Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU

Patrícia Helena Maia Alves de Andrade

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcelo Ladeira

Coorientador
Prof. Dr. Rommel Novaes Carvalho

Brasília
2015

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

MP314a

Maia Alves de Andrade, Patrícia Helena
Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU / Patrícia Helena Maia Alves de Andrade; orientador Marcelo Ladeira; co-orientador Rommel Novaes Carvalho. -- Brasília, 2015.
65 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2015.

1. Mineração de textos. 2. Classificação de textos. 3. Árvore de Huffman . 4. Triagem automática de denúncias. I. Ladeira, Marcelo, orient. II. Novaes Carvalho, Rommel, co-orient. III. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU

Patrícia Helena Maia Alves de Andrade

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcelo Ladeira (Orientador)
Departamento de Ciência da Computação/UnB

Prof. Dr. Thiago Veiga Marzagão
Conselho Administrativo de Defesa Econômica

Prof. Dr. Hércules Antônio Prado
Universidade Católica de Brasília

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 29 de setembro de 2015

Dedicatória

Dedico este trabalho a todos aqueles que se aventuram a buscar o conhecimento e a aprendizagem.

Agradecimentos

Agradeço à minha família pelo apoio. Agradeço ao Professor Marcelo Ladeira pelo esforço e paciência dedicados à este trabalho. Agradeço ao Rommel Carvalho, ao Henrique Rocha e ao Libório pelo incentivo e apoio. Agradeço aos colegas de mestrado e aos colegas da Controladoria Geral da União. Agradeço à CGCID pelos esclarecimentos e parceria neste trabalho. Agradeço pelas contribuições, pela força e pela ajuda neste caminho, ao Márcio Otta, Carla Maia, Bárbara Pessoa, Cíntia Maia, Ticiane Ottoni, Ana Paula Oliveira, Sarah Lima e Leonardo Sales.

Resumo

A Controladoria Geral da União é o órgão do Poder Executivo responsável pelas atividades de controle interno, auditoria pública, correição, prevenção e combate à corrupção e ouvidoria dos gastos públicos do Poder Executivo. Por meio do menu "Denuncias e Manifestações", no portal da CGU, se tem acesso a um formulário para apresentação de denúncias por parte da sociedade. Após cadastradas pelo cidadão as mesmas devem ser triadas e encaminhadas para a coordenação temática da CGU com competência para realizar a apuração. Atualmente essa triagem é feita de forma manual e a denúncia encaminhada para uma dentre as 91 opções de destino pré-determinadas. Essa grande quantidade de categorias é um fator que dificulta a classificação automática de textos. Considerando o acúmulo de denúncias existentes na base atualmente e a chegada de novas denúncias, aliadas ao tempo gasto com a triagem manual, torna-se cada vez mais difícil a análise tempestiva das ocorrências reportadas. Esse contexto pode causar prejuízos financeiros para a Administração Pública Federal além de desmotivar a utilização do canal pelo cidadão. As denúncias cadastradas são provenientes de municípios presentes em todas as Unidades da Federação gerando assim um grande impacto em todo o território nacional. Esta pesquisa tem como objetivo elaborar uma prova de conceito para um modelo para a triagem automática de denúncias na CGU, utilizando mineração de textos. Os melhores resultados foram alcançados utilizando classificação por ranking baseada em Árvore de Huffman. Esta prova de conceito demonstrou a viabilidade de uma triagem de denúncias de forma automática na CGU, sem perda de qualidade em comparação à triagem manual.

Palavras-chave: Mineração de textos, Árvore de Huffman, Triagem de Documentos

Abstract

The Office of the Comptroller General (CGU) is the agency of the Federal Government in charge of assisting the President of the Republic in matters related to internal control activities, public audits, corrective and disciplinary measures, corruption prevention and combating and coordinating ombudsman's activities. Through a complaints link of the CGU site, citizens have access to a form to file their complaints. These complaints must be screened and delivered to the coordination of CGU by subject. Nowadays the complaints screening is done manually and they are delivered to one of the 91 coordinating units of CGU. This large amount of categories is more complex in automatic text classification. Considering the complaints storage on the database now and the arrival of new complaints, combined with the time spent on manual sorting, the timely analysis of the reported occurrences it becomes increasingly difficult. This context can cause financial losses to Federal Public Administration as well as discouraging the use of the channel by the citizen. Complaints registered origins are municipalities present in all Brazilian states, generating a great impact on the entire national territory. This research intends to develop a proof of concept for an automatic model of complaints screening, using text mining. The best results were achieved using ranking based on the Huffman Tree algorithm. This proof of concept demonstrated the feasibility of automatical sorting without the loss of quality compared to manual sorting.

Keywords: Text mining, Huffman Tree, sorting of documents

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Definição do Problema

1.1 Introdução

A Controladoria-Geral da União (CGU) é o órgão de controle e transparência da Administração Pública Federal (APF) e atua na prevenção e no combate à corrupção. Visando atender esse objetivo, a CGU abre espaço para que o cidadão colabore com a fiscalização do uso do dinheiro público disponibilizando o acesso ao formulário Denúncias e Manifestações, por meio do portal deste órgão. Este canal de comunicação tem o intuito de estimular a denúncia de atos relativos à defesa do patrimônio público, ao controle sobre a aplicação dos recursos públicos federais, à correição, à prevenção e ao combate à corrupção, às atividades de ouvidoria e ao incremento da transparência da gestão no âmbito da APF.

Apesar do intuito de atrair o controle e a participação popular, a CGU não consegue dar vazão e analisar todo o quantitativo das denúncias enviadas. Isso se deve ao grande volume de denúncias recebidas e a dificuldade de tratamento das mesmas em tempo hábil. Diariamente chegam, em média, 32 novas denúncias e a capacidade de análise diária da CGU é em torno de 20 denúncias. Aproximadamente 4 mil denúncias deixarão de ser apuradas por ano. Este fato cria uma defasagem e, em alguns casos, a perda do prazo de apuração por ter se tornado intempestivo.

Uma das formas de combater a corrupção é a verificação tempestiva de licitações e contratações fraudulentas utilizando dinheiro público. Estas contratações, denunciadas por meio deste canal de comunicação com o cidadão, podem indicar fatos a serem apurados antes ou durante o processo de licitações podendo evitar danos ou prejuízos ao erário. Este fator possibilita uma gestão de gastos públicos mais consciente, transparente e com a participação da população no controle e fiscalização.

Um maior número de denúncias analisadas bem como melhor direcionamento das fiscalizações e auditorias podem proporcionar maior efetividade no combate à corrupção.

No cenário atual, um processo automatizado de triagem de denúncias pode trazer vários benefícios como uma maior efetividade no combate à corrupção; uma melhor aplicação da força de trabalho na apuração da denúncia, o aumento do número de denúncias analisadas; maior transparência dos gastos públicos; dentre outros.

Um modelo para classificar automaticamente as denúncias pode possibilitar um ganho efetivo de tempo e recursos, tornando mais eficiente e ágil o serviço prestado à comunidade além de impulsionar o combate à corrupção. Essa pesquisa tem como objetivo realizar uma prova de conceito para a construção de um modelo automático de triagem de denúncias recebidas pela Controladoria-Geral da União (CGU).

1.2 Definição do Problema

O uso de classificação automática de textos tem se tornado cada vez mais comum nos últimos anos. Contudo, poucos trabalhos têm focado a classificação em domínios com grande número de classes. Em muitos casos, o uso dessa técnica é aplicado a textos formatados e semi-estruturados, como classificação de artigos acadêmicos, sites ou *spam*, facilitando a aplicação de algoritmos de aprendizagem de máquina.

O problema focado nessa pesquisa é a triagem automática realizada com a aplicação de algoritmos de classificação de documentos textuais em formato livre e redigidos em português informal, considerando um grande número das classes possíveis. Esses documentos constituem denúncias encaminhadas à CGU, em geral, por meio do preenchimento de formulário disponível on-line no portal dessa Controladoria. O formulário possui um campo texto principal para o denunciante apresentar as informações que tem a respeito do fato que está denunciando.

Atualmente essas denúncias são triadas manualmente e encaminhadas para uma dentre as 91 possíveis opções de destino pré-determinadas, incluindo arquivamento, descartar como lixo, ou ser direcionada para apuração por um dos órgãos internos da CGU. O fluxo de novas denúncias é superior à atual capacidade de triagem manual das denúncias, resultando em acúmulo crescente de denúncias não triadas.

1.3 Justificativa

Do ponto de vista científico, a pesquisa de classificadores com grande número de classes constitui uma área ativa e relevante. Foi realizado um levantamento e foram encontrados 103 artigos em 2015 e 145 artigos em 2014 referindo-se ao tema *large scale classification*. Além disso, existem inúmeros trabalhos que se referem ao tema em anos anteriores a

2014 e 2015 citados. Dessa forma, verifica-se que a área em questão possui relevância acadêmica.

Do ponto de vista da CGU, a proposição de um método para triagem automática de denúncias pode contribuir para reduzir o estoque de denúncias não triadas e para que a triagem e a averiguação dos fatos denunciados seja feita tempestivamente.

1.4 Objetivo

O objetivo geral dessa pesquisa é estudar os algoritmos atuais de classificação com grande número de classes de documentos textuais. O objetivo secundário é propor um algoritmo automático para a triagem de documentos textuais e realizar uma prova de conceito considerando a triagem das denúncias recebidas pela CGU.

1.5 Organização do Trabalho

Este trabalho está dividido em 5 capítulos. O Capítulo 2 descreve o estado da arte e os trabalhos correlatos ao tema. O Capítulo 3 descreve a solução proposta. O Capítulo 4 apresenta os resultados obtidos. E, finalizando, o Capítulo 5 aponta as conclusões e os trabalhos futuros.

Capítulo 2

Fundamentação Teórica e Revisão do Estado da Arte

Este capítulo apresenta os principais conceitos de mineração de textos abrangendo o processamento e preparação dos textos, os algoritmos de classificação utilizados e os métodos de avaliação dos resultados gerados. O levantamento do estado da arte abrange artigos envolvendo classificação automática de textos e classificação com grande número de classes.

2.1 Mineração de Textos

Mineração de textos é o processo de descoberta de conhecimento que utiliza técnicas de análise e extração de dados a partir de textos, frases ou palavras. É o processo de extrair padrões interessantes e não triviais ou conhecimento a partir de documentos em textos não estruturados [12]. Esta descoberta de conhecimento envolve diversas aplicações tais como análise de textos, extração de informações, sumarização, classificação, agrupamentos, linguística computacional, dentre outras.

Segundo Jurasfsky [19], classificação de textos é uma das aplicações da mineração de textos e pode ser usada em alguns contextos como detecção de spam, identificação dos autores, identificação de gênero (usando pronomes ou outros determinantes), análise de sentimentos, definição de categorias e identificação da linguagem em que foi escrito o texto.

No processo de classificação de textos tem-se como entrada um documento d e um conjunto fixo de classes C = c1,c2,...,cj. A saída será determinar a classe sobre à qual o documento d está semanticamente relacionado, ou seja, dado um documento, será assinalada a classe à que o mesmo pertence [19].

Os modelos de classificação podem ser divididos em single-label ou multi-label [1]. No single-label, cada documento pode pertencer a apenas uma classe. Nos modelos de classificação multi-label, um documento pode ser associado a uma ou mais categorias [8]. Esse tipo de classificação pode ser uma solução para classificação de textos em larga escala.

2.2 Etapas da Mineração de Textos

A classificação de textos é constituída pelas seguintes etapas: coleta de dados, definição da abordagem utilizada, a etapa de pré-processamento, um mecanismo de indexação, a aplicação do algoritmo de aprendizagem de máquina e a análise dos resultados [20].

Os textos podem ser extraídos das mais variadas fontes, tais como emails, campos textuais em banco de dados, páginas web, textos eletrônicos digitalizados, etc. Baseado no que se pretende analisar, define-se qual será abordagem a ser utilizada: semântica ou estatística [4]. A abordagem estatística é baseada na frequência dos termos encontrados no texto. A abordagem semântica procura identificar a importância das palavras dentro do contexto, utilizando, para isto, as relações morfológicas e sintáticas do idioma. Essas abordagens podem ser utilizadas separadamente ou em conjunto.

O processo de mineração de textos, apesar de assemelhar-se ao processo de mineração de dados propriamente dito, trabalha com dados não estruturados. Ao extrair informações de texto em linguagem natural, o mesmo deve ser normalizado, removendo o que for desnecessário para o entendimento do texto e preparando o mesmo para a análise a ser realizada. Essa normalização é feita através de mecanismos de tokenização, stemming, e outras técnicas utilizadas na etapa de pré-processamento[19].

A *tokenização* é a primeira etapa do pré-processamento e tem como objetivo dividir o texto em unidades, mais conhecidas como *tokens*. Essas unidades podem ser números, espaços, palavras ou termos compostos por mais de uma palavra. Quando uma dessas unidades é formada por uma palavra, chamamos de unigrama. Quando é formada por duas palavras, bigrama [31], e assim por diante. O processo de tokenização é feito, em geral, identificando-se espaços em branco e pontuações que costumam delimitar os termos. [29]

Depois de marcados estes termos, são aplicadas técnicas como a remoção de espaços em branco, a transformação de letras maiúsculas para minúscula, a retirada de números e de pontuação, bem como o uso de *stemming* e a remoção de *stopwords*. O *stemming* é uma técnica utilizada para reduzir um termo ao seu radical, eliminando as variações morfológicas como prefixos, sufixos, vogais temáticas e desinências. Desse modo, palavras como *documentação* e *documentos* seriam ambas transformadas em *document*, possibilitando a

geração de um menor número de dimensões. Outro tratamento efetuado é a retirada de termos considerados irrelevantes, como as *stopwords* e as palavras pouco frequentes. As *stopwords* são artigos, preposições, conjunções, bem como outras palavras auxiliares que não agregam valor ao texto. Já as palavras pouco frequentes são retiradas com o intuito de otimizar o processamento dos textos. Cheng at all [10] mostrou que termos medianamente frequentes tem maior probabilidade de acerto, alcançando assim maior eficácia na classificação se comparado com termos muito ou pouco frequentes. Portanto, palavras pouco frequentes, que aparecem em um número muito pequeno de documentos, ou palavras muito frequentes, presentes na maioria dos documentos, não permitem a identificação da classe a que pertence o mesmo, sendo recomendada a sua retirada.

O resultado do pré-processamento é um conjunto de termos independentes. A BOW é uma representação matricial desse conjunto de termos em que um documento é representado por uma linha e os termos por colunas nessa matriz, desconsiderando a ordem ou a classificação gramatical. Opcionalmente, pode conter a quantidade de vezes que o termo ocorre no documento. Esse tipo de classificação pode ser bastante eficiente e simples, no entanto, não considera relações semânticas e sintáticas entre os termos [3].

Dependendo do contexto, o resultado pode ser uma BOW bastante extensa. Entretanto, deve-se avaliar se a quantidade de termos gerado é realmente necessária. Uma grande quantidade de termos pode afetar o desempenho computacional. Outro ponto a ser considerado é o fato de nem sempre uma grande quantidade de termos (colunas) significa maior precisão nos resultados. Essa questão tem suscitado o uso de uma técnica conhecida na literatura como redução de dimensionalidade [14]. A mesma consiste em eliminar dimensões que podem afetar o desempenho computacional sem agregar eficácia aos classificadores [24]. A utilização de todos os termos de um documento não traz resultados significativos. Portanto, é importante calcular a relevância dos termos de um documento [18].

Os métodos mais conhecidos para calcular a relevância dos termos são: frequência absoluta (TF), frequência relativa, e a relação entre a frequência do termo e a frequência inversa do documento (TF-IDF).

Frequência absoluta ou *term frequency* (TF) [16] é a quantidade de ocorrências de um termo em um determinado documento. Esta pode ser considerada uma medida simples mas tem algumas desvantagens. A primeira delas é o fato de não levar em conta a quantidade de palavras existentes no documentos. Com isso, um termo que ocorre poucas vezes em um documento pequeno poderá ter a mesma frequência de um termo que ocorre muitas vezes em documentos com grande quantidade de palavras. Outra desvantagem é o fato de não ser possível identificar se um termo ocorre em poucos ou em muitos documentos. Essa identificação pode ser importante para definir com mais precisão a classe

em que um documento será encaminhado, pois termos raros podem ser mais informativos que termos mais frequentes.

A frequência calculada como TF-IDF [5] [30] utiliza no seu cálculo a frequência do termo e a quantidades de documentos presentes na coleção. Essa medida possibilita indicar a quantidade de documentos em que um termo aparece. Ela é calculada usando-se a frequência em que o termo ocorre e a remoção de termos onde a frequência nos documentos é inferior a um determinado parâmetro. Esse processo resulta na diminuição da importância dada aos termos que aparecem em muitos documentos e no aumento da importância daqueles que aparecem em poucos documentos. Uma das maneiras de se identificar o peso do termo pode ser dividindo o número de vezes que o termo (TF) ocorre em um documento d pelo número de documentos em que o termo t aparece (DF):

$$TF - IDF = TF/DF \tag{2.1}$$

Esse cálculo deve ser efetuado novamente com a entrada de novos documentos, pois a dimensão pode variar com a entrada de novos documentos e novos termos.

Para otimizar o tratamento dos textos, uma técnica utilizada é a indexação. O objetivo da indexação é facilitar a identificação e busca dos termos. Nesta fase, serão criadas estruturas de dados conhecidas como índices, que serão utilizados para organizar a base. Assim, quando for necessário realizar buscas de um documento, não será preciso varrer toda a base.

2.3 Algoritmos de Indução de Classificadores

Após o pré-processamento aplicam-se as técnicas de mineração de texto, como, por exemplo, a classificação, para extrair padrões dos dados trabalhados. A classificação consiste na utilização de algoritmos de aprendizagem para atribuir um documento a uma classe, considerando os dados de entrada do documento. Árvores de Decisão (Tree), Naive Bayes, Random Forest e SVM são exemplos de modelos de classificação. Esse tipo de aprendizado, em que se utiliza exemplos anteriormente rotulados, é chamado de aprendizado supervisionado.

Support Vector Machine ou Máquinas de Vetores Suporte (SVM), criado por Vapnik [11], é um método matemático capaz de classificar documentos de forma supervisionada. O SVM utiliza hiperplanos para separar os dados, deixando a maior margem possível na separação das classes. Ao se encontrar um hiperplano que separe as classes, a classificação de um novo ponto será trivial [37]. Em um problema de classificação binária, por exemplo, o SVM irá separar as duas classes através de um hiperplano maximizando a distância entre elas.

A ideia básica do SVM pode ser observada na Figura 2.1. Esta representa um espaço dimensional onde cada dimensão corresponde a um termo: saúde, educação e convênios. Dessa forma, todos os documentos estarão mapeados em algum lugar nesse espaço vetorial. Pode-se observar que o documento d1 provavelmente cobre os conceitos de saúde e convênios mas não diz nada sobre educação. Já o documento d2, provavelmente cobre assuntos relacionados a convênio e educação, mas não aborda o tema saúde. Portanto, todos os documentos serão verificados através da perspectiva do espaço vetorial, ignorando outras informações presentes no documento que não estejam no vetor.

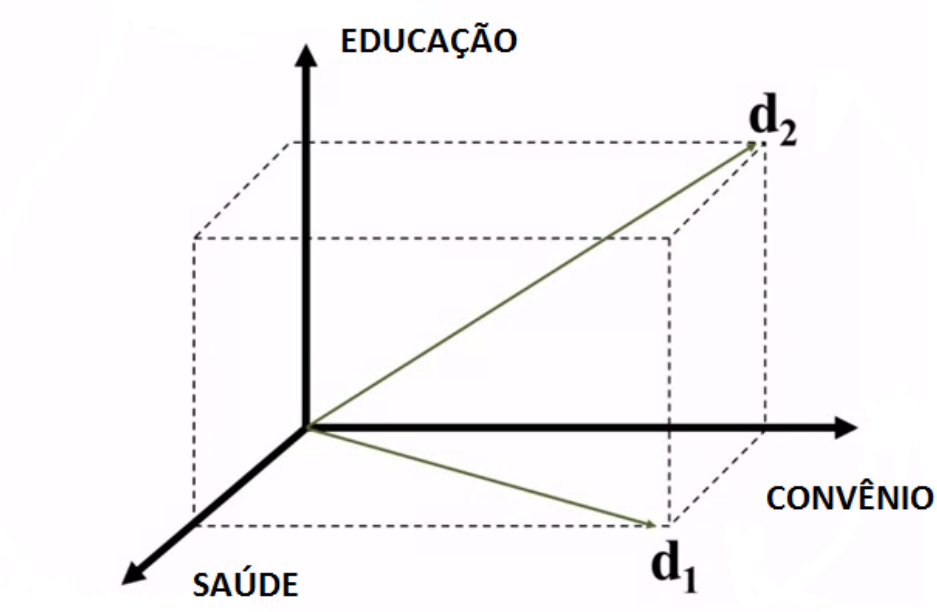

Figura 2.1: SVM

Naive bayes [22] é um algoritmo probabilístico de classificação que produz estimativas de probabilidade para as classes. O objetivo é calcular a probabilidade de um dado documento para escolher a melhor classe em que o mesmo se encaixa. Ele é chamado de classificador ingênuo, pois assume que os atributos são independentes um do outro, dado um valor da classe. Dentro dos modelos de classificação bayesiana, dois são mais conhecidos: o modelo binário e o modelo multinomial. No modelo binário, também chamado de básico, cada atributo pode indicar ou não a ocorrência de um evento no documento. O modelo multinomial trabalha com o número de vezes que cada evento ocorre no documento.

Árvores de Decisão ou *Decision Tree* [35] é um algoritmo usado para regressão ou classificação, de modo supervisionado. O mesmo é responsável por criar uma estrutura de árvore. Essa estrutura irá mapear as observações de cada classe de acordo com os valores apresentados pelos seus atributos. As Árvores de Decisão buscam encontrar uma

solução dividindo o problema em subproblemas de menor tamanho e aplicando-se, de forma recursiva, a solução em cada subproblema. Os nós da árvore montada pelo algoritmo deverão indicar o teste de um determinado atributo, partindo da raiz para as folhas. Sendo assim, um nó testa um atributo e o ramo corresponde ao valor do atributo que o nó testa. Esse processo será repetido para cada nó até chegar ao nó folha, que fornecerá a classe. Esse algoritmo pode ser usado com atributos de entrada discretos ou contínuos. No entanto, ele trabalha melhor com valores discretos pois valores contínuos podem gerar árvores muito grandes e de difícil compreensão.

O *Random Forest* é um algoritmo desenvolvido por Breiman [6] para classificação de dados. Esse algoritmo gera várias árvores de decisão construídas simultaneamente e considerando todas as variáveis selecionadas para a análise. Essas árvores serão utilizadas conjuntamente para classificação de novos objetos através da agregação dos resultados. Para isto, cada árvore irá fornecer um voto indicando a sua decisão sobre a classe à que o objeto pertence. O algoritmo escolhe a classificação com o maior número de votos.

Além dos algoritmos de classificação descritos anteriormente, na literatura são encontrados outros algoritmos ou técnicas aplicados à problemas de classificação como a Análise Discriminante Linear (LDA), a Codificação adaptativa de Huffman em conjunto com o Minimum Description Length (CAH+MDL) e a Similaridade de Cossenos na modelagem VSM (Vector Space Model).

A codificação de Huffman foi desenvolvida por David A. Huffman [17] e consiste em um código de tamanho variável para compressão de dados, onde caracteres que aparecem com mais frequência recebem representações com menos bits que caracteres menos frequentes, gerando assim uma árvore binária menor. Esse método utiliza probabilidades das ocorrências dos caracteres, associando cada um deles a um peso. A princípio, cada árvore tem um nó com um caracter e o peso associado ao mesmo. A cada iteração, o algoritmo junta duas árvores criando uma nova. Depois de juntar todas as árvores, o algoritmo percorre a árvore atribuindo valor zero para os caracteres à esquerda e 1 para os da direita, calculando depois qual a taxa de compressão para aquela árvore.

A frase "Controladoria Geral da União" sem considerar os espaços em branco, requer 200 bits para ser representada com a codificação ASCII, pois é composta por 28 caracteres de 8 bits cada. Na codificação de Huffman é atribuído uma codificação com menor quantidade de bits para os caracteres mais frequentes. Após a construção da árvore, o código de um caracter é formada por uma representação binária percorrendo a árvore a partir da raiz até o nó folha representado o caracter, onde acrescenta-se zero ao percorrer um link à esquerda e um ao percorrer um link à direita. A frase exemplo requer um total de 85 bits para ser representada. Por exemplo, a codificação para o caracter A é 0 0. A Figura 2.2 ilustra a Árvore de Huffman criada para essa frase, resultando em uma economia de 115

bits. Nesta compressão não foram considerados os espaços em branco.

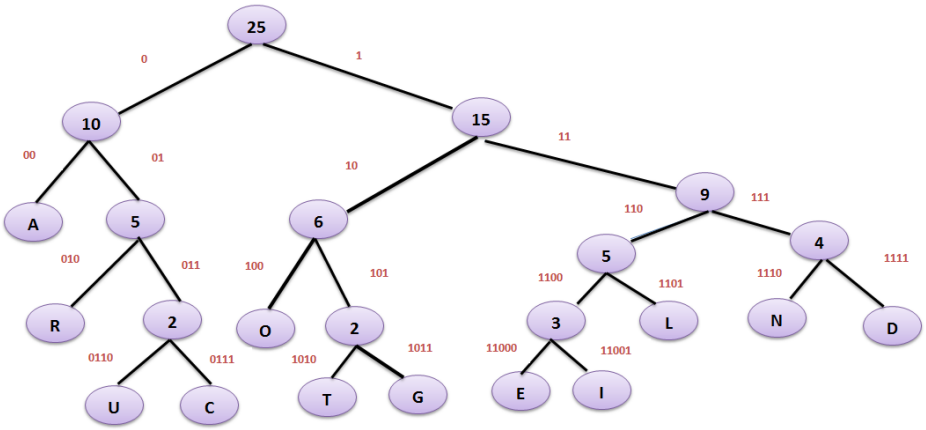

Figura 2.2: Árvore de Huffman

Minimum Description Length (MDL) trabalha com a compressão de dados basea-das em suas regularidades, possibilitando assim que um determinado dado seja descrito utilizando-se menos símbolos do que seriam necessários para a sua representação literal. Quanto mais regularidades forem encontradas mais os dados serão comprimidos. Segundo Witten et al. [36] o MDL leva em consideração o princípio de que o melhor modelo para definir uma massa de dados é aquele que minimiza o número de bits requeridos para definir este modelo. O MDL calcula o comprimento do modelo e dos dados. Dessa maneira, para determinado conjunto de dados, seleciona-se o modelo que possua a menor soma dos seus dados com o próprio modelo.

A técnica de representação de documentos utilizando VSM (Vector Space Model) utiliza o cosseno do ângulo entre dois documentos como uma medida da similaridade entre eles. Essa abordagem pode ser utilizada para classificação de textos. Caso a similaridade (cosseno) seja próxima de 1, indicará que os documentos são muito similares, pertencendo, provavelmente à mesma classe ou categoria. No entanto, se o resultado for próximo de 0, indicará que os documentos são pouco similares, não pertencendo à mesma classe. Resultados médios, nem próximos de 0 e nem próximos de 1, podem não ser conclusivos, não sendo capazes de identificar a classe a que o documento pertence [21].

2.4 Avaliação da Performance dos Classificadores

Para validar o resultado gerado pelos classificadores, torna-se necessária a aplicação de métricas de avaliação. Segundo Sebastiani [12], a avaliação experimental de um classificador geralmente mede sua efetividade, isto é, sua capacidade de tomar as decisões corretas de classificação. Em mineração de textos e aprendizagem de máquinas, a avaliação dos

resultados é feita utilizando métricas de desempenho, como acurácia, precisão, recall, kappa, dentre outras. Portanto, qualidade de recuperação e acerto dos classificadores é baseada em noções de relevância. Acurácia mede os acertos realizados pelo classificador. Kappa [7] é uma métrica utilizada para avaliar o índice de concordância de uma tarefa de classificação, indicando a coesão dos dados classificados. Os valores do Kappa ficam entre 0 e 1, sendo que o 0 representa a falta de concordância, considerado puro acaso e 1 representa a concordância perfeita. Além dessas métricas, é possível avaliar os resultados utilizando uma matriz de confusão. Essa matriz compara os valores classificados, verificando, para cada valor previsto, se o mesmo corresponde ao valor real.

Precisão é a quantidade de itens selecionados que estão corretamente classificados. Mede, dentre todos os documentos julgados, a quantidade de documentos classificados corretamente como positivos, sendo portanto, a proporção do número de itens selecionados que foram recuperados corretamente. Em um contexto com um total de 1000 denúncias, por exemplo, caso a precisão seja igual a 0.87, quer dizer que 870 das 1000 denúncias foram classificadas corretamente e 130 foram classificadas incorretamente.

Recall é a porcentagem de itens relevantes que foi recuperada [25]. Mede, dentre todos os documentos que são realmente da classe de positivos, a relação entre a quantidade de documentos classificados como positivos e os itens recuperados, ou seja, a proporção entre o número de itens relevantes recuperados e o número total de itens relevantes. F-Measure é uma média harmônica ponderada entre o recall e a precisão.

Os classificadores podem retornar o resultado indicando uma categoria a que os documentos pertencem, mais de uma categoria ou um ranking de probabilidades. Ao indicar uma categoria, os resultados serão validados de acordo com os índices de acerto para cada categoria e em relação ao conjunto todo. As medidas acima descritas, geralmente, calculam a taxa de efetividade baseada nas taxas de acertos e quantidades recuperadas. No entanto, podemos analisar o valor de confiança retornado por um classificador. Essa informação é necessária na criação de ranking de classificação, definindo, a partir daí, quais são as probabilidades de um documento pertencer à categoria determinada. Esse tipo de métrica é realizada através da Curva ROC.

Gráfico de ROC é um gráfico em que os eixos X e Y representam a taxas de verdadeiros positivos (TVP) e a taxa de falsos positivos (TFP) respectivamente.

2.5 CRISP-DM

Ao se trabalhar com mineração de textos deve-se levar em consideração não somente a busca por padrões e descoberta de conhecimento mas também todo o pré-processamento e limpeza dos dados que antecedem a mineração propriamente dita. Uma forma consa-

grada na literatura de abordar as etapas envolvidas na mineração de dados é o CRISP-DM (Cross Standard Process for Data Mining) [2]. Este modelo é dividido em 6 fases e tem como finalidade definir os passos a serem seguidos em um projeto de mineração de dados e que pode ser aplicado a mineração de textos desde que as fases de entendimento dos dados e de pré-processamento sejam adaptadas para dados textuais. Conforme mostrado na Figura 2.3, adaptada de [2], o CRISP-DM possui 6 fases: entendimento do negócio, entendimento dos dados, pré-processamento dos dados, modelagem, avaliação e implantação.

Figura 2.3: CRISP-DM

- Entendimento do Negócio

  Esta é a primeira fase do ciclo e de primordial importância. Foca o entendimento dos objetivos, sob a perspectiva do negócio, e os requerimentos do projeto, a relevância do conhecimento prévio e os objetivos do usuário final. Os objetivos específicos a serem alcançados e os critérios para saber se foram ou não alcançados devem ser formalizados e proposta uma tarefa de mineração de dados que possa contribuir para a consecução desses objetivos. Os recursos (dados, infraestrutura, software, recursos humanos, etc.) necessários para a realização da tarefa de mineração devem ser mapeados e avaliado se estão disponíveis ou não. Nessa etapa são elaborados o

plano do projeto, especificando os passos a serem executados no resto do projeto e a definição do problema.

- Entendimento dos Dados

  Nesta fase, realiza-se a seleção dos dados disponíveis e a análise dos dados, visando maior familiaridade com os mesmos. A análise dos dados consiste em: identificar problemas de qualidade nos dados; descobrir os primeiros conhecimentos; descrever os dados em termos de formato, quantidade de registros e campos (atributos); estimar a distribuição dos atributos; verificar a existência de relacionamentos entre pares de atributos; e identificar agrupamentos ou subconjuntos existentes nos dados.

  Na mineração de textos, deve-se levar em consideração algumas peculiaridades relacionadas ao entendimento dos dados, tais como: erros de ortografia, erros de digitação, uso de palavras em diferentes línguas, gírias, etc. Deve-se avaliar também se os dados são semi estruturados ou não estruturados com a finalidade de determinar a melhor abordagem a ser aplicada.

- Pré-processamento dos dados

  Nesta fase os dados são tratados visando torná-los adequados à aplicação dos algoritmos que serão utilizados para a indução de modelos. As principais tarefas a serem executadas são: seleção dos atributos relevantes, limpeza dos dados, construção, integração e formatação dos dados para entrada nos algoritmos de indução a serem utilizados. São tomadas decisões relativas à aplicação de técnicas para remoção de ruído ou de dados espúrios, estratégias para lidar com valores faltantes, criação de atributos derivados e de novos registros, integração de tabelas se existirem, e discretização dos dados numéricos, se necessário.

  No caso da mineração de textos são executadas tarefas específicas de pré-processamento do texto. São aplicadas técnicas de remoção de *stopwords*, remoção de caracteres estranhos, números, pontuação, acentuação, espaços em branco, etc. Após essa limpeza dos dados podem ser aplicadas técnicas de redução de dimensionalidade e o resultado desse processo será a matriz termo documento (BOW - Bag of Words, em inglês) onde cada termo (coluna) representa um atributo e cada linha representa um documento.

- Modelagem

  A modelagem consiste na aplicação de técnicas que objetivam encontrar padrões e descoberta do conhecimento. Aqui são utilizados algoritmos de aprendizagem de máquina, sendo que estes terão como entrada os dados tratados na etapa anterior.

- Avaliação

A avaliação consiste em testar a efetividade do modelo aplicado. Normalmente essa análise é baseada em indicadores e métricas comparativas. É uma validação da adequação dos tratamentos aplicados aos dados e da modelagem escolhida.

- Implantação

  Na implantação os resultados alcançados são colocados à disposição do usuário com a finalidade de melhorar os processos de negócio.

2.6 Trabalhos Correlatos

Ormonde [28] aborda o problema da classificação automática de páginas Web, utilizando a Codificação Adaptativa de Huffman (CAH) na classificação de sites Web em português e em inglês. No trabalho em questão, foi considerado o conteúdo textual das páginas a serem classificadas e informações não disponíveis nos problemas de classificação de textos em geral, como metadados e informações semi-estruturadas, com o objetivo de melhorar a precisão do classificador. A extensão do algoritmo CAH+MDL (anteriormente utilizado para classificação binária ou multi classe) possibilitou tratar problemas de classificação multi-label com ou sem hipóteses de mundo fechado, ou seja, os documentos podem ou não ser classificados em ao menos uma categoria. A performance do algoritmo foi comparada com a performance do algoritmo SVM, muito utilizado para classificação de textos. O algoritmo CAH+MDL alcançou uma acurácia quase tão boa quanto a produzida pelo SVM e um Recall maior, mas a precisão foi menor. Esses valores foram considerados na classificação usando a hipótese de mundo fechado. A vantagem desse algoritmo é o fato do mesmo permitir treinamento em tempo real e incremental. Além disso, a complexidade do mesmo cresce linearmente de acordo com as categorias conhecidas independente das suas combinações. Já os outros algoritmos de classificação têm complexidade exponencial, proporcional ao número de categorias, porque decompõem o problema original em diversos problemas menores a serem tratados por classificadores binários ou multi classe.

Byron et al. [34] utilizaram aprendizagem de máquinas para triagem de citações em revisões sistemáticas. As revisões sistemáticas avaliam e analisam a literatura pertinente. Triagem de Citações é um passo demorado e crítico neste tipo de processo. Normalmente, os usuários devem avaliar milhares de citações para identificar artigos elegíveis. Semi-automatizar o processo de triagem de citação é difícil porque tal estratégia deve identificar todas as citações elegíveis para a revisão sistemática. Esta exigência é ainda mais difícil devido ao desequilíbrio de classe; há muito menos “relevantes“ do que as citações “irrelevantes“ para qualquer revisão sistemática. O trabalho em questão explora a aplicação de técnicas de aprendizado de máquina para semi-automatizar a triagem de citação, reduzindo assim a carga de trabalho dos colaboradores, sem perder a qualidade e

a abrangência da revisão. Foi utilizado SVM para classificação on-line de citação de triagem, discriminando automaticamente as "relevantes" de "irrelevantes". Segundo o autor, o resultado experimental de três conjuntos de dados de revisão sistemática demonstrou a diminuição do número de citações que devem ser rastreadas manualmente pela metade em dois conjuntos e cerca de 40% no terceiro, sem excluir quaisquer das citações elegíveis para a revisão sistemática.

Qiwei He [15] trabalhou com técnicas de aprendizagem de máquina para triagem e diagnóstico para transtorno de estresse pós-traumático (PTSD). O autor acredita que a saúde física e mental das pessoas pode ser prevista pelas palavras que usam. No entanto, o procedimento para lidar com essas palavras é bastante difícil com métodos quantitativos tradicionais. O primeiro desafio seria extrair informações robustas de padrões de expressão diversificados, o segundo seria transformar o texto não estruturado em um conjunto de dados estruturado. O estudo desenvolveu um novo método de análise textual usando 300 narrativas auto coletados on-line, extraindo palavras-chave altamente discriminativas com o algoritmo de Qui-quadrado e construiu um modelo de avaliação textual para classificar os indivíduos com a presença ou ausência de PTSD. O estudo revelou algumas características expressivas nos escritos de pacientes com PTSD e alcançou uma acurácia de 81,6 quando usado com unigrams mas a acurácia decresce quando usada em bigrams. Embora os resultados da análise de texto não tenham sido completamente análogos com os resultados de entrevistas estruturadas em diagnóstico de PTSD, a aplicação de mineração de texto é um complemento promissor para avaliar PTSD em ambientes clínicos e de pesquisa.

Chen et al.[9] aplicou Análise Discriminante Linear (LDA) e multi granularidade ao trabalhar com classificação de textos pequenos, o que reduziu o erro na classificação para 16.68 %. Textos curtos são caracterizados pela falta de tamanho dos textos e por possuir *sparcity* dos termos apresentados, os quais dificultam o gerenciamento e análise baseada na BOW. Primeiramente foram gerados tópicos utilizando LDA. Em seguida, foi escolhido um subconjunto de todos os tópicos gerados automaticamente para criar os tópicos de múltipla granularidade, através de hierarquia. Os classificadores foram aplicados combinando-se as características dos conjuntos de tópicos de múltipla granularidade com as características da palavras dos textos.

Tie-Yan Liu et al.[32] utilizou dois tipos de implementação SVM (flat e hierárquico) para classificação de páginas web em toda a taxonomia de categorias do Yahoo!. SVM flat são classificadores que não tiram vantagem da estrutura de uma árvore de taxonomia. SVM hierárquico são métodos que decompõem as tarefas de treinamento de acordo com a estrutura da taxonomia. Segundo o autor, o número máximo de categorias testadas usando SVM não chega a 5000. Essa quantidade é menos que 2% do número de categorias do Yahoo! Diretórios. O trabalho teve como objetivo avaliar a escalabilidade e

a efetividade do SVM aplicado à classificação em larga escala utilizando 792.601 documentos classificados em 292.216 categorias, em 6 níveis hierárquicos. O resultado mostra que em termos de escalabilidade o flat SVM possui uma complexidade muito alta sendo o SVM hierárquico mais eficiente para classificações em larga escala no mundo real. Em termos de efetividade, nenhum dos dois preenche as necessidades de um classificador em larga escala. A grande quantidade de taxonomias associadas a categorias extremamente raras nos diretórios do Yahoo tornam a performance dos classificadores inaceitável.

Capítulo 3

Solução Proposta

Este capítulo descreve a metodologia utilizada e a solução proposta para a prova de conceito de classificador com grande número de classes para triagem de documentos semi-estruturados escritos em português.

3.1 Metodologia

Essa pesquisa foi realizada segundo a metodologia a seguir. O modelo de referência CRISP-DM foi utilizado na etapa de mineração de textos.

- Revisão bibliográfica
- Entendimento dos dados
- Revisão do processo atual de triagem utilizado pela CGU
- Proposição e implementação de modelos de mineração de textos
- Análise de resultados e testes
- Realização de prova de conceito do modelo proposto para triagem automática de denúncias recebidas pela CGU
- Validação com a Unidade da CGU encarregada da triagem

3.1.1 Visão Geral

Primeiramente foi feito levantamento do estado da arte em mineração de textos e classificação objetivando encontrar possíveis soluções para o problema aqui abordado. Realizou-se também levantamento dos dados e do processo de triagem com a área responsável.

O levantamento do estado da arte realizado não localizou nenhuma abordagem para classificação com grande número de classes de documentos textuais compostos por reduzido número de palavras, sem a ocorrência de metadados, e com número reduzido de documentos por classe. Desta forma se partiu para a experimentação visando o desenvolvimento de uma abordagem específica para o contexto da CGU.

Os estudos iniciais para indução de classificadores com grande número de classes foram realizados com o uso da linguagem R e da interface R Studio para indução de classificadores utilizando os algoritmos SVM, Random Forest, Naive Bayes e Árvore de Decisão (C4.5). Os resultados preliminares foram publicados no ENIAC 2014 [23]. Essa abordagem foi realizada com um pequeno número de classes (apenas quatro classes) e não apresentou bons resultados ao se aumentar o número de classes. Por não ter se mostrado escalável, foi abandonada.

A seguir foi testado o uso do Classificador CAH+MDL onde se constrói uma árvore de Huffman para cada uma das classes possíveis. Essa abordagem apresentou métricas de precisão e recall da ordem de 0.80, considerando classificação multi-label baseada em ranking, sendo um documento classificado em até três classes.

Durante o processo de preparação dos dados para o classificador, os dados são divididos em 2 conjuntos: treinamento e teste, com 70% dos dados reservados para treinamento e os outros 30% para testes. Essa divisão tem como objetivo separar os dados que o algoritmo irá usar para treinar dos dados de teste.

Os resultados foram avaliados utilizando as métricas Precisão e Recall. Essas métricas são utilizadas para avaliar a efetividade dos modelos de classificação e aprendizagem de máquina.

Não foi desenvolvida nenhuma interface gráfica para facilitar a utilização do protótipo construído pelos servidores da CGU pois nesse órgão um sistema de denúncias e a triagem automática de denúncias ao qual o protótipo desenvolvido será incorporado.

3.2 Entendimento do Negócio

A CGU está dividida nas quatro grandes áreas: Secretaria Federal de Controle (SFC), Secretaria de Transparência e Prevenção da Corrupção (STPC)[33],Corregedoria Geral da União (CRG) e Ouvidoria Geral da União (OGU)[27], as quais são estruturadas em coordenações. A SFC é responsável por avaliar a execução de programas de Governo, comprovar a legalidade e avaliar os resultados, quanto à eficácia e eficiência, da gestão dos administradores públicos federais, exercer o controle das operações de crédito, além de exercer atividades de apoio ao controle externo. À CRG cabem as atividades relacionadas à apuração de possíveis irregularidades cometidas por servidores públicos e à aplicação

das devidas penalidades. A STPC desenvolve mecanismos de prevenção à corrupção. A OGU é responsável por receber, examinar e encaminhar denúncias, reclamações, elogios, sugestões e solicitações.

A Figura 3.1 apresenta o organograma da CGU.

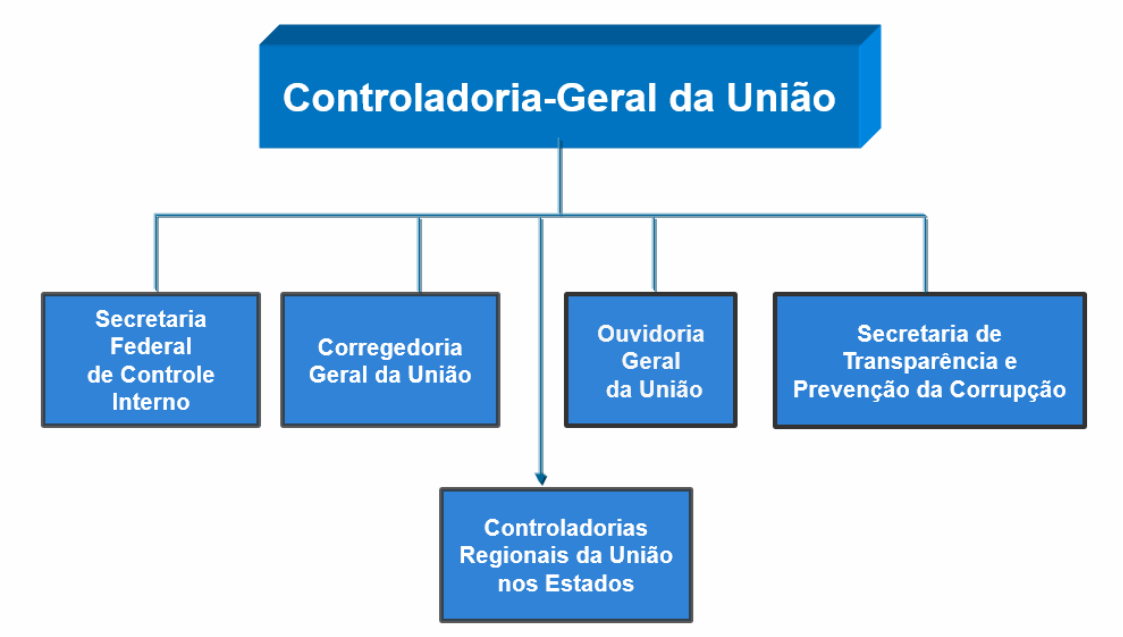

Figura 3.1: Organograma da CGU

A SFC é dividida em coordenações temáticas de acordo com o assunto a ser tratado. Assim, a Coordenação Geral de Auditoria na Área de Saúde, da Secretaria Federal de Controle (SFC/DS/DSSAU), por exemplo, realiza auditorias relativas à área de saúde, bem como aos programas de governo e aos órgãos ligados ao tema saúde. A CRG também segue o mesmo padrão relacionado à SFC mas realiza auditorias de assuntos relacionados aos servidores públicos.

3.2.1 Processo de triagem

As denúncias podem ser cadastradas no site da CGU (através do menu "Denúncias e Manifestações") ou recebidas por carta ou envidas por e-mail. O e-ouv é um sistema de ouvidorias do Poder Executivo Federal o qual consolida todas as denúncias recebidas pela CGU. Elas são recebidas pelo Protocolo (unidade da CGU) e digitalizadas caso a mesma venha por e-mail ou quando recebidas por carta. No Protocolo é feita uma primeira verificação, e o cadastramento da denúncia no Sistema de Gestão de Informações (SGI) da CGU, podendo ser aceita ou rejeitada. Caso a mesma seja rejeitada, ela será transformada em lixo eletrônico. As denúncias são encaminhas para o Lixo Eletrônico se: os dados são

insuficientes; a demanda é repetida; trata-se de SPAM (correntes, propagandas, etc.); a mensagem é sem propósito; possui conteúdo pornográfico; textos que utilizam palavras ou expressões de baixo calão sem conteúdo pertinente. Essa primeira verificação é realizada por usuários não especializados, sendo considerada somente uma verificação superficial.

A Figura 3.2 é um exemplo de uma denúncia enviada para Lixo Eletrônico que contém queixa que não faz parte da competência da CGU averiguar.

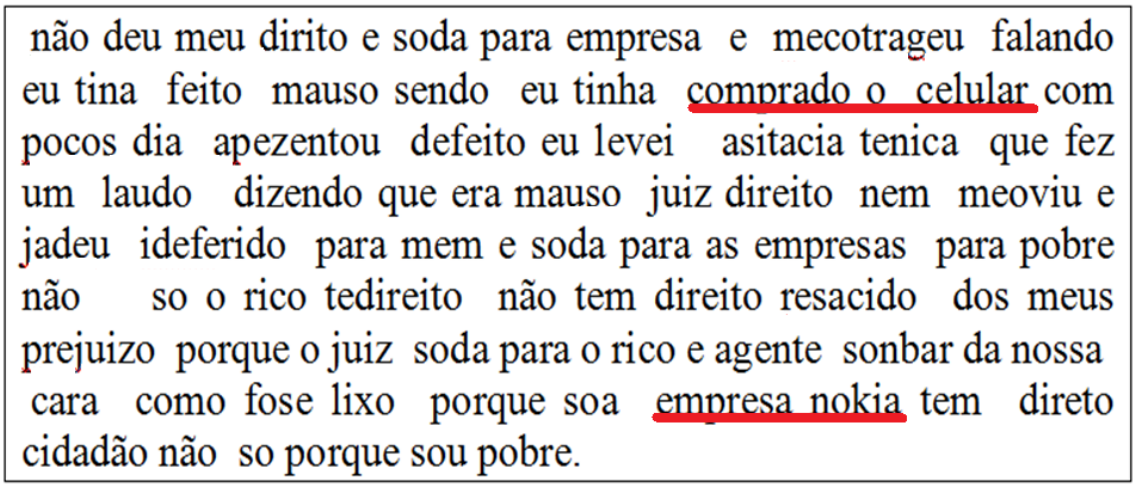
não deu meu dirito e soda para empresa e mecotrageu falando eu tina feito mauso sendo eu tinha comprado o celular com pocos dia apezentou defeito eu levei asitacia tenica que fez um laudo dizendo que era mauso juiz direito nem meoviu e jadeu ideferido para mem e soda para as empresas para pobre não so o rico tedireito não tem direito resacido dos meus prejuizo porque o juiz soda para o rico e agente sonbar da nossa cara como fose lixo porque soa empresa nokia tem direto cidadão não so porque sou pobre.

Figura 3.2: Exemplo de Lixo Eletrônico

Após este primeiro passo, o processo de triagem propriamente dito é realizado pela Coordenação de Atendimento ao Cidadão (CGCID), da Ouvidoria-Geral da União. Os funcionários responsáveis pela triagem identificam a área de destino da denúncia de acordo com o Manual de Triagem. O Manual de Triagem indica as principais coordenações da CGU às quais poderão ser encaminhadas as denúncias bem como os assuntos relacionadas a cada área. Este manual detalha as opções de arquivamento e todos os passos para triar uma denúncia dentro do sistema SGI, módulo, Demandas Externas.

Além da equipe responsável pela triagem da denúncia na CGCID, em alguns casos, as denúncias são enviadas às unidades regionais da CGU nos Estados da Federação para que as mesmas ajudem no processo de triagem, já que o percentual de denúncias recebidas é maior do que a capacidade da coordenação de análise das mesmas.

A triagem de denúncias tem por finalidade selecionar aquelas que apresentam elementos sobre uso irregular de recursos públicos federais ou má conduta de servidor público federal. No processo de triagem, a equipe responsável irá ler cada denúncia, avaliar a pertinência e a materialidade da mesma, além de verificar se é competência da CGU. Caso seja, ao final deste processo, será encaminhada à unidade responsável pela apuração. Segundo o Manual de Triagem, as denúncias devem ser habilitadas ou inabilitadas. As habilitadas são as que reunirem elementos consistentes relativos à irregularidades na

aplicação de recursos públicos federais ou irregularidades em órgãos ou entidades da Administração Pública Federal direta ou indireta, ou ainda, a conduta de servidor público federal. As denúncias inabilitadas são as denúncias "vazias" ou "superficiais" que não ofereçam detalhes sobre as irregularidades que possam sustentar um trabalho de apuração pela SFC ou CRG. Também são consideradas inabilitadas as denúncias que não envolvam a aplicação de recursos federais ou que tratem tão somente de irregularidades em órgãos estaduais ou municipais, bem como aquelas que, embora sejam de matéria federal, não sejam da competência da CGU.

As habilitadas serão encaminhadas, conforme a matéria, à unidade competente da CGU para apuração, normalmente, uma das coordenações da Secretaria Federal de Controle Interno (SFC) [13] ou da Corregedoria Geral da União (CRG)[26], e serão cadastradas no SGI, recebendo um número de protocolo.

As denúncias inabilitadas são arquivadas. São seis os motivos pelos quais uma denúncia pode ser arquivada: por perda de objeto; por não ser competência da CGU; por insuficiência de elementos; por já ter sido objeto de fiscalização; ou arquivar com ciência de órgão externo.

Denúncias são inabilitadas por perda do objeto quando o fato ou situação não existe mais no momento em que ocorre o processo de triagem. São arquivadas por não ser competência da CGU as denúncias que se referem a irregularidades nos órgãos do Poder Judiciário ou Legislativo ou dos Estados e Municípios, estes últimos não relacionados à aplicação de recursos federais ou transferência de recursos federais, mas não sujeitos ao controle da CGU. Denúncias que se referem a fatos já fiscalizados ou conhecidos e investigados pela CGU serão arquivadas por já ter sido objeto de fiscalização mas será necessário informar o número do relatório de fiscalização/auditoria. Neste caso, deverá ser feita uma pesquisa nos relatórios de auditoria, fiscalização e ações na CGU. Denúncias consistentes e relevantes mas não inseridas na competência da CGU serão arquivadas e dada ciência ao órgão externo a que se refere. No caso da denúncia arquivada com ciência de órgão externo, deve-se ter cuidado com os dados do cidadão. Se o mesmo solicitou sigilo, deve ocorrer a desidentificação do mesmo ao encaminhar a denúncia para o órgão devido.

As denúncias habilitadas e triadas serão classificadas como procedimento ordinário ou simplificado. Será procedimento simplificado quando a mesma possuir elementos suficientes e estiver dentro da competência da CGU, mas não se justifica uma apuração específica. Será procedimento ordinário quando estiver devida e objetivamente fundamentada e possuir materialidade econômica ou relevante interesse/repercussão social, ou ainda quando tratar de irregularidades praticadas por altos dirigentes - DAS 4 ou maior.

Outra validação feita é a busca de precedentes. Ocorre, muitas vezes, de um cidadão

enviar a mesma denúncia mais de uma vez ou denúncias relativas ao mesmo tema serem encaminhadas por pessoas diferentes. Essa busca serve para indicar outra denúncia já existente na Casa que trate do mesmo objeto da denúncia sob análise. Essa verificação pode ser determinante na decisão de habilitar a denúncia em análise caso houver novo fato relevante, de modo a complementar a denúncia anterior. Entretanto, se a nova denúncia for mera definição do conteúdo já denunciado em precedente, deve-se indicar o precedente e arquivar a nova denúncia.

A busca de precedentes é realizada através do texto da denúncia e retorna o NUP (número do protocolo dentro do SGI). Esse número deve ser informado ao arquivar a denúncia para que fique clara a justificativa do arquivamento e para posterior consulta.

Após identificar qual unidade irá receber a denúncia ou optar pelo recebimento da mesma, o funcionário responsável redige um pequeno texto resumindo os dados informados pelo denunciante e encaminha a denúncia para validação. Essa validação é feita por outro funcionário diferente do que fez a primeira triagem. Este texto será encaminhado junto com a denúncia para a área de destino para facilitar a análise ou dar um pequeno direcionamento do assunto tratado.

Mesmo as denúncias triadas pelas regionais passam novamente pela Ouvidoria para uma validação final. As regionais realizam a primeira triagem sendo necessário que a denúncia retorne para a CGCID para validação.

A Figura 3.3 detalha o fluxo da denúncia desde o momento em que a mesma é recebida em papel pelo protocolo ou cadastrada no sistema.

Após a realização da triagem, a denúncia é encaminhada a uma das coordenações responsáveis para a apuração da mesma. Esta apuração não foi escopo deste trabalho.

3.2.2 Sistema

O sistema de denúncias atual está passando por um processo de mudança, convivendo temporariamente o sistema existente e o novo sistema. A implantação do novo sistema está prevista para dezembro de 2015. Serão descritos aqui os dados e a estrutura do sistema atual bem como as mudanças do novo sistema a ser implementado e que possam ter impacto sobre este trabalho.

No contexto do sistema atual, existem dois bancos de dados utilizados para o tratamento de denúncias. Um banco que recebe as denúncias cadastradas no formulário eletrônico ou digitalizadas pelo protocolo, o banco Denúncia e o banco de dados do SGI que irá fazer o trâmite de triagem dentro deste sistema. O SGI é de uso interno da CGU, que contempla um sistema de protocolo e módulos especializados para cada área da CGU, entre eles o módulo de Demandas externas, responsável pela triagem e categorização das denúncias.

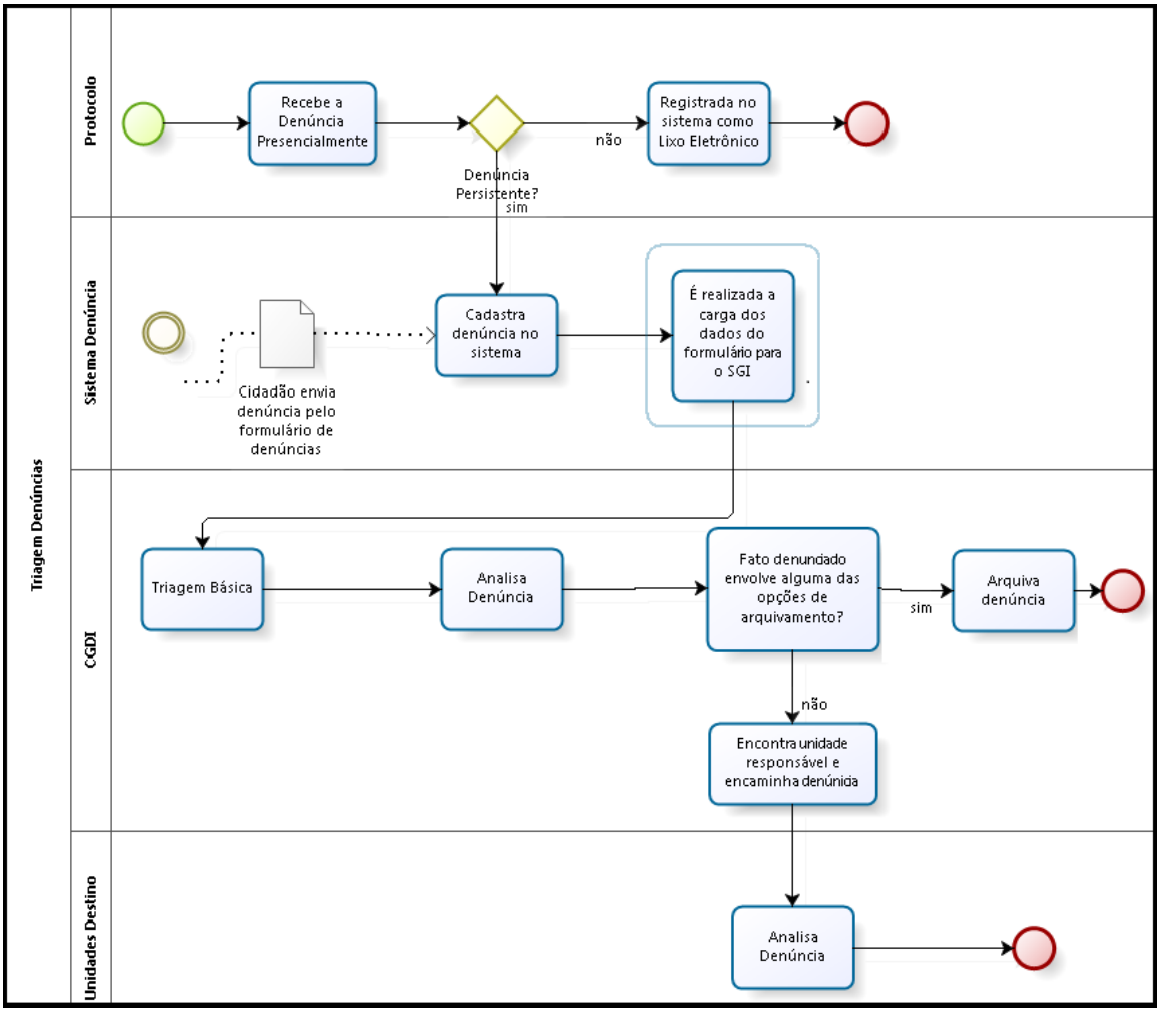

Figura 3.3: Fluxo da Denúncia

Quando as denúncias são cadastradas no formulário disponível no portal da CGU, o texto delas é gravado no tabela FatoDenunciado no banco de dados Denúncia. Este banco armazena todos os dados relativos à denúncia, como o texto, o denunciante, os arquivos de texto anexos, etc. As denúncias recebidas por carta também serão inseridas nesta mesma tabela. A Figura 3.4 exibe o modelo de dados do banco de denúncias. Os dados relativos ao denunciante estão armazenados na tabela Denunciante. No entanto, a denúncia pode ser feita de forma anônima, não sendo obrigatório ao denunciante informar seus dados. Os arquivos anexos como fotos, documentos, notas, etc, serão armazenados na tabela ArquivoFatoDenunciado.

Todos os dias é realizada uma carga destes dados inserindo as denúncias cadastradas no banco de denúncias para o sistema SGI, na tabela Manifestação. Todas as denúncias recebidas, independente de serem consideradas lixo eletrônico ou não, serão gravadas no sistema SGI, na tabela Manifestação e classificadas no grupos de assunto: Denúncia/Representação, com o campo idorigemmanifestação igual a 1, indicando que tratar-se

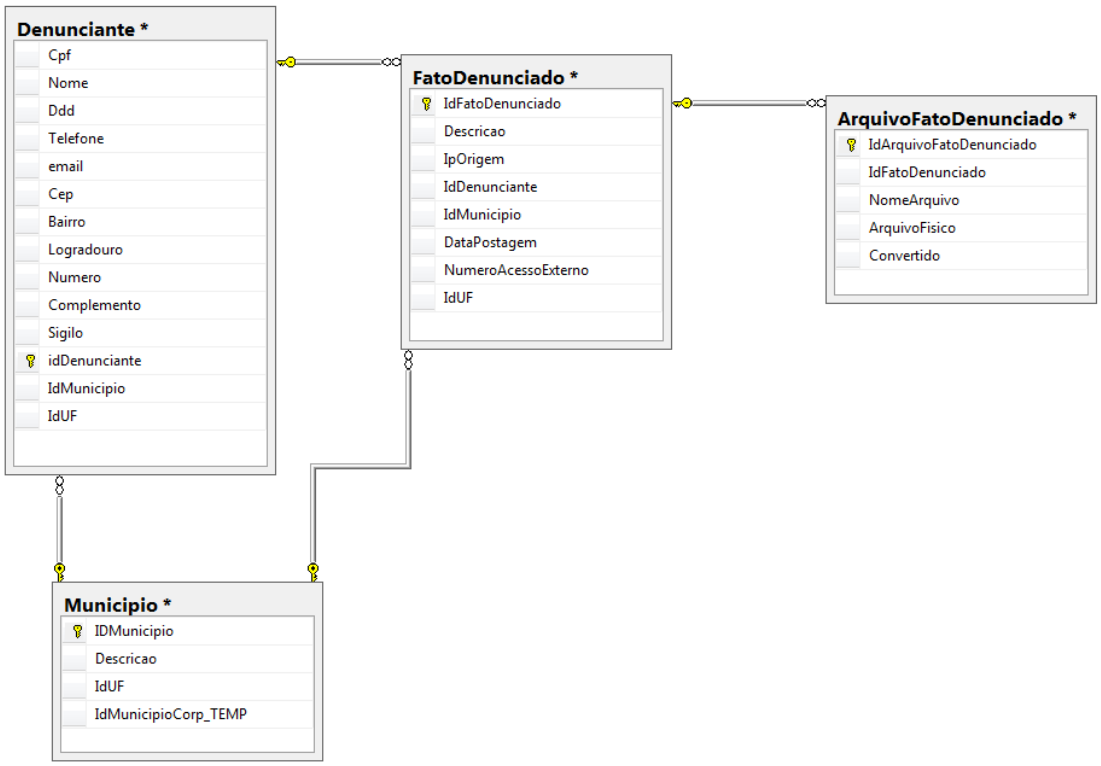

Figura 3.4: Modelo do Banco de Denúncias

de um formulário de denúncia. Isso porque existem outros tipos de manifestação, tais como formulário de ouvidoria, solicitação de registro, etc. Aquelas que forem consideradas lixo eletrônico serão inseridas na tabela LixoEletronico e não receberão um número de protocolo (NUP). Já aquelas que forem aceitas, receberão um número de protocolo e, depois de triadas é inserido um registro na tabela Denúncia, com um resumo a respeito do assunto da denúncia e um registro da tabela UnidadeDenuncia com o iddenuncia e o idunidade, indicando para qual unidade a denúncia será triada. Ao final deste processo, a denúncia estará na carga da unidade de destino. A Figura 3.5 exibe uma versão simplificada das tabelas do SGI utilizadas para o tratamento das denúncias.

As buscas de precedentes são feitas, geralmente, baseadas no nome do município, para tentar identificar se já existe denúncia relativa ao mesmo município que trate do assunto abordado ou simplesmente do assunto definido na denúncia. Um exemplo disso é a grande quantidade de denúncias relacionadas a um assunto recorrente na mídia. Assim, ao chegar uma nova denúncia relacionada a este fato, é feita uma busca na base de dados tentando encontrar possíveis denúncias idênticas ou muito parecidas.

Quando a denúncia se refere a alguns assuntos como convênios, Bolsa Família, dentre outros, o funcionário da triagem realiza uma pesquisa no Portal da Transparência tentando validar os dados apresentados. Por exemplo, se uma denúncia for relativa ao Bolsa Família

e dados como nomes ou CPF, eles devem ser verificados no Portal da Transparência.

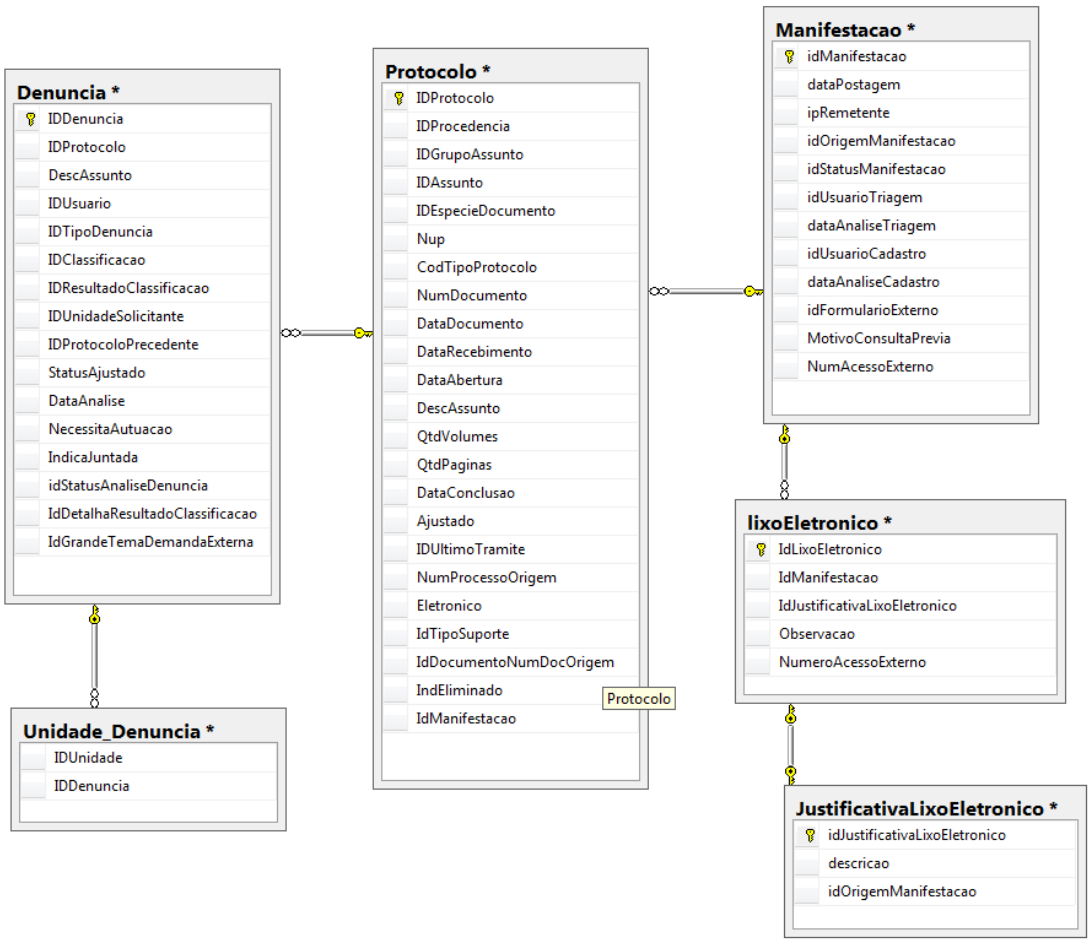

Figura 3.5: Modelo do SGI - Denúncias

O formulário de denúncias disponível no site possui um campo texto com 2048 caracteres, em formato livre. Como é possível verificar na Figura 3.6, além deste campo, o formulário possui outros campos mas nenhum deles é obrigatório. Possui também uma parte destinada a informações sobre os denunciados, podendo o denunciante informar dados como nome, cpf e órgão do denunciado. Outros dois campos permitem o cidadão marcar o município e a unidade da federação correspondente ao fato denunciado. A última opção é a referente à identificação do denunciado. A denúncia pode ser feita de forma anônima. No entanto, o denunciante pode informar os seus dados, como nome, cpf, email, telefone, etc. Essas informações facilitam entrar em contato como o denunciante caso seja necessário agilizar a apuração ou esclarecer algum fato, bem como informar ao mesmo o andamento da apuração da denúncia.

A ideia de realizar a denúncia em formato simplificado visa facilitar para o cidadão. No entanto, dado o problema da triagem aqui apresentado e a dificuldade de trabalhar

Figura 3.6: Formulário de Denúncias Disponível no Site da CGU

com textos livres, está sendo desenvolvido um novo banco de denúncias onde o formulário será mais interativo, sendo possível escolher algumas subáreas de acordo com o tipo da denúncia.

Os campos relativos a identificação da denúncia, como dados do denunciado, município e UF, poderão ser incluídos para auxiliar a classificação automática. Já os dados relativos ao denunciante, não serão escopo deste trabalho, tendo em vista que os mesmos não fazem parte da triagem.

Anexo às denúncias existem textos, fotos ou qualquer tipo de documento que possa fundamentar o fato relatado. Esses anexos têm como objetivo fundamentar o fato relatado e são usados para auxiliar a investigação. Os documentos também são cadastrados no SGI com a finalidade de auxiliar e enriquecer a fiscalização e apuração dos fatos relatados na denúncia. Esses anexos também não foram escopo do trabalho em questão.

O novo sistema irá contar com mais informações mas ainda conterá um campo de texto livre. Sendo assim, a triagem ainda precisará ser realizada.

Outra mudança no novo sistema é que os dados, possivelmente, não serão carregados mais no sistema SGI. Isso significa que a forma como a triagem é feita atualmente será afetada. No entanto, o propósito deste trabalho não será consideravelmente afetado pelas mudanças aqui descritas.

As novas informações existentes no novo sistema poderão ser acopladas ao modelo proposto caso os metadados possam contribuir na definição do modelo apresentado. En-

tretanto, para que este fato ocorra, é preciso primeiro que o novo sistema esteja em funcionamento e com um número considerável de denúncias triadas para que os algoritmos possam aprender com as mesmas.

3.3 Entendimento dos Dados

Em fevereiro de 2015 existiam 50.551 denúncias recebidas pelo formulário do Portal ou diretamente no Protocolo. Destas, 17.077 não continham dados ou informações que demandassem uma investigação e foram transformadas em lixo eletrônico. As outras 33.226 foram cadastradas como denúncia, recebendo um número de protocolo e uma entrada no sistema SGI. Considerando as denúncias recebidas, 29.398 já foram triadas, sendo que 5.844 foram habilitadas, transformadas em procedimento ordinário ou simplificado e 23.554 foram arquivadas. Dessa maneira, restam 248 denúncias aguardando para serem triadas. O quantitativo em estoque avaliado na data em que os dados foram extraídos do banco se deve ao fato de ter sido realizada uma força tarefa no final do ano, sendo enviadas todas as denúncias em estoque para serem triadas pelas Unidades Regionais da CGU.

A Tabela 3.1 exibe a distribuição das denúncias na base de dados em fevereiro de 2015.

Tabela 3.1: Distribuição das Denúncias na Base de Dados

Origem das Denúncias	Qtde Denúncias
Recebidas	50.551
Lixo Eletrônico	17.077
Viraram Denúncia Efetivamente	33.226
Denúncias Triadas	29.398
Arquivadas	23.554
Habilitadas	5.844
Esperando Triagem	248

Na Tabela 3.2 podem ser observadas denúncias enviadas para Lixo Eletrônico bem como o quantitativo por Justificativa, sendo que a maior quantidade de denúncias que são encaminhadas para Lixo Eletrônico não possuem dados suficientes para serem apuradas.

Conforme levantamento feito pela Ouvidoria da CGU (Figura 3.7), grande parte das denúncias enviadas à CGU são arquivadas. Em 2013, das 5.889 denúncias recebidas, 1.200 denúncias foram habilitadas e 4.889 foram inabilitadas e arquivadas. Portanto, 82% das denúncias triadas em 2013 foram desconsideradas.

A Tabela 3.3 demonstra os quantitativos de denúncias arquivadas pelos motivos anteriormente expostos ou triadas para alguma coordenação, mas considerando todas as

Tabela 3.2: Distribuição das Denúncias Consideradas Lixo Eletrônico

Justificativa	Qtde
Dados Insuficientes	9.898
Demanda Repetida	5.259
Mensagem sem Propósito	1.854
Palavras de Baixo Calão	44
Conteúdo Pornográfico	12
SPAM (Correntes, Propaganda)	10
Total	17.077

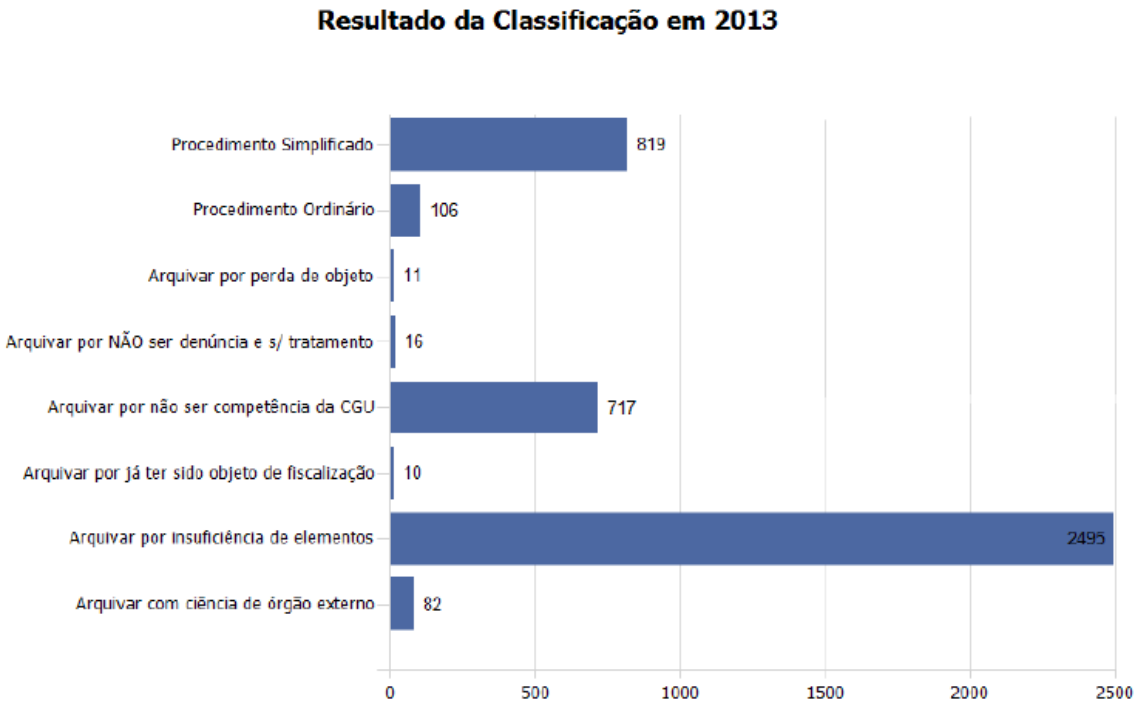

Figura 3.7: Distribuição das Denúncias Recebidas em 2013

denúncias existentes na base até fevereiro de 2015. Como é possível notar, das .5844 denúncias triadas, 780 foram convertidas como procedimento ordinário e 4.815 como procedimento simplificado. Além disso, a maior parte das denúncias arquivadas referem-se ao fato de as mesmas não possuírem dados suficientes para a realização de uma fiscalização.

Das denúncias que já foram triadas e não foram arquivadas, a maioria delas foi encaminhada para uma das coordenações da SFC. A Tabela 3.8 exibe a distribuição das quantidades de denúncias triadas por unidades da CGU.

O sistema de denúncia hoje existente na CGU guarda o histórico de toda a tramitação

Tabela 3.3: Distribuição das Denúncias Arquivadas

Motivo do Arquivamento	Qtde
Ciência de Órgão Externo	427
Insuficiência de Elementos	17.200
Objeto de Fiscalização Anterior	65
Não é Competência da CGU	4.046
Não é Denúncia/ Sem Tratamento	110
Perda de Objeto	191
Procedimento Ordinário	780
Procedimento Simplificado	4.815

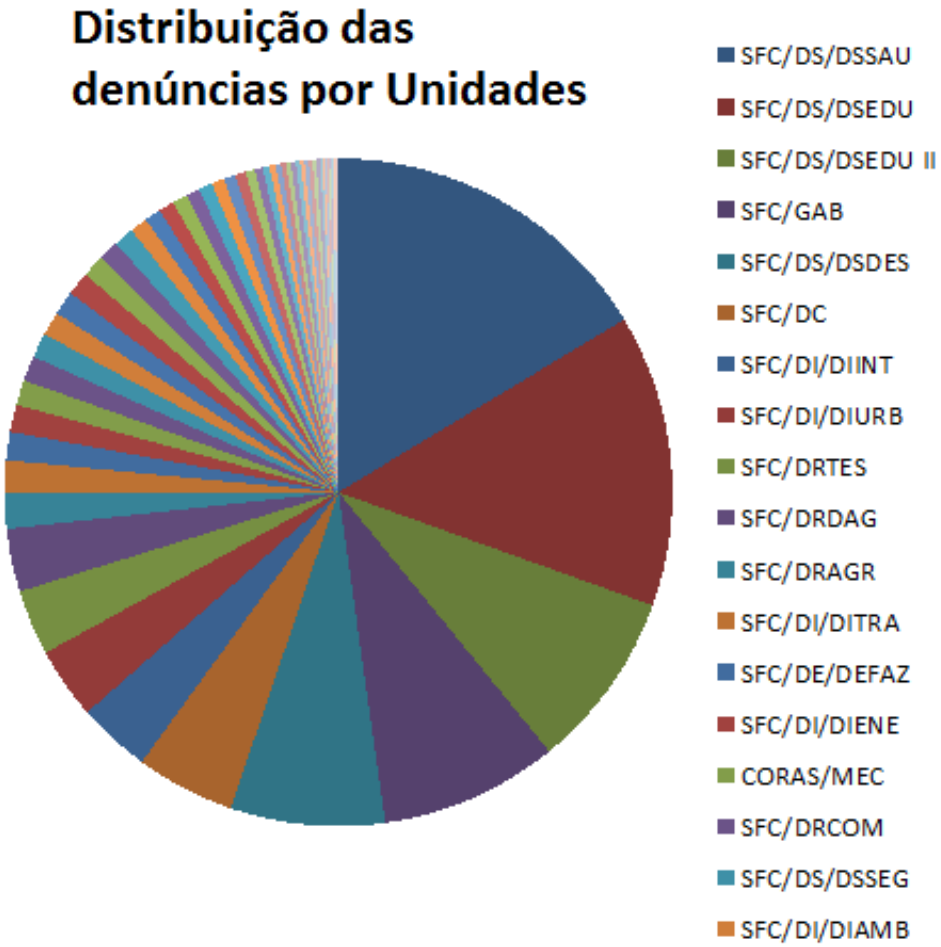

Figura 3.8: Distribuição de Denúncias Triadas por Coordenações

de uma denúncia, possibilitando identificar se a mesma foi encaminhada para mais de uma área, se foi arquivada, etc. Dessa forma, é possível identificar a quantidade de denúncias que foram enviadas corretamente para cada área, bem como a quantidade de denúncias

tramitadas de forma errada e que precisaram ser tramitadas novamente.

Avaliando-se os dados acima expostos, verificou-se que o índice de acerto da denúncia triada de forma manual é maior do que 90%. Este dados foram levantados através da quantidade de denúncias que foram retriadas para mais de uma área. Foram utilizados os dados da tabela UnidadeDenuncia, verificando-se as denúncias que possuem mais de uma Unidade cadastrada. Verificou-se também que o tempo em que a denúncia leva para começar a ser triada é bastante elevado, em virtude da falta de pessoal para a realização do processo. Algumas denúncias demoram meses para começarem a ser triadas.

A Figura 3.9 demonstra as quantidades de denúncias cadastradas por ano na ouvidoria. Esses dados, bem como todos os valores aqui utilizados foram extraídos da base em fevereiro de 2015. Razão pela qual a quantidade cadastrada no ano de 2015 é de somente 1.070 denúncias.

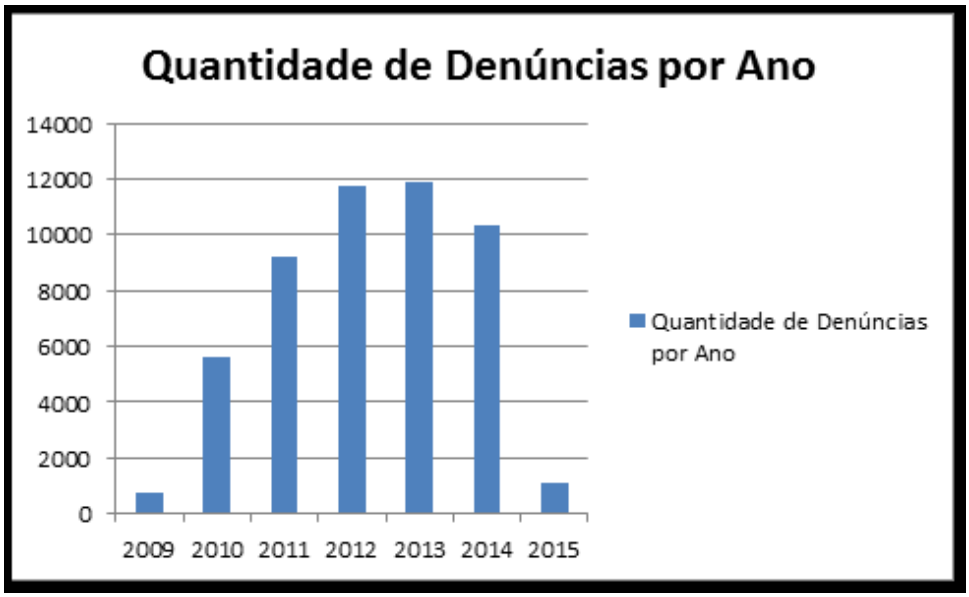

Figura 3.9: Distribuição das Denúncias Triadas

Outro fator a ser considerado, refere-se às denúncias anônimas. Na base de dados constam 33.443 denúncias anônimas. Caso o autor se identificasse, poderiam ser obtidas estatísticas relativas aos denunciantes. Essa identificação poderia também permitir o envio uma resposta do andamento do processo de apuração da denúncia ao autor. Como não é caso para a totalidade das denúncias, não se utilizou o campo autor ou outros campos do formulário de denúncias, na indução de classificadores nessa prova de conceito.

Além dos problemas acima levantados existem diversos tipos de textos cadastrados, desde aqueles bem escritos até texto com muitos problemas de digitação ou mesmo erros. Em algumas denúncias, verificou-se a ocorrência de erros graves de português ou digitação como no texto exposto na Figura 3.10.

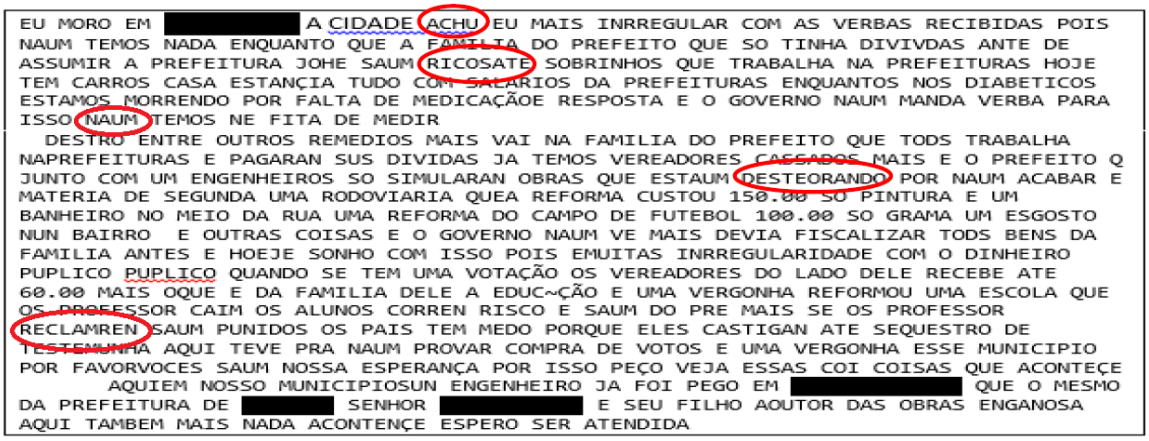
EU MORO EM [redacted] A CIDADE ACHU EU MAIS INRREGULAR COM AS VERBAS RECIBIDAS POIS NAUM TEMOS NADA ENQUANTO QUE A FAMILIA DO PREFEITO QUE SO TINHA DIVIVDAS ANTE DE ASSUMIR A PREFEITURA JOHE SAUM RICOSATE SOBRINHOS QUE TRABALHA NA PREFEITURAS HOJE TEM CARROS CASA ESTANÇIA TUDO COM SALARIOS DA PREFEITURAS ENQUANTOS NOS DIABETICOS ESTAMOS MORRENDO POR FALTA DE MEDICAÇÃOE RESPOSTA E O GOVERNO NAUM MANDA VERBA PARA ISSO NAUM TEMOS NE FITA DE MEDIR
DESTRO ENTRE OUTROS REMEDIOS MAIS VAI NA FAMILIA DO PREFEITO QUE TODS TRABALHA NAPREFEITURAS E PAGARAN SUS DIVIDAS JA TEMOS VEREADORES CAESADOS MAIS E O PREFEITO Q JUNTO COM UM ENGENHEIROS SO SIMULARAN OBRAS QUE ESTAUM DESTEORANDO POR NAUM ACABAR E MATERIA DE SEGUNDA UMA RODOVIARIA QUEA REFORMA CUSTOU 150.00 SO PINTURA E UM BANHEIRO NO MEIO DA RUA UMA REFORMA DO CAMPO DE FUTEBOL 100.00 SO GRAMA UM ESGOSTO NUN BAIRRO E OUTRAS COISAS E O GOVERNO NAUM VE MAIS DEVIA FISCALIZAR TODS BENS DA FAMILIA ANTES E HOEJE SONHO COM ISSO POIS EMUITAS INRREGULARIDADE COM O DINHEIRO PUPLICO PUPLICO QUANDO SE TEM UMA VOTAÇÃO OS VEREADORES DO LADO DELE RECEBE ATE 60.00 MAIS OQUE E DA FAMILIA DELE A EDUC~ÇÃO E UMA VERGONHA REFORMOU UMA ESCOLA QUE OS PROFESSOR CAIM OS ALUNOS CORREN RISCO E SAUM DO PRE MAIS SE OS PROFESSOR RECLAMREN SAUM PUNIDOS OS PAIS TEM MEDO PORQUE ELES CASTIGAN ATE SEQUESTRO DE TESTEMUNHA AQUI TEVE PRA NAUM PROVAR COMPRA DE VOTOS E UMA VERGONHA ESSE MUNICIPIO POR FAVORVOCES SAUM NOSSA ESPERANÇA POR ISSO PEÇO VEJA ESSAS COI COISAS QUE ACONTEÇE AQUIEM NOSSO MUNICIPIOSUN ENGENHEIRO JA FOI PEGO EM [redacted] QUE O MESMO DA PREFEITURA DE [redacted] SENHOR [redacted] E SEU FILHO AOUTOR DAS OBRAS ENGANOSA AQUI TAMBEM MAIS NADA ACONTENÇE ESPERO SER ATENDIDA

Figura 3.10: Exemplo de Texto de Denúncia

3.4 Pré-processamento dos Dados

Os dados relativos ao texto da denúncia e a unidade de destino foram extraídos do banco de dados no SQL Server e carregados no R. Os textos passaram por um processamento usando o pacote do R chamado *tm*. Este pacote possui várias funções pré-definidas para tratamento de textos, dentre as quais foram utilizadas:

- *removeNumbers* para remoção de números
- *removestripWhitespace* para remoção de espaços em branco
- *removePunctuation* para remoção de pontuação
- *contenttransformer(tolower)* para transformação das letras maiúsculas em minúsculas
- *removeWords, Stopwords("portuguese")* para remoção de *stopwords*
- *stemDocument* para redução das palavras aos seus radicais

Essas funções foram aplicadas com a finalidade de excluir palavras e caracteres indesejados ou que tem pouca importância para a identificação das classes. Além disso, a lista de *stopwords* utilizada no R foi alterada, incluindo palavras relativas ao contexto e que não agregam valor na classificação.

Aplicou-se também a função *removeSparseTerms* com o objetivo de reduzir o número de dimensões. A redução de dimensionalidade, desde que não impacte na eficácia do classificador, gera ganhos na medida em que melhora o desempenho computacional.

Foi realizado um tratamento também para a retirada de palavras que aparecem muito pouco no texto, pois as mesmas geram um aumento de dimensões e não contribuem para

classificação. Este processo foi realizado utilizando a função *removeSparseTerms*, primeiramente com um parâmetro de 0.99, gerando aproximadamente 6 mil termos e com um parâmetro de 0.96 gerando aproximadamente 1.200 termos. Os dois conjuntos foram testados e como o resultado dos classificadores foi o mesmo independente do tamanho conjunto, utilizamos o menor pois este possibilitou uma performance melhor da ferramenta.

3.5 Modelagem

Esta fase consiste na construção do modelo baseado nos algoritmos de classificação SVM, Random Forest, Naive Bayes e Árvore de Decisão (C4.5). Dessa forma, após efetuados os devidos tratamentos dos textos, os dados foram divididos em conjuntos de treinamento e teste. Os modelos foram testados utilizando algoritmos de classificação.

Após avaliar a viabilidade de modelo usando um número reduzido de unidades (apenas quatro classes), foi feita uma tentativa de aplicar os modelos de classificação a todas as 91 áreas. No entanto, os resultados alcançados não foram satisfatórios, já que a F-Measure e a Precisão ficaram em torno de 0.4. Realizou-se então um estudo mais aprofundado dos dados resultando no agrupamento das denúncias por diretorias, reduzindo o conjunto para 27 áreas. Este agrupamento resultou em uma F-Measure e uma Precisão ainda menor que a primeira. Outro problema encontrado na base foi que algumas classes possuem poucos registros, tendo em alguns casos somente uma denúncia, optou-se por limitar essa pesquisa à triagem de denúncias encaminhadas às unidades da CGU com mais de 30 denúncias já triadas.

O modelo proposto foi desenvolvido considerando 58 unidades de destino mais 6 unidades de arquivamento, resultando em um total de 64 classes. Os resultados dos testes efetuados demonstraram que o classificador induzido com o algoritmo SVM foi o que apresentou melhor resultado, alcançando uma precisão de apenas 0,591. Esse resultado foi considerado muito baixo. Por esse motivo, essa abordagem de indução de classificador com número muito grande de classes foi abandonada e se partiu para o estudo da abordagem de indução de um classificador baseado em árvore de Huffman onde se constrói uma árvore para cada classe considerada. O procedimento de classificação consiste em submeter o documento a classificar a cada uma das árvores de Huffman e escolher como classe do documento aquela correspondente à árvore cuja codificação do documento seja a mais curta possível. Por isso essa classificação foi chamada de CAH+MDL.

3.5.1 Experimentos CAH+MDL

O classificador CAH+MDL utilizado foi derivado a partir dos fontes em Java implementados por Ormonde [28] em sua pesquisa de mestrado na qual focou a classificação

automática de páginas Web utilizando dados semi-estruturados.

CAH+MDL é uma combinação da aplicação do princípio MDL e do classificador CAH, sendo que a árvore de Huffman foi adaptada para trabalhar com palavras em vez de caracteres. O CAH+MDL possui como vantagens a exigência de pouco recurso computacional para realizar a classificação de textos e permite treinamento incremental. Ormonde implementou um classificador CAH+MDL que pode trabalhar com hipótese de mundo fechado ou de mundo aberto. Na hipótese de mundo fechado, cada documento deve pertencer a pelo menos uma categoria. Na hipótese de mundo aberto, ao contrário da anterior, um documento não precisa pertencer a uma determinada categoria, sendo neste caso marcado como indefinido. O classificador foi desenvolvido em Java e implementada uma versão que também utiliza o SVM para comparação dos resultados.

O classificador CAH+MDL de Ormonde atuava da seguinte maneira: os dados eram lidos dos sites escolhidos e realizava-se um tratamento dos textos contidos nestes sites. Neste pré-processamento, são normalizadas as palavras para letras minúsculas, retirada a acentuação e os números, espaços em branco, e pontuação. Também são extraídas as *stopwords*.

No classificador proposto como prova de conceito para a CGU, é adotada a hipótese de mundo fechado. Os dados das denúncias são obtidos diretamente do banco de dados da CGU. A etapa de pré-processamento do programa de Ormonde não foi utilizada e passou-se a utilizar o tratamento de textos efetuado pela ferramenta R Studio, com as funções do pacote tm. A saída dessa fase de pré-processamento não gerou uma matriz termo documento única, sendo que foram criados arquivos individuais para cada texto de denúncia processado. Cada arquivo contém uma lista com os termos resultantes do pré-processamento seguidas da frequência em que eles aparecem na denúncia.

Estes arquivos são gravados em diretórios com o nome da classe a que o mesmo pertence. Em razão dos textos de denúncias terem sido digitados em um campo com tamanho máximo de 2.048 caracteres, além de muitos desses caracteres se referem a palavras consideradas como *stopwords*, os arquivos gerados a partir deste processamento são, em geral, pequenos.

Em algumas denúncias os textos são bem pequenos, sem conter muita informação relevante. Nesses casos, após o tratamento dos textos, resultavam poucos ou nenhum termo. Portanto, só foram considerados arquivos que possuíssem pelo menos cinco termos ao final desse processo.

Após essa primeira etapa, o classificador inicia chamando a classe Java *cria categorias*, informando o nome e número correspondente a cada uma das categorias existentes. Essa classe irá definir as categorias que poderão ser utilizadas pelo classificador CAH+MDL. Depois da criação de todas as categorias, a fase seguinte é o treinamento das mesmas.

Nesta fase, o classificador lê cada arquivo com os termos mais frequentes, aprende a categoria daquele arquivo e grava a saída em um arquivo .vsm. Este arquivo será único para todas as categorias. Para realizar este processo, o caminho da pasta contendo os arquivos de cada categoria e o arquivo vsm, são passados como parâmetro para o programa. Assim, são lidos todos os arquivos separados para treinamento por diretório e cada um irá gerar uma linha no arquivo vsm. Este arquivo só contém valores numéricos. Cada termo é transformado em um número que irá corresponder à dimensão referente ao termo na lista de todos os termos obtidos a partir de todas as denúncias. As dimensões começam de 0 e são ordenadas de forma crescente. A linha conterá a categoria a que pertence a denúncia, seguida da dimensão e de um valor que representa a quantidade de vezes que aquela dimensão aparece no documento. A Figura 3.11 ilustra linhas que compõem o arquivo .vsm. O número 1072, presente em todas as linhas, refere-se à categoria da denúncia, e é seguido por pares de números separados por dois pontos. O valor antes do dois pontos indica a dimensão (isto é, o termo) e o valor depois dos dois pontos refere-se à quantidade de ocorrências desse termo na denúncia representada por uma linha nesse arquivo.

```
1 1072 0:1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 10:1 11:1 12:1 13:1 14:1 15:1 16:1 17:1 18:1 19:1 20:1 21:1 22:1
2 1072 4:1 5:1 8:1 9:1 10:1 13:2 17:1 30:2 35:1 37:1 38:1 39:1 40:1 41:1 42:1 43:1 44:1 45:1 46:1 47:1 48:1 49:1
3 1072 3:1 4:1 8:1 9:1 12:1 17:1 18:1 19:1 30:1 31:1 34:2 35:1 52:1 54:1 121:1 133:2 138:1 150:1 152:1 153:1 154:1
4 1072 8:1 12:1 18:1 19:1 30:2 35:2 67:1 85:1 89:1 116:1 118:1 138:1 166:1 178:1 182:2 192:1 193:1 194:1 195:1
5 1072 8:1 19:1 27:1 29:1 35:2 71:1 74:1 116:1 138:1 167:2 178:1 182:1 220:1 228:1 229:1 230:1 231:1 232:1 233:1
6 1072 8:1 35:1 138:1 145:1 148:1 169:1 178:1 181:1 182:1 220:1 227:2 242:1 256:1 257:1 258:1 259:1 260:1 261:1
7 1072 8:1 19:1 35:1 41:1 51:1 72:1 116:1 138:1 150:2 181:1 220:1 228:1 254:2 286:1 287:1 288:1 289:1 290:1 291:1
8 1072 7:1 8:1 19:1 34:2 35:1 118:1 150:1 175:1 180:1 194:1 236:1 308:1 309:1 310:1 311:1 312:1 313:1 314:1 315:1
9 1072 8:1 19:1 35:1 51:1 54:1 147:1 149:1 150:1 156:1 159:1 175:1 180:1 289:1 310:1 330:1 331:1 332:1 333:1 334:1
10 1072 3:1 8:1 9:1 12:1 19:1 24:1 31:1 35:1 48:1 51:1 54:1 98:1 150:1 159:1 180:1 328:2 336:1 337:1 338:1 339:1
```

Figura 3.11: Exemplo do Arquivo VSM Gerado

Após criar o arquivo .vsm, é necessário induzir a árvore de Huffman. Serão criadas árvores de Huffman considerando as frequências dos termos nas denúncias para cada uma das categorias. No exemplo citado, todas as linhas apresentadas na Figura 3.11 serão utilizadas para a montagem da árvore de Huffman que representa a classe cujo identificador é 1072. Depois disso, o MDL irá escolher o modelo que melhor representa cada classe e guardar essas informações em um arquivo com terminação mdl.

O passo seguinte é o teste desses dados. Originalmente o algoritmo utilizava as mesmas categorias de entrada do treinamento para testar. Ele recebia o mesmo arquivo de treinamento mas ignorava a categoria da linha especificada. Essa categoria era utilizada somente para verificar o número de acertos do classificador. Nesta fase, cada categoria irá receber uma variável de contagem de bits começando com 0 e sendo aumentada de acordo com a presença ou não das dimensões em categorias. O objetivo é contar o peso de todas as dimensões encontradas em cada categoria para encontrar o modelo em que a contagem de bits mais se assemelhe à contagem dos dados de teste. O classificador lê cada uma das dimensões e atribui um peso a essas dimensões de acordo com a presença ou

não delas em cada classe. Aqui ele não sabe a classe que irá incluir, por isso ele verifica a dimensão em todas as categorias. Caso a dimensão não exista em determinada categoria, esta categoria receberá o valor 79 como peso. Este é o maior valor que pode ser atribuído a uma dimensão, porque quando uma palavra não é encontrada em determinada classe, aumenta a probabilidade desta classe não ser a classe destino. E como este algoritmo trabalha com árvores de compressão, quanto maior a árvore, menor a probabilidade do documento pertencer àquela categoria.

No final, o classificador tem uma lista com a contagem de bits total, considerando todas as dimensões do arquivos de entrada. Contém também uma contagem de bits para cada uma das categorias formadas pela presença ou não da dimensão naquela categoria. A classificação ocorrerá na classe ou classes que tiverem a contagem mais parecida com a contagem da linha do arquivo de entrada.

Para realizar o treinamento e o teste acima especificados com o banco de denúncias, foram selecionadas 58 categorias descritas anteriormente mais 6 categorias de arquivamento. Os textos dessas áreas foram carregados no R e foi realizado o pré-processamento de textos de cada um deles. O resultado foi gravado em arquivos individuais, separados em pastas de acordo com a categoria das unidades. Os dados foram treinados utilizando-se 70% das denúncias escolhidas aleatoriamente. O restante, 30% das denúncias, foram separadas para testar a efetividade do algoritmo. O melhor desempenho foi obtido com o classificador CAH+MDL que alcançou precisão de 0,41.

Com o intuito de melhorar os índices encontrados, foram aplicadas algumas técnicas adicionais no pré-processamento como a utilização de funções para reduzir as palavras ao seu radical e alteração da lista de *stopwords* utilizadas como descrito a seguir.

As métricas TF-IDF e TF trabalham utilizando parâmetros como a quantidade de palavras de um documento em relação a todos os documentos do conjunto, o tamanho variável dos textos em cada documento, dentre outras características. O textos de denúncia têm um tamanho de no máximo 2.048 caracteres. Além disso, aqui os arquivos de saída são gerados separadamente para cada texto, não sendo criada uma matriz esparsa da mesma forma que geralmente ocorre ao final do processamento de textos. Portanto, tiveram que ser utilizadas outras opções para tratar o fato de algumas palavras estarem presentes na maioria dos documentos de todas as classes, agregando pouco ou nenhum valor na classificação.

Para resolver o problema apresentado acima, foi realizado, no R Studio, um pré-processamento de textos considerando todo o conjunto de dados, ou seja, foram selecionados todos os textos de todas as categorias, inclusive das denúncias arquivadas. Foi criada uma matriz termo documento DTM com as palavras resultantes do pré-processamento. Na criação desta matriz foi utilizada a função *removeSparseTerms* com o parâmetros de

0.98. Como uma matriz esparsa é preenchida com 0 na maioria das suas posições, a DTM inicial possuía 165.203 termos. Essa quantidade de termos é desnecessária para o objetivo desta lista além de dificultar o processamento da mesma. Como a ideia aqui é encontrar as palavras mais frequentes, usamos essa função para eliminar posições menos significativas e o resultado foram 520 termos. O mesmo processo foi executado com o *removeSparseTerms* de 0.993 e foram retornados 1.350 termos. Ao final, verificou-se que para encontrar as palavras mais frequentes, presentes na maioria dos documentos, poderia ser utilizada o parâmetros de 0,993, com 1.350 termos. Para geração desses arquivos não foi utilizada nenhuma função de *stemming* ou lematização porque estava-se buscando verificar como as palavras foram cadastradas no formulário.

Baseado neste termos, foi gerada uma lista com as palavras mais frequentes. Observou-se que as palavras que estavam no topo da lista eram, normalmente, relacionadas a todas as classes, como as palavras denúncia, irregularidade, desvio, município, união, recursos, etc. Essas palavras foram acrescentadas à lista de *stopwords* para serem excluídas dos arquivos, evitando-se assim que os documentos contivessem palavras que não agregariam valor à classificação.

Para certificar se o problema anterior ocorre ou não na maioria das denúncias ou em uma quantidade pequena do conjunto, foi executado novamente o processo descrito acima mas utilizando o parâmetro de 0,998 da função removesparseterms. Neste caso, o objetivo era avaliar se ocorrem muitos erros de português nas palavras mais frequentes, e, avaliar assim, a necessidade do uso de um corretor ortográfico. Foi possível verificar que a maioria das palavras avaliadas não apresenta erro grave de português ou digitação. Portanto, pode-se concluir que o caso apresentado não é representativo e optou-se pela não utilização de um corretor ortográfico.

Outro fator avaliado e que dificultava a comparação das palavras era o fato de alguns textos apresentarem problemas de acentuação. A solução encontrada para tratar esses textos foi a remoção de acentos. Mas, ao aplicar a remoção de acentos no texto, algumas palavras como *não*, apesar de constarem na lista de *stopwords*, não eram retiradas, porque o R comparava a palavra nao sem acento com a palavra não acentuada presente na lista de *stopwords*. Para padronizar os resultados, foram então removidos acentos de todos os textos processados e também da lista de *stopwords*.

Após a aplicação de todas as alterações acima descritas (isto é, a criação de uma nova lista de *stopwords* e a aplicação de processo de *stemming* utilizando a função *stemDocument* do R studio), os dados foram novamente submetidos ao classificador CAH+MDL. O resultado foi uma precisão de 0,54, ou seja, o pré-processamento aplicado gerou uma melhora considerável.

Outro teste realizado foi a aplicação de bigramas. Ao invés de criar listas de termos

únicos, criou-se uma lista com os bigramas encontrados. No entanto, dado o tamanho reduzido do texto da denúncia, os arquivos gerados ao final do pré-processamento continham poucos termos, além de muitos deles não terem significado. Optou-se por utilizar unigramas ao invés de bigramas ou trigramas, tendo em vista os motivos expostos anteriormente.

Os testes até aqui realizados no classificador CAH+MDL, haviam contemplado somente as categorias triadas, totalizando 5.844 denúncias. No entanto, além das denúncias triadas, existe uma grande quantidade de denúncias arquivadas (mais de 20 mil denúncias). Essas denúncias arquivadas são divididas em 6 categorias, de acordo com os motivos do arquivamento. Foram então incluídas as denúncias arquivadas juntamente com as triadas, aumentando de 58 para 64 áreas de destino ou categorias possíveis. O total de denúncias utilizadas aqui ficou em torno de 30 mil. Todavia, ao incluir essas denúncias arquivadas, a performance dos classificadores piorou, encontrando uma precisão de 0,38 com o CAH+MDL.

Uma nova tentativa de resolver o problema então foi a classificação multi-label. O algoritmo original utilizava essa classificação com hipótese de mundo aberto e com várias categorias. Neste trabalho foi utilizada somente hipótese de mundo fechado (os documentos devem necessariamente pertencer a uma das 64 categorias) e classificação multi-label baseada em três categorias apenas.

O classificador anterior treinava a base informando se o documento pertencia a uma ou mais categorias e quais eram essas categorias. Na base do sistema SGI não existe qualquer critério que pode informar se o classificador pertence a mais de uma categoria, não sendo possível treinar os classificadores com esse tipo de informação. Portanto, a classificação multi-label utilizada originalmente no CAH+MDL foi alterada para classificação por ranking de probabilidades, informando ao usuário as três possíveis categorias em que o documento teria mais chances de pertencer. Para o usuário, ao invés de o mesmo escolher entre as 64 categorias disponíveis, ele terá as três mais prováveis para avaliar, resultando em minimização do trabalho e ganho de tempo.

Sendo assim, a primeira categoria em que o algoritmo irá classificar um documento será aquela em que ele encontrar a árvore com quantidade de bits mais próxima do resultado retornado. A segunda categoria será a segunda árvore com resultado mais próximo da primeira categoria, e assim sucessivamente. Esse tipo de classificação minimiza o erro retornado pelo classificador já que este terá mais chances de acertar.

Como o algoritmo original já calcula a quantidade de bits gerada para cada categoria e depois compara com as árvores originais, esses cálculos foram utilizados para identificar quais seriam as pontuações mais próximas da categoria. Desse modo, para classificar utilizando multi-label, foram selecionadas as três categorias com a quantidade de bits

mais próximos da primeira.

Resultado da classificação multi-label para o CAH+MDL está apresentado na Tabela 3.4.

Tabela 3.4: Precisão do CAH+MDL Utilizando Multi-label

Categorias	Precisão
1 Categoria	0,554
2 Categorias	0,778
3 Categorias	0,842

Além de aumentar as chances de acerto quando se escolhe mais de uma categoria, observa-se que em muitos casos o classificador erra a primeira categoria que seria uma unidade de arquivamento e acerta na segunda quando escolhe uma unidade da SFC e vice versa. Isso pode ocorrer, por exemplo, quando uma denúncia que seria arquivada por já ter sido objeto de fiscalização, se encaixa na classe que seria apurada caso ainda não tivesse sido objeto de fiscalização. Outro fator que contribui para o uso de classificação multi-label é o fato de algumas denúncias envolverem mais de uma área de apuração na CGU. Também ocorre de um mesmo tema aparecer em uma denúncia relativa à SFC e à CRG. Por exemplo, uma denúncia envolvendo a área de educação poderia estar relacionada à SFC se envolver algum órgão ou programa de governo. No entanto, se envolver servidor público, a mesma deve ser encaminhada à CRG. Com a classificação multi-label, o classificador tem chances de retornar as duas unidades (SFC e CRG), e seria mais fácil um servidor da CGU validar para qual ou quais unidades a denúncia deve ser enviada.

Um outro problema encontrado foi a quantidade de denúncias repetidas existentes na base. Em alguns casos o mesmo cidadão informa os fatos denunciados mais de uma vez, inclusive com o texto idêntico. Em outros casos pessoas diferentes denunciam o mesmo fato. O texto pode ser um pouco diferente na forma de apresentação mas se referem aos mesmos fatos. Para tratar esses casos foi utilizada a técnica de similaridade baseada no cosseno do ângulo entre as denúncias representadas segundo o modelo VSM. A mesma identifica denúncias repetidas ou denúncias parecidas que versem sobre o mesmo assunto. Desse modo, quando as denúncias eram idênticas, o ângulo de similaridades dos cossenos calculados retornou igual a 1. Nos casos das denúncias parecidas o mesmo também foi eficiente, retornando próximo de 1. Essa técnica passou a ser aplicada antes do uso dos classificadores, eliminando as denúncias idênticas ou referentes ao mesmo assunto da triagem. Este tipo de consulta é constantemente feita pela CGCID buscando identificar precedentes, isto é, denúncias relacionadas ao mesmo fato ou mesmo denúncias idênticas.

Apesar dessa denúncia idêntica ou similar não precisar ser triada novamente, a mesma irá constar na base para outras finalidades, como avaliação das denúncias mais recorrentes,

avaliação de possível spam, entre outros. Outra aplicação do uso da similaridade pode ser para avaliar o peso do que deve ser fiscalizado primariamente. Considerando o quantitativo de denúncias que se repetem, esse fato pode ser um indicador relevante na decisão de quais denúncias devem ser apuradas em ordem de prioridade.

As verificações de dados da denúncia utilizando outras bases de dados ou o Portal da Transparência não foram aqui tratadas em um primeiro momento. Como a busca de dados relativos a um convênio denunciado ou aos dados relacionados a uma pessoa física que receba recursos de programas de governo como Bolsa Família.

Capítulo 4

Resultados

O modelo proposto foi desenvolvido considerando 58 unidades de destino mais 6 unidades de arquivamento, resultando em um total de 64 classes.

Os resultados alcançados foram validados de acordo com algumas das métricas consagradas pela literatura. A Tabela 4.1 ilustra os valores da precisão encontrados durante o decorrer do trabalho. Os primeiros conjuntos de dados trabalhados obtiveram uma precisão de 0,41. Após este resultado, o pré-processamento foi alterado, incluindo novas palavras na lista *Stopwords*, aplicando-se a técnica de *Stemming* e o resultado atingiu uma melhora de mais de 20%. Por outro lado, ao incluir as categorias de arquivamentos, constatou-se uma diminuição na precisão, chegando a mesma a 0,38. Por fim, percebe-se que os melhores resultados foram alcançados utilizando-se o algoritmo CAH+MDL e a classificação multi-label, com uma precisão de 0,84.

Tabela 4.1: Comparativo dos Resultados Alcançados

Classificador	Precisão
CAH+MDL com StopWords Padrão (58 classes)	0,41
CAH+MDL com StopWords e Stemming (58 classes)	0,55
CAH+MDL com StopWords e Stemming (64 classes)	0,38
CAH+MDL com StopWords e Stemming Multi-label (64 classes)	0,84

Apesar da melhora alcançada no modelo proposto, os índices de acerto da triagem manual não foram superados. Isso porque, ao avaliar a quantidade de reencaminhamentos de denúncias triadas para uma nova área, percebe-se que esta taxa de erro fica próxima de 10%. Considerando-se a taxa de erro do classificador CAH+MDL com classificação multi-label em torno de 16%, concluímos que este apresentou um resultado pouco abaixo da triagem manual se comprarmos somente as taxas de acerto, conforme ilustrado na Tabela 4.2.

Tabela 4.2: Comparativo entre a Triagem Manual e Automática

Processos	Taxa de Acerto
Triagem Manual	0,90
CAH+MDL Multi-label	0,84

Não obstante, alguns fatores devem ser levados em consideração para ponderação dos resultados. Um deles é o tempo decorrido entre o cadastramento da denúncia pelo cidadão e o começo da triagem. Ao receber uma nova denúncia, o sistema grava automaticamente a data da postagem. Quando um usuário responsável pela triagem começa a trabalhar na mesma, é incluída uma data de triagem. Comparando-se essas duas datas, averiguou-se que uma grande quantidade de denúncias demoravam meses, chegando a alcançar mais de um ano entre o cadastramento e o começo da triagem.

O tempo da triagem também deve ser ponderado. Esse tempo é calculado comparando-se a data de começo da triagem e a data em que a denúncia foi encaminhada para a unidade de destino. O tempo médio de triagem é de 4 horas.

Portanto, os indicadores anteriormente descritos medem o tempo que uma denúncia demorou a ser triada e o tempo que a mesma demorou para começar a ser triada. Esses valores permitem comparar o tempo resultante da triagem manual com o tempo da triagem automatizada.

O tempo da triagem automatizada, uma vez realizado o treinamento e aplicação dos modelos, será instantâneo. Consequentemente, a despeito de a triagem manual apresentar um índice de acerto um pouco melhor que a triagem automatizada, esta é capaz de agilizar o processo. Assim, tão logo seja cadastrada uma nova denúncia, ela será triada imediatamente.

A busca de precedentes também tem que ser levada em consideração na comparação entre o modelo automático e o manual. Atualmente, esta busca é executada de forma manual, tentando encontrar palavras chave no texto de denúncias anteriores ou pesquisando por denúncia com textos parecidos. A implantação da técnica VSM mostrou bons resultados quando aplicada a este contexto, retornando de forma quase imediata as denúncias similares ou idênticas.

4.1 Implantação

Até o momento não foi desenvolvido um sistema capaz de fazer a interface entre o usuário e a triagem automática. Conforme conversado com a CGCID, a ideia é acoplar a técnica aqui desenvolvida ao novo banco de denúncias. No entanto, como a princípio não irá ocorrer a carga dos dados do sistema atual para o novo banco de denúncias, ao

ser implementada, esta solução deverá acessar as duas bases de dados. Isso porque, ao realizar buscas para encontrar precedentes, deve-se efetuar uma busca nos dois sistemas, possibilitando-se assim um retorno de todas as denúncias que envolvam assuntos similares ao procurado. No caso de treinamento do classificador, poderá ser efetuado uma única vez para aprendizado utilizando os dados do sistema atual. E, cada vez que for cadastrada uma nova denúncia, a mesma será comparada com as árvores geradas pelo classificador no treinamento, com a finalidade de encontrar as unidade de destino.

Ao exibir os resultados para o usuário final, estes devem ser retornados mostrando as três principais categorias que foram identificadas pelo classificador. Na implementação da busca por precedentes ou denúncias idênticas, deve ser retornado o texto das denúncias encontradas bem como o NUP, já que todo o trâmite da denúncia no SGI é realizado através do NUP. Este NUP permitirá rastrear e agrupar denúncias para tratamento em conjunto, além de possibilitar indicar o NUP da denúncia idêntica, no caso de arquivamento da mais recente.

Capítulo 5

Conclusões e Trabalhos Futuros

O trabalho apresentado realizou estudo e aplicação das técnicas de mineração de textos com o objetivo de criar um modelo de classificação automática de triagem de denúncias na CGU. A prova de conceito aqui efetuada demonstrou a viabilidade de aplicação desta solução, evidenciando que a triagem de denúncias pode ser semi-automatizada sem perda de qualidade.

Primeiramente foram realizados levantamento dos dados e entendimento do processo de triagem de denúncias. Após esta fase, foram executadas as etapas de pré-processamento e limpeza dos dados, com o objetivo de transformar os textos em uma representação capaz de ser utilizada pelos algoritmos de aprendizagem de máquina. Foram utilizados algoritmos conhecidos na literatura para classificação de textos (random forest, decision tree, naive bayes e svm). Entretanto nenhum deles alcançou resultados satisfatórios. Optou-se por utilizar o classificador CAH+MDL baseado na Árvore de Huffman. Os primeiros testes realizados com este classificador apresentaram performance ainda pior que os classificadores anteriores. No entanto, ao implementar técnicas de classificação multi-label utilizando o CAH+MDL, os índices de acerto aumentaram consideravelmente, chegando a uma precisão de 0.84.

Cabe ressaltar que durante o desenvolvimento deste trabalho o formulário de denúncias existente no site migrou para o e-ouv, sistema que visa interligar as ouvidorias da Administração Pública. Os campos continuaram sendo os mesmos, apenas com a diferença do órgão de destino da denúncia que deve ser preenchido antes do campo texto. Se anteriormente denúncias de vários órgãos eram enviadas à CGU, agora as mesmas são encaminhadas ao real destinatário do assunto a ser apurado. Essa mudança foi implementada nos últimos dois meses e reduziu em 30% a quantidade de denúncias enviadas à CGU. Uma consequência lógica dessa alteração, provavelmente, será a diminuição das denúncias arquivadas com ciência de órgão externo. No entanto, não se tem ainda o quantitativo de denúncias triadas o suficiente para constatar essa afirmação.

Ponderando os resultados expostos, podemos concluir que o classificador automático é uma solução plausível para o problema de triagem de denúncias. O CAH+MDL, utilizando classificação multi-label, apresentou índices consideráveis de acerto além de reduzir o tempo da triagem. Desse modo, com a utilização do método aqui descrito é possível reduzir o tempo gasto na triagem manual. Outro benefício a ser gerado na utilização dessa solução é a redução do tempo entre o cadastramento da denúncia pelo cidadão e o começo da triagem. Este fator dará celeridade ao processo, resultando em benefícios para a CGU e para a sociedade, na medida em que proporciona melhor aplicação dos gastos públicos.

A agilidade na busca de precedentes também é um fator determinante na implementação do modelo proposto, visto que possibilita encontrar denúncias parecidas e evitar retrabalho no caso de denúncias idênticas.

5.1 Trabalhos Futuros

Como trabalhos futuros pretende-se abordar os seguintes temas:

- Resumo automático das denúncias
- Inclusão dos metadados na classificação
- Utilização de entidades nomeadas para auxiliar a busca em outros sistemas
- A classificação dos dados destinados ao lixo eletrônico

Ao final da triagem, o servidor responsável redige um pequeno resumo da denúncia informando o tema a que ela se refere e os principais tópicos abordados. Apesar do trabalho aqui desenvolvido ter realizado a classificação automática de acordo com a área de destino, o servidor ainda precisará ler a denúncia e criar um resumo da mesma. A ideia seria então criar este texto automaticamente através de recursos de mineração de textos.

Os campos relativos a identificação do denunciante, os dados de municípios e órgãos, bem como os textos anexos à denúncia podem auxiliar na triagem ou mesmo na apuração da denúncia. Pretende-se desenvolver trabalhos futuros com o objetivo de explorar melhor estes campos e estes anexos, seja para triagem ou seja durante o trabalho de apuração da denúncia.

Ao observar as denúncias analisadas, é possível constatar que muitas possíveis informações são relacionadas a pessoas, órgãos, localidades, etc. O uso de técnicas de entidades nomeadas poderia ajudar a extrair informações relevantes dentro dos textos e comparar essas informações com outras bases de dados da CGU. Esse cruzamento de dados pode agilizar o processo de apuração da denúncia.

Os dados relativos às denúncias que não são aceitas, ou seja, o lixo eletrônico, não fizeram parte deste trabalho pois essa separação é realizada pela unidade responsável pelo protocolo. No entanto, em uma segunda etapa, pretende-se estender a classificação para a primeira etapa de aceitação ou não da denúncia, objetivando classificar automaticamente as denúncias em aceitas ou em lixo eletrônico

Referências

[1] Mohamed Aly. Survey on multiclass classification methods. *Neural Netw*, pages 1–9, 2005. 5

[2] Azevedo, Ana Isabel Rojão Lourenço. KDD, SEMMA and CRISP-DM: a parallel overview. 2008. 12

[3] Michael W. Berry and Malu Castellanos. *Survey of text mining II*. Springer, 2008. 6

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly, Beijing; Cambridge [Mass.], 2009. 5

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 7

[6] Leo Breiman. Random forests. Technical report, Technical Report 567, Department of Statistics, UC Berkeley, 1999. 31, 2001. 11939. 9

[7] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996. 11

[8] Ricardo Cerri, André Carlos PLF de Carvalho, and Alex A. Freitas. Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification. *Intelligent Data Analysis*, 15(6):861–887, 2011. 5

[9] Mengen Chen, Xiaoming Jin, and Dou Shen. Short text classification improved by learning multi-granularity topics. In *IJCAI*, pages 1776–1781. Citeseer, 2011. 15

[10] Hong Cheng, Xifeng Yan, Jiawei Han, and Philip S. Yu. Direct discriminative pattern mining for effective classification. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 169–178. IEEE, 2008. 6

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 7

[12] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. 2002. 4, 10

[13] Federal Internal Control Secretariat. Audit and Inspection, 2014. Available: http://cgu.gov.br/assuntos/auditoria-e-fiscalizacao/. 21

[14] Yoni Halpern, Steven Horng, Larry A. Nathanson, Nathan I. Shapiro, and David Sontag. A comparison of dimensionality reduction techniques for unstructured clinical text. In *ICML 2012 Workshop on Clinical Data Analysis*, 2012. 6

[15] Qiwei He. *Text Mining and IRT for Psychiatric and Psychological Assessment*. University of Twente [Host], 2013. 15

[16] Djoerd Hiemstra. A probabilistic justification for using tf$\times$ idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139, 2000. 6

[17] David A Huffman. Method for the Construction of Minimum-Redundancy Codes. *PROCEEDINGS OF THE I.R.E.*, pages 1098–1101, 1952. 9

[18] M. Ikonomakis, S. Kotsiantis, and V. Tampakas. Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8):966—-974, 2005. 6

[19] Daniel Jurasfsky and James H Martin. *Speech and Language Processing*. Stuart Russell and Peter Norvig, 1998. 4, 5

[20] Vandana Korde. Text Classification and Classifiers:A Survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85–99, March 2012. 5

[21] Eugene F. Krause. *Taxicab Geometry: an adventure in non-Euclidean geometry*. Dover Publications, New York, 1987. 10

[22] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. 1998. 8

[23] Patrícia Maia, Rommel N. Carvalho, Marcelo Ladeira, Henrique Rocha, and Gilson Mendes. Application of text mining techniques for classification of documents: a study of automation of complaints screening in a Brazilian Federal Agency. 18

[24] Claudia Aparecida Martins, Maria Carolina Monard, and Edson Takashi Matsubara. Reducing the dimensionality of bag-of-words text representation used by learning algorithms. In *Proceedings of The Third IASTED International Conference on Artificial Intelligence and Applications (AIA 2003), Benalmádena, Espanha.(to be published)*, volume 38, 2003. 6

[25] Michael Gordon and Manfred Kochen. Recall-precision trade-off: A derivation, 1988. 11

[26] National Disciplinary Board. Disciplinary Action, 2014. Available: http://cgu.gov.br/assuntos/atividade-disciplinar/. 21

[27] Ombudsman's Office. Ombudsman's Office, 2014. Available: http://cgu.gov.br/assuntos/ouvidoria/. 18

[28] Rodrigo de La Rocque Ormonde. Classificação automática de páginas web multi-label via MDL e support vector machines. 2009. 14, 32

[29] Terry Padgett, Angelo Maniquis, Mark Hoffman, Will Miller, and Jennifer Lautenschlager. A semantic visualization tool for knowledge discovery and exploration in a collaborative environment. In *International Conference on Intelligence Analysis: The Office of the Assistant Director of Central Intelligence*, 2005. 5

[30] Juan Ramos. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003. 7

[31] Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee. The use of bigrams to enhance text categorization. *Information processing & management*, 38(4):529–546, 2002. 5

[32] Liu Tie-Yan, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter - Natural language processing and text mining*, Volume 7(Issue 1):36–43, June 2005. 15

[33] Transparency and Corruption Prevention Secretariat. Preventing Corruption, 2014. Available: http://cgu.gov.br/assuntos/. 18

[34] Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182. ACM, 2010. 14

[35] Hans Friedrich Witschel. Using decision trees and text mining techniques for extending taxonomies. In *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by Using Machine Learning Methods*, 2005. 8

[36] I. H Witten, Eibe Frank, and Mark A Hall. *Data mining practical machine learning tools and techniques*. Morgan Kaufmann, Burlington, MA, 2011. 10

[37] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, January 2008. 7

Anexo I

Regras usadas pelo Manual de Denúncias da CGU

A seguir são exibidas as coordenações da CGU as quais recebem a maior quantidade de denúncias triadas bem como as palavras chave utilizadas para identificação dessas como unidade de destino. As palavras chave ou Programas de Governo aqui exibidos foram retirados do manual de triagem produzido pela CGCID.

Diretoria de Auditoria da Área Econômica	Coordenação-Geral de Auditoria da Área Fazendária I-**DEFAZ I**	-Responsável por demandas envolvendo assuntos como BACEN, BB, CEF,instituições financeiras
	Coordenação-Geral de Auditoria da Área Fazendária II –**DEFAZ II**	-Denúncias envolvendo órgãos da adm direta e indireta vinculadas ao Ministério da Fazenda como Receita Federal do Brasil, SERPRO etc...
	Coordenação-Geral de Contas do Governo-**DECON**	-Dificilmente são encaminhadas denúncias para lá
	Coordenação-Geral de Auditoria da Área de Planejamento, Orçamento e Gestão-**DEPOG**	-Denúncias que tiverem relação com a Secretaria de Patrimônio da União, terras públicas, de marinha etc...
	Coordenação-Geral de Auditoria das Áreas de Desenvolvimento, Indústria e Comércio Exterior-**DEDIC**	-Denúncias relacionadas ao BNDES, INMETRO, INPI e SUFRAMA
Diretoria de Auditoria da Área Social	Coordenação-Geral de Auditoria da Área de Justiça e Segurança Pública-**DSSEG**	-Denúncias que envolvem a Polícia Federal, Polícia Rodoviária Federal, programas do Ministério da Justiça, como PRONASCI
	Coordenação-Geral de Auditoria da Área de Desenvolvimento Social-**DSDES**	-Denúncias sobre Bolsa Família, PETI, FNAS-Fundo Nacional de Assistência Social, programa para construção de cisternas, PAA-Programa de Aquisição de Alimentos, e ainda Restaurantes Populares, Cozinhas Comunitárias e Bancos de Alimentos, denúncias sobre irregularidades no CRAS (CRAS-Centro de Referência e Assistência Social, as prefeituras devem observar normas para fazer jus aos recursos do programa da União)
	Coordenação -Geral de Auditoria da Área de Saúde- **DSSAU**	-Geralmente denúncias sobre o SUS. Contudo, a habilitação de uma denúncia sobre recursos do SUS é bem criteriosa. Isso porque, em parte, há órgãos especializados como o DENASUS e a Ouvidoria do SUS. Normalmente, são arquivadas ou encaminhadas a esses órgãos (arquivar com ciência de órgão externo). Para habilitar, tem

Figura I.1: Unidades SFC - Econômica e Social

		que estar muito, muito bem fundamentada. Ex: informa vagamente que não tem remédio etc... Já quando a denúncia vem bem detalhada e/ou cita EXPRESSAMENTE um programa federal, como **Farmácia Básica, é habilitada . Os programas de saúde mais freqüentes nas denúncias são: Saúde da Família, UPA (geralmente na aquisição das instalações, envolve quantias vultosas e transferências fundo a fundo), SAMU e Farmácia Básica.** Convênios : normalmente aqueles que se destinam a saneamento básico, melhorias habitacionais para controle de doenças e agravos. Também pode haver, em raros casos, contratos de repasse do MS.
	Coordenação -Geral de Auditoria da Área de Educação I-**DSEDU I**	Denúncias que envolvam as universidades FEDERAIS e bolsas da CAPES
	Coordenação -Geral de Auditoria da Área de Educação II-**DSEDU II**	Denúncias sobre PNAE-Programa Nacional de Alimentação Escolar, PNATE- Programa Nacional de Transporte Escolar, IFETS (os antigos CEFET'S) , Dinheiro Direto na Escola/Caixa Escolar. Tem também como exemplos clássicos denúncias que envolvem convênios para construção de creche (programa Proinfância), aquisição de ônibus escolar.
	Coordenação-Geral de Auditoria da Área do Meio Ambiente-**DIAMB**	Denúncias que envolvam IBAMA e suas autarquias, normalmente relacionadas a convênios
Diretoria de Auditoria da Área de Infraestrutura	Coordenação-Geral de Auditoria da Área de Minas e Energia-**DIENE**	-Programas: BIODIESEL Luz Para Todos PROINFA PROMINP Ônibus a Hidrogênio META Deve-se verificar também a existência de convênios no Portal da Transparência relacionados a um dos programas.
	Coordenação-Geral de Auditoria das Áreas de Ciência e Tecnologia-**DICIT**	-Clássicas são as denúncias que envolvam as bolsas do CNPq

Figura I.2: Unidades SFC - Infraestrutura

	Coordenação-Geral de Auditoria da Área de Transportes-**DITRA**	-São comuns denúncias de irregularidades em obras do DNIT para pavimentação de rodovias. Deve-se pesquisar no consulta convênios do Portal da Transparência o número de um possível convênio relacionado ao fato. Ultimamente, o próprio DNIT tem diretamente acordado com o Exército Brasileiro obras de pavimentação de rodovias ou fazendo licitações por trechos de rodovias. Para denúncias nesse estilo deve-se pesquisar o site do DNIT.
	Coordenação-Geral de Auditoria da Área de Cidades-**DIURB**	-Normalmente são as denúncias que envolvem obras para construção de casas populares/produção de unidades habitacionais, pavimentação urbana, drenagem de águas pluviais.
	Coordenação-Geral de Auditoria da Área de Integração Nacional-**DIINT**	-Normalmente as denúncias da DIINT relacionam-se a obras para construção de barragens, adutoras, passagens molhadas, de infraestrutura hídrica em estados do NE. Consulta-se o Portal da Transparência para encontrar o número do convênio.
Diretoria de Auditoria das Áreas de Produção e Comunicações	Coordenação-Geral de Auditoria das Áreas de Agricultura, Pecuária e Abastecimento- **DRAGR**	-Denúncias sobre irregularidades em convênios para aquisição de trator agrícola, patrol. Também relacionadas a irregularidades na CONAB nos Estados
	Coordenação-Geral de Auditoria da Área de Desenvolvimento Agrário-**DRDAG**	-Denúncias que envolvem PRONAF, ATES (termos de cooperação em assentamentos, que oferecem assistência técnica agrícola a assentados), ,
	Coordenação-Geral de Auditoria da Área de Turismo e Esportes-**DRTES**	-Essa coordenação é responsável pelos programas do Ministério do Turismo e do Ministério do Esporte. O Ministério do Turismo se utiliza muito de contratos de repasse com municípios para obras de pavimentação de ruas, embelezamento de orla, construção de Portal de Entrada, realização de festas típicas locais,organização de eventos, festas. Já o Ministério do Esporte tem o programa Segundo Tempo, que pode funcionar tanto celebrando convênios para construção de quadra de esporte, como também convênios para o Município organizar turmas de jovens, com o pagamento de professores que executam atividades com os adolescentes e oferta de

Figura I.3: Unidades SFC - Comunicações

		lanches. Às vezes os Municípios contratam ONG's para executar o programa Segundo Tempo.
	Coordenação-Geral de Auditoria da Área de Cultura-**DRCULT**	-Denúncias sobre irregularidades em convênios com ONGs para execução de projetos. Consultar convênios ou ONGs no Portal da Transparência.
	Coordenação-Geral de Auditoria da Área de Comunicações-**DRCOM**	-irregularidades em agências de correios/ECT, na ANATEL.
Diretoria de Planejamento e Coordenação das Ações de Controle	Coordenação-Geral de Planejamento e Avaliação-**DCPLA**	-Geralmente não são encaminhadas denúncias para essas unidades
	Coordenação-Geral de Técnicas, Procedimentos e Qualidade-**DCTEQ**	
	Coordenação-Geral de Operações Especiais-**DCOPE**	-Muitas vezes essa unidade tramita denúncias já classificadas como procedimento simplificado para a carga OGU/CGCid solicitando a alteração da classificação para procedimento ordinário.
	Coordenação-Geral de Recursos Externos-**DCREX**	-Geralmente denúncias sobre irregularidades com recursos do PNUD ou organismos internacionais.
Diretoria de Auditoria das Áreas de Previdência, Trabalho, Pessoal, Serviços Sociais e Tomada de Contas Especial	Coordenação-Geral de Auditoria da Área de Pessoal e Benefícios e de Tomada de Contas Especial-**DPPCE**	-Dificilmente são encaminhadas denúncias para lá.
	Coordenação-Geral de Auditoria da Área de Previdência Social-**DPPAS**	-Denúncias sobre irregularidades em agências do INSS seja em relação a patrimônio, ou em conduta de servidor. Nesse último caso, se bem embasada, pode-se habilitar para procedimento ordinário e encaminhar para CRG/CORAS/CSMPS
	Coordenação-Geral de Auditoria da Área de Serviços Sociais-**DPSES**	-São as denúncias sobre irregularidades no Sistema "S", com exceção do SEBRAE (que vai para de DEDIC)
	Coordenação-Geral de Auditoria das Áreas de Trabalho e Emprego-**DPTEM**	-Programas envolvendo recursos do FAT

Figura I.4: Unidades SFC - Açoes de Controle

CORAS	1 Corregedoria Setorial do Ministério da Justiça CORAS/MJ
	2 Corregedoria Setorial do Ministério da Previdência Social CORAS/MPS
	3 Corregedoria Setorial do Ministério da Saúde CORAS/MS
	4 Corregedoria Setorial do Ministério do Trabalho e Emprego CORAS/MTE
	5 Corregedoria Setorial do Ministério da Cultura e do Esporte CORAS/MINC
	6 Corregedoria Setorial do Ministério do Desenvolvimento Social e Combate à Fome CORAS/MDS
	7 Corregedoria Setorial do Ministério da Educação CORAS/MEC
COREC	1 Corregedoria Setorial do Ministério do Desenvolvimento, Indústria e Comércio Exterior–COREC/MDIC
	2 Corregedoria Setorial do Ministério do Desenvolvimento Agrário COREC/MDA
	3 Corregedoria Setorial do Ministério do Ministério da Fazenda CRG/COREC/MF
	4 Corregedoria Setorial do Ministério do Planejamento, Orçamento e Gestão COREC/MPOG
	5 Corregedoria Setorial do Ministério das Relações Exteriores COREC/MRE
CORIN	1 Corregedoria Setorial do Ministério das Cidades CORIN/MCID
	2 Corregedoria Setorial do Ministério das Comunicações CORIN/MC
	3 Corregedoria Setorial do Ministério dos Transportes CORIN/MT
	4 Corregedoria Setorial do Ministério da Defesa e da Ciência e Tecnologia CORIN/MD
	5 Corregedoria Setorial do Ministério do Meio Ambiente CORIN/MMA
	6 Corregedoria Setorial do Ministério de Minas e Energia CORIN/MME
	7 Corregedoria Setorial do Ministério da Integração Nacional
	8 Comissão Conjunta de Apuração SUDAM SUDENE-CORIN/CCASS

Figura I.5: Unidades da Corregedoria