00	0,00	0,04	1,00	0,61
Kroton	0,07	1,00	1,00	1,00	0,07	0,63
Lojas Americanas	1,00	1,00	1,00	1,00	0,11	0,82
Lojas Renner	0,00	1,00	1,00	1,00	1,00	0,80
Marfrig	1,00	1,00	1,00	1,00	1,00	1,00
Pão de Açúcar	0,05	0,08	0,13	0,13	1,00	0,28
Petrobras	0,04	0,03	0,07	0,09	0,09	0,06
RaiaDrogasil	0,03	1,00	0,00	0,34	0,05	0,28
SABESP	1,00	1,00	1,00	1,00	1,00	1,00
Suzano Papel	0,00	1,00	0,06	1,00	0,06	0,42
Telefonica	0,03	0,61	0,08	0,61	0,03	0,27
TIM Participações	0,13	0,43	0,43	1,00	0,13	0,42
ULTRAPAR	0,00	1,00	0,02	0,06	0,00	0,22
USIMINAS	1,00	1,00	0,09	0,09	1,00	0,64
Vale	0,00	1,00	1,00	1,00	0,00	0,60
WEG	1,00	1,00	0,00	0,00	0,00	0,40
Média geral						0,47
Desvio padrão						0,23

naiveSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	sumBasic	simpleSum	
AMBEV	0,02	0,06	0,80	0,08	0,08	0,21
BRF	0,00	0,00	0,00	0,07	0,00	0,01
CEMIG	0,00	0,00	0,00	0,00	0,00	0,00
COPEL	1,00	1,00	0,00	0,00	0,00	0,40
Cosan	0,00	0,00	0,00	0,00	0,00	0,00
Embraer	0,02	0,07	0,07	1,00	0,10	0,25
Energias BR	1,00	0,00	0,00	0,00	0,00	0,20
Natura	0,00	0,00	0,00	1,00	1,00	0,40
Sid Nacional	0,00	0,03	0,68	0,10	0,04	0,17
Fibria	0,00	0,00	0,00	0,00	0,00	0,00
Gerdau Met	0,00	0,10	0,00	0,10	0,10	0,06
Gerdau	0,00	0,10	0,00	0,10	0,10	0,06
Hypermarcas	0,00	0,00	1,00	0,04	1,00	0,41
Itau	0,00	0,04	0,03	0,08	0,04	0,04
JBS	0,00	0,00	0,00	1,00	0,00	0,20
Klabin	1,00	1,00	0,00	0,04	1,00	0,61
Kroton	1,00	0,13	0,13	0,13	0,13	0,31
Lojas Americanas	0,17	0,17	0,17	0,17	0,17	0,17
Lojas Renner	0,00	1,00	1,00	1,00	1,00	0,80
Marfrig	1,00	1,00	1,00	1,00	1,00	1,00
Pão de Açúcar	0,05	0,08	0,13	0,13	1,00	0,28
Petrobras	0,15	0,11	0,78	1,00	0,10	0,43
RaiaDrogasil	0,24	0,14	0,06	0,10	0,14	0,13
SABESP	1,00	1,00	1,00	1,00	1,00	1,00
Suzano Papel	0,00	0,05	1,00	0,05	0,05	0,23
Telefonica	1,00	0,04	0,04	0,04	0,06	0,23
TIM Participações	1,00	0,00	0,00	0,07	0,07	0,23
ULTRAPAR	0,73	0,00	0,73	0,00	0,00	0,29
USIMINAS	1,00	1,00	0,09	0,09	1,00	0,64
Vale	0,00	0,00	0,00	0,00	0,00	0,00
WEG	0,00	0,00	1,00	1,00	0,00	0,40
Média geral						0,29
Desvio padrão						0,27

APÊNDICE E – RESULTADO DAS AVALIAÇÕES NA NOTA EXPLICATIVA DO IMOBILIZADO

LSA

Empresa	ROUGE com base no resumo de					Média
	LexRank	KLSum	sumBasic	simpleSum	naiveSum	
AMBEV	0,33	0,03	0,50	0,19	0,01	0,21
BM&F Bovespa S,A,	0,05	0,00	0,03	1,00	0,00	0,22
Bradesco	0,03	0,03	0,02	0,03	0,02	0,02
BRF	0,66	0,04	0,03	0,37	0,04	0,23
CCR	0,32	0,01	0,10	0,76	0,42	0,32
CEMIG	0,88	0,04	0,28	0,45	0,41	0,41
CESP	0,37	0,37	0,03	0,01	0,51	0,26
CETIP	0,00	1,00	0,00	0,00	0,00	0,20
CIELO	0,48	0,09	1,00	1,00	0,09	0,53
COPEL	0,02	1,00	0,02	0,02	0,02	0,22
Cosan	0,09	0,87	0,24	0,34	0,54	0,41
CPFL	0,00	0,39	0,02	0,39	0,02	0,16
CYRELA	1,00	0,00	0,08	1,00	0,00	0,42
Embraer	0,30	0,28	0,03	0,30	0,04	0,19
Estacio	0,00	0,00	0,00	0,00	0,00	0,00
Natura	0,00	0,40	0,00	0,47	0,00	0,17
Sid Nacional	0,73	0,05	0,12	0,46	0,08	0,29
Fibria	0,05	0,00	0,04	0,00	0,51	0,12
Gerdau Met	0,07	0,07	0,12	0,07	0,07	0,08
Gerdau	0,07	0,07	0,12	0,07	0,07	0,08
Hypermarcas	0,00	0,74	0,04	0,01	0,57	0,27
Itau	0,32	0,04	0,30	0,00	0,01	0,14
ItauUnibanco	1,00	1,00	0,00	0,00	0,00	0,40
JBS	0,18	0,03	0,02	0,05	0,45	0,15
Klabin	1,00	0,16	0,06	0,06	0,06	0,27
Kroton	0,69	1,00	0,03	0,03	0,03	0,36
Lojas Americanas	0,47	0,02	0,31	0,00	0,03	0,16
Lojas Renner	1,00	1,00	0,14	0,00	1,00	0,63
Marfrig	1,00	1,00	0,00	0,00	1,00	0,60
MRV	1,00	0,07	1,00	1,00	1,00	0,81
Pão de Açúcar	0,02	0,03	0,03	0,48	0,04	0,12
Petrobras	0,28	0,51	0,30	0,48	0,31	0,37
Qualicorp	0,26	0,07	0,14	0,46	0,47	0,28
RaiaDrogasil	0,38	0,00	0,38	0,37	0,00	0,22
Rumo Log	0,52	0,26	0,03	0,07	0,08	0,19
SABESP	0,65	0,00	0,00	0,63	0,00	0,25
Santander	0,17	0,69	0,17	0,69	0,17	0,38
Suzano Papel	0,00	0,12	0,13	0,12	0,12	0,10
Telefonica	0,35	0,26	0,37	0,05	0,45	0,29
TIM Participações	0,35	0,01	0,03	0,21	0,04	0,13
ULTRAPAR	0,00	1,00	0,19	0,00	0,00	0,24
USIMINAS	0,55	0,53	0,55	0,07	1,00	0,54
Vale	0,00	0,01	0,02	0,02	0,02	0,01
WEG	0,49	0,09	0,47	0,36	0,09	0,30
Média geral						0,27
Desvio padrão						0,17

LexRank

Empresa	ROUGE com base no resumo de					Média
	LSA	KLSum	sumBasic	simpleSum	naiveSum	
AMBEV	0,29	0,64	0,16	0,84	0,64	0,51
BM&F Bovespa S,A,	0,05	0,08	0,47	0,05	0,08	0,14
Bradesco	0,04	1,00	0,02	1,00	0,02	0,41
BRF	0,58	0,35	0,07	0,41	0,07	0,29
CCR	0,36	0,83	0,02	0,40	1,00	0,52
CEMIG	0,82	0,04	0,40	0,44	0,41	0,42
CESP	0,32	0,56	0,08	0,27	0,04	0,25
CETIP	0,00	0,00	0,09	1,00	1,00	0,42
CIELO	0,43	0,00	1,00	1,00	0,00	0,49
COPEL	0,02	0,02	0,39	1,00	1,00	0,49
Cosan	0,06	0,06	0,15	0,70	0,28	0,25
CPFL	0,00	0,00	0,56	0,00	0,00	0,11
CYRELA	1,00	0,00	0,08	1,00	0,00	0,42
Embraer	0,28	0,27	0,42	0,30	0,07	0,27
Estacio	0,00	0,79	0,84	0,05	1,00	0,54
Natura	0,00	0,58	1,00	0,00	0,02	0,32
Sid Nacional	0,77	0,06	0,12	0,59	0,28	0,36
Fibria	0,04	0,00	0,56	0,00	0,00	0,12
Gerdau Met	0,04	0,14	1,00	0,07	0,07	0,26
Gerdau	0,04	0,14	1,00	0,07	0,07	0,26
Hypermarcas	0,00	0,00	0,42	0,32	0,00	0,15
Itau	0,36	0,62	0,21	0,66	0,57	0,49
ItauUnibanco	1,00	1,00	0,00	0,00	0,00	0,40
JBS	0,20	0,42	0,11	0,40	0,06	0,24
Klabin	0,44	0,02	0,00	0,00	0,00	0,09
Kroton	0,30	0,30	0,00	0,00	0,00	0,12
Lojas Americanas	0,63	0,02	0,21	0,00	0,03	0,18
Lojas Renner	1,00	1,00	0,14	0,00	1,00	0,63
Marfrig	1,00	1,00	0,00	0,00	1,00	0,60
MRV	1,00	0,07	1,00	1,00	1,00	0,81
Pão de Açúcar	0,04	0,26	0,74	0,06	0,02	0,22
Petrobras	0,35	0,37	0,31	0,30	0,11	0,29
Qualicorp	0,23	0,19	0,21	0,13	0,38	0,23
RaiaDrogasil	0,49	0,09	1,00	0,26	0,07	0,38
Rumo Log	0,62	0,34	0,41	0,32	0,12	0,36
SABESP	0,60	0,00	0,27	0,61	0,00	0,30
Santander	0,16	0,09	1,00	0,09	1,00	0,47
Suzano Papel	0,00	0,12	0,00	0,12	0,12	0,07
Telefonica	0,33	0,63	0,47	0,42	0,42	0,45
TIM Participações	0,98	0,17	0,05	0,22	0,04	0,29
ULTRAPAR	0,00	0,00	0,00	1,00	1,00	0,40
USIMINAS	0,45	0,53	1,00	0,07	0,45	0,50
Vale	0,00	0,03	0,00	0,00	0,00	0,01
WEG	0,59	0,09	0,47	0,00	0,09	0,25
Média geral						0,34
Desvio padrão						0,17

KLSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	sumBasic	simpleSum	naiveSum	
AMBEV	0,03	0,65	0,02	0,82	0,64	0,43
BM&F Bovespa S,A,	0,00	0,08	0,05	0,00	1,00	0,23
Bradesco	0,04	1,00	0,02	1,00	0,02	0,41
BRF	0,03	0,34	0,37	0,66	0,41	0,36
CCR	0,01	0,66	0,12	0,01	0,56	0,27
CEMIG	0,03	0,04	0,32	0,06	0,38	0,17
CESP	0,32	0,56	0,48	0,46	0,51	0,46
CETIP	1,00	0,00	0,00	0,00	0,00	0,20
CIELO	0,08	0,00	0,00	0,00	1,00	0,22
COPEL	1,00	0,02	0,02	0,02	0,02	0,22
Cosan	0,70	0,06	0,03	0,33	0,53	0,33
CPFL	0,36	0,00	0,00	1,00	1,00	0,47
CYRELA	0,00	0,00	0,04	0,00	1,00	0,21
Embraer	0,27	0,27	0,26	0,83	0,50	0,43
Estacio	0,00	1,00	0,84	0,05	1,00	0,58
Natura	1,00	0,78	0,78	0,47	0,00	0,60
Sid Nacional	0,04	0,06	0,32	0,03	0,54	0,20
Fibria	0,00	0,00	0,43	1,00	0,47	0,38
Gerdau Met	0,07	0,27	0,45	1,00	1,00	0,56
Gerdau	0,07	0,27	0,45	1,00	1,00	0,56
Hypermarcas	0,68	0,00	0,05	0,00	0,55	0,26
Itau	0,05	0,69	0,13	0,66	0,57	0,42
ItauUnibanco	1,00	1,00	0,00	0,00	0,00	0,40
JBS	0,02	0,25	0,71	0,03	0,56	0,31
Klabin	0,18	0,05	0,06	0,06	0,06	0,08
Kroton	1,00	0,69	0,03	0,03	0,03	0,36
Lojas Americanas	0,02	0,02	0,56	0,43	1,00	0,40
Lojas Renner	1,00	1,00	0,14	0,00	1,00	0,63
Marfrig	1,00	1,00	0,00	0,00	1,00	0,60
MRV	0,05	0,05	0,05	0,05	0,05	0,05
Pão de Açúcar	0,08	0,29	0,02	0,06	1,00	0,29
Petrobras	0,38	0,23	0,27	0,53	0,34	0,35
Qualicorp	0,08	0,24	0,64	0,59	0,12	0,33
RaiaDrogasil	0,00	0,09	0,09	0,07	0,63	0,18
Rumo Log	0,30	0,34	0,09	0,43	0,53	0,34
SABESP	0,00	0,00	0,65	0,00	1,00	0,33
Santander	0,71	0,10	0,10	1,00	0,10	0,40
Suzano Papel	0,19	0,11	0,29	1,00	1,00	0,52
Telefonica	0,20	0,53	0,32	0,40	0,39	0,37
TIM Participações	0,02	0,11	0,72	0,78	0,39	0,40
ULTRAPAR	1,00	0,00	0,19	0,00	0,00	0,24
USIMINAS	0,45	0,55	0,55	0,07	0,45	0,41
Vale	0,01	0,10	0,86	1,00	1,00	0,59
WEG	0,09	0,07	0,07	0,00	1,00	0,25
Média geral						0,36
Desvio padrão						0,14

sumBasic

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	simpleSum	naiveSum	
AMBEV	0,51	0,19	0,03	0,20	0,37	0,26
BM&F Bovespa S,A,	0,02	0,47	0,05	0,02	0,05	0,13
Bradesco	0,04	0,03	0,03	0,03	1,00	0,22
BRF	0,03	0,07	0,40	0,37	0,98	0,37
CCR	0,11	0,02	0,15	0,33	0,03	0,13
CEMIG	0,24	0,37	0,32	0,09	0,05	0,22
CESP	0,02	0,07	0,46	0,48	0,51	0,31
CETIP	0,00	0,05	0,00	0,05	0,05	0,03
CIELO	0,43	0,48	0,00	1,00	0,00	0,38
COPEL	0,02	0,41	0,02	0,41	0,41	0,25
Cosan	0,18	0,16	0,03	0,12	0,08	0,11
CPFL	0,02	0,63	0,00	0,00	0,00	0,13
CYRELA	0,08	0,08	0,03	0,08	0,03	0,06
Embraer	0,03	0,41	0,25	0,22	0,22	0,23
Estacio	0,00	1,00	0,79	0,05	1,00	0,57
Natura	0,00	1,00	0,58	0,00	0,02	0,32
Sid Nacional	0,12	0,11	0,33	0,16	0,07	0,16
Fibria	0,04	0,77	0,45	0,45	0,47	0,44
Gerdau Met	0,04	0,55	0,13	0,07	0,07	0,17
Gerdau	0,04	0,55	0,13	0,07	0,07	0,17
Hypermarcas	0,03	0,42	0,04	0,11	0,04	0,13
Itau	0,36	0,23	0,13	0,45	0,14	0,26
ItauUnibanco	0,00	0,00	0,00	1,00	1,00	0,40
JBS	0,01	0,06	0,62	0,07	0,55	0,26
Klabin	0,04	0,00	0,04	1,00	1,00	0,42
Kroton	0,05	0,00	0,05	1,00	1,00	0,42
Lojas Americanas	0,46	0,23	0,58	0,01	1,00	0,45
Lojas Renner	0,17	0,17	0,17	0,10	0,17	0,16
Marfrig	0,00	0,00	0,00	0,09	0,00	0,02
MRV	1,00	1,00	0,07	1,00	1,00	0,81
Pão de Açúcar	0,08	0,69	0,01	0,31	0,02	0,22
Petrobras	0,38	0,32	0,45	0,65	0,47	0,45
Qualicorp	0,17	0,28	0,70	0,48	0,39	0,41
RaiaDrogasil	0,49	1,00	0,09	0,26	0,07	0,38
Rumo Log	0,03	0,37	0,08	0,28	0,49	0,25
SABESP	0,00	0,35	1,00	0,00	1,00	0,47
Santander	0,16	1,00	0,09	0,09	1,00	0,47
Suzano Papel	0,12	0,00	0,17	0,17	0,17	0,12
Telefonica	0,34	0,46	0,38	0,14	0,07	0,28
TIM Participações	0,05	0,04	0,83	0,81	0,40	0,42
ULTRAPAR	0,15	0,00	0,15	0,00	0,00	0,06
USIMINAS	0,45	1,00	0,53	0,07	0,45	0,50
Vale	0,01	0,00	0,64	1,00	1,00	0,53
WEG	0,59	0,49	0,09	0,00	0,09	0,25
Média geral						0,29
Desvio padrão						0,17

simpleSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	sumBasic	naiveSum	
AMBEV	0,17	0,83	0,79	0,17	0,64	0,52
BM&F Bovespa S,A,	1,00	0,05	0,00	0,03	0,00	0,22
Bradesco	0,04	1,00	1,00	0,02	0,02	0,41
BRF	0,34	0,43	0,71	0,37	0,41	0,45
CCR	0,68	0,32	0,01	0,27	0,42	0,34
CEMIG	0,39	0,41	0,06	0,09	0,39	0,27
CESP	0,01	0,27	0,46	0,51	0,49	0,35
CETIP	0,00	1,00	0,00	0,09	1,00	0,42
CIELO	0,43	0,48	0,00	1,00	0,00	0,38
COPEL	0,02	1,00	0,02	0,39	1,00	0,49
Cosan	0,27	0,76	0,32	0,13	0,09	0,31
CPFL	0,36	0,00	1,00	0,00	1,00	0,47
CYRELA	1,00	1,00	0,00	0,08	0,00	0,42
Embraer	0,27	0,29	0,79	0,22	0,50	0,41
Estacio	0,00	0,03	0,02	0,02	0,03	0,02
Natura	1,00	0,00	0,40	0,00	0,00	0,28
Sid Nacional	0,50	0,60	0,03	0,17	0,28	0,32
Fibria	0,00	0,00	1,00	0,43	0,47	0,38
Gerdau Met	0,06	0,13	0,92	0,21	1,00	0,47
Gerdau	0,06	0,13	0,92	0,21	1,00	0,47
Hypermarcas	0,01	0,42	0,00	0,14	0,43	0,20
Itau	0,00	0,66	0,59	0,42	0,60	0,45
ItauUnibanco	0,00	0,00	0,00	1,00	1,00	0,40
JBS	0,04	0,29	0,04	0,09	0,04	0,10
Klabin	0,04	0,00	0,04	1,00	1,00	0,42
Kroton	0,05	0,00	0,05	1,00	1,00	0,42
Lojas Americanas	0,00	0,00	0,44	0,01	0,03	0,10
Lojas Renner	0,00	0,00	0,00	0,06	0,00	0,01
Marfrig	0,00	0,00	0,00	0,07	0,00	0,01
MRV	1,00	1,00	0,07	1,00	1,00	0,81
Pão de Açúcar	1,00	0,05	0,04	0,28	0,04	0,28
Petrobras	0,57	0,29	0,84	0,61	0,60	0,58
Qualicorp	0,55	0,17	0,65	0,48	0,45	0,46
RaiaDrogasil	0,57	0,31	0,08	0,31	0,42	0,34
Rumo Log	0,07	0,27	0,36	0,25	0,04	0,20
SABESP	0,62	0,65	0,00	0,00	0,00	0,25
Santander	0,71	0,10	1,00	0,10	0,10	0,40
Suzano Papel	0,19	0,11	1,00	0,29	1,00	0,52
Telefonica	0,04	0,39	0,45	0,13	0,59	0,32
TIM Participações	0,40	0,15	0,83	0,75	0,39	0,50
ULTRAPAR	0,00	1,00	0,00	0,00	1,00	0,40
USIMINAS	0,06	0,07	0,07	0,07	0,06	0,07
Vale	0,01	0,00	0,62	0,84	1,00	0,50
WEG	0,38	0,00	0,00	0,00	0,00	0,08
Média geral						0,35
Desvio padrão						0,17

naiveSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	sumBasic	simpleSum	
AMBEV	0,01	0,66	0,64	0,33	0,65	0,46
BM&F Bovespa S,A,	0,00	0,08	1,00	0,05	0,00	0,23
Bradesco	0,04	0,03	0,03	1,00	0,03	0,22
BRF	0,03	0,07	0,40	0,90	0,37	0,35
CCR	0,36	0,74	0,53	0,02	0,40	0,41
CEMIG	0,40	0,43	0,43	0,06	0,44	0,35
CESP	0,41	0,04	0,48	0,49	0,46	0,37
CETIP	0,00	1,00	0,00	0,09	1,00	0,42
CIELO	0,08	0,00	1,00	0,00	0,00	0,22
COPEL	0,02	1,00	0,02	0,39	1,00	0,49
Cosan	0,42	0,30	0,51	0,08	0,09	0,28
CPFL	0,01	0,00	0,61	0,00	0,61	0,25
CYRELA	0,00	0,00	1,00	0,04	0,00	0,21
Embraer	0,04	0,08	0,53	0,25	0,56	0,29
Estacio	0,00	1,00	0,79	0,84	0,05	0,54
Natura	0,00	0,05	0,00	0,05	0,00	0,02
Sid Nacional	0,08	0,29	0,60	0,07	0,28	0,26
Fibria	0,53	0,00	0,45	0,43	0,45	0,37
Gerdau Met	0,06	0,13	0,92	0,21	1,00	0,47
Gerdau	0,06	0,13	0,92	0,21	1,00	0,47
Hypermarcas	0,69	0,00	0,74	0,07	0,54	0,41
Itau	0,02	0,66	0,59	0,15	0,69	0,42
ItauUnibanco	0,00	0,00	0,00	1,00	1,00	0,40
JBS	0,32	0,04	0,62	0,70	0,03	0,34
Klabin	0,04	0,00	0,04	1,00	1,00	0,42
Kroton	0,05	0,00	0,05	1,00	1,00	0,42
Lojas Americanas	0,02	0,02	0,56	0,54	0,01	0,23
Lojas Renner	1,00	1,00	1,00	0,14	0,00	0,63
Marfrig	1,00	1,00	1,00	0,00	0,00	0,60
MRV	1,00	1,00	0,07	1,00	1,00	0,81
Pão de Açúcar	0,08	0,02	0,74	0,02	0,04	0,18
Petrobras	0,39	0,11	0,56	0,47	0,64	0,43
Qualicorp	0,57	0,52	0,13	0,39	0,46	0,41
RaiaDrogasil	0,00	0,09	0,79	0,09	0,46	0,29
Rumo Log	0,08	0,10	0,48	0,48	0,04	0,24
SABESP	0,00	0,00	1,00	0,65	0,00	0,33
Santander	0,16	1,00	0,09	1,00	0,09	0,47
Suzano Papel	0,19	0,11	1,00	0,29	1,00	0,52
Telefonica	0,40	0,40	0,45	0,07	0,60	0,38
TIM Participações	0,09	0,03	0,43	0,39	0,41	0,27
ULTRAPAR	0,00	1,00	0,00	0,00	1,00	0,40
USIMINAS	1,00	0,55	0,53	0,55	0,07	0,54
Vale	0,01	0,00	0,62	0,84	1,00	0,50
WEG	0,09	0,07	1,00	0,07	0,00	0,25
Média geral						0,38
Desvio padrão						0,14

APÊNDICE F – RESULTADO DAS AVALIAÇÕES NA NOTA EXPLICATIVA DO RECONHECIMENTO DA RECEITA

LSA

Empresa	ROUGE com base no resumo de					Média
	LexRank	KLSum	sumBasic	simpleSum	naiveSum	
AMBEV	0,37	0,23	0,01	0,69	0,00	0,26
BB Seguridade	0,10	0,70	0,44	0,15	0,59	0,40
BM&F Bovespa S,A,	1,00	1,00	0,00	1,00	1,00	0,80
BRF	0,07	0,07	1,00	0,07	1,00	0,44
CCR	0,79	0,10	0,19	0,79	0,00	0,37
CESP	0,55	0,09	0,11	0,56	0,09	0,28
CETIP	0,82	0,19	0,19	0,04	0,32	0,31
CIELO	0,29	0,59	0,54	0,05	0,55	0,40
COPEL	1,00	0,23	0,23	1,00	1,00	0,69
Cosan	0,17	0,08	0,17	1,00	1,00	0,48
CPFL	0,73	0,05	0,05	0,73	0,65	0,44
Embraer	0,11	0,10	0,09	0,10	0,07	0,10
Estacio	0,17	0,08	0,17	0,44	0,02	0,18
Natura	0,39	0,66	0,39	0,41	0,14	0,40
Sid Nacional	0,09	0,26	0,26	0,09	1,00	0,34
Fibria	0,73	0,04	0,01	1,00	1,00	0,56
Gerdau Met	0,07	0,09	0,07	0,10	1,00	0,26
Gerdau	0,07	0,09	0,07	0,10	1,00	0,26
Hypermarcas	0,48	0,00	0,00	0,04	0,00	0,10
JBS	0,00	0,67	0,00	0,00	0,02	0,14
Klabin	1,00	1,00	1,00	1,00	1,00	1,00
Kroton	0,26	0,09	0,11	0,35	0,28	0,22
Lojas Americanas	0,54	0,02	0,39	0,54	0,02	0,30
MRV	1,00	0,04	0,29	1,00	1,00	0,67
Pão de Açúcar	0,01	1,00	1,00	0,01	0,01	0,41
Petrobras	0,02	0,82	0,63	0,04	0,04	0,31
RaiaDrogasil	0,09	0,25	0,25	0,09	1,00	0,34
SABESP	0,00	1,00	0,38	0,00	1,00	0,48
Santander	0,87	0,48	0,51	0,57	0,12	0,51
Telefonica	0,04	1,00	0,02	0,61	0,06	0,35
USIMINAS	0,23	0,01	0,21	0,54	0,33	0,27
Vale	0,72	0,16	0,07	0,37	0,08	0,28
WEG	0,00	0,00	0,00	0,00	1,00	0,20
Média geral						0,38
Desvio padrão						0,20

LexRank

Empresa	ROUGE com base no resumo de					Média
	LSA	KLSum	sumBasic	simpleSum	naiveSum	
AMBEV	0,40	0,04	0,34	0,46	0,06	0,26
BB Seguridade	0,08	0,00	0,13	0,58	0,45	0,25
BM&F Bovespa S,A,	1,00	1,00	0,00	1,00	1,00	0,80
BRF	0,03	1,00	0,03	1,00	0,03	0,42
CCR	0,63	0,23	0,43	1,00	0,00	0,46
CESP	0,43	0,06	0,44	0,53	0,06	0,30
CETIP	1,00	0,33	0,33	0,06	0,33	0,41
CIELO	0,31	0,37	0,67	1,00	0,48	0,57
COPEL	1,00	0,23	0,23	1,00	1,00	0,69
Cosan	0,04	0,04	1,00	0,04	0,04	0,23
CPFL	1,00	0,05	0,05	1,00	0,65	0,55
Embraer	0,11	0,70	0,55	0,70	0,56	0,52
Estacio	0,16	0,24	1,00	0,41	0,10	0,38
Natura	0,35	0,41	1,00	0,47	0,13	0,47
Sid Nacional	0,03	0,05	0,05	1,00	0,03	0,23
Fibria	0,78	0,00	0,03	0,78	1,00	0,52
Gerdau Met	0,01	0,65	1,00	0,50	0,01	0,44
Gerdau	0,01	0,65	1,00	0,50	0,01	0,44
Hypermarcas	1,00	0,05	0,05	1,00	0,05	0,43
JBS	0,00	0,31	0,34	1,00	0,00	0,33
Klabin	1,00	1,00	1,00	1,00	1,00	1,00
Kroton	0,26	0,54	0,55	0,41	0,77	0,51
Lojas Americanas	0,40	0,04	0,67	1,00	0,04	0,43
MRV	1,00	0,04	0,29	1,00	1,00	0,67
Pão de Açúcar	0,05	0,05	0,05	1,00	1,00	0,43
Petrobras	0,02	0,02	0,02	0,01	0,01	0,02
RaiaDrogasil	0,08	0,17	0,17	1,00	0,08	0,30
SABESP	0,00	0,00	0,41	1,00	0,00	0,28
Santander	0,74	0,57	0,50	0,42	0,10	0,47
Telefonica	0,04	0,04	0,72	0,06	0,03	0,18
USIMINAS	0,21	0,04	0,20	0,36	0,31	0,23
Vale	1,00	0,49	0,07	0,40	0,11	0,41
WEG	0,00	1,00	0,17	1,00	0,00	0,43
Média geral						0,43
Desvio padrão						0,19

KLSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	sumBasic	simpleSum	naiveSum	
AMBEV	0,24	0,04	0,68	0,03	1,00	0,40
BB Seguridade	0,65	0,00	0,15	0,09	0,54	0,29
BM&F Bovespa S,A,	1,00	1,00	0,00	1,00	1,00	0,80
BRF	0,03	1,00	0,03	1,00	0,03	0,42
CCR	0,11	0,32	0,43	0,32	1,00	0,44
CESP	0,09	0,08	0,39	0,54	1,00	0,42
CETIP	0,26	0,38	1,00	1,00	0,23	0,57
CIELO	0,53	0,31	0,31	0,42	0,53	0,42
COPEL	0,13	0,13	1,00	0,13	0,13	0,30
Cosan	0,04	0,08	0,08	0,04	0,04	0,06
CPFL	0,06	0,04	0,39	0,04	0,09	0,12
Embraer	0,09	0,58	0,54	1,00	0,55	0,55
Estacio	0,06	0,19	0,19	0,01	1,00	0,29
Natura	0,61	0,43	0,43	0,45	0,14	0,41
Sid Nacional	0,15	0,09	1,00	0,09	0,15	0,30
Fibria	0,03	0,00	0,41	0,03	0,00	0,10
Gerdau Met	0,03	1,00	1,00	0,50	0,03	0,51
Gerdau	0,03	1,00	1,00	0,50	0,03	0,51
Hypermarcas	0,00	0,07	1,00	0,13	1,00	0,44
JBS	1,00	1,00	0,34	1,00	0,02	0,67
Klabin	1,00	1,00	1,00	1,00	1,00	1,00
Kroton	0,09	0,53	0,39	0,32	0,77	0,42
Lojas Americanas	0,02	0,04	0,03	0,04	1,00	0,23
MRV	0,03	0,03	0,32	0,03	0,04	0,09
Pão de Açúcar	1,00	0,01	1,00	0,01	0,01	0,41
Petrobras	1,00	0,02	0,77	0,04	0,04	0,38
RaiaDrogasil	0,13	0,09	1,00	0,09	0,13	0,29
SABESP	1,00	0,00	0,38	0,00	1,00	0,48
Santander	0,49	0,70	0,51	0,18	0,31	0,44
Telefonica	1,00	0,04	0,02	0,61	0,06	0,35
USIMINAS	0,01	0,04	0,35	0,03	0,40	0,16
Vale	0,18	0,39	0,57	0,13	0,10	0,27
WEG	0,00	1,00	0,17	1,00	0,00	0,43
Média geral						0,39
Desvio padrão						0,20

sumBasic

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	simpleSum	naiveSum	
AMBEV	0,01	0,36	0,75	0,11	1,00	0,45
BB Seguridade	0,31	0,13	0,12	0,53	0,12	0,24
BM&F Bovespa S,A,	0,00	0,00	0,00	0,00	0,00	0,00
BRF	1,00	0,07	0,07	0,07	1,00	0,44
CCR	0,11	0,32	0,23	0,32	0,00	0,20
CESP	0,10	0,53	0,36	0,54	0,36	0,38
CETIP	0,26	0,38	1,00	1,00	0,23	0,57
CIELO	0,59	0,69	0,38	0,59	0,78	0,61
COPEL	0,13	0,13	1,00	0,13	0,13	0,30
Cosan	0,04	1,00	0,04	0,04	0,04	0,23
CPFL	0,04	0,03	0,25	0,03	0,06	0,08
Embraer	0,10	0,59	0,70	0,70	0,55	0,53
Estacio	0,16	1,00	0,24	0,41	0,10	0,38
Natura	0,35	1,00	0,41	0,47	0,13	0,47
Sid Nacional	0,15	0,09	1,00	0,09	0,15	0,30
Fibria	0,02	0,03	0,51	0,02	0,02	0,12
Gerdau Met	0,01	1,00	0,65	0,50	0,01	0,44
Gerdau	0,01	1,00	0,65	0,50	0,01	0,44
Hypermarcas	0,00	0,07	1,00	0,13	1,00	0,44
JBS	0,00	1,00	0,31	1,00	0,02	0,47
Klabin	1,00	1,00	1,00	1,00	1,00	1,00
Kroton	0,10	0,50	0,37	0,20	0,36	0,31
Lojas Americanas	0,43	1,00	0,04	1,00	0,04	0,50
MRV	0,28	0,28	0,38	0,28	0,14	0,27
Pão de Açúcar	1,00	0,01	1,00	0,01	0,01	0,41
Petrobras	1,00	0,02	1,00	0,09	0,09	0,44
RaiaDrogasil	0,13	0,09	1,00	0,09	0,13	0,29
SABESP	0,36	0,44	0,36	0,44	0,05	0,33
Santander	0,51	0,58	0,49	0,11	0,09	0,36
Telefonica	0,02	0,62	0,02	0,39	0,06	0,22
USIMINAS	0,21	0,21	0,39	0,20	0,15	0,23
Vale	0,08	0,06	0,56	0,06	0,08	0,17
WEG	0,00	0,10	0,10	0,10	0,00	0,06
Média geral						0,35
Desvio padrão						0,19

simpleSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	sumBasic	naiveSum	
AMBEV	0,75	0,47	0,03	0,11	0,04	0,28
BB Seguridade	0,10	0,51	0,06	0,49	0,47	0,33
BM&F Bovespa S,A,	1,00	1,00	1,00	0,00	1,00	0,80
BRF	0,03	1,00	1,00	0,03	0,03	0,42
CCR	0,63	1,00	0,23	0,43	0,00	0,46
CESP	0,45	0,53	0,41	0,45	0,41	0,45
CETIP	0,04	0,05	0,70	0,70	0,03	0,30
CIELO	0,04	0,74	0,37	0,43	0,48	0,41
COPEL	1,00	1,00	0,23	0,23	1,00	0,69
Cosan	1,00	0,17	0,08	0,17	1,00	0,48
CPFL	1,00	1,00	0,05	0,05	0,65	0,55
Embraer	0,09	0,58	1,00	0,54	0,55	0,55
Estacio	0,46	0,47	0,02	0,47	0,02	0,29
Natura	0,41	0,51	0,49	0,51	0,13	0,41
Sid Nacional	0,03	1,00	0,05	0,05	0,03	0,23
Fibria	1,00	0,73	0,04	0,01	1,00	0,56
Gerdau Met	0,04	1,00	0,65	1,00	0,04	0,55
Gerdau	0,04	1,00	0,65	1,00	0,04	0,55
Hypermarcas	0,05	0,52	0,05	0,05	0,05	0,15
JBS	0,00	1,00	0,31	0,34	0,00	0,33
Klabin	1,00	1,00	1,00	1,00	1,00	1,00
Kroton	0,36	0,42	0,33	0,22	0,32	0,33
Lojas Americanas	0,40	1,00	0,04	0,67	0,04	0,43
MRV	1,00	1,00	0,04	0,29	1,00	0,67
Pão de Açúcar	0,05	1,00	0,05	0,05	1,00	0,43
Petrobras	0,07	0,02	0,06	0,09	1,00	0,25
RaiaDrogasil	0,08	1,00	0,17	0,17	0,08	0,30
SABESP	0,00	1,00	0,00	0,41	0,00	0,28
Santander	0,57	0,49	0,17	0,11	0,32	0,33
Telefonica	0,33	0,04	0,33	0,28	0,09	0,22
USIMINAS	0,57	0,41	0,03	0,21	0,35	0,32
Vale	0,49	0,38	0,16	0,07	0,66	0,35
WEG	0,00	1,00	1,00	0,17	0,00	0,43
Média geral						0,43
Desvio padrão						0,18

naiveSum

Empresa	ROUGE com base no resumo de					Média
	LSA	LexRank	KLSum	sumBasic	simpleSum	
AMBEV	0,00	0,04	0,75	0,68	0,03	0,30
BB Seguridade	0,48	0,51	0,47	0,14	0,60	0,44
BM&F Bovespa S,A,	1,00	1,00	1,00	0,00	1,00	0,80
BRF	1,00	0,07	0,07	1,00	0,07	0,44
CCR	0,00	0,00	0,74	0,00	0,00	0,15
CESP	0,09	0,08	1,00	0,39	0,54	0,42
CETIP	0,38	0,33	0,20	0,20	0,04	0,23
CIELO	0,54	0,44	0,58	0,70	0,59	0,57
COPEL	1,00	1,00	0,23	0,23	1,00	0,69
Cosan	1,00	0,17	0,08	0,17	1,00	0,48
CPFL	1,00	0,73	0,11	0,12	0,73	0,54
Embraer	0,08	0,59	0,70	0,54	0,70	0,52
Estacio	0,01	0,06	0,82	0,06	0,01	0,19
Natura	0,13	0,13	0,13	0,13	0,11	0,12
Sid Nacional	1,00	0,09	0,26	0,26	0,09	0,34
Fibria	0,76	0,72	0,00	0,01	0,76	0,45
Gerdau Met	1,00	0,07	0,09	0,07	0,10	0,26
Gerdau	1,00	0,07	0,09	0,07	0,10	0,26
Hypermarcas	0,00	0,07	1,00	1,00	0,13	0,44
JBS	0,04	0,00	0,02	0,03	0,00	0,02
Klabin	1,00	1,00	1,00	1,00	1,00	1,00
Kroton	0,27	0,74	0,76	0,37	0,30	0,49
Lojas Americanas	0,02	0,04	1,00	0,03	0,04	0,23
MRV	0,82	0,82	0,04	0,12	0,82	0,53
Pão de Açúcar	0,05	1,00	0,05	0,05	1,00	0,43
Petrobras	0,07	0,02	0,06	0,09	1,00	0,25
RaiaDrogasil	1,00	0,09	0,25	0,25	0,09	0,34
SABESP	0,66	0,00	0,66	0,03	0,00	0,27
Santander	0,14	0,14	0,36	0,11	0,40	0,23
Telefonica	0,04	0,02	0,04	0,05	0,10	0,05
USIMINAS	0,40	0,41	0,54	0,18	0,41	0,39
Vale	0,12	0,11	0,12	0,11	0,71	0,23
WEG	1,00	0,00	0,00	0,00	0,00	0,20
Média geral						0,37
Desvio padrão						0,21