5-2015

Comparing Text Mining Algorithms for Predicting Irregularities in Public Accounts

Breno Santana Santos
UFS - Federal University of Sergipe, breno1005@hotmail.com

Methanias Colaço Júnior
UFS – Federal University of Sergipe, mjrse@hotmail.com

Bruno Cruz da Paixão
UFS - Federal University of Sergipe, brunopaixao87@gmail.com

Rafael M. Santos
UFS - Federal University of Sergipe, rafaelmsse@gmail.com

André Vinicius Rodrigues P Nascimento
UFS – Federal University of Sergipe, andreviniciusnascimento@gmail.com

See next page for additional authors

Authors
Breno Santana Santos, Methanias Colaço Júnior, Bruno Cruz da Paixão, Rafael M. Santos, André Vinicius Rodrigues P Nascimento, Hallan Cosmo dos Santos, Wallace H. L. Filho, and Arquimedes S. L. de Medeiros

Análise Comparativa de Algoritmos de Mineração de Texto Aplicados a Históricos de Contas Públicas

Alternative Title: Comparing Text Mining Algorithms for Predicting Irregularities in Public Accounts

Breno Santana Santos[a], Methanias Colaço Júnior[a, b], Bruno Cruz da Paixão[a], Rafael M. Santos [a, b], André Vinícius R. P. Nascimento[a], Hallan Cosmo dos Santos[c], Wallace H. L. Filho[c], Arquimedes S. L. de Medeiros[c]

[a] Competitive Intelligence Research and Practice Group - NUPIC
UFS - Federal University of Sergipe
Itabaiana/SE – Brazil

[b] Postgraduate Program in Computer Science - PROCC
UFS - Federal University of Sergipe
São Cristóvão/SE – Brazil

[c] Court of Accounts of Sergipe - TCE-SE
Aracaju/SE – Brazil

{breno1005, mjrse}@hotmail.com, {brunopaixao87, rafaelmsse, andreviniciusnascimento}@gmail.com, {hallan.santos, wallace.lessafilho, arquimedes.medeiros}@tce.se.gov.br

RESUMO

Grandes massas de dados são geradas pelas aplicações que apoiam as atividades rotineiras dos órgãos públicos. Uma parcela significativa destes dados está em formato textual, sendo cabível o uso da Mineração de Texto, para extrair conhecimento potencialmente útil e previamente desconhecido. O objetivo deste artigo é avaliar o desempenho e qualidade de 3 algoritmos de mineração de texto aplicados à classificação de irregularidades em históricos de contas públicas, custodiadas pelo Tribunal de Contas de Sergipe. Para realizar a avaliação, foi desenvolvida uma ferramenta que implementa os algoritmos, bem como foi realizado um estudo de caso que avaliou métricas de desempenho e qualidade, tais como: Tempo Médio de Execução, Acurácia, Precisão, Cobertura e Medida F. Os resultados evidenciaram que o algoritmo Naïve Bayes Multinomial, com Frequência Inversa, foi a melhor abordagem para detectar evidências de irregularidades em pagamentos de diárias.

Palavras-Chave

Mineração de Texto, Históricos de Contas Públicas, Tribunal de Contas, Auditoria.

ABSTRACT

Information systems that support public sector daily activities generate large data sets. As a large proportion of the data in these data sets are text, Text Mining can play an important role in deriving potentially useful and previously unknown information. The overall goal of this paper is evaluate the performance and quality of three text mining classification algorithms applied to detect irregularities in public sector records. To evaluate the algorithms, a tool was designed and a case study was carried out at the Court of Accounts of Sergipe. Performance and Quality metrics were evaluated: mean execution time, accuracy, precision, coverage and F-measure. The results show that the multinomial naive bayes algorithm using inverse document frequency was the best approach to find evidences of travel reimbursement irregularities.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Systems – *Relational databases*, *Textual databases*.
H.2.7 [**Database Management**]: Database Administration – *Data warehouse and repository*.
H.2.8 [**Database Management**]: Database Applications – *Data mining*.
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistic processing*.
I.5.3 [**Pattern Recognition**]: Clustering – *Similarity measures*.
I.5.4 [**Pattern Recognition**]: Applications – *Text processing*.
J.1 [**Computer Applications**]: Administrative Data Processing – *Government*, *Law*.

General Terms

Algorithms, Management, Measurement, Performance, Economics, Experimentation.

Keywords

Text Mining, Public Accounts History, Court of Accounts, Audit.

1. INTRODUÇÃO

Na era da informação, esta passou a ser um dos maiores bens de uma organização, tendo o poder de influenciar o processo de tomada de decisão. Grandes massas de dados são geradas diariamente pelos sistemas que apoiam as atividades rotineiras das organizações, dificultando a tarefa analítica dos gestores. Diante dessa necessidade, surgiram os Sistemas de Apoio à Decisão (SADs) que permitem apoiar, contribuir e influenciar o processo de tomada de decisão [6]. Os SADs permitem, a partir dos dados

transacionais da organização, gerar informações gerenciais que facilitam o referido processo.

Como grande parte dos dados manipulados pelas organizações está em formato textual, torna-se fundamental o uso da técnica de Mineração de Texto (também conhecido por *Knowledge Discovery in Texts*, KDT, em inglês) para identificar padrões e conhecimentos para auxiliar nas decisões.

KDT é o processo de extração de informações, úteis e não-triviais, e conhecimento em texto desestruturado [21]. O processo de Mineração de Texto é dividido em quatro etapas bem definidas: Seleção, Pré-processamento, Mineração e Assimilação [11].

Na *Seleção*, os documentos relevantes devem ser definidos para serem processados. No *Pré-processamento*, os documentos selecionados sofrerão um tratamento especial, para que seja definida uma estrutura, a qual será utilizada na próxima etapa, *Mineração*. Nesta, serão utilizadas técnicas para detectar os padrões não visíveis nos dados. Por fim, na *Assimilação*, os usuários irão utilizar o conhecimento gerado para apoiar as suas decisões [3, 11, 18].

O conhecimento gerado pode ser avaliado para determinar se o mesmo é relevante ou não para o usuário [24], ou seja, avaliar o desempenho do processo de mineração para a geração do conhecimento. Neste contexto, existem várias métricas, sendo as principais relacionadas ao desempenho, à acurácia, à precisão e à cobertura.

De forma análoga, os órgãos governamentais geram um imenso volume de dados provenientes dos sistemas de informações utilizados para apoiar suas atividades rotineiras. Esses dados são utilizados pelos órgãos de auditoria governamental para planejamento e execução de auditorias e fiscalizações dos recursos públicos. Analisá-los em busca de irregularidades sem o uso de mecanismos de análise de dados e extração de conhecimento é uma tarefa árdua para esses órgãos fiscalizadores.

Em um levantamento realizado em 2007, constatou-se que 27 dos Tribunais de Contas existentes no Brasil, apenas um utilizava técnicas de extração de conhecimento em grandes massas de dados [17]. Fica clara a necessidade de uma maior utilização de mecanismos que tornem efetivas as atividades de auditoria.

Diante dessa situação e objetivando atender às necessidades do Tribunal de Contas do Estado de Sergipe (TCE-SE), foi desenvolvida uma aplicação, nomeada AccountMiner, que realiza a mineração de texto em qualquer campo descritivo de um sistema de informação de auditoria.

A aplicação permite classificar registros como evidências de irregularidades, ou seja, se uma descrição está ou não de acordo com a lei e com o que se espera dos jurisdicionados, nos casos em que tudo parece correto, mas o histórico evidencia o contrário. A ferramenta usa 2 algoritmos base de mineração de texto e tem como objetivo tornar efetivo o trabalho do auditor, na identificação de irregularidades. Um dos algoritmos, após parametrização, para este artigo, será considerado como uma terceira opção.

Este artigo avalia o desempenho e qualidade dos 3 algoritmos da ferramenta supracitada, na detecção de irregularidades nos pagamentos de diárias, contidos nos históricos de contas públicas custodiadas pelo TCE-SE. Os resultados mostraram que o algoritmo Naïve Bayes Multinomial, com Frequência Inversa, foi a melhor abordagem para detectar irregularidades nos pagamentos de diárias.

O restante do trabalho está estruturado como segue. A seção 2 apresenta a ferramenta AccountMiner. Na seção 3, são apresentados os conceitos necessários para o entendimento deste trabalho. Na seção 4, encontra-se a definição e o planejamento do estudo de caso. A seção 5 apresenta a operação do estudo de caso. A seção 6 contém os resultados do estudo de caso. Na seção 7, são apresentados os trabalhos relacionados. Por fim, na seção 8, a conclusão e trabalhos futuros.

2. FERRAMENTA ACCOUNTMINER

Nesta seção, são apresentados os principais módulos da ferramenta AccountMiner.

2.1 Gerenciamento de Dicionários

Dicionários são os modelos de conhecimentos que servem de base para tornar possível a descoberta de evidências de fraudes semelhantes em toda base de dados ou em unidades e cidades específicas.

Um dicionário é criado por meio da seleção de amostras, as quais são dados selecionados pelo auditor como "Evidência" (possível evidência de irregularidade) e "Em Conformidade" (descrição que está de acordo com a lei), bem como o auditor pode informar amostras avulsas, as quais são especificadas manualmente e classificadas como "Evidência" ou "Em Conformidade", como é mostrado na Figura 1.

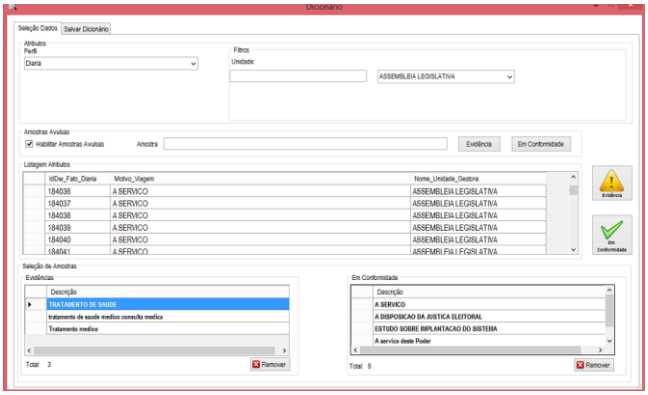

Figura 1. Tela Dicionário (Ferramenta AccountMiner)

2.2 Gerenciamento de Classificações

Após a definição do dicionário, o auditor poderá escolher os dados a serem classificados pela aplicação, ou seja, local em que serão buscadas novas evidências semelhantes às do dicionário criado.

Para classificar uma descrição, a ferramenta dispõe de dois algoritmos base, Naïve Bayes e Similaridade, bem como parametrizações para o algoritmo Bayes, as quais permitem mais 2 possibilidades de algoritmos, como é visto na Figura 2.

Os algoritmos foram escolhidos mediante pesquisa bibliográfica sobre o uso dos mesmos em campos descritivos (texto). Naïve Bayes foi escolhido pela existência de evidências de que esse é a melhor abordagem para classificação de texto [10, 19].

Já Similaridade, hipoteticamente, seria adequado para o contexto do TCE-SE, pelo fato ter um bom desempenho proporcional à alta qualidade do conjunto de treinamento [22, 23]. No nosso caso, o fato de um especialista (auditor) supervisionar o modelo garante uma boa qualidade. Vale ressaltar que o conjunto de treinamento é composto por dados reais, selecionados ou informados pelo especialista.

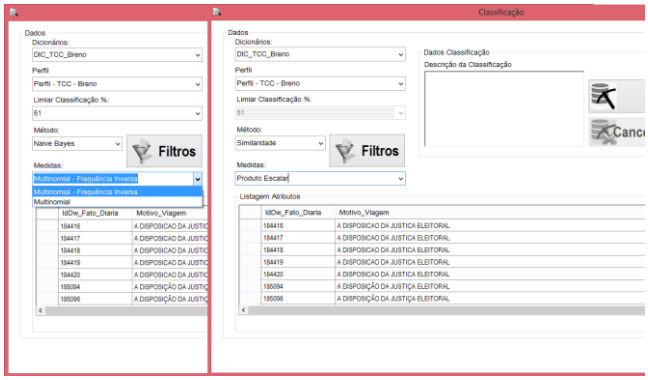

Figura 2. Telas Classificação sobrepostas, mostrando os algoritmos disponíveis (Ferramenta AccountMiner).

Para realizar a classificação de um registro, Naïve Bayes calcula a probabilidade desse registro ser ou não uma evidência de irregularidade [9, 12]. Na ferramenta, Bayes foi implementado de forma parametrizada, tanto para especificar um limiar mínimo para auxiliar na classificação, quanto para determinar o uso das frequências dos termos. As parametrizações do uso das frequências são:

- **Multinomial (M.N.):** considera o número de ocorrências de um termo no texto, para o cálculo da probabilidade do termo pertencer a uma determinada classe [12];
- **Multinomial – Frequência Inversa (M.N.F.I.):** considera o grau de representatividade do termo, *tfidf*, para o cálculo da probabilidade do termo pertencer a uma determinada classe (ver seção 3.3) [9].

Já o algoritmo de Similaridade calcula o grau de similaridade entre um novo documento (uma sentença) e um dicionário (um conjunto de amostras), por meio dos termos que ambos possuem em comum, para classificar o novo documento [22]. Em nosso contexto, determinar se a sentença é ou não uma evidência.

3. BASE CONCEITUAL

Nesta seção, são apresentados alguns conceitos necessários para o entendimento desse trabalho.

3.1 Auditoria Governamental

Auditoria é um conjunto de técnicas aplicadas sobre determinadas ações, objetivando constatar se elas estão em conformidade com as normas, regras, orçamentos e objetivos [2]. Existem diversas classificações da Auditoria. Dentre elas, quanto ao campo de atuação, podem ser classificadas em governamental e privada [2].

Auditoria Governamental é o tipo de auditoria que atua diretamente sobre a administração pública [2]. Ela está diretamente relacionada ao acompanhamento das ações efetuadas pelos órgãos e entidades que compõem a administração direta e indireta das três esferas de governo. Constitui-se num importante instrumento de controle para garantir uma melhor alocação de recursos públicos, transparência e prevenção e combate à corrupção [2].

Tribunais de contas são órgãos de auditoria governamental com autonomia administrativa e financeira e que não possuem nenhuma subordinação com outros órgãos ou Poderes da Administração Pública [5, 15]. Em outras palavras, os tribunais de contas são órgãos fiscalizadores dos recursos públicos utilizados na Administração Pública, podendo responsabilizar os administradores pelos atos administrativos.

O processo de auditoria se dá em três etapas: Plano, Planejamento e Execução [15]. Na etapa Plano, é definido um documento, Plano de Auditoria, no qual deve constar o que deverá ser auditado, o objetivo a ser alcançado e os motivos para a realização da auditoria [15]. No Planejamento, as tarefas a serem realizadas para o alcance dos trabalhos, tempo necessário, metodologia a ser aplicada e os recursos a serem utilizados no processo de auditoria [1]. Por fim, na Execução, é realizada a auditoria, por meio da aplicação dos procedimentos, objetivando a obtenção de provas e evidências [1], ou seja, é a inspeção do auditor para encontrar fraudes e erros nas demonstrações contábeis, nos históricos de receitas e despesas, nas prestações de contas e em outras fontes de informações que achar necessário investigar.

Evidência é toda prova obtida pelo auditor mediante aplicação de procedimentos de auditoria, para avaliar se os critérios estabelecidos estão sendo atendidos ou não [1].

Ao final da execução, todos os fatos observados pelo auditor devem estar descritos formalmente em um relatório de auditoria [1], isto é, todo o parecer do auditor com base nos indícios encontrados.

É dever do auditor, tanto planejar seu trabalho de forma a detectar fraudes e erros que impliquem em efeitos relevantes nas demonstrações contábeis, bem como de comunicar a administração da entidade e sugerir medidas corretivas [13].

3.2 Classificação de Documentos

Mineração de Texto ou KDT é o processo de descoberta de conhecimento, potencialmente útil e previamente desconhecido, em bases de dados desestruturadas, ou seja, extração de conhecimento útil para o usuário em bases textuais [7]. Neste artigo, utilizamos técnicas de mineração de texto para auxiliar auditores na detecção de irregularidades.

Classificação de documento é um dos campos de atuação da mineração de texto. Classificação ou categorização é o processo de atribuição automática de uma categoria pré-definida a um novo documento de acordo com seu tema [22]. Formalmente, um classificador é uma função $f : Documento \mapsto \{c_1,\ldots,c_n\}$ que mapeia um documento (em nosso caso, uma sentença) para uma certa categoria ou classe contida em $\{c_1,\ldots,c_n\}$ (em nosso contexto, {*Evidência, Em Conformidade*}) [22].

Existem muitos algoritmos no campo da mineração de texto para classificar um texto, cada um com suas particularidades. Neste artigo, comparamos três algoritmos bem conhecidos (Naïve Bayes Multinomial, Naïve Bayes Multinomial – Frequência Inversa e Similaridade) em nosso problema para encontrar o algoritmo mais indicado para classificar sentenças como evidências de irregularidades ou não.

3.3 Pré-processamento de Sentenças

Cada termo que ocorre em um documento é considerado como uma dimensão do texto, então, documentos podem ter alta dimensionalidade [3, 11, 18, 24]. As etapas de pré-processamento resolvem parcialmente esse problema, reduzindo o número de termos [3, 11, 18, 24].

As principais etapas de pré-processamento são:

- **Tokenização**: consiste da divisão de um texto em um conjunto de termos [18, 22]. Nesse passo, são removidos caracteres especiais e pontuações, pois esses não contribuem para classificação. Além disso, os caracteres maiúsculos são convertidos para minúsculos.

- **Remoção de *StopWords***: os termos com pouca ou nenhuma relevância serão removidos [18]. São palavras auxiliares ou conectivas, isto é, não são discriminantes para o conteúdo do documento [11, 24]. Estas palavras são, em sua maioria, pronomes, preposições, artigos, numerais e conjunções [14].
- **Radicalização (*Stemming*)**: os termos são reduzidos ao seu radical, ou seja, as palavras variantes morfologicamente serão combinadas em uma única representação, o radical [11, 14]. Por exemplo, os termos *aglomerar* e *aglomeração* são reduzidas para o radical *aglomer*.
- **Indexação**: atribui-se uma pontuação para cada termo, garantindo uma única instância do termo no documento [18]. De acordo com [3, 11, 18, 24], os principais métodos de pontuação são:
    - Frequência do Termo (*Term Frequency* ou *tf*): consiste na razão entre a quantidade de vezes que o termo apareceu no documento e a quantidade total de termos do documento, como é mostrado na Figura 3, onde n_i é a quantidade de ocorrências do termo t_i no documento e $|D|$ a quantidade total de termos no documento.

$$tf(t_i) = \frac{n_i}{|D|}$$

Figura 3. Fórmula para calcular o *tf* do termo.

    - *tfidf* (*Term Frequency – Inverse Document Frequency*): obtém o índice de maior representatividade do termo (ver Figura 4), onde $tf(t_i)$ é a frequência do termo t_i, $|N|$ a quantidade total de documentos e $|\{d \in N : t_i \in d\}|$ o número de documentos que possuem o termo t_i.

$$tfidf(t_i) = tf(t_i) \times log \frac{|N|}{|\{d \in N : t_i \in d\}|}$$

Figura 4. Fórmula para calcular o *tfidf* do termo.

3.4 Matriz de Confusão

Dentre as diversas formas de avaliar a capacidade de predição de um classificador para determinar a classe de vários registros, a matriz de confusão é a mais simples dessas formas [4, 8].

Para *n* classes, a matriz de confusão é uma tabela de dimensão *n* x *n*. Para cada classificação possível, existe uma linha e coluna correspondente, ou seja, os valores das classificações serão distribuídos na matriz de acordo com os resultados, assim gerando a matriz de confusão para as classificações realizadas [4, 8]. As linhas correspondem às classificações corretas e as colunas representam as classificações realizadas pelo classificador [4].

Quando existem apenas duas classes, uma é considerada como *positive* (em nosso contexto, "Evidência") e a outra como *negative* ("Em Conformidade") [4]. Assim, podemos ter quatro resultados possíveis:

- *True Positive* (*TP*): uma instância de classe *positive* (uma evidência) é classificada corretamente como *positive* ("Evidência");
- *False Negative* (*FN*): uma instância de classe *positive* é classificada incorretamente como *negative* ("Em Conformidade");
- *True Negative* (*TN*): uma instância de classe *negative* pode ser classificada corretamente como *negative*;
- *False Positive* (*FP*): uma instância de classe *negative* é classificada incorretamente como *positive*.

As predições corretas e errôneas para as duas classes podem ser dispostas em uma única matriz, conforme é vista na Tabela 1 [4].

Tabela 1. Matriz de confusão.

Actual class	**Predicted class**	
	Positive	***Negative***
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positive (FP)*	*True Negative (TN)*

3.5 Métricas de Desempenho e Qualidade

Com a matriz de confusão apresentada na seção 3.4, podemos utilizar as principais métricas de desempenho e qualidade: acurácia, cobertura, medida F e precisão [4, 8]. Além disso, também consideraremos o tempo médio de execução.

3.5.1 Acurácia

É o percentual de instâncias (sentenças) classificadas corretamente (ver Figura 5).

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

Figura 5. Fórmula da acurácia.

3.5.2 Cobertura

É o percentual de instâncias que foram classificadas corretamente como *positive* ("Evidência") (ver Figura 6).

$$cobertura = \frac{TP}{TP + FN}$$

Figura 6. Fórmula da cobertura.

3.5.3 Precisão

É o percentual de instâncias classificadas como *positive* (evidências) que são realmente *positive* (evidências de irregularidades) (ver Figura 7).

$$precisão = \frac{TP}{TP + FP}$$

Figura 7. Fórmula da precisão.

3.5.4 Medida F

É a média harmônica da precisão e cobertura, ou seja, é medida que combina a precisão e cobertura (ver Figura 8).

$$Medida\,F = \frac{2 \times precisão \times cobertura}{precisão + cobertura}$$

Figura 8. Fórmula da medida F.

3.5.5 Tempo Médio de Execução

É a média aritmética dos tempos de execução para um determinado algoritmo. O tempo de execução é a duração de uma

classificação, isto é, a diferença entre o tempo de término e de início da classificação (ver Figura 9).

$$T = \frac{\sum_{k=1}^{n} \left(T_{fk} - T_{ik} \right)}{n}$$

Figura 9. Fórmula do tempo médio de execução.

4. DEFINIÇÃO E PLANEJAMENTO DO ESTUDO DE CASO

Nesta e na próxima seção, o estudo de caso é apresentado. Esta seção terá como foco a definição do objetivo e o planejamento do referido estudo.

4.1 Definição do Objetivo

O objetivo deste estudo é avaliar os algoritmos de mineração de texto da ferramenta AccountMiner, na detecção de irregularidades nos pagamentos de diárias contidos nos históricos de contas públicas sob custódia do TCE-SE.

4.2 Planejamento

As questões de pesquisa que precisam ser respondidas, por meio do estudo de caso, são:

- Qual o melhor algoritmo em termos de desempenho?
- Qual o melhor algoritmo em termos de acurácia?
- Qual o melhor algoritmo em termos de precisão?
- Qual o melhor algoritmo em termos de cobertura?
- Qual o melhor algoritmo em termos de medida F?
- Qual o melhor algoritmo em termos de tempo médio de execução?

Serão utilizadas cinco métricas para avaliar estas questões: (1) Acurácia; (2) Precisão; (3) Cobertura; (4) Medida F e (5) Tempo médio de execução.

4.2.1 Seleção de Participantes

O processo de seleção de participantes deve atender dois critérios: os participantes devem ser unidades gestoras cadastradas no sistema SISAP[1] e devem possuir uma quantidade considerável (mínimo de 500) de registros no histórico de pagamento de diárias. Foram avaliadas irregularidades relativas ao pagamento de diárias para tratamento de saúde. Por questão de sigilo, os nomes das unidades gestoras não foram revelados.

Conforme a cópia do DW[2] do SISAP, o qual foi cedido pelo TCE-SE, existem 481 unidades cadastradas, sendo necessário escolher, aleatoriamente, três unidades para a realização desse estudo de caso. As unidades escolhidas serão nomeadas como Unidade A, Unidade B e Unidade C. Para estas unidades, a quantidade de registros das amostras aleatórias utilizadas para teste são, respectivamente, 500, 625 (este era o total desta unidade) e 500.

4.2.2 Dicionário (Modelo de Conhecimento)

O modelo de conhecimento utilizado para detectar evidências na concessão de diárias, para tratamento de saúde, será composto por 60 sentenças, definidas com o auxílio de um especialista, sendo distribuídas da seguinte forma:

- Amostras da Própria Base:
  - 40 sentenças escolhidas da Unidade A, sendo 20 classificadas como "Em Conformidade" e 20 classificadas como "Evidência";
- Amostras Avulsas (inseridas manualmente):
  - 10 sentenças classificadas como "Em Conformidade", as quais são similares às da própria base, formuladas por meio de uma análise dos dados das unidades envolvidas;
  - 02 sentenças classificadas como "Evidência", as quais são similares às da própria base e formuladas por meio de uma análise dos dados da Unidade A;
  - 08 sentenças classificadas como "Evidência", as quais foram formuladas por meio de pesquisas de termos da área de Medicina como, por exemplo, profissionais da saúde, tratamentos médicos, doenças, exames, etc.

Para definição das sentenças escolhidas dentro da própria base (Unidade A), foi utilizada a abordagem *3-Fold Cross-validation*. Neste caso, uma parte dos dados da Unidade A foi utilizada para treinamento e testes, até a escolha do subconjunto de dados que gerou o modelo com melhor acurácia para os algoritmos. Outros dados sorteados da Unidade A, não utilizados para o treinamento, bem como os dados sorteados das Unidade B e C, foram utilizados para os testes finais com o melhor modelo, os quais geraram os resultados aqui apresentados.

4.2.3 Instrumentação

O processo de instrumentação se dará inicialmente com a configuração do ambiente para o estudo de caso e planejamento de coleta de dados. Os materiais / recursos utilizados foram:

- Ferramenta AccountMiner;
- SQL Server 2012, SGBD que servirá como base para o armazenamento dos dados do DW do SISAP e da aplicação AccountMiner;
- Cópia do DW do SISAP;
- Dicionário discutido na seção 4.2.2.

5. OPERAÇÃO DO ESTUDO DE CASO

Nesta seção, está descrita a preparação e execução do estudo em questão.

5.1 Preparação

Em síntese, foi preparado o ambiente para a realização do estudo de caso, ou seja, o carregamento dos dados do DW no SQL Server e criação do esquema de armazenamento dos dados da ferramenta AccountMiner.

5.2 Execução

Consistiu na realização do processo classificatório nos dados dos participantes envolvidos, para cada algoritmo de mineração de texto, utilizando o dicionário discutido na seção 4.2.2.

Foram efetuadas três classificações nos dados dos participantes envolvidos.

5.2.1 Coleta dos Dados

Ao término da execução, os resultados das classificações foram obtidos para que, a partir desses, fossem geradas as matrizes de confusão de cada algoritmo em cada unidade envolvida. Após a definição das matrizes de confusão, foram coletadas as métricas

[1] SISAP: Sistema de Auditoria Pública, um banco de dados de informações orçamentárias, financeiras e administrativas dos órgãos sob jurisdição do TCE-SE [20].

[2] DW: do inglês *Data Warehouse* (Armazém de Dados), é um banco de dados histórico que auxilia nas tarefas analíticas [6].

de acurácia, cobertura, medida F, precisão e tempo médio de execução.

O resultado desses dados coletados será apresentado na próxima seção.

6. RESULTADOS

Após a execução, foram obtidos os resultados das classificações para que fossem definidas as matrizes de confusão. Nas Tabelas 2 e 3, são apresentados os valores das matrizes de confusão. De posse desses valores, podemos coletar as métricas de desempenho e qualidade utilizadas.

Antes da análise dos resultados das classificações, foram verificadas as quantidades de evidências existentes nas unidades A (parte utilizada para testes), B e C. Foi constatado que o número de evidências foram, respectivamente, 137, 1 e 3.

Tabela 2. Valores da Diagonal Principal por Algoritmo e Unidade Gestora.

Valores da Diagonal Principal – Matriz de Confusão						
Unidades	***True Positive (TP)***			***True Negative (TN)***		
	M.N. F.I.	**M.N.**	**SIM.**	**M.N. F.I.**	**M.N.**	**SIM.**
Unidade A	136	137	137	363	363	362
Unidade B	0	1	0	613	603	588
Unidade C	3	3	2	457	449	430

Tabela 3. Valores da Diagonal Secundária por Algoritmo e Unidade Gestora.

Valores da Diagonal Secundária – Matriz de Confusão						
Unidades	***False Positive (FP)***			***False Negative (FN)***		
	M.N. F.I.	**M.N.**	**SIM.**	**M.N. F.I.**	**M.N.**	**SIM.**
Unidade A	0	0	1	1	0	0
Unidade B	11	21	36	1	0	1
Unidade C	40	48	67	0	0	1

6.1 Acurácia

Conforme pode ser visto na Tabela 4, é perceptível o ótimo desempenho do algoritmo Naïve Bayes Multinominal – Frequência Inversa (M. N. F. I.), pois o mesmo possui, em média, a maior porcentagem de acurácia. Vale ressaltar que Naïve Bayes Multinomial (M. N.) superou Naïve Bayes Multinominal – Frequência Inversa apenas na Unidade A e que Similaridade (SIM.) obteve, em média, o menor desempenho dentre as três abordagens.

Tabela 4. Comparativo da acurácia dos algoritmos.

Unidades	**Algoritmos de Mineração de Texto**		
	M. N. F. I.	**M. N.**	**SIM.**
Unidade A	99,80 %	**100 %**	99,80 %
Unidade B	**98,08 %**	96,64 %	94,08 %
Unidade C	**92,00 %**	90,40 %	86,40 %
Média	**96,63 %**	95,68 %	93,43 %

6.2 Precisão

De acordo com a Tabela 5, em média, Naïve Bayes Multinomial obteve um desempenho melhor do que Naïve Bayes Multinomial – Frequência Inversa em termos de precisão. Similaridade obteve o menor desempenho dentre as três abordagens.

Tabela 5. Comparativo da precisão dos algoritmos.

Unidades	**Algoritmos de Mineração de Texto**		
	M. N. F. I.	**M. N.**	**SIM.**
Unidade A	**100 %**	**100 %**	99,28 %
Unidade B	0,00 %	**4,55 %**	0,00 %
Unidade C	**6,98 %**	5,88 %	2,90 %
Média	35,66 %	**36,81 %**	34,06 %

É notável o péssimo desempenho das três abordagens nas unidades B e C. As baixas porcentagens de precisão nas unidades, principalmente na Unidade B, se deram pelo fato das unidades B e C possuírem apenas, respectivamente, 1 e 3 evidências de irregularidades, gerando também um número considerável de *false positive* (ver Tabela 3).

No intuito de analisar a possibilidade de aumento da precisão com o aumento do limiar para os algoritmos Bayes, foi executado novamente o processo classificatório, para os algoritmos Bayes, nas unidades envolvidas, com o limiar de 75 %.

Tabela 6. Comparativo da precisão dos algoritmos Bayes com os limiares 51 % e 75%.

Unidades	**Algoritmos de Mineração de Texto**			
	M. N. F. I.		**M. N.**	
	51 %	**75 %**	**51 %**	**75 %**
Unidade A	100 %	100 %	100 %	100 %
Unidade B	0,00 %	0,00 %	4,55 %	□
Unidade C	6,98 %	8,33 %	5,88 %	3,33 %
Média	35,66 %	36,11 %	36,81 %	34,44 %

Conforme é visto na Tabela 6, com o aumento do limiar para 75 %, é notável o aumento da precisão do Naïve Bayes Multinomial – Frequência Inversa (M.N.F.I.) e queda da precisão do Naïve Bayes Multinomial (M.N.), nas unidades B e C. **Vale ressaltar que os percentuais nas Unidades B e C ainda estão baixos por causa da pequena quantidade de evidências (1 e 3, respectivamente)** e pela consequente existência de *false positive*. Também foi constatado que, na unidade A (parte para testes), o algoritmo Naïve Bayes Multinomial classificou duas evidências como "Em Conformidade", enquanto o Naïve Bayes Multinomial – Frequência Inversa classificou corretamente 136 evidências, das 137 existentes. Na Unidade B, o algoritmo M.N.F.I. classificou erroneamente 4 conformidades como evidências e a única evidência como conformidade, enquanto o M.N. classificou a evidência como conformidade. E, por fim, na Unidade C, o M.N.F.I. classificou corretamente todas as três evidências e 33 *false positive*, enquanto o M.N. classificou apenas uma das três evidências e 29 *false positive*.

No geral, considerando os percentuais de precisão e quantidade de *false positive* de todos os algoritmos, Naïve Bayes Multinomial – Frequência Inversa foi a melhor abordagem em termos de precisão.

6.3 Cobertura

De acordo com a Tabela 7, Naïve Bayes Multinomial obteve um ótimo desempenho em relação às demais abordagens, considerando a cobertura. Similaridade obteve o menor desempenho dentre as três abordagens.

É perceptível o péssimo desempenho das abordagens M.N.F.I. e SIM., na Unidade B. As baixas porcentagens de cobertura na Unidade B foram ocasionadas pelo fato dessas abordagens terem

classificado erroneamente a única evidência de irregularidade como uma conformidade (ver Tabela 3).

Tabela 7. Comparativo da cobertura dos algoritmos.

Unidades	Algoritmos de Mineração de Texto		
	M. N. F. I.	**M. N.**	**SIM.**
Unidade A	99,27 %	**100 %**	**100 %**
Unidade B	0,00 %	**100 %**	0,00 %
Unidade C	**100 %**	**100 %**	66,67 %
Média	66,42 %	**100 %**	55,56 %

6.4 Medida F

Conforme é visto na Tabela 8, Naïve Bayes Multinomial obteve desempenho melhor do que as demais abordagens em termos de medida F. Similaridade obteve o menor desempenho dentre as três abordagens.

Tabela 8. Comparativo da medida F dos algoritmos.

Unidades	Algoritmos de Mineração de Texto		
	M. N. F. I.	**M. N.**	**SIM.**
Unidade A	99,63 %	**100 %**	99,64 %
Unidade B	∄	**8,70 %**	∄
Unidade C	**13,04 %**	11,11 %	5,56 %
Média	37,56%	**39,94%**	35,07%

Novamente é perceptível o péssimo desempenho das três abordagens nas Unidade B e C. Isso se deu pelo fato das Unidades B e C possuírem apenas, respectivamente, 1 e 3 evidências de irregularidades, bem como pelos baixos percentuais de precisão (ver Tabela 5). Como a Unidade B só possuía uma evidência, desconsideramos esta vitória do algoritmo Multinomial e optamos por uma análise do contexto geral, emitida na conclusão deste artigo.

6.5 Tempo Médio de Execução

De acordo com a Tabela 9, Similaridade obteve um ótimo desempenho em relação às demais abordagens, considerando o tempo médio de execução. Naïve Bayes Multinomial obteve o menor desempenho dentre as três abordagens.

Tabela 9. Comparativo do tempo médio de execução dos algoritmos.

Unidades	Algoritmos de Mineração de Texto		
	M. N. F. I.	**M. N.**	**SIM.**
Unidade A	135,25 s	152,37 s	**83,46 s**
Unidade B	256,77 s	278,04 s	**158,95 s**
Unidade C	194,93 s	271,26 s	**117,03 s**
Média	195,65 s	233,89 s	**119,81 s**

Após a análise das métricas e dos valores da matriz de confusão, para todos os algoritmos (ver Tabelas 2, 3, 4, 5, 6, 7, 8 e 9), o algoritmo Naïve Bayes Multinomial – Frequência Inversa foi a melhor abordagem para detectar evidências no pagamento de diárias, pelo fato de possuir, em média, as melhores porcentagens. Além disso, classificou erroneamente menos sentenças.

7. TRABALHOS RELACIONADOS

Diante dos poucos estudos nacionais encontrados, no contexto da utilização de Mineração de Texto nas atividades de auditoria em históricos de contas públicas, destaca-se o trabalho de [16]. O autor definiu um processo de Mineração de Texto com o intuito de classificar despesas públicas por objeto de gasto, por meio dos campos descritivos de notas de empenho nos históricos de contas públicas, sob custódia do Tribunal de Contas dos Municípios do Estado do Ceará (TCMCE). Para implementar a solução, o mesmo utilizou o modelo de projeto CRISP-DM, o SGBD PostgreSQL e a ferramenta Weka, bem como utilizou os dados armazenados no SIM (Sistema de Informações Municipais), para realizar a mineração nos documentos de nota de empenho.

No contexto de pesquisas relacionadas à comparação de algoritmos de mineração de texto, destacam-se os trabalhos de [10], [19] e [21].

Em [10], os autores realizaram um processo de mineração de texto para classificar erros reportados em dois projetos *open source*, Eclipse e GNOME, por meio do Bugzilla, sistema de rastreamento de erros. Para prever o tipo de gravidade do erro, foram utilizados diferentes algoritmos do ambiente Weka (Naïve Bayes, Naïve Bayes Multinomial, *K-Nearest Neighbor* e *Support Vector Machines*). Além disso, foi realizada uma análise comparativa em relação à acurácia e tamanho da base de treinamento. Após a análise, foi constatado que o Naïve Bayes Multinomial obteve melhor performance em relação aos outros algoritmos.

Em [19], os autores realizam uma análise comparativa dos algoritmos Naïve Bayes, *SVM* (*SMO*), *K-Nearest Neighbour* (*lazy IBk*) e árvore de decisão J48, todos disponíveis na ferramenta Weka. Para tal análise, foi utilizado um conjunto de dados de 4000 documentos classificados em quatro diferentes classes: *business*, *politic*, *sports* e *travel*. O conjunto de treinamento era constituído de 1200 documentos (30% do total de documentos), já o de teste era composto pelos 2800 documentos restantes. Ao final da análise, foi comprovado que o Naïve Bayes era o melhor algoritmo em termos de acurácia, precisão, cobertura e medida F.

Em [21], as autoras analisaram o desempenho dos classificadores bayesianos e *lazy* para classificar arquivos que estão armazenados em um disco rígido de um computador. Foram escolhidos cinco algoritmos, sendo dois classificadores bayesianos, BayesNet e Naïve Bayes, e três classificadores *lazy*, *IBL* (*Instance Based Learning*), *IBK* (*K-Nearest Neighbour*) e *Kstar*, todos disponíveis na ferramenta Weka. Esta foi utilizada para analisar o desempenho dos algoritmos em um conjunto de dados, o qual possui 80000 instâncias e quatros atributos (nome, tamanho, extensão e caminho do arquivo). Inicialmente, foram analisados os desempenhos de BayesNet e Naïve Bayes, sendo que o primeiro obteve os melhores resultados. Da mesma forma, os classificadores *lazy* foram avaliados. Foi constatado que o algoritmo *IBK* foi a melhor abordagem *lazy*. Por fim, foi realizada a análise comparativa entre BayesNet e *IBK*. Após a verificação dos resultados, os classificadores *lazy* são mais eficientes, em termos de métricas de acurácia (precisão, cobertura, medida F, curva ROC e *Kappa Statistic*), do que os bayesianos, sendo o *IBK* a melhor técnica dentre as demais analisadas.

8. CONCLUSÃO E TRABALHOS FUTUROS

A principal contribuição deste trabalho foi a avaliação dos algoritmos de mineração de texto disponíveis na ferramenta AccountMiner, em termos de desempenho e qualidade, para detectar irregularidades em históricos de contas públicas. O trabalho foi consolidado pela realização de um estudo de caso, o qual, a partir das unidades escolhidas, determinou Naïve Bayes Multinomial – Frequência Inversa como a melhor abordagem para identificação de evidências, no pagamento de diárias. De posse do

melhor algoritmo, este pode ser utilizado para tornar mais efetivo o trabalho de auditores de diversos órgãos de auditoria municipais, estaduais e federais, na identificação de irregularidades e na tomada de decisão.

Por fim, como trabalhos futuros, implementações de outras funcionalidades serão feitas, no intuito de tornar a aplicação AccountMiner mais robusta e efetiva, considerando o uso de informações textuais da Web e das redes sociais, as quais podem indicar irregularidades nos gastos públicos.

9. REFERÊNCIAS

[1] Araújo, I. da P. S. 2006. *Introdução à Auditoria Operacional* (3rd. ed.). FGV Editora, Rio de Janeiro, RJ.

[2] Araújo, I. da P. S. 1998. *Introdução à auditoria: breves apontamentos de aula aplicáveis à área governamental*. Egba, Salvador, BA.

[3] Balinski, R. 2002. *Filtragem de informações no ambiente do direito*. Master's thesis. Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil.

[4] Bramer, M. 2007. *Principles of data mining*. Springer London, New York, NY.

[5] Castro, D. P. de. 2009. *Auditoria e controle interno na administração pública: evolução do controle interno no Brasil: do código de contabilidade de 1992 até a criação da CGU em 2003: guia para atuação das auditorias e organização dos controles internos nos estados, municípios e ONGs* (2nd. ed.). Atlas, São Paulo, SP.

[6] Colaço Jr., M. 2004. *Projetando sistemas de apoio à decisão baseados em data warehouse*. Axcel Books, Rio de Janeiro, RJ.

[7] Feldman, R. and Dagan, I. 1995. Knowledge Discovery in Textual Databases (KDT). (1995). Retrieved February 1, 2015 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.7462&rep=rep1&type=pdf.

[8] Han, J., Kamber, M. and Pei, J. 2011. *Data mining: concepts and techniques* (3rd. ed.). Morgan Kaufmann Publishers, San Francisco, CA.

[9] Kibriya, A. M., Frank, E., Pfahringer, B. and Holmes, G. 2005. Multinomial naive bayes for text categorization revisited. (2005). Retrieved January 3, 2015 from http://link.springer.com/content/pdf/10.1007%2F978-3-540-30549-1_43.pdf.

[10] Lamkanfi, A., Demeyer, S., Soetens, Q. D. and Verdonck, T. 2011. Comparing mining algorithms for predicting the severity of a reported bug. (2011). Retrieved January 3, 2015 from http://ieeexplore.ieee.org/ielx5/5740650/5741244/05741332.pdf?tp=&arnumber=5741332&isnumber=5741244.

[11] Magalhães, C. C. 2008. *MinerJur: uma ferramenta para mineração de bases de jurisprudência*. Master's thesis. Salvador University (UNIFACS), Salvador, Brazil.

[12] Mccallum, A. and Nigam, K. 1998. A comparison of event models for naive bayes text classification. (1998). Retrieved January 5, 2015 from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=6D492C0CABE07EEE0E3BF2DCD8DC1628?doi=10.1.1.46.1529&rep=rep1&type=pdf.

[13] Pinho, R. C. de S. 2007. *Fundamentos de auditoria: auditoria contábil: outras aplicações de auditoria*. Atlas, São Paulo, SP.

[14] Sá, H. R. de. 2008. *Seleção de características para classificação de texto*. Federal University of Pernambuco, Recife, PE.

[15] Silva, M. M. da. 2012. *Curso de auditoria governamental: de acordo com as normas internacionais de auditoria pública aprovadas pela INTOSAI* (2nd. ed.). Atlas, São Paulo, SP.

[16] Soares, A. M. 2010. *A mineração de texto na análise de contas públicas municipais*. Master's thesis. State University of Ceará, Fortaleza, Brazil.

[17] Sousa, M. U. R. dos S. de. 2009. *Mineração de dados aplicada à celeridade processual do tribunal de contas do estado de Pernambuco (TCE-PE)*. Master's thesis. Federal University of Pernambuco, Recife, Brazil.

[18] Souza, J. G. de. 2011. *Uma aplicação de mineração de texto para descoberta de características psicológicas de indivíduos*. Federal University of Sergipe, Itabaiana, SE.

[19] Ting, S. L., Ip, W. H. and Tsang, A. H. C. 2011. Is Naïve Bayes a Good Classifier for Document Classification?. (July 2011). Retrieved January 10, 2015 from http://www.sersc.org/journals/IJSEIA/vol5_no3_2011/4.pdf.

[20] Tribunal de Contas de Sergipe. 2009. SISAP. (2009). Retrieved January 10, 2015 from http://www.tce.se.gov.br/sitev2/sisap.php.

[21] Vijayarani, S. and Muthulakshmi, S. 2013. Comparative analysis of bayes and lazy classification algorithms. (August 2013) Retrieved January 10, 2015 from http://www.ijarcce.com/upload/2013/august/34-h-Uma%20Gopalakrishnan%20-Comparative%20Analysis%20of%20Bayes%20and%20Lazy%20classification%20algorithms.pdf.

[22] Weiss, S. M., Indurkhya, N. and Zhang, T. 2010. *Fundamentals of predictive text mining*. Springer London, New York, NY.

[23] Weiss, S. M., Indurkhya, N., Zhang, T and Damerau, F. J. 2005. *Text mining: predictive methods for analyzing unstructured information*. Springer Science+Business Media, New York, NY.

[24] Wives, L. K. 2002. Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva. (January 2002). Retrieved January 1, 2015 from http://www.leandro.wives.nom.br/pt-br/publicacoes/eq.pdf.