UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO TECNOLÓGICO

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

HENRIQUE GOMES BASONI

MODELO ASSISTENTE PARA CLASSIFICAÇÃO DE DADOS PROVENIENTES DE REDES SOCIAIS: UM ESTUDO DE CASO COM DADOS DO TWITTER

VITÓRIA-ES

2015

HENRIQUE GOMES BASONI

MODELO ASSISTENTE PARA CLASSIFICAÇÃO DE DADOS PROVENIENTES DE REDES SOCIAIS: UM ESTUDO DE CASO COM DADOS DO TWITTER

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para a obtenção de Título de Mestre em Informática.

Orientador: Prof. Dr. Elias de Oliveira

VITÓRIA-ES

2015

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)

Basoni, Henrique Gomes, 1990-
B313m Modelo assistente para classificação de dados provenientes de redes sociais : um estudo de caso com dados do twitter / Henrique Gomes Basoni. – 2015.
81 f. : il.

Orientador: Elias Silva de Oliveira.
Dissertação (Mestrado em Informática) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Twitter (Rede social on-line). 2. Redes sociais on-line. 3. Recuperação da informação. 4. Algoritmos. 5. Aprendizado do computador. 6. Modelos e construção de modelos. I. Oliveira, Elias Silva de. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 004

HENRIQUE GOMES BASONI

MODELO ASSISTENTE PARA CLASSIFICAÇÃO DE DADOS PROVENIENTES DE REDES SOCIAIS: UM ESTUDO DE CASO COM DADOS DO TWITTER

Dissertação submetida ao programa de Pós-Graduação em Informática do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para a obtenção do Grau de Mestre em Informática.

Aprovada em 14 de Abril de 2015.

COMISSÃO EXAMINADORA

Prof. Dr. Elias de Oliveira
Universidade Federal do Espírito Santo
Orientador

Prof. Dr. Patrick Marques Ciarelli
Universidade Federal do Espírito Santo

Prof. Dr. Ricardo Bastos Cavalcante Prudencio
Universidade Federal de Pernambuco

Prof. Dr^{a}. Maria Claudia Silva Boeres
Universidade Federal do Espírito Santo

À Deus, minha noiva, minha família, amigos, colegas de trabalho e orientador pelo apoio, força, incentivo, companheirismo e amizade. Sem eles nada disso seria possível.

Agradecimentos

Agradeço a Deus em primeiro lugar, por Seu fiel apoio em todo momento permitindo que eu chegasse até o fim desta jornada.

A meus pais Jonas e Isabel, que me deram suporte desde o início mesmo em momentos de dificuldades.

A minha noiva Tatiane, que sempre me apoiou mesmo tendo eu estado muitas vezes distante nesse período.

A meus tios Marcos, Delza e sua maravilhosa família, que me acolheu em sua casa com muito carinho todo este tempo de dedicação aos estudos.

A meu orientador, pela paciência e dedicação me guiando nesta empreitada, pelas suas correções e incentivos.

Também sou grato aos demais colegas de mestrado, aos professores do PPGI e aos profissionais da secretaria. Obrigado pelo companheirismo, pela contribuição nas diversas disciplinas oferecidas, e pelo auxílio em momentos diversos, durante este tempo.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

“Só sei que nada sei, e o fato de saber isso, me coloca em vantagem sobre aqueles que acham que sabem alguma coisa..”

(Sócrates)

Publicações

Como parte deste trabalho, foram desenvolvidos e publicados os seguintes trabalhos, no período do mestrado, que apresentam, em maior ou menor grau, relação com o tema proposto.

Publicação de trabalhos em anais de congressos:

- Saúde, M. R.; Soares, M. M.; Basoni, H. G.; Ciarelli, P. M.; Oliveira, E. **A Strategy of Automatic Moderation of a Large Data Set of Users Comments**. XL Conferencia Latinoamericana en Informática (CLEI). Edición 40. Montevideo, Uruguay. 2014.
- Oliveira, E.; Basoni, H. G.; Saúde, M. R.; Ciarelli, P. M. **Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification**. 6th Conference Internacional of Knowledge Discovery and Information Retrieval (KDIR). Rome, Italy. 2014.

Resumo

Desde seu surgimento, as redes sociais virtuais como Twitter têm alcançado exorbitante quantidade de usuários em todo o mundo, tornando-se um ambiente de imensurável potencial para pesquisas sociais, econômicas, culturais e etc. Cada vez mais pesquisadores têm voltado sua atenção para a grande massa de dados gerada diariamente nesse meio. Entretanto, lidar com grandes quantidades de dados é uma tarefa custosa quando realizada manualmente. O objetivo desta pesquisa é propor um conjunto de ferramentas e metodologia tal que possa diminuir o esforço humano gasto na organização de grandes massas de dados provenientes de redes sociais. Para atingir tal objetivo, é proposto um modelo de trabalho iterativo, que explora ao máximo o conhecimento existente em uma pequena porção de dados manualmente analisada por especialistas. O modelo de trabalho combina técnicas de recuperação de informação como algoritmos de classificação e agrupamento com objetivo de tornar o resultado do processo mais parecido ao que o especialista obteria caso o realiza-se completamente manualmente. O modelo proposto foi colocado a prova com uso de dois conjuntos de dados extraídos do Twitter e manualmente classificado muito antes da realização desta pesquisa. Os resultados mostraram-se promissores.

Palavras-chave: Redes Sociais; Recuperação da Informação; Aprendizado; Algoritmo; Modelo.

Abstract

Since its inception, virtual social networks like Twitter have reached exorbitant amount of users worldwide, making it an immeasurable potential environment for social research, economic, cultural and etc. Increasingly researchers have turned their attention to the great mass of data generated daily in this environment. However, handling large amounts of data is a costly task when performed manually. The objective of this research is to propose a set of tools and methodology that it can reduce the human effort spent in the organization of large masses of data from social networks. To achieve this goal, we propose an iterative work model that makes the most of existing knowledge in a small amount of data manually analyzed by experts. The working model combines information retrieval techniques such as classification and clustering algorithms in order to make the result of the most similar process to what the expert would get if carried out completely manually. The proposed model was put to the test with use of two sets of extracted data from Twitter and manually classified before this research. The results were promising.

Keywords: Social networks; Information Retrieval; Learning; Algorithm; Model.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Este capítulo apresenta o contexto, motivação e objetivos deste trabalho, bem como os métodos de pesquisa aplicados e sua organização textual, com o intuito de proporcionar ao leitor melhor compreensão do universo no qual se encontra o problema que buscamos solucionar: classificar grandes massas de dados provenientes de redes sociais com o menor esforço humano possível nesse processo.

1.1 Contexto

Com a modernidade vivida nesta era, principalmente com o advento da Internet, as pessoas têm gerado informações, em grande quantidade e a todo momento. Foi estimado que, no ano de 2012, por dia, cerca de 2,5 Exabytes (2,5 bilhões de Gigabytes) de novas informações foram geradas, a cada dia, pelos usuários da grande rede (WALL, 2014). Como o número de usuários tende a crescer, com a chegada da Internet nas áreas ainda sem acesso, a quantidade de informações compartilhadas também aumentará.

As redes sociais virtuais, impulsionadas pela popularização da Internet, ajudaram a aumentar essa grande massa de dados gerada constantemente. Em (ELLISON et al., 2007)

essa tecnologia relativamente nova é definida como um serviço web que (1) permite a seus usuários construir um perfil público ou semi-público, (2) manusear uma lista de conexões com outros usuários que queiram compartilhar informações, (3) visualizar suas conexões e a dos demais. A nomenclatura dada às conexões varia de acordo com o sistema em questão. Como exemplo de redes sociais virtuais podemos citar: Facebook, Youtube e Twitter, sendo a última o foco deste trabalho.

Fundado em 21 de março de 2006 nos Estados Unidos, o Twitter é uma rede social que funciona como um microblog no qual seus usuários seguem e são seguidos por outros usuários que compartilham suas opiniões através de tuítes, um pequeno texto formado por, no máximo, 140 caracteres. Conhecida em nível mundial, esta rede social atingiu a expressiva marca de 271 milhões de usuários, que geram cerca de 500 milhões de tuítes todos os dias. No Brasil, os usuários dessa ferramenta já passam de 40 milhões. Devido à quantidade de usuários, o Twitter é um ambiente ímpar que pode fornecer opiniões sobre os mais variados assuntos e de forma rápida. Segundo a própria administração da ferramenta, nas últimas eleições presidenciais, mais de 40 milhões de tuítes foram publicados, evidenciando assim seu potencial informativo (G1, 2014).

Devido às características anteriores, muitos especialistas têm voltado sua atenção para essas ferramentas, crendo ser fonte vital para responder algumas questões, tais como: O que as pessoas querem? Do que elas gostam? Qual a sua opinião sobre um dado assunto? O que está acontecendo agora em um determinado lugar? Para capturar dados do Twitter e responder tais questionamentos, os especialistas têm feito uso de técnicas como a descrita em (BRUNS; LIANG, 2012). Os autores estão interessados em técnicas que capturam dados do Twitter a fim de identificar desastres naturais. Objetivando identificar tuítes relevantes para o assunto em questão, os autores propõem também o uso de palavras-chave ou as famosas *hashtags*, que no Twitter, é uma forma de indexar e explicitar um tuíte, como indício de que é relevante, caso contenha uma destas ou ambas no seu corpo de texto.

Com mostrado acima, existem métodos na literatura propostos para obter dados e identi-

ficar tuítes relevantes. Especialistas têm usado tais métodos para capturar dados de interesse e, montar suas bases de dados a fim de analisar e responder questões como as já ditas anteriormente. Segundo (BORTOLON; REGATTIERI; MALINI, 2013), uma vantagem que o uso do Twitter apresenta, em detrimento de outras redes sociais, diz respeito à pessoalidade das falas dentro dele, ou seja, dos tuítes. Como o site funciona como um microblog, o discurso muitas vezes carrega maior carga de sentimentalismo, sendo assim mais espontâneo.

A análise dos dados capturados frequentemente está ligada à classificação dos tuítes. Classificar manualmente uma base, como normalmente é feito, tem se mostrado um árduo processo, devido ao fato de que o especialista passa maior quantidade de tempo classificando a base do que analisando os dados, ou seja, o custo de pré-processamento dos dados é alto. Este trabalho está no contexto de classificação dos dados provenientes de redes sociais virtuais, em que especialistas têm gastado um considerável tempo classificando dados manualmente.

1.2 Motivação

A Recuperação de Informação promove a representação, armazenamento, organização e acesso a itens de informação tais como documentos, paginas Web, catálogos *online*, registros estruturados e semi-estruturados e objetos multimídia. A representação e organização de tais itens de informação deve ser tal que possa prover acesso fácil ao que seja relevante para o usuário (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 1).

A fim de atender à demanda do parágrafo anterior, a Recuperação de Informação faz uso de técnicas de Aprendizado de Máquina, uma área de estudo contida na Inteligência Artificial (IA), que visa desenvolver métodos computacionais a respeito do aprendizado, bem como a modelagem de sistemas que possam, de modo automático, adquirir conhecimento. Alguns sistemas aprendem tendo como base um conjunto de informações avaliadas previamente e de modo bem sucedido (REZENDE, 2003, p. 89).

As técnicas de Aprendizado de Máquina, podemos dizer, tentam imitar a forma como um especialista trabalha com base em exemplos gerados pelo próprio especialista. Dada uma tarefa a ser realizada, utilizando tais técnicas, faz-se necessária a análise do mundo no qual as mesmas serão empregadas, essa compreensão possibilitará a adaptação da técnica para o problema em questão, tornando o sistema mais parecido com o especialista.

A estatística, ainda que com uma visão diferente da computação, fornece técnicas que possibilitam a análise de grandes conjuntos de informações. Algumas dessas técnicas estão contidas na Análise Multivariada de Dados. As técnicas de Análise Multivariada são métodos presentes na estatística que possibilitam a análise concorrente de múltiplas medidas a respeito de cada objeto de investigação, como tuítes por exemplo (JUNIOR et al., 2005, p. 23).

Técnicas Estatísticas e Aprendizado de Máquina, podem ajudar a diminuir o esforço do especialista quanto ao problema apresentado na Seção 1.1 e diminuir o tempo gasto pelo mesmo no processo de classificação de uma base de dados proveniente de redes sociais virtuais usada para análise. A motivação desta pesquisa está na necessidade de ajustar tais técnicas a um modelo de trabalho para uso em um ambiente específico.

1.3 Objetivos

Este trabalho tem como objetivo elaborar um conjunto de ferramentas e uma metodologia com o auxilio de técnicas de Aprendizado de Máquina e estatísticas que possibilite ao especialista diminuir os esforços gastos no processo de classificação de um conjunto de dados provenientes de redes sociais virtuais. Tal objetivo geral pode ser decomposto como os seguintes objetivos específicos:

- Elaborar uma metodologia que nos possibilite atingir o objetivo geral deste trabalho (OLIVEIRA et al., 2014);

- Pontuar e descrever os detalhes de cada componente da metodologia proposta;
- Desenvolver a arquitetura da metodologia proposta a fim de analisar a viabilidade de utilização da mesma;
- Avaliação da proposta.

1.4 Metodologia do Trabalho

O passo inicial foi a revisão de literatura a fim de encontrar estudos recentes que abordem assuntos relacionados às áreas de pesquisa abrangidas neste trabalho como: Aprendizado de Máquina, Recuperação Inteligente da Informação, Estatística e outras.

Realizada a revisão, um modelo para solucionar o problema foi proposto e, algumas técnicas que possibilitam a construção do mesmo foram analisadas, o que acarretou em várias versões do modelo compostas por diferentes técnicas.

Para a realização dos experimentos, foram disponibilizadas algumas bases de dados compostas por comentários reais obtidos do Twitter, bases estas formadas por 2080 tuítes únicos a respeito do Marco Civil da Internet no Brasil, classificados manualmente pelos especialistas muito antes da realização deste estudo.

Com objetivo de avaliar os resultados obtidos foram utilizados os resultados dos classificadores *K-Nearest Neighbors* (*K*NN) (BAEZA-YATES; RIBEIRO-NETO et al., 2011; SOUCY; MINEAU, 2001) e *Centroid-Based Classifier* (CBC) (HAN; KARYPIS, 2000; SOUZA; CIARELLI; OLIVEIRA, 2014).

A métrica F_1 (SOUZA; CIARELLI; OLIVEIRA, 2014) de classificação foi utilizada para avaliação de cada versão do modelo. Tal métrica possibilita observar a qualidade do trabalho realizado com uso do modelo, se comparada a base manualmente classificada pelo especialista.

O texto da base passou por um pré-processamento onde foram utilizadas técnicas de Análise Léxica, retirada de *Stopwords*, Lematização, ou seja, redução de termos à sua forma canônica, retirada de termos com baixa frequência nos documentos da base, além de uma medida de ponderação de termos.

1.5 Estrutura do Trabalho

Este capítulo contextualizou a problemática abordada neste trabalho, descreveu a motivação para realização desta pesquisa, os objetivos e a metodologia aplicada. Além desta introdução, a organização deste trabalho se dá do seguinte modo:

- **Capítulo 2 (Recuperação Inteligente da Informação)**: apresenta os principais conceitos de Aprendizado de Máquina e Estatística relacionados à Classificação, Agrupamento e Análise Fatorial.
- **Capítulo 3 (Trabalhos Relacionados)**: neste capítulo são mencionados alguns trabalhos interessantes, relacionados aos assuntos abordados neste trabalho.
- **Capítulo 4 (O Modelo)**: descreve todo o modelo proposto para solução do problema. Neste capítulo cada parte do modelo é apresentada e discutida.
- **Capítulo 5 (Experimentos e Resultados)**: são descritos e discutidos os resultados obtidos nos experimentos com uso das várias versões da metodologia proposta.
- **Capítulo 6 (Conclusões e Trabalhos Futuros)**: apresenta as conclusões do trabalho, suas contribuições e propostas futuras de aprimoramento do trabalho.

Capítulo 2

Recuperação da Informação

Este capítulo apresenta uma sucinta revisão literária dos conceitos necessários para melhor compreensão deste trabalho. Nesta seção, serão abordadas técnicas e ferramentas utilizadas na construção do modelo indicado por esta pesquisa como solução factível ao problema em questão.

Através dos séculos, o homem tem aplicado tempo na organização de informações a fim de facilitar sua recuperação, quando necessário. Devido a isso, surgiram as bibliotecas (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 2). Apesar das facilidades trazidas por essa instituição, o crescimento da população e a evolução da tecnologia impulsionaram o aumento da quantidade de informações geradas, fazendo com que as bibliotecas crescessem e se tornasse cada vez mais árduo o trabalho de encontrar algo relevante para o usuário.

Com o advento da Internet o problema se agravou, pois esta fez crescer ainda mais a quantidade de informações onde o usuário pode encontrar o assunto desejado. Muitas das páginas criadas na Internet tem informações úteis, entretanto, estão armazenadas de modo desestruturado, dificultando o trabalho de encontrar as de interesse em meio a tantas outras não interessantes.

Nos últimos 10 anos, o interesse em técnicas de gerenciamento para grandes massas de

documentos, seja qual for a sua natureza: áudio, vídeo, texto e etc, cresceu expressivamente, elevando o status dos sistemas de Recuperação de Informação, visto que estes têm como objetivo prover acesso simplificado e flexível a documentos (SEBASTIANI, 2002). Nas próximas sub-seções, deste capítulo, são descritas algumas técnicas utilizadas para alcançar os objetivos dos sistemas de recuperação de informação, citados anteriormente.

2.1 Modelo Espaço Vetorial

Uma tarefa presente no processo de modelagem de sistemas de Recuperação da Informação (RI) é a representação de documentos. Um método muito presente na literatura é a representação de documentos em vetores de índices.

Dado um conjunto $\mathcal{D}$ de documentos, cada documento d_i do conjunto pode ser representado por um ou mais termos (informações encontradas no conteúdo dos documentos como palavras, datas e etc) ou índices t_i. Os termos podem receber pesos $w_i \geq 0$ que evidenciem o nível de importância de cada um para cada documento. Tais pesos podem ser adquiridos por uma simples contagem de ocorrência dos termos (SALTON; WONG; YANG, 1975) dentro de cada documento, por exemplo. Este processo é chamado Indexação de Documentos.

Assim sendo, um conjunto de documentos (ou base de dados) é representado por uma matriz termo-documentos, onde cada documento é uma linha dessa matriz, como mostrado na Equação 2.1 (SALTON; WONG; YANG, 1975; BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 62). A Figura 2.1 apresenta um conjunto de três documentos, representados de modo tridimensional, ou seja, cada documento é formado por três índices no espaço.

$$d_i = \{w_{1i}, w_{2i}, ..., w_{ni}\} \tag{2.1}$$

O modelo vetorial está baseado em um conceito presente nas tarefas de mineração de dados do tipo texto (tipo abrangido por essa pesquisa) denominado Bolsa de Palavras (*Bag*

of Words). Neste caso, cada índice t_i representa uma palavra presente no conjunto de documentos $\mathcal{D}$ (SCOTT; MATWIN, 1999). Mais uma vez olhando para a Figura 2.1 e vendo cada documento d_i como do tipo texto, os índices t_1, t_2 e t_3 são pesos dados as palavras presentes no texto de um ou mais documentos.

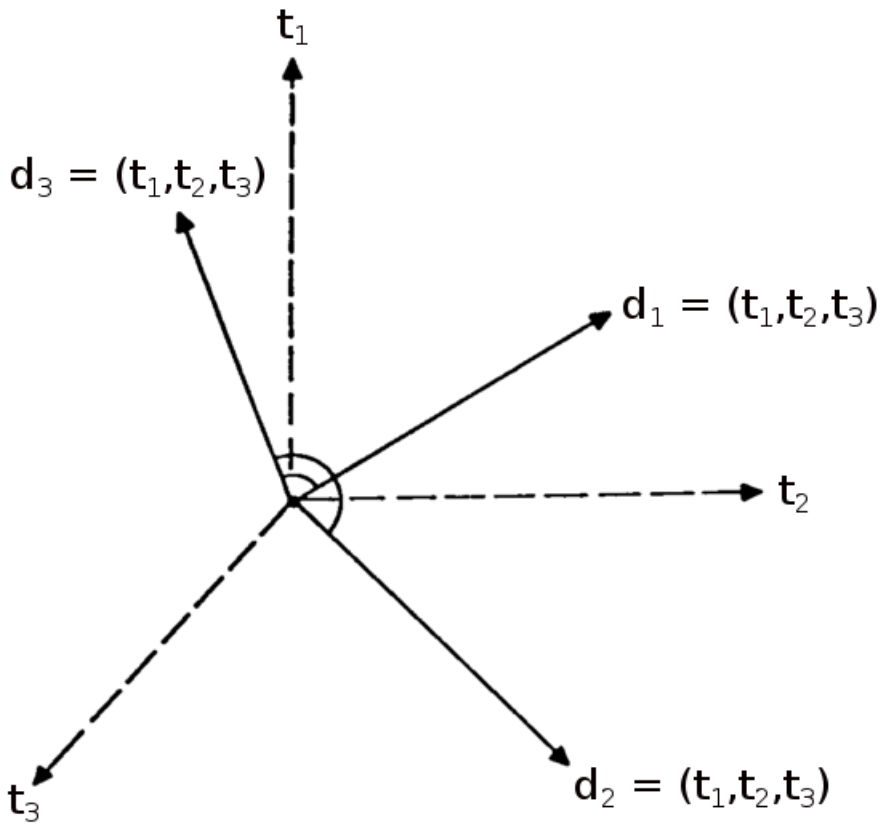

Figura 2.1: Representação vetorial de documentos no espaço (SALTON; WONG; YANG, 1975).

2.1.1 Cálculo de Similaridade

Devido ao método usado pelo modelo vetorial para representação de documentos, a similaridade entre documentos de um mesmo conjunto pode ser obtida através de um cálculo matemático vetorial. O cálculo de cosseno entre pares de documentos é a estratégia utilizada, neste trabalho, para obter tal similaridade, a fórmula do cosseno é exibida na Equação 2.2.

$$cos(d_i, d_j) = \frac{d_i.d_j}{||d_i||.||d_j||} = \frac{\sum_{k=1}^{n} w_{k,i}.w_{k,j}}{\sqrt{\sum_{k=1}^{n}(w_{k,i})^2}\sqrt{\sum_{k=1}^{n}(w_{k,j})^2}} \tag{2.2}$$

O resultado esperado para a similaridade entre dois documentos quaisquer é um número entre 0 e 1, sendo que a similaridade 1 é obtida para documentos cuja suas representações vetoriais são idênticas, enquanto 0 para documentos que não compartilham nehum termo em comum.

2.1.2 Caracterização de Bases de Dados

A qualidade de uma base de dados, ou seja, aquela com a qual um algoritmo de aprendizado (Seção 2.4) obtêm melhores taxas, quanto à qualidade de aprendizado, está relacionada ao conceito de densidade da base, dado que este trabalho adota as medidas de caracterização propostas por (SALTON; WONG; YANG, 1975).

Para melhor compreensão do conceito de densidade, será utilizada a Figura 2.2. Um espaço $\mathcal{D}$ qualquer de documentos é representado na figura; cada um nos n documento pertencente ao conjunto $\mathcal{D}$ é representado pela letra x e pertencente a um agrupamento $\mathcal{K}_u$; tais agrupamentos são representados pelas linhas que separam os conjuntos de documentos na figura; os centroides $\mathcal{C}$, círculos em cor preta relativamente centralizados em cada agrupamento, têm cada termo c_i (Seção 2.1) obtido a partir da média dos pesos de cada termo dos documentos pertencentes a um mesmo agrupamento (Equação 2.3) (SALTON; WONG; YANG, 1975).

$$c_i = \frac{1}{n} \sum_{\substack{j=1 \\ d_j \in \mathcal{D}}}^{n} d_{j,i} \tag{2.3}$$

Na Figura 2.2, também existe um retângulo em cor preta, o qual representa o centroide principal $\mathcal{C}_p$. Este pode ser obtido de duas maneiras: (1) através das médias de cada termo

dos d documentos do espaço D; (2) através das médias de cada termo c_i dos centroides. No primeiro método, o centroide principal é polarizado pelo agrupamento com maior quantidade de documentos. A polarização ocorre apenas se o agrupamento não for balanceado, ou seja, se os agrupamentos não tiverem a mesma quantidade de documentos. No segundo método, não existe diferença quanto à contribuição de cada agrupamento na formação do centroide (SALTON; WONG; YANG, 1975; SOUZA; CIARELLI; OLIVEIRA, 2014).

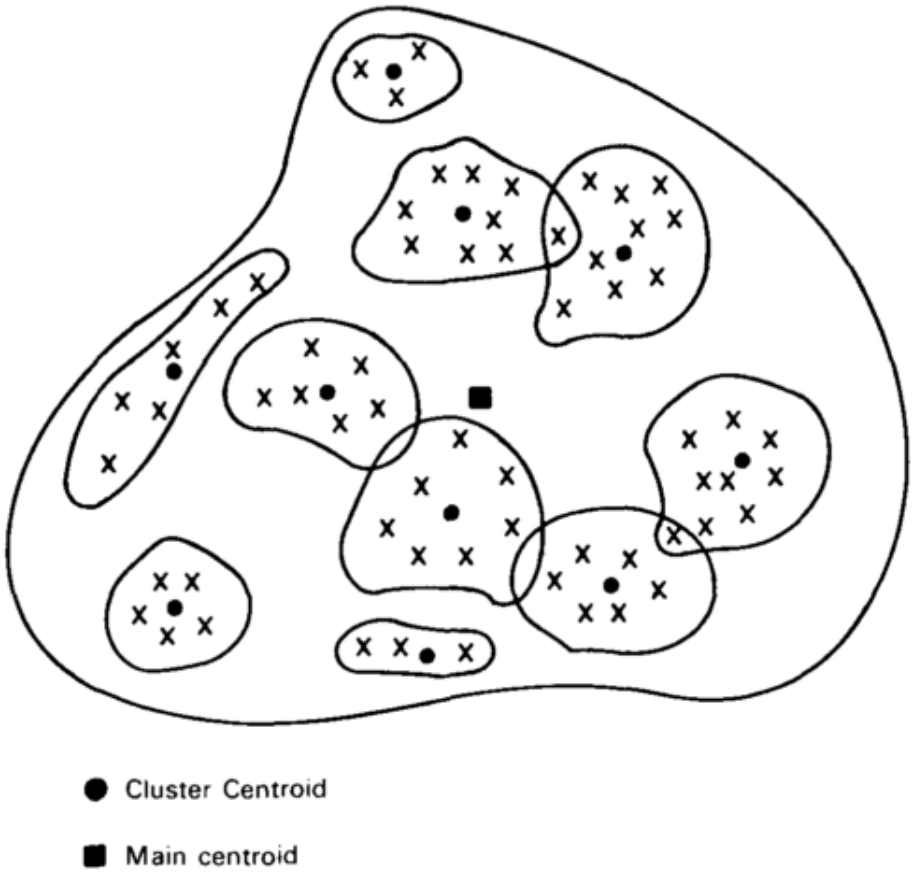

Figura 2.2: Espaço de documentos agrupados (SALTON; WONG; YANG, 1975).

Obtidos os centroides, pode-se calcular as medidas necessárias para caracterização do espaço de documentos. A Média de Similaridade entre Documentos e sua Centroide (MSDC) mostra o quão próximos entre si os documentos de um mesmo agrupamento estão, ou seja, o quão denso é um agrupamento. O MSDC é um valor entre 0 e 1, em que quanto mais próximo de 1 mais denso é o agrupamento.

Já na Média de Similaridade entre os Pares de Centroides (MSPC), avaliada também

como um valor entre 0 e 1, quanto mais próximo de 0, mais distantes estão os agrupamentos uns dos outros, ou seja, menor a densidade entre os agrupamentos.

A Média de Similaridade entre os Centroides de cada agrupamento e o Centroide Principal (MSCCP) é uma medida que indica se as médias de similaridade dos documentos de um agrupamento estão próximas da média geral de similaridade dos agrupamentos. Quanto mais próximo de 1 for o valor do MSPPC, maior a proximidade.

Para identificar a densidade geral do espaço de documentos, utilizamos a métrica *razão* (Equação 2.4). Quanto mais próximo de 0 o valor dessa métrica, melhor a distribuição do espaço de documentos, isto significa que os agrupamentos estão mais afastados uns dos outros, enquanto os documentos de um agrupamento estão mais próximos entre si.

O fato dos agrupamentos estarem mais distantes melhora a qualidade do trabalho a ser realizado com o conjunto de documentos, visto que a identificação dos membros de cada agrupamento é mais simples. Na prática, isso significa que a probabilidade de agrupar erroneamente um novo documento é menor do que se os agrupamentos estivessem mais próximos uns dos outros. Olhando novamente para a Figura 2.2, pode-se observar a densidade do espaço de documentos percebendo a distância entre os agrupamentos.

$$\text{razão} = \frac{MSPC}{MSDC} \tag{2.4}$$

2.2 Processamento de Documentos

Um conjunto de documentos não é formado apenas por informação útil. Existem também os ruídos, que podem ser compreendidos como termos de documentos que prejudicam os algoritmos de aprendizado utilizados em sistemas de Recuperação de Informação (RI), como classificadores e agrupadores (Seção 2.4). Além dos ruídos, a alta dimensionalidade do

espaço de termos (quanto maior a quantidade de palavras, maior o espaço de termo) pode causar dificuldades para o aprendizado de algoritmos (YANG; PEDERSEN, 1997).

Neste trabalho são utilizadas técnicas visando diminuir os efeitos dos ruídos presentes em documentos do tipo texto, tal tipo de documento é foco desta pesquisa. Na Figura 2.3, tais técnicas estão posicionadas tal qual sua ordem de execução nas bases em estudo neste trabalho. Cada uma das técnicas presentes na figura serão esclarecidas nas seções que se seguem.

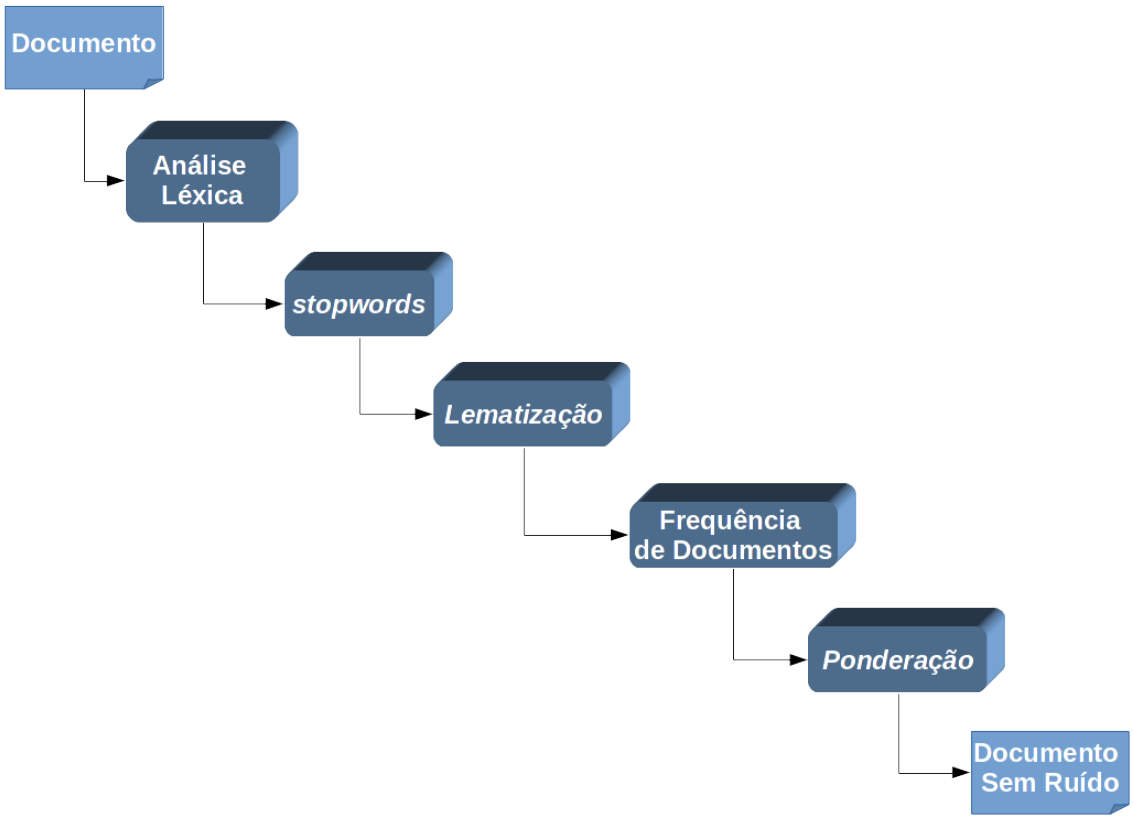

Figura 2.3: Esquema de processamento de documentos.

2.2.1 Análise Léxica

O principal objetivo da Análise Léxica é a identificação do que são ou não são palavras em um documento texto. Com isso, o texto deixa de ser um conjunto de caracteres, passando a ser um conjunto de palavras (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 224).

Em sua maioria, termos não considerados palavras são (FRAKES; BAEZA-YATES, 1992, p. 117):

- Dígitos: Salvo em casos de exceção, como bases de dados bancárias ou algo do tipo, números não são bons termos para indexação. Horas, datas, valores, dentre outros, normalmente não agregam grande peso semântico ao texto;
- Hífens: Remover ou não hífens é uma decisão um pouco mais complexa. Com a permanência do hífen, um documento com a palavra "afro-americano", e outro com a palavra "afro-brasileiro", são considerados completamente diferentes, porém, ambos os documentos podem estar falando de um mesmo assunto. Entretanto, palavras como "F-16" ou "MS-DOS" perdem sua carga semântica sem o hífen;
- Pontuações ou caracteres especiais: Em sua maioria, pontuações e caracteres especiais prejudicam a indexação. Eles podem fazer com que uma mesma palavra seja indexada como diferentes palavras, caso uma delas seja sucedida por um ponto de final de frase, por exemplo;
- *Case*: Letras maiúsculas ou minúsculas não são significantes para a indexação, assim sendo, normalmente o processo de Análise Léxica converte todos os caracteres presentes em um conjunto de documentos para maiúsculo ou minúsculo.

Apesar de sua simples implementação computacional, a Análise Léxica é custosa computacionalmente, pois exige a examinação de cada caractere presente no conjunto de documentos (FRAKES; BAEZA-YATES, 1992, p. 119).

2.2.2 Eliminação de *Stopwords*

80% das palavras presentes nos documentos de um conjunto de dados não são consideradas discriminativas, ou seja, não contribuem de forma considerável para identificação do

assunto ao qual um documento está relacionado. Tais palavras são, em sua maioria, pertencentes ao conjunto denominado Lista de *Stopwords*. Artigos, preposições e conjunções, normalmente, fazem parte desse grupo.

A eliminação de *Stopwords* traz importante benefício, pois reduz, consideravelmente, o tamanho da estrutura de indexação e remove termos de baixo valor semântico que podem se caracterizar como ruído (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 226). A Tabela 2.1 é composta por alguns exemplos de *stopwords* existentes na língua portuguesa.

Lista de *Stopwords*						
rt	marco	civil	marcocivil	de	é	depois
a	com	dos	já	também	sem	estão
o	não	como	está	só	mesmo	você
que	uma	mas	seu	pelo	aos	tinha
e	os	foi	sua	pela	ter	foram
do	no	ao	ou	até	seus	essa
da	se	ele	ser	isso	quem	num
em	na	das	quando	ela	nas	nem
um	por	tem	muito	entre	me	suas
para	mais	à	há	era	esse	meu
qual	essas	tu	minhas	nossa	estes	isto
será	esses	te	teu	nossos	estas	aquilo
nós	pelas	vocês	tua	nossas	aquele	havia
tenho	este	vos	teus	dela	aquela	seja
lhe	fosse	lhes	tuas	delas	aqueles	pelos
deles	dele	meus	nosso	esta	aquelas	elas
numa	têm	minha	às	as	nos	eu
eles						

Tabela 2.1: Lista de *Stopwords* (BETTIO et al., 2007).

2.2.3 Lematização (*Stemming*)

Uma técnica que visa aumentar a performance dos algoritmos de aprendizado é a Lematização. Essa técnica objetiva identificar as variantes morfológicas de uma palavra, mantendo no conjunto de documentos apenas seu radical (*stem*) (FRAKES; BAEZA-YATES, 1992, p. 141).

Como exemplo de Lematização podem ser observadas as palavras "representação", "representado" e "representando", pois podem ser representadas de um modo comum como "represent". Com isso, três documentos contendo cada uma das três formas citadas passarão a conter um mesmo radical que represente todas elas (ORENGO; HUYCK, 2001).

As variações de uma palavra também podem prejudicar o processo de aprendizado de algoritmos. Um conjunto de documentos conterá não apenas uma forma de uma palavra, mas muitas, como plurais, formas de gerúndio, variações verbais entre outras. O problema ocorre devido ao fato de o processo de indexação identificar tais palavras como diferentes, causando aumento do espaço de indexação (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 226).

O algoritmo utilizado neste trabalho para Lematização é o Redutor de Sufixos da Língua Portuguesa (RSLP), proposto em (ORENGO; HUYCK, 2001). Como pode ser visto na Figura 2.4, o algoritmo executa a extração de radicais em oito passos (representados como retângulo), obedecendo ao fluxo exibido na figura.

O algoritmo RSLP, atualmente, possui 199 regras, cada uma composta por 4 partes: (1) o sufixo, que será removido; (2) o tamanho mínimo do radical; (3) o sufixo que deverá substituir o sufixo atual da palavra (caso necessário) e; (4) uma lista de exceções. A Figura 2.5 mostra uma regra do RSLP para o sufixo "inho", com radical de tamanho mínimo formado por 3 letras, nenhuma substituição para o sufixo e uma lista de exceções.

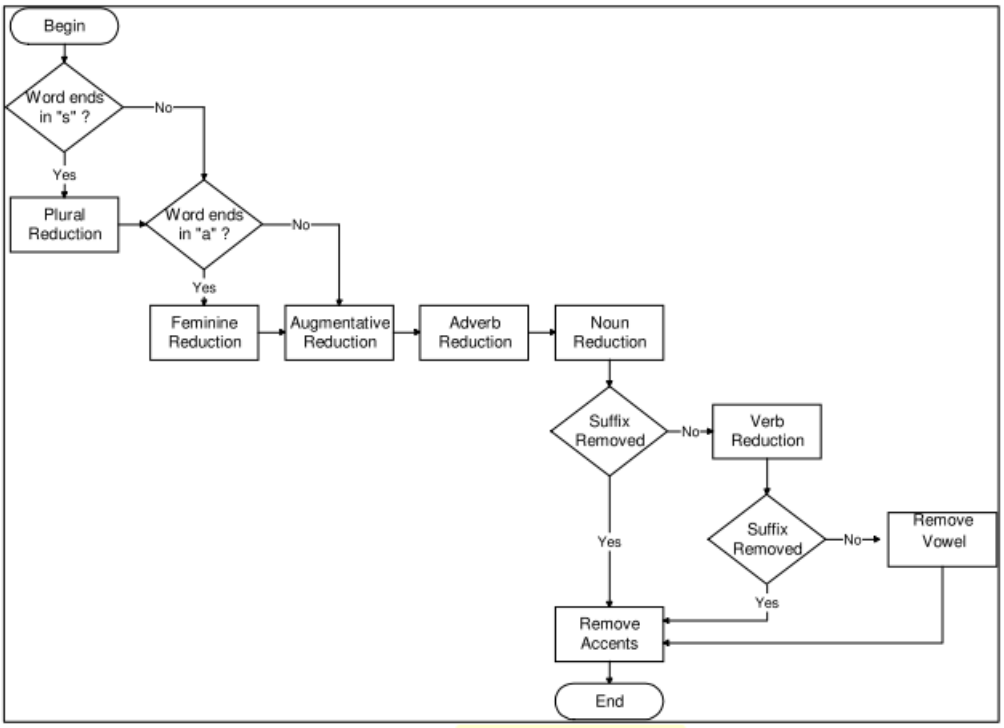

Figura 2.4: Passos do algoritmo RSLP para Lematização (ORENGO; HUYCK, 2001).

```
"inho", 3, "", {"caminho", "carinho",
"cominho", "golfinho", "padrinho",
"sobrinho", "vizinho"}
```

Figura 2.5: Modelo de regra do RSLP (ORENGO; HUYCK, 2001).

2.2.4 Frequência de Documentos

A análise da Frequência de Documentos (FD) é uma técnica bem simples de redução de dimensionalidade do espaço de termos. Termos que possuam baixa frequência não contribuem para o aumento na performance global de algoritmos de aprendizado. Neste caso, termos com baixa frequência podem ser removidos do espaço de termos, sem causar prejuízos, pois estes termos não são discriminantes para identificação de qualquer grupo de documentos (YANG; PEDERSEN, 1997). Neste trabalho, termos que não se repitam em mais do que dois documentos são removidos do espaço de termos.

2.2.5 Ponderação de Termos

Ao olhar para um conjunto de documentos, é possível notar que os termos deste conjunto não têm a mesma importância na discriminação de um documento. No entanto, verificar os termos com maior ou menor peso discriminativo não é tarefa simples, pois em um conjunto de dados, com centenas de milhares de documentos, existem muitos termos a serem observados. Para caracterizar o quão discriminativo um termo é, um peso numérico $w_{i,j} > 0$ é associado a cada termo k_i de um documento d_j pertencente ao conjunto. Tal peso, então, quantifica a importância de um termo na discriminação de um documento (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 66).

Os pesos estão relacionados à frequência de ocorrência dos termos no conjunto. Seja $f_{j,i}$ a frequência de ocorrência do termo k_i em um documento d_j, ou seja, o número de vezes em que o termo k_i aparece no texto do documento d_j; e N o número total de documentos no conjunto, a frequência F_i de cada termo no conjunto de documentos é obtida pelo somatório da ocorrência do mesmo em cada documento como mostrado na Equação 2.5 (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 67).

$$F_i = \sum_{i=1}^{N} f_{j,i} \tag{2.5}$$

TF-IDF

Duas das formas mais populares utilizadas para ponderação de termos são *Term Frequency* (TF) e *Inverse Document Frequency* (IDF). A TF é definida como o valor ou peso de um termo k_i com ocorrência no documento d_j, proporcional à frequência $f_{j,i}$ do termo. Isto significa que quanto mais o termo k_i aparecer no texto do documento d_j, maior será o peso $TF_{j,i}$ do mesmo. Dada a definição da TF, pode-se obtê-la apenas calculando a frequência $f_{j,i}$ de cada termo k_i presente no documento d_j , como mostrado na Equação 2.6 (LUHN, 1957; BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 68).

$$TF_{j,i} = f_{j,i} \tag{2.6}$$

A IDF é uma medida baseada na contagem do número de documentos em que o termo k_i se repete, com o objetivo de observar se um termo está presente em muitos documentos do conjunto. Em caso afirmativo, tal termo não é considerado como bom discriminador (ROBERTSON, 2004; JONES, 1972; SALTON; WONG; YANG, 1975).

Como pode ser observado na Equação 2.7, o IDF do termo k_i é obtido pelo logaritmo da razão entre N, quantidade total de documentos do conjunto, e n_i, quantidade de documentos em que k_i ocorre.

$$IDF_i = \log \frac{N}{n_i} \tag{2.7}$$

Uma combinação de TF e IDF é muito utilizada na literatura (Equação 2.8). Em (SALTON; WONG; YANG, 1975), os experimentos mostraram que tal combinação pode diminuir a densidade entre os agrupamentos de documentos e, consequentemente, aumentar a qualidade de aprendizado dos algoritmos utilizados em sistemas de RI. Como os já citados classificadores e agrupadores.

$$TF * IDF_i = f_{i,j} \times \log \frac{N}{n_i} \tag{2.8}$$

2.3 *Latent Semantic Indexing* (LSI)

O tradicional modelo de indexação utilizado nos sistemas de RI, denominado Bolsa de Palavras (*Bag of Words*), ignora qualquer relação semântica entre as palavras (SCOTT; MATWIN, 1999). Este método não tem a capacidade de identificar casos de sinônimos, em

que um mesmo significado pode ser obtido através de várias palavras; ou polissemia, os vários sentidos ou significados de uma palavra (DEERWESTER et al., 1990). Neste caso, a *Latent Semantic Indexing* (Indexação por Semântica Latente) foi proposta para solucionar os problemas apresentados.

Sendo t o número de termos no conjunto de documentos, N o número total de documentos, e $M = \left[m_{ij}\right]$ uma matriz termo-documentos com t linhas e N colunas, o processo de indexação por semântica latente decompõe a matriz M em três componentes usando a *singular value decomposition* (SVD), como pode ser visto na Equação 2.9 (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 101).

$$M = K \cdot S \cdot D^T \tag{2.9}$$

As componentes K e D são matrizes de vetores singulares à esquerda e à direita. A componente S é diagonal com elementos ordenados, assim sendo, as matrizes podem ser reduzidas determinando um limiar para os elementos de S e zerando todos os que não atingem este limiar. Todas as colunas de K e D correspondentes às linhas com valor zero em S podem ser ignoradas. A multiplicação das matrizes reduzidas resultará em uma matriz $\hat{X}$ aproximada da matriz termo-documentos original. (ZELIKOVITZ; HIRSH, 2001).

Realizadas a transformação descrita no parágrafo anterior, as matrizes K e D são matrizes termos-fatores e documentos-fatores, respectivamente. Fatores são uma representação "resumida" de um conjunto de variáveis ou termos fortemente relacionados. Neste trabalho, o interesse está na matriz D. Nessa matriz, cada documento passa a ser representado pela quantidade de fatores resultantes do processo de redução. Além de diminuir o espaço de indexação, os fatores passam a representar conjuntos de palavras relacionadas.

2.4 Aprendizado de Máquina

A Inteligência Artificial (IA), área de estudo que visa compreender a inteligência (RUSSELL; NORVIG, 1995, p. 3), contém uma linha de pesquisa denominada Aprendizado de Máquina que visa a criação de algoritmos computacionais capazes de imitar tarefas realizadas por seres humanos (DOMINGOS, 2012).

O aprendizado, trabalho que torna possível uma máquina realizar tarefas humanas, obtendo um resultado final próximo do realizado manualmente pelo homem, pode ser dividido em três partes (DOMINGOS, 2012):

- Representação: um algoritmos de aprendizado é representado através de linguagem computacional formal como C ou Java.
- Avaliação: faz-se necessário saber o quão bem feito foi o trabalho do algoritmo, se comparado ao realizado pelo humano. Sendo assim, há necessidade de uma ou mais funções que retornem métricas, possibilitando a avaliação.
- Otimização: visa evoluir a performance do algoritmo.

Dentre os tipos de aprendizado podemos citar o supervisionado, não-supervisionado e semi-supervisionado. Cada um dos métodos de aprendizado citados serão descritos nas subseções que se seguem.

2.4.1 Aprendizado Supervisionado

O aprendizado supervisionado recebe este nome devido à necessidade de participação de um agente externo (especialista), que apresente ao algoritmo um conjunto de exemplos do trabalho a ser realizado automaticamente (HAYKIN, 1999, p. 63).

A classificação é uma tarefa de aprendizado supervisionado. Esta recebe um conjunto de dados pré-classificados pelo especialista e, ao receber um conjunto de dados de classificação desconhecida, relacionado ao tema do conjunto conhecido, realiza a classificação baseada nos exemplos.

Seja D um conjunto de documentos, $C = \{c_1, ..., c_{|L|}\}$, um grupo previamente conhecido de classes, e $\Omega = \{d_1, ..., d_{|\Omega|}\}$, um subconjunto de D, previamente classificado por um especialista nas classes C, o processo automático de classificação é uma função binária $\mathcal{F} : D \times C \rightarrow \{0,1\}$, em que o par $[d_j, c_p]$ recebe valor 1 caso $d_j \in D$, $c_p \in C$ e $d_j \in c_p$. Caso contrário, o par recebe valor 0. Em outras palavras, o processo de classificação extrai o conhecimento representado em Ω, ou seja, o algoritmo aprende com os exemplos dados pelo especialista e busca classificar corretamente, nas classes de C, os documentos de classificação desconhecida (SEBASTIANI, 2002; BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 291).

Existem dois tipos de classificação, a *Single-Label* (único rótulo) e *Multi-Label* (Múltiplos Rótulos). No primeiro caso, cada documento a ser classificado pertencerá apenas a uma única classe; no segundo caso, um documento pode pertencer a mais de uma classe. (SEBASTIANI, 2002).

Avaliação

Uma vez construído o classificador, será necessária uma efetiva avaliação de seu desempenho. Para realizar a avaliação, o conjunto Ω, formado por documentos de classificação conhecida, é dividido em duas partes, não necessariamente de tamanhos iguais, são elas (SEBASTIANI, 2002):

- um conjunto de treino $Tr = \{d_1, ..., d_{|Tr|}\}$, do qual o classificador construído extrairá o conhecimento a respeito do conjunto;

- um conjunto de teste $Te = \{d_{|Tr|+1}, ..., d_{|\Omega|}\}$, utilizado para testar efetivamente o classificador. Cada documento $d_j \in Te$ é passado ao classificador, que baseado nos exemplos dados pelo especialista, no conjunto de treino Tr, atribui 0 ou 1 para cada par $\{d_j, c_p\}$. Na prática, a avaliação do classificador vem da quantidade de vezes em que o par $d_j \in c_p$ recebe valor 1 tanto para o classificador automático, quanto para o especialista.

O método de avaliação descrito nos parágrafos anteriores é chamado abordagem *train-and-test*. Um método alternativo é o *k-fold cross-validation*. Neste método, k distintas tarefas de classificação $\Phi_1, ..., \Phi_k$ são criadas e o conjunto Ω é dividido em sub-conjuntos $Te_1, ..., Te_k$. Isto feito, um processo iterativo do método *train-and-test* pode ser realizado, em que cada tarefa Φ_j utiliza a partição Te_i como conjunto de teste, e os demais como treino (MITCHELL, 1996; SEBASTIANI, 2002).

kNN (*K-Nearest Neighbors*)

Como já dito anteriormente, uma das partes que torna o aprendizado de máquina possível é sua representação computacional através de uma linguagem formal. Chamamos tal representação de algoritmo. Quanto à tarefa de classificação, o algoritmo kNN é um dos mais presentes na literatura. Ele é considerado um classificador preguiçoso, pelo fato de não construir um modelo de classificação *a priori* (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 299).

O kNN realiza sua classificação baseado nos k "vizinhos mas próximos" (documentos matematicamente similares) a d_j da seguinte forma: utilizando uma medida de similaridade (Seção 2.1.1), (1) os k documentos mais próximos a d_j, pertencentes ao conjunto Ω de documentos pré-classificados são encontrados, (2) e a classe atribuída a d_j será baseada em algum critério relacionado a eles, como a classe que mais se repetir, por exemplo. Na Figura 2.6, de forma análoga a explicação anterior, o círculo em cor verde representa o documento

d_j, a circunferência não tracejada é composta por d_j e seus $k = 3$ "vizinhos mais próximos", um pertencente a classe Retângulo e outros dois a Triângulo. No caso da figura, o kNN classificaria o documento d_j como Triângulo (YANG; LIU, 1999).

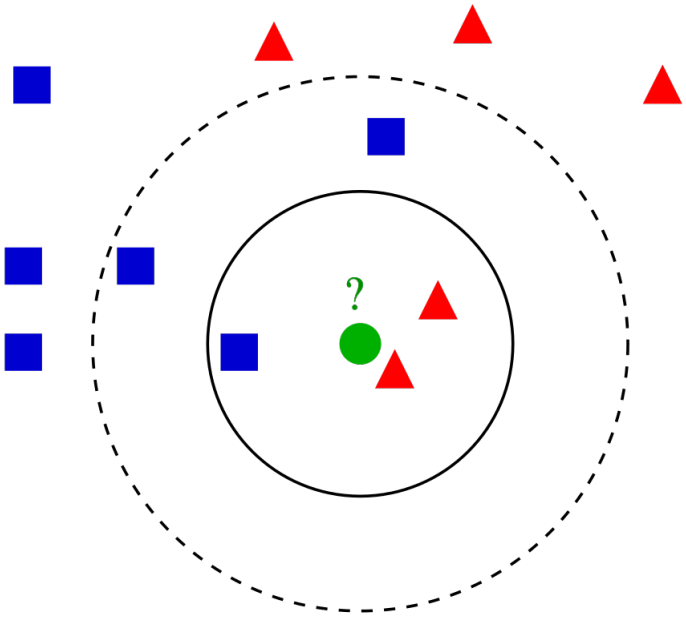

Figura 2.6: Modelo de classificação do kNN.

CBC (Centroid Based Classifier)

O CBC é um classificador baseado nos conceitos do modelo vetorial (SALTON; WONG; YANG, 1975). A ideia por trás do algoritmo é relativamente simples. Para cada conjunto de documentos de uma mesma classe pertencentes a Ω, conjunto previamente rotulado, o centroide é calculado. Sendo assim, L quantidade de classes geram L vetores de centroides $\left\{\vec{C}_1, \vec{C}_2, ..., \vec{C}_{|L|}\right\}$, onde $\vec{C}_i$ é o centroide da i-ésima classe (HAN; KARYPIS, 2000; BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 300).

Um documento x a ser classificado é representado no modelo vetorial e tem os valores de similaridade calculados para todos os centroides. Realizados os cálculos, a classe representada pelo centroide mais próximo de x será atribuída a ele (Equação 2.10) (HAN; KARYPIS, 2000).

$$\arg \max_{j=1,\ldots,L} \left(cos \left(\vec{x}, \vec{C}_j \right) \right) \tag{2.10}$$

Métricas de Avaliação

As métricas de avaliação auxiliam na análise da qualidade de classificação de um conjunto de dados. Normalmente para isso são utilizadas as clássicas medidas *Precision*, apresentada na Equação 2.11 e *Recall*, Equação 2.12.

Nas equações citadas, *True Positive* (TP) representa a quantidade de documentos que foram classificados corretamente, pelo classificador automático, de acordo com o realizado pelo especialista; *False Positive* (FP), a quantidade de documentos que o classificador automático atribuiu a classe C_i, diferente da sugerida pelo especialista; *False Negative* (FN), a quantidade de documentos que o especialista classificou em C_i, porém o classificador automático não (SEBASTIANI, 2002).

A métrica *Precision* mostra a probabilidade de um documentos d_i, aleatório, ser classificado corretamente pelo classificador automático. A *Recall*, por sua vez, indica a probabilidade de que o mesmo documento d_i, que deve ser classificado como C_i, o seja. A fim de obter uma ponderação entre *Precision* e *Recall*, a métrica F_1 (Equação 2.13) pode ser utilizada.

$$Precision(C_p) = \frac{TP(C_p)}{TP(C_p) + FP(C_p)} \tag{2.11}$$

$$Recall(C_p) = \frac{TP(C_p)}{TP(C_p) + FN(C_p)} \tag{2.12}$$

$$F_1(C_p) = \frac{2Precision(C_p)Recall(C_p)}{(Precision(C_p) + Recall(C_p))} \tag{2.13}$$

2.4.2 Aprendizado Não-Supervisionado

O aprendizado não-supervisionado não exige qualquer conjunto de exemplos rotulado por especialistas. O Agrupamento (*Clustering*) se encaixa nessa categoria. Algoritmos de agrupamento buscam descobrir uma forma de dividir um conjunto de dados automaticamente, baseado em um critério pré-definido (WAGSTAFF; CARDIE, 2000; FISHER, 1987; BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 286).

Um processo de agrupamento, normalmente, é formado por um padrão de representação para um conjunto de dados que se deseja agrupar (modelo vetorial); uma métrica de similaridade entres os documentos do conjunto (cosseno); o processo de agrupamento em si (algoritmo de aprendizado); abstração e avaliação dos resultados, quando necessário (JAIN; MURTY; FLYNN, 1999; JAIN; DUBES, 1988).

Agrupamento Hierárquico Aglomerativo

Como o próprio nome sugere, métodos de agrupamento hierárquico criam relações de hierarquia entre os grupos. A hierarquia pode ser gerada a partir de um grande grupo que se decompõem em grupos menores, ou através da aglomeração de grupos definidos anteriormente em um grande agrupamento (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 289).

No processo de agrupamento hierárquico aglomerativo, inicialmente, cada documento do conjunto representa um agrupamento. Em uma árvore que represente a hierarquia entre os agrupamentos, tais documentos representam folhas. Para cada agrupamento deve ser encontrado outro que seja o mais similar a ele no conjunto, cada par de agrupamento forma um novo agrupamento, diminuindo o número de agrupamentos e aumentando um nível na árvore. O processo de fundir pares de agrupamentos ocorre até que todos os documentos estejam em um mesmo agrupamento (BAEZA-YATES; RIBEIRO-NETO et al., 2011, p. 289). A Figura 2.7 apresenta um dendrograma (árvore binária) gerado após o processo de agru-

pamento para um conjunto de 12 documentos. Dendrogramas têm sido usados na literatura para navegação em conjuntos de dados, de forma que o usuário possa influenciar o processo de agrupamento em tempo real.

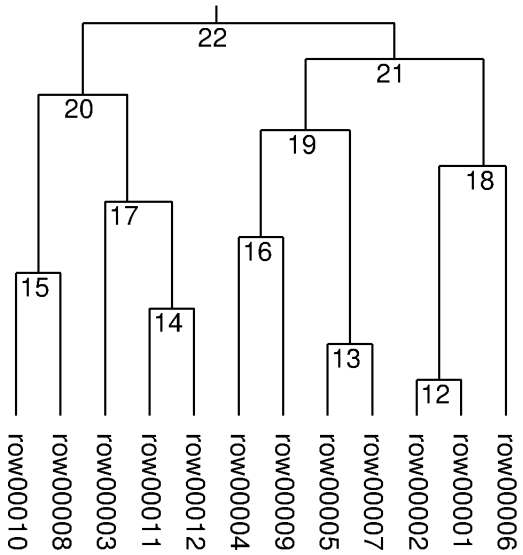

Figura 2.7: Árvore de agrupamentos.

Agrupamento Baseado em Conectividade de Grafo

Algoritmos de agrupamento são baseados na similaridade entre os documentos de um conjunto. No agrupamento baseado em conectividade de grafo, a similaridade entre os elementos é utilizada na formação de um grafo de similaridades, no qual os vértices representam os documentos e as arestas os valores de similaridade (HARTUV; SHAMIR, 2000).

Visualizando o conjunto de documentos como um grafo, os subgrafos existentes representam os agrupamentos de documentos. A fim de identificar tais subgrafos em G, o algoritmo em questão executa os seguintes passos (HARTUV; SHAMIR, 2000):

- calcula o valor de C, número mínimo de arestas a serem removidas, objetivando dividir G em dois sub-grafos H e $\bar{H}$;
- caso G seja considerado um grafo com alta conectividade, ou seja, $C > \frac{n}{2}$, em que n

é o número de arestas de G, então este é um agrupamento existente no conjunto de documentos;

- em caso negativo do item anterior, passos descritos nos dois primeiros itens são realizados para os subgrafos H e $\bar{H}$.

Como pode ser notado, o processo pontuado no parágrafo anterior descreve uma função recursiva. A Figura 2.8 exibe um exemplo da execução desta função. As linhas pontilhadas na figura representam as arestas removidas, separando grafos com baixa conectividade em subgrafos.

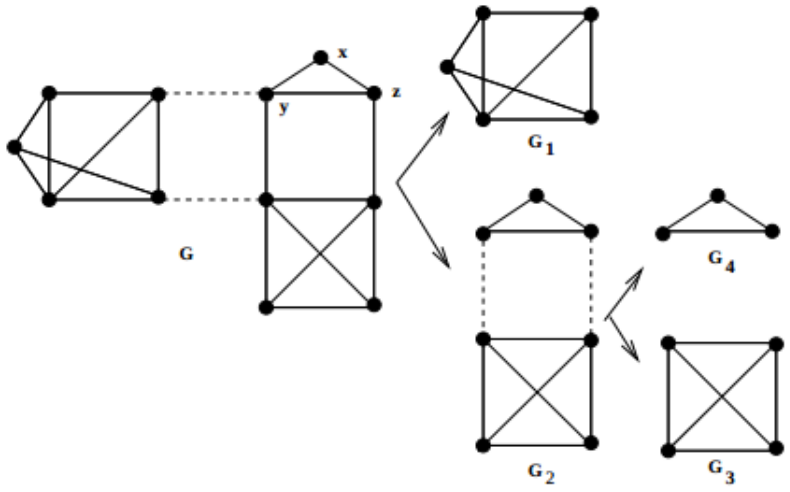

Figura 2.8: Exemplo de aplicação de algoritmo de agrupamento baseado em conectividade de grafos (HARTUV; SHAMIR, 2000).

2.4.3 Aprendizado Semi-supervisionado

Apesar da inegável eficiência dos métodos de aprendizado supervisionado, estes exigem grande porção de dados previamente analisados por especialistas. Classificadores de texto não são diferentes, para que boas taxas, quanto à qualidade de classificação, sejam alcançadas, boa parte do conjunto de dados deve ser classificada manualmente. O problema é que o processo de classificação manual é extremamente custoso, tornando quase inviável fazê-lo.

Algoritmos de aprendizado não-supervisionado são aparentemente uma boa opção para conjuntos de dados com pequena porção analisada previamente por especialistas. Porém, tais algoritmos não farão uso do conhecimento presente na pequena porção previamente analisada, podendo diminuir a qualidade da classificação. Um método de aprendizado que extraia o conhecimento da porção analisada previamente e faça uso de tal conhecimento para influenciar o processo de aprendizado não-supervisionado (agrupamento) é denominado aprendizado semi-supervisionado.

Em uma definição mais formal, seja D_t um pequeno conjunto de dados previamente classificados, e $D_u = D - D_t$ um conjunto de dados de classificação desconhecida, o aprendizado semi-supervisionado utiliza ambos para construção de um modelo de classificação (ZENG et al., 2003).

Esse método de aprendizado é o "coração" do modelo proposto neste trabalho. O modelo faz uso de uma mínima quantidade de documentos possível a ser rotulada por especialistas, e o conhecimento presente nessa pequena porção será utilizado na tentativa de identificar as classes da imensa quantidade de dados com classificação desconhecida.

Capítulo 3

Trabalhos Relacionados

Este capítulo tem como objetivo apresentar alguns trabalhos recentemente publicados que abordam temas relacionados à proposta deste trabalho.

3.1 Análise de Dados de Redes Sociais

Um método muito comum para coleta e análise de dados em redes sociais, no caso deste trabalho o Twitter, está ligado a uma cuidadosa seleção de *hashtags* ou palavras-chave (BRUNS; LIANG, 2012).

O experimento realizado em (LEE et al., 2011) fez uso do método citado no parágrafo anterior. No artigo, os autores monitoraram a lista de *trending topic*, composta por pequenas frases, palavras ou *hashtags* mais comentadas no momento. Para cada *trending topic* da lista foi gerado um documento com nome da própria *trending topic*, contendo todos os tuítes relacionados. Um tuíte contendo mais do que uma *trending topic* foi armazenado em todos os documentos referentes a cada *trending topic* nele presente. Das mais de 23000 *trending topics* coletadas, 768 foram selecionadas, aleatoriamente, para compor o conjunto de documentos utilizado nos experimentos.

Analisando os tuítes relacionados às 768 *trending topic* selecionadas, foram identificadas 18 classes, são elas: arte & *design*, livros, caridade & promoções, moda, comida & bebida, saúde, humor, música, política, religião, feriados e datas, ciência, esportes, tecnologia, negócio, tv & cinema, outras notícias, e outros. A classe outras notícias está relacionada a tuítes de notícias sem classe definida. A classe outros recebe todos os tuítes que não se encaixam em nenhuma das classes definidas. O conjunto de documentos rotulado manualmente é passado como conjunto de treino para para classificação automática dos documentos. No melhor caso, o classificador atingiu 70,96% de acurácia.

Apesar da clara diminuição de esforços provida pela metodologia proposta em (LEE et al., 2011), classificar tuítes baseado em *trending topic*, em alguns casos, pode não ser uma boa opção, pois uma *hashtag* ou palavra-chave pode sofrer mudança de contexto (HADGU; GARIMELLA; WEBER, 2013).

Pode-ser encontrar em (BORTOLON; REGATTIERI; MALINI, 2013) e (REGATTIERI et al., 2014) exemplos de análises realizadas por especialistas com dados extraídos do Twitter. Em (BORTOLON; REGATTIERI; MALINI, 2013), procura-se entender o fenômeno que ocorreu nas redes sociais causado pela exibição da telenovela "Avenida Brasil". A base de dados capturada foi formada por mais de 150 mil tuítes e, através da análise destes, buscaram entender quais perfis de usuários lideravam a rede durante o fenômeno. Em (REGATTIERI et al., 2014), o objetivo foi mapear a controvérsia ao redor do Marco Civil da Internet no Brasil. A base de dados capturada foi formada por 21997 tuítes.

O trabalho realizado em (AGARWAL et al., 2011) se assemelha aos anteriores, o objetivo neste foi analisar o sentimento das pessoas com relação a um dado produto, classificando a base de tuítes capturada em Positivos, Negativos ou Neutros. Um dos pontos interessantes do artigo foi o uso dos chamados *emoticons* (desenhos de expressões faciais) como uma das características dos tuítes. Os *emoticons* foram expressados em palavras que definiam a expressão desenhada através de um dicionário previamente criado.

O trabalho desenvolvido em (RECUERO; BASTOS; ZAGO, 2014) analisa a cobertura

dos protestos ocorridos no Brasil em 2013. Um total de 2852 tuítes foram capturados de dez perfis do Twitter pertencentes a veículos jornalísticos brasileiros. Esses tuítes tiveram seus conteúdos analisados com objetivo de identificar tuítes relacionados à violência ocorrida nos protestos, além de os dados oficiais a respeito de mortos, feridos e presos com o divulgado oficialmente após os atos. Os resultados mostraram divergências entre as fontes oficiais e a impressa no Twitter com amplificação da violência por parte dos perfis jornalísticos devido ao foco dado a esse assunto.

3.2 Classificação Baseada em Agrupamento

Como mencionado anteriormente, conjuntos de dados de redes sociais podem ser coletados de forma relativamente fácil, rápida. Devido a grande quantidade em que os dados são coletados, a classificação manual é quase impraticável.

A classificação automática (Seção 2.4.1), possível solução para classificação de grandes massas de documentos de modo rápido e barato, exige grande quantidade de dados analisados previamente por especialistas (conjunto de treino). A criação desse conjunto exije menor esforço do que a classificação completa da base, assim como menor custo.

Diferente da classificação, o agrupamento (Seção 2.4.2) não depende de um conjunto de treino. No entanto, não separa os grupos de documentos baseado no conhecimento do especialista humano. Tal conhecimento é importante para tornar o resultado do processo automático o mais próximo possível ao realizado manualmente.

A pesquisa realizada em (VENS; VERSTRYNGE; BLOCKEEL, 2013) propõe um método semi-supervisionado combinando o processo de agrupamento com pequena porção de classificação conhecida. O CLUE (do inglês *Clustering Using Examples*) recebe como entrada um conjunto de documentos D, e E, um pequeno conjunto de documentos classificados, sendo que $E \subseteq D$; no dendrograma gerado pela clusterização hierárquica, o método CLUE

busca um intervalo de partições que melhor reconstrua a divisão de classes existente no conjunto E.

Um outro trabalho relevante para esta pesquisa é encontrado em (DUARTE; FRED; DUARTE, 2013). No artigo, os autores propõem um método de agrupamento no qual o especialista influencia em tempo real no resultado. A lógica do método é de fácil percepção estatística. Um par de documentos mais dissimilares (Seção 2.1.1) pertencentes a cada grupo, e outro de documentos mais similares de diferentes grupos são selecionados; isto feito, o especialista é questionado se cada par deve pertencer a um mesmo grupo. O objetivo do método é criar restrições que influenciem no processo de agrupamento. Os resultados mostraram que o uso de tais restrições torna o resultado do agrupamento mais próximo do desejado.

O CBC do inglês *Clustering Based text Classification* (apesar da semelhança no nome não é o algoritmo de classificação citado na Seção 2.4.1), classificação baseada em agrupamento, proposto em (ZENG et al., 2003), tem como objetivo aumentar a acurácia do processo de classificação.

Para conjunto de dados com pequena porção de documentos previamente classificados, a metodologia do CBC proporciona a evolução da qualidade de classificação. As três imagens exibidas na Figura 3.1 ilustram o efeito, no resultado de classificação, proporcionado pelo método, para um conjunto de documentos com apenas duas classes, representados pelos pontos em preto e cinza. Os sinais de + e - representam os documentos de classificação conhecida.

A linha não tracejada na primeira imagem (Figura 3.1(a)), mostra como o classificador classificou os documentos. Na mesma imagem, a linha tracejada representa a classificação ideal. Pode ser facilmente notado que vários documentos foram classificados erroneamente, devido à pequena quantidade de documentos do conjunto de treino.

Na segunda imagem (Figura 3.1(b)), visando aumentar a quantidade de documentos do conjunto de treino e, consequentemente, a qualidade da classificação, o processo de agrupa-

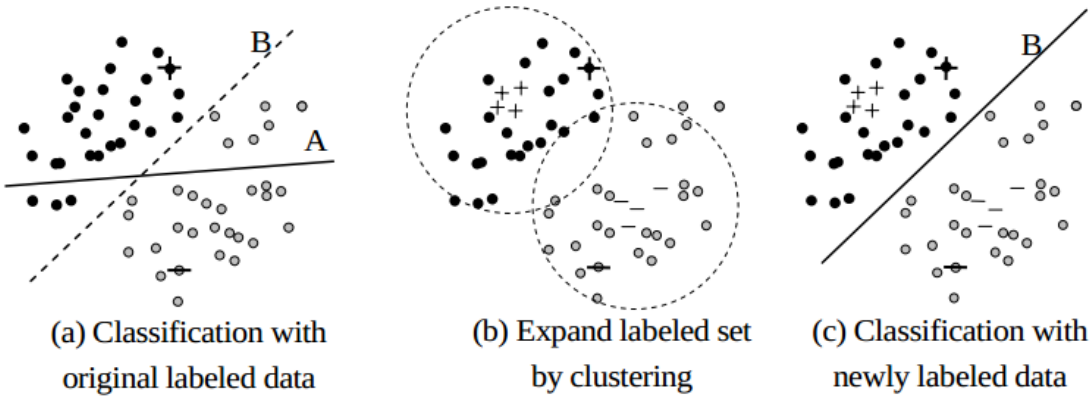

Figura 3.1: Ilustração da classificação baseada em agrupamento (ZENG et al., 2003).

mento é realizado com todos os documentos do conjunto. Para cada grupo de documentos G_i é calculado o centroide C_i. Os k documentos mais próximos a C_i são classificados na classe mais frequente entre os documentos de classificação conhecida, agrupados em G_i.

A terceira imagem (Figura 3.1(c)) ilustra o resultado ótimo de classificação, atingido com o crescimento do conjunto de treino, proporcionado pelo método. Os resultados dos experimentos mostraram que um pequeno aumento do conjunto de documentos classificados, realizado pelo CBC, já proporciona significativa melhora da qualidade de classificação.

Capítulo 4

O Modelo

Como mencionado no Capítulo 1, o objetivo desta pesquisa é: encontrar um método de trabalho que diminua o esforço do especialista no processo de classificação de grandes massas de dados. A solução proposta para atingir o objetivo citado é um modelo de trabalho combinando técnicas de recuperação da informação e o conhecimento humano especializado. Neste capítulo, este modelo de trabalho é apresentado e discutido.

4.1 Combinando Agrupamento e Classificação

Dado um conjunto de documentos D, em que cada documento d_i tem classificação desconhecida, a solução proposta neste trabalho quer explorar o conhecimento existente de uma pequena porção Ω, extraída de D, rotulada pelo especialista. O conhecimento extraído de Ω proporciona a diminuição do esforço no processo de classificação, pois auxilia na descoberta automática da classe de muitos documentos presentes na base.

A Figura 4.1 apresenta os passos e fluxo da solução proposta. O modelo de trabalho funciona do seguinte modo:

- um conjunto de documentos D passa pelo processo de agrupamento. Nesta primeira

versão do modelo, o algoritmo de agrupamento é baseado em conectividade de grafos (Seção 2.4.2). A escolha por tal algoritmo é devido a sua característica de decidir, baseado na quantidade de subgrafos fortemente conectados, a quantidade de grupos em que D deve ser dividido;

- o próximo passo é a escolha dos documentos a compor o conjunto Ω. Para tal são selecionados dois documentos dentro de cada grupo gerado pelo processo de agrupamento. A fim de escolher bem o par a representar cada grupo, maximizando a certeza de que os documentos de um mesmo grupo tenham o mesmo rótulo, são selecionados os dois documentos mais dissimilares (2.1.1) entre si;

- realizada a seleção de Ω, o especialista é questionado a respeito dos rótulos (ou classes) de cada par de documentos mais dissimilares. Caso os dois documentos do mesmo par recebam o mesmo rótulo, todos os documentos do grupo ao qual o par pertence também o recebem;

- o processo é iterativo e reagrupa $\hat{D}$, conjunto formado por todos os documentos pertencentes a grupos que o especialista não tenha atribuído mesmo rótulo ao par mais dissimilar;

- o processo iterativo é executado até que todos os documentos de D sejam rotulados, ou não seja mais possível rotular qualquer grupo de documentos. Ocorrendo o segundo caso, o especialista deverá rotular todos os documentos pertencentes a $\hat{D}$.

4.2 Agrupando por Fatores

O modelo descrito na seção anterior é flexível, ou seja, pode ser utilizado com diferentes combinações de técnicas de recuperação de informação. Sendo assim, uma versão do modelo que faz uso da indexação por semântica latente (Seção 2.3) é proposto.

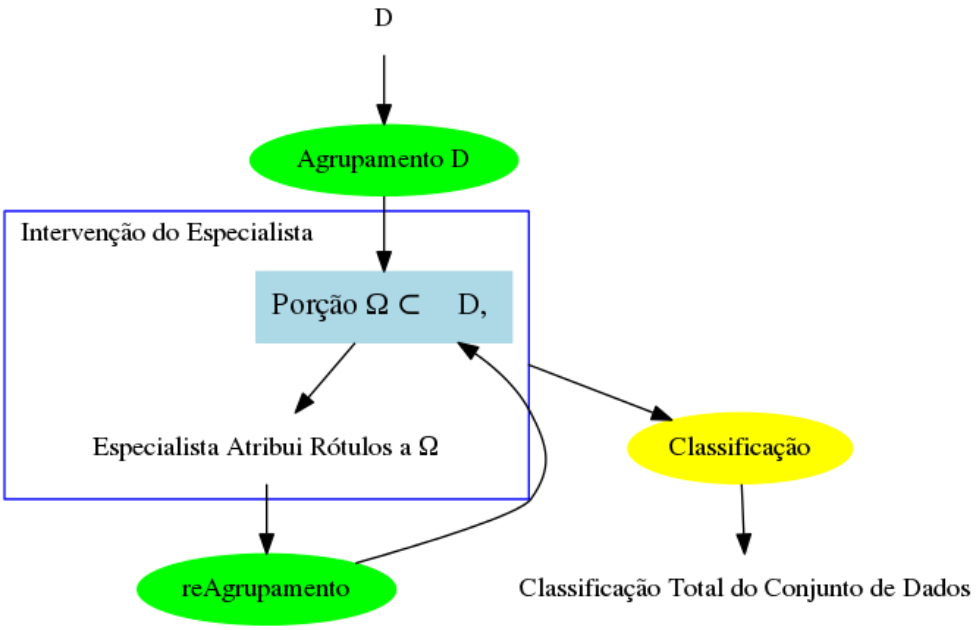

Figura 4.1: Primeira versão do modelo proposto.

A Figura 4.2 apresenta cada passo dessa versão do modelo. Quase todo o processo de funcionamento do modelo descrito na seção anterior se aplica a essa nova versão, portanto, apenas as alterações serão pontuadas aqui. São elas:

- antes do processo de agrupamento, um conjunto de documentos D submetido ao processo sofrerá a transformação matricial SVD (Seção 2.3). Através desse processo os documentos são representados como vetor de fatores (Seção 2.3) $d_i = \{f_1, ..., f_n\}$;
- um outro ponto divergente está relacionado ao método de agrupamento. Cada grupo é formado por documentos em que o fator f_i tem maior peso, ou seja, todos os documentos em que o fator f_1 tiver valor maior que os outros fatores formarão um grupo. Esse processo ocorre para i quantidade de fatores obtidas pela transformação SVD;
- a cada iteração do processo, o conjunto $\hat{D}$, citado na seção anterior, sofre novamente a transformação SVD. A transformação a cada iteração é necessária para que os fatores estejam baseados apenas nos índices dos documentos restantes no conjunto.

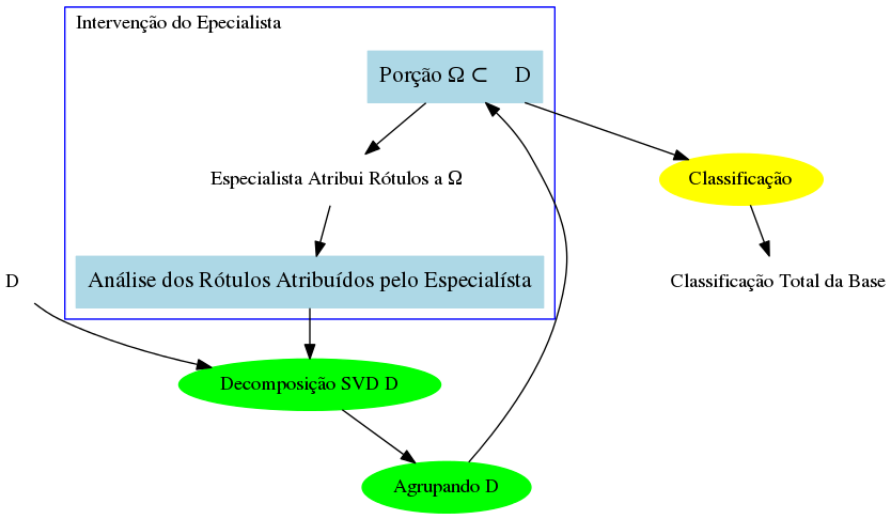

Figura 4.2: Segunda versão do modelo proposto.

4.3 Os Efeitos da Clusterização Hierárquica

Uma interessante versão do modelo é obtida realizando o agrupamento de modo hierárquico (Seção 2.4.2), e explorar o conhecimento a respeito do conjunto de dados proporcionado pelo dendrograma (Seção 2.4.2).

A Figura 4.3 apresenta graficamente o método de trabalho dessa versão. Mais uma vez, devido a semelhança entre as versões, apenas pontos divergentes entre elas serão pontuados. São eles:

- o processo de agrupamento é realizado com algoritmo hierárquico aglomerativo. O resultado do agrupamento proporciona a construção de um dendrograma que mostre as relações hierárquicas entre os grupos de documentos;
- o dendrograma pode ser visto como uma árvore binária em que a cada passo do processo iterativo no modelo um nível é analisado, neste caso cada nó da árvore é entendido como um grupo composto pelas folhas (que representam os documentos na árvore) que possuam alguma relação hierárquica aos nós de um nível qualquer da árvore;

- cada nó onde o par mais dissimilar de documentos (folhas) não receber o mesmo rótulo do especialista terá seu nível imediatamente abaixo explorado, ou seja, os próximos dois nós hierarquicamente ligados a ele serão assumidos como agrupamentos de documentos e, consequentemente, terão seu par mais dissimilar analisado pelo especialista. Caso contrário, o conjunto de documentos hierarquicamente relacionados ao nó em questão serão rotulados e retirados do processo.

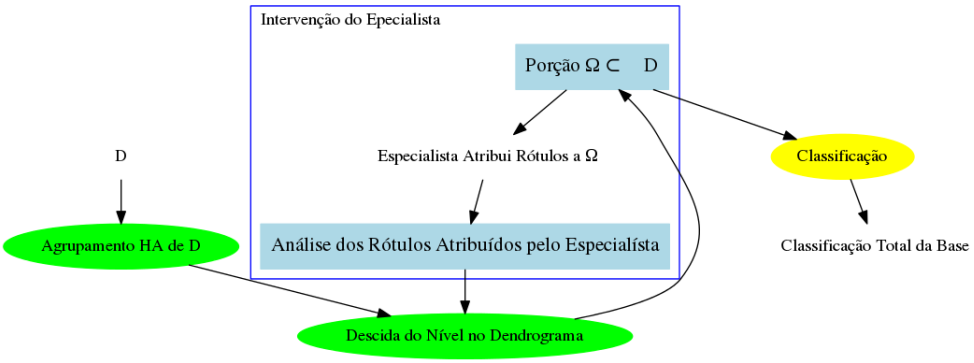

Figura 4.3: Terceira versão do modelo proposto.

Capítulo 5

Experimentos e Resultados

Este capítulo tem como objetivo apresentar os resultados obtidos para as versões do modelo proposto no Capítulo 4. Serão apresentadas as bases de dados utilizadas para avaliação da proposta; os experimentos realizados, bem como as ferramentas usadas para sua execução. Por fim, algumas especulações e discussões a respeito dos resultados serão apresentadas.

5.1 Bases de Dados

As bases de dados utilizadas neste trabalho foram coletadas do Twitter entre Agosto de 2012 e Dezembro de 2013, somando uma quantia total superior a 30000 (trinta mil) tuítes relacionados ao projeto de lei popularmente conhecido como Marco Civil da Internet no Brasil. Removendo todos os tuítes idênticos ou que (por algum motivo) não podiam ser lidos, foi obtida uma base de dados composta por 2080 tuítes.

Cada tuíte, muito antes da realização dos experimentos aqui apresentados, foi manualmente classificado em duas meta-categorias, que são conjuntos de classes relacionadas a uma mesma visão da base). A primeira foi denominada Posição Política, composta por três classes: Conservador, Neutro e Progressista. Como já sugerido pelos nomes das classes, tuítes

em que não estivesse claro o apoio ou rejeição ao Marco Civil foram classificados como Neutros. Os que não apoiassem as mudanças foram considerados Conservadores e Progressistas foi a classificação dada aos tuítes em que estivesse claro o apoio ao Marco Civil.

A segunda meta-categoria foi denominada Opinião. Essa meta-categoria tem cada tuíte atribuído a uma das nove classes que a compõem, são elas: Alerta, Antagonismo, Apoio, Complacência, Explicação, Indignação, Informação, Mobilização e Observação. A descrição de cada uma das classes pode ser encontrada no campo Conteúdo das Tabelas A.1 e A.2.

Os resultados para as meta-categorias serão apresentados como Marco Civil I e II para Posição Política e Opinião, respectivamente. A Tabela 5.1 apresenta os resultados de caracterização das bases. A métrica *razão* apresenta um alto valor, indicando que as classes estão espacialmente sobrepostas, o que torna mais difícil o aprendizado por parte das técnicas de aprendizado de máquina. Quanto mais próximo de 0 o valor dessa métrica, ou seja, quanto maior a similaridade entre os documentos de uma classe e sua centroide (MSDC) e menor a similaridade entre os pares de centroides (MSPC), menor será a probabilidade de erro do algoritmo.

Base de Dados	MSDC (x)	MSCCP	MSPC (y)	*Razão* (y/x)
Marco Civil I	0,561809	0,989217	0,967745	1,722553
Marco Civil II	0,568465	0,962814	0,918002	1,614879

Tabela 5.1: Caracterização dos conjuntos de dados usados nos experimentos.

5.2 Experimentos de Processamento de Documentos

O processamento de documentos (Seção 2.2) pode influenciar na qualidade da classificação automática de um conjunto de documentos, pois busca remover todos os ruídos de

uma base de dados. Além disso, também pode diminuir o tempo gasto para realização da classificação, porque proporciona redução de dimensionalidade.

Uma preocupação inicial deste trabalho foi entender como as ferramentas de processamento de documentos influenciam na classificação das bases utilizadas na avaliação do modelo. O objetivo era descobrir quais delas proporcionariam as vantagens descritas no parágrafo anterior.

Observe as Figuras 5.1 e 5.2, a primeira barra vertical, identificada como " Básica", na legenda de ambas, indica o resultado da classificação quando utilizadas apenas a análise léxica, a frequência de documentos e a ponderação de termos, na fase de processamento de documentos. As outras duas barras, identificadas como "Stopwords" e "Stopwords/Lematização" mostram o resultado da classificação com uso apenas da remoção de *stopwords* e da Lematização com remoção de *stopwords*, respectivamente. Os dois conjuntos de barras presentes nas duas figuras exibem os resultados da classificação com uso dos classificadores CBC e kNN quanto à métrica F_1 obtida para cada um deles.

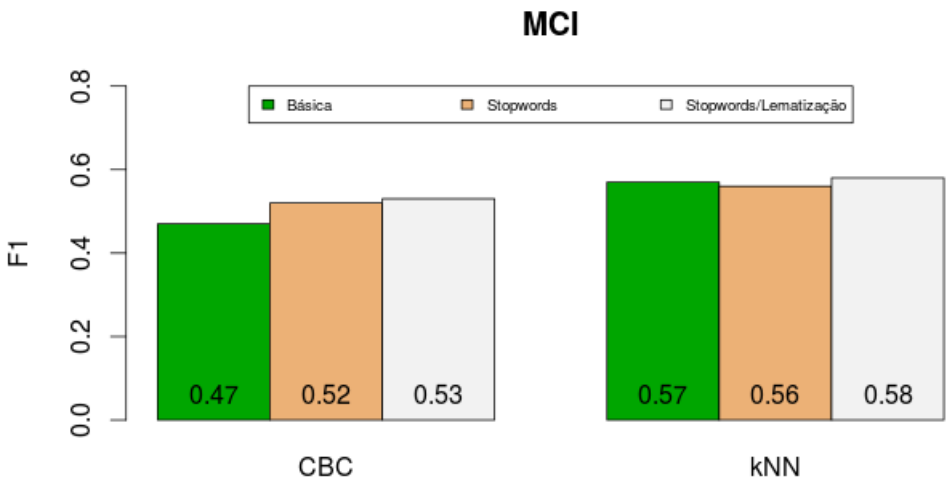

Figura 5.1: Resultados dos experimentos de processamento de documentos para Marco Civil I.

Os resultados obtidos mostram que o classificador kNN apresentou melhor qualidade de

classificação para as duas bases, por isso os resultados apresentados nas próximas seções são referentes ao uso desse algoritmo de classificação. Além disso, o processamento de documentos com remoção de *stopwords* e Lematização proporcionou uma pequena melhora no resultado. Apesar da melhora citada não ser muito significativa em alguns casos, o uso das técnicas de processamento de documento se justifica também pela redução de dimensionalidade que elas proporcionam, tornando mais rápidas as iterações do modelo proposto.

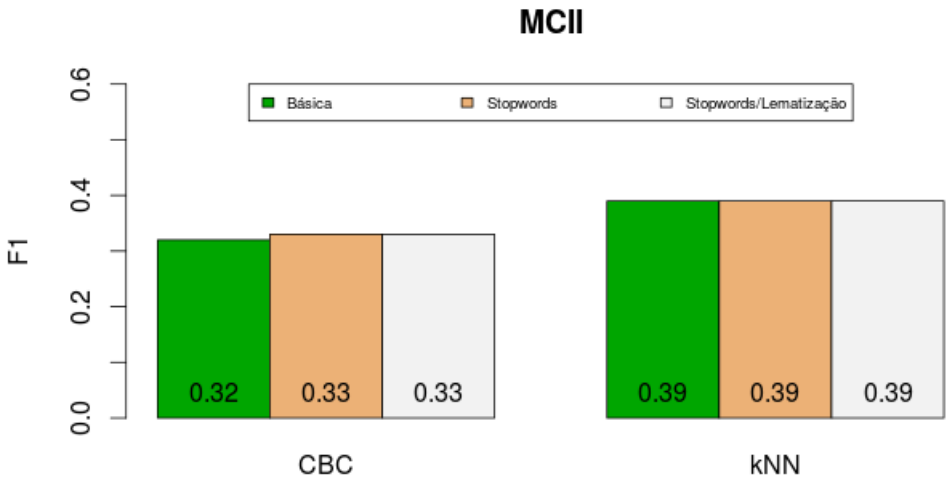

Figura 5.2: Resultados dos experimentos de processamento de documentos para Marco Civil II.

5.3 Experimentos de Agrupamento

A fase de agrupamento tem como objetivo auxiliar o modelo a classificar um conjunto de dados com a menor quantidade de esforço possível do especialista, ou seja, com menor número de documentos classificados manualmente. Nos parágrafos seguintes serão apresentadas as técnicas de agrupamento utilizadas em cada versão do modelo descrito no Capítulo 4. Os algoritmos de agrupamento adotados neste trabalho estão implementados na ferramenta *Clustering Toolkit* CLUTO^{TM} (KARYPIS, 2002).

A primeira versão do modelo faz uso de um algoritmo de agrupamento baseado em conectividade de grafo (Seção 2.4.2). Para a execução do agrupamento é necessário definir um valor entre 0 e 1 para ρ, métrica utilizada para eliminar certas arestas entre documentos que, matematicamente, estão próximos, porém, tendem a estar em grupos de documentos diferentes.

Encontrar o valor ideal para ρ é importante pois este influencia no esforço humano empregado na classificação. Um valor de ρ mais próximo a 1 torna os grupos de documentos mais homogêneos, entretanto pode acarretar em uma grande quantidade de documentos sem agrupamento. Um valor de ρ mais próximo de zero gera agrupamentos com menor homogeneidade, porém, agrupa todos ou a maior parte dos documentos2. A execução do modelo com essa versão se deu para $\rho = \{0,6; 0,75; 0,8; 0,9; 0,95\}$.

A segunda versão do modelo realiza o processo de agrupamento baseado nas cargas fatoriais resultantes da transformação matricial SVD (Seção 2.3). Neste caso, uma quantidade x de fatores devem ser escolhidos para representar os índices da base, cada documento d do conjunto passa a ser representado como $d_i = \{f_1, ..., f_x\}$. Todos os documentos em que f_i apresentar maior carga fatorial pertencerão ao mesmo grupo de documentos.

Semelhante a primeira versão, escolher um valor ótimo para x é importante para o desempenho do modelo. A fim de descobrir o valor ótimo para x, o modelo foi executado com $x = \{c, ..., 50\}$, onde c é a quantidade de classes existentes no conjunto de dados. Sendo x, quantidade de fatores, o que determina a quantidade de agrupamentos, não é de interesse obter uma solução de agrupamento com menor número de grupos que a quantidade de classes que buscamos encontrar, por esse motivo, c é iniciado com o valor da quantidade de classes da base.

Diferente das duas versões anteriores, o processo de agrupamento da terceira versão do modelo não tem qualquer parâmetro a ser ajustado. Baseado no resultado do agrupamento hierárquico aglomerativo de uma base de dados, é construída uma árvore hierárquica (ou dendrograma) completa. A cada iteração do modelo, mais níveis da árvore serão explorados,

iniciando na raiz e descendo até as folhas, caso necessário.

A Tabela 5.2 apresenta o esforço máximo, médio e mínimo gasto pelo especialista no processo de classificação com auxílio de cada versão do modelo. Visto que a terceira versão não necessita de ajuste de parâmetro, seu resultado será sempre o mesmo, por isso possui apenas valor médio apresentado na tabela.

Versão	Esforço Máximo	Esforço Médio	Esforço Mínimo
Total de 2080 Tuítes			
Marco Civil I			
1ª (Grafos)	1888	1329	327
2ª (Fatores)	821	308	40
3ª (Hierárquico)	-	73	-
Marco Civil II			
1ª (Grafos)	2000	1424	538
2ª (Fatores)	1349	867	272
3ª (Hierárquico)	-	880	-

Tabela 5.2: Quantidade de passos máxima, mínima e média exigida por cada versão do modelo.

5.4 Experimentos de Classificação

Os tuítes que formam as bases de dados em estudo, neste trabalho, foram classificados em duas meta-categorias como já dito anteriormente, sendo assim, estes dados podem, então, ser vistos como um problema de classificação *Multi-Label* (Seção 2.4.1). Entretanto, atacaremos este como um problema de classificação *Single-Label*.

Para realização dos experimentos de classificação escolhemos dois algoritmos muito utilizados na literatura: o kNN (Seção 2.4.1) e o CBC (Seção 2.4.1). Esses algoritmos foram

selecionados com base nos objetivos deste trabalho em comparação aos resultados encontrados na literatura.

O uso do kNN requer o ajuste do parâmetro k, número de documentos matematicamente mais próximos. Para tal, executamos a classificação das bases com valores ímpares dentre $k = \{1, ..., 50\}$. A opção por números ímpares objetiva, simplesmente, evitar empates no processo de decisão de classificação. Como mostram os gráficos das Figuras 5.3 e 5.4, $k = 3$ e $k = 1$ retornaram as melhores métricas de classificação para Marco Civil I e II, respectivamente.

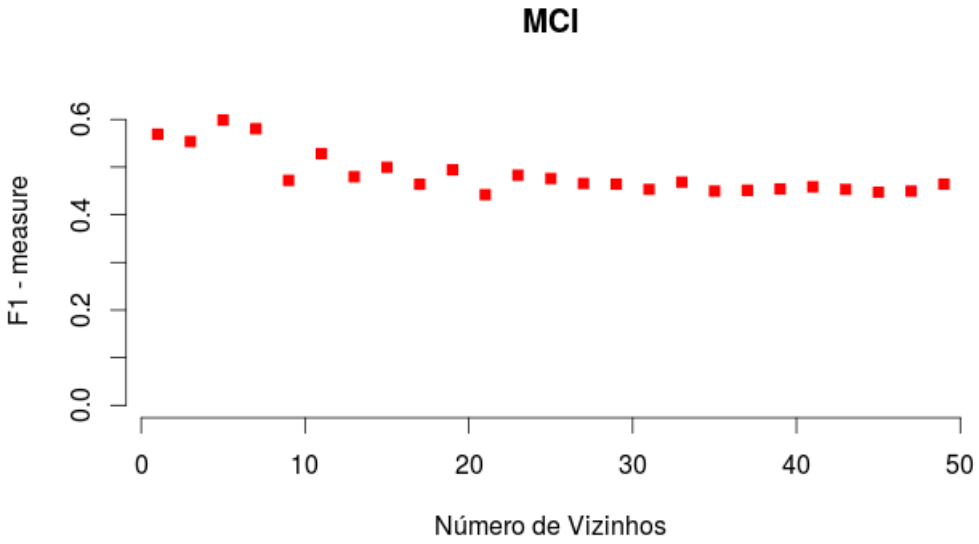

Figura 5.3: Melhor k para Marco Civil I.

A avaliação da classificação é baseada na métrica F_1 (Seção 2.4.1). A fim de validar, estatisticamente, os resultados da classificação automática, o *k-fold cross-validation* (Seção 2.4.1) é utilizado com $k = 3$. Com intuito de melhorar a validação, o processo de classificação é executado 30 vezes para cada k número de vizinhos, sendo a mediana das métricas F_1 obtidas assumida como resultado final da classificação.

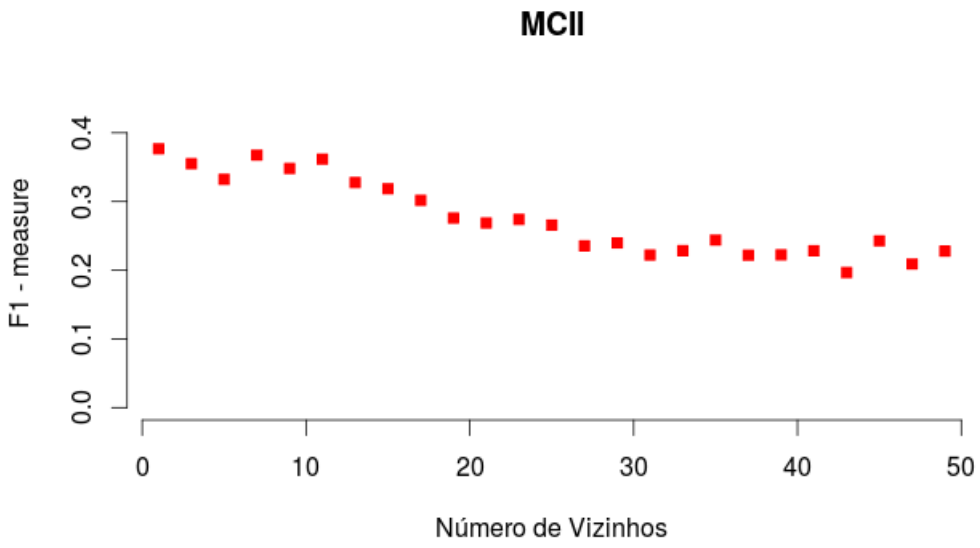

Figura 5.4: Melhor *k* para Marco Civil II.

5.5 Análise de Resultados

A qualidade de classificação do conjunto de dados, resultante da execução do modelo, auxilia na avaliação da eficácia de suas versões propostas neste trabalho. Quanto maior a qualidade de classificação, maior a probabilidade de que os documentos foram classificados corretamente pelo modelo. O gráfico da Figura 5.5 apresenta a qualidade de classificação citada.

Como dito em parágrafos anteriores, a primeira versão do modelo faz uso do agrupamento baseado em conectividade de grafos; a segunda indexa os documentos, com auxílio da análise de semântica latente, e os agrupa, observando as cargas fatoriais de cada documento; a terceira versão faz uso do agrupamento hierárquico aglomerativo e navega, a cada passo da iteração do modelo, os ramos da árvore (ou dendrograma) que representa o resultado do agrupamento.

Cada uma das versões citadas, no parágrafo anterior, foi submetida ao processo de classificação do algoritmo kNN 30 (trinta) vezes. Os resultados apresentados na Figura 5.5 mostram que para Marco Civil I, em média, a primeira versão do modelo apresentou melhor

qualidade de classificação. Um outro ponto importante, quanto aos resultados da figura para Marco Civil I, é que todas as versões retornaram qualidade de classificação aceitáveis, ou seja, métricas $F_1 \geq 0,6$.

Com relação a Marco Civil II, a terceira versão do modelo apresentou melhor qualidade de classificação, entretanto nenhuma das versões apresentou, na média, valor de F_1 aceitável.

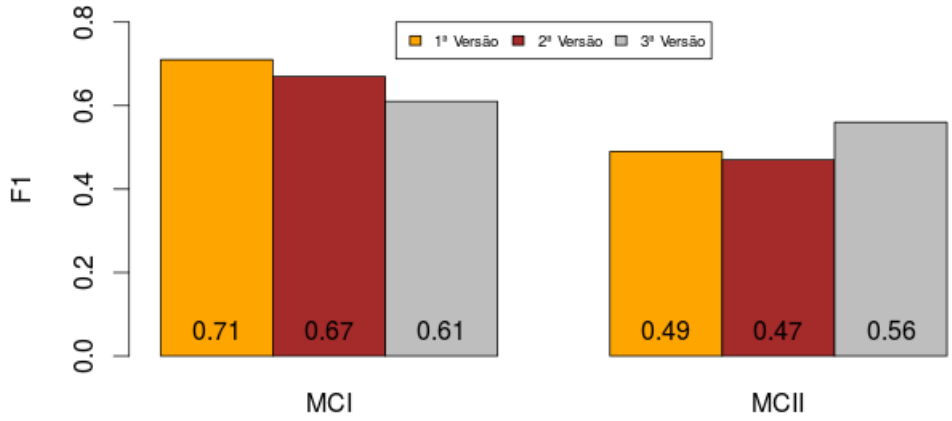

Figura 5.5: Avaliação das versões do modelo quanto a qualidade de classificação das bases de dados.

Outra forma de avaliação pode ser realizada visualizando a variância do conjunto de métricas F_1 obtidas através das 30 execuções da classificação para cada versão do modelo. O gráfico da Figura 5.6 apresenta tais variâncias. Pode-se observar que para ambas as bases, Marco Civil I e II, a terceira versão do modelo apresentou mais estabilidade com relação à qualidade de classificação, ou seja, o valor da pior e a melhor métrica F_1 obtidas estão mais próximas que nos outros casos.

Os dados em negrito, apresentados na Tabela 5.3, representam o melhor modelo em cada quesito avaliado anteriormente. É interessante ressaltar que a terceira versão, para Marco Civil I, apresentou melhores resultados quanto aos quesitos Variância e Esforço, ou seja, essa versão diminuiu o trabalho manual do especialista e obteve maior estabilidade na

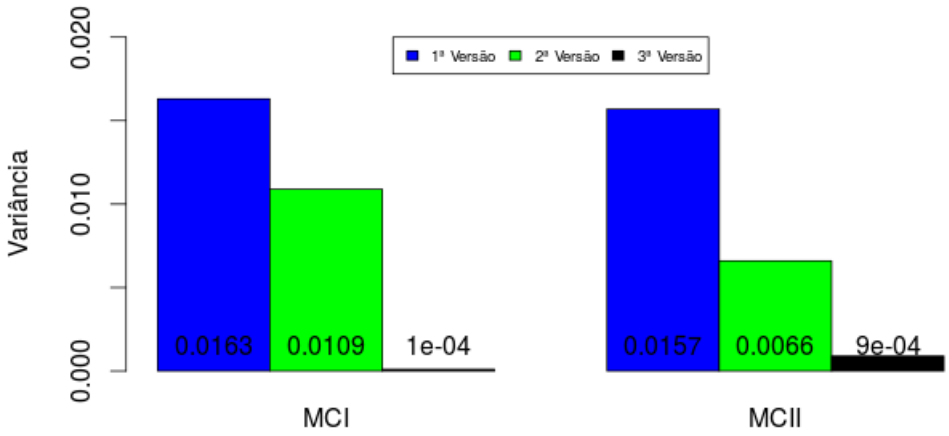

Figura 5.6: Avaliação das versões do modelo quanto a variância dos resultados de classificação das bases de dados.

execução da classificação. Para Marco Civil II, essa mesma versão também foi melhor em dois quesitos, entretanto, desta vez, quanto à qualidade de classificação (F_1) e variância.

A 2ª versão do modelo mostrou-se melhor que as outras apenas no quesito Esforço com relação à Marco Civil II. Assim como a segunda versão, a primeira apresentou melhores resultados apenas no quesito qualidade de classificação (F_1) para Marco Civil I.

Versão	Esforço	F_1	Variância
Marco Civil I			
1ª (Grafos)	1329	**0,71**	0,0163
2ª (Fatores)	308	0,67	0,0109
3ª (Hierárquico)	**73**	0,61	**0,0001**
Marco Civil II			
1ª (Grafos)	1424	0,49	0,0157
2ª (Fatores)	**867**	0,47	0,0066
3ª (Hierárquico)	880	**0,56**	**0,0009**

Tabela 5.3: Tabela de avaliação geral de cada versão do modelo.

Suponha que um especialista necessite, em média, de 30 (trinta) segundos para realizar a classificação manual de um único tuíte. A classificação total das bases Marco Civil I ou II custaria, aproximadamente, 18 horas. A Tabela 5.4 apresenta os tempos necessários para realização da classificação das bases com auxílio do modelo. Pode-se observar que, para ambas as bases, mesmo a versão do modelo que apresentou os piores resultados, quando à diminuição de esforços, proporcionaria uma economia de, no mínimo, 6 horas do trabalho do especialista.

Versão	Esforço	Tempo (horas)
Marco Civil I		
1ª (Grafos)	1329	11
2ª (Fatores)	308	3
3ª (Hierárquico)	73	0,6
Marco Civil II		
1ª (Grafos)	1424	12
2ª (Fatores)	867	7,2
3ª (Hierárquico)	880	7,3

Tabela 5.4: Tabela de avaliação da diminuição de tempo gasto na classificação proporcionada pelo modelo.

5.6 Tabela de Categorias Emergentes

No trabalho realizado em (RECUERO; BASTOS; ZAGO, 2014), os autores apresentam uma tabela de considerável relevância para o trabalho de análise do conjunto de dados em evidência no artigo. A tabela citada objetiva apresentar de modo estruturado as categorias (ou classes) emergentes da análise de conteúdo do conjunto de dados. Os campos que compõem a tabela são:

- Categoria: o nome de cada categoria emergente na análise de conteúdo;

- Conteúdo: este campo apresenta um resumo informado, pelo especialista, a respeito dos documentos que compõem uma determinada categoria, ou seja, o assunto abordado por seus documentos;

- Palavras frequentes: apresenta as palavras que mais se repetem nos documentos de uma determinada categoria;

- Exemplos: apresenta documentos que exemplifiquem bem o conteúdo da categoria.

Submetendo uma base de dados qualquer ao modelo proposto neste trabalho, apenas o campo Conteúdo precisa ser manualmente informado pelo especialista. Dos outros três, automaticamente gerados, apenas o campo Exemplos apresenta maiores desafios para ser obtido. A estratégia utilizada, neste trabalho, para obtê-lo foi selecionar três documentos exemplos de cada categoria, um primeiro documento mais similar ao centroide da categoria, e outros dois mais distantes dentre todos os documentos da categoria.

Os métodos de escolha dos três documentos exemplos proporcionará ao leitor da tabela ampla compreensão do assunto abordado pelos documentos de cada categoria. As Tabelas A.1 e A.2 foram geradas pelo método semiautomático citado no parágrafo anterior para Marco Civil I e II, respectivamente. Os pontos em destaque são: na Tabela A.1 a categoria Conservador possui apenas 1 documento exemplo. Isso ocorreu porque, após a submissão da base ao método, apenas este documento foi classificado nessa categoria; dos exemplos escolhidos, para cada categoria, apenas 5 exemplos foram selecionados erroneamente. O modelo de classificação não é perfeito, ou seja, existe probabilidade de que ele erre ao decidir a categoria de um documento, por esse motivo um documento exemplo pode ser escolhido erroneamente.

Dentre os documentos erroneamente selecionados podem ser encontrados exemplos controversos. Observe os tuítes "*RT @meioemensagem: Facebook, Google e MercadoLivre manifestam apoio ao Marco Civil em carta aberta http://t.co/v7mLpZ0s*" e "*RT @hashonomy_rafa: Google, Facebook e MercadoLivre declaram apoio ao Marco Civi... ♯google*

♯*facebook* ♯*internet (vi* ...", o primeiro foi classificado como Informação em Marco Civil II, entretanto, o segundo tuíte, bem semelhante ao primeiro, foi classificado como Apoio pelo especialista humano. Exemplos como estes prejudicam o aprendizado e, consequentemente, os resultados, tanto da classificação quanto da criação da tabela de categorias emergentes.

A construção semiautomática da tabela de categorias emergentes também diminui o esforço do especialista. Caso necessária a construção manual dessa tabela para grandes bases de dados, o especialista teria dificuldades de escolher bem os documentos exemplos, pois isso pode exigir a leitura de grande parte dos documentos de cada categoria.

5.7 Discussão

Nesta subseção serão discutidos os resultados obtidos com os experimentos apresentados anteriormente; há também a tentativa de entender os motivos que levaram o modelo proposto a retornar resultados não satisfatórios em alguns casos; por último, pensando no processo de continuação do desenvolvimento e aprimoramento do modelo, serão apresentadas técnicas que podem tornar os resultados do modelo ainda mais parecidos com o realizado pelo especialista.

5.7.1 Conjunto Léxico

Como já mencionado anteriormente, os resultados dos experimentos para Marco Civil I se mostraram aceitáveis; diferente do obtido com Marco Civil II, em que o resultado da classificação obtido foi $F_1 < 0,6$. Mesmo com todo o esforço realizado no pré-processamento das bases, algumas características do texto que as compõem podem diminuir a qualidade do aprendizado e, consequentemente, influenciar negativamente a qualidade dos resultados.

Classificar documentos provenientes de redes sociais tem seus desafios: restrição quanto à quantidade de caráteres (no caso do Twitter em 140 caracteres) que torna o texto sucinto,

porém, pobre em relação à informação textual necessária para tornar melhor o aprendizado por parte dos algoritmos; os vários erros gramaticais, uso de gírias, abreviações e diferentes jargões para diferentes eventos; a ampla cobertura do assunto representado pelo tema (VARGA et al., 2014). Esses pontos aliados a subjetividade do processo humano de classificação e, a existência de caracteres acentuados corrompidos nas bases Marco Civil I e II, tornam ainda mais complexo o processo automático de classificação.

As bases Marco Civil I e II, como esperado, apresentam problemas como os descritos no parágrafo anterior. Nos tuítes "*RT @f_trad: Bom dia a todos! Quarta-feira promissora na* ***Camara*** *Federal. Muitos desafios: novo CPC, Marco Civil, Fator Previd., CCJC e prefeitos de MS.*" e "*RT @folha_tec: Votao do Marco Civil da Internet emperra na* ***Cmara*** *dos Deputados. http://t.co/7tnTt8Yx.*" observe como a palavra *câmara* foi escrita em ambos. A diferença na escrita acarretará na diminuição da similaridade entre estes documentos, pois o processo de indexação as compreenderá como duas palavras distintas.

Outro caso importante está relacionado à língua de escrita dos tuítes. Observe os tuítes "*RT @ifikra: the world_s first bill of internet rights http://t.co/uAr4u2Th ♯MarcoCivil ♯Brazil*" e "*RT @igorsoares: Congresso brasileiro erra feio ao aprovar primeiro um projeto criminal, ao inv_s de aprovar o Marco Civil que garante os direitos b_sicos.*", ambos pertencem à classe progressista de Marco Civil I, entretanto, o cálculo de similaridade retornará valor 0 pois seus termos são completamente diferentes, o que pode acarretar no aumento de erros nos processos de agrupamento e classificação executados pelo modelo proposto.

O fato das bases de dados estarem desbalanceadas, ou seja, existirem classes com quantidade de documentos consideravelmente maior que outras, pode provocar a polarização do aprendizado pelas grandes classes (CHAWLA; JAPKOWICZ; KOTCZ, 2004). A Tabela 5.5 apresenta a proporção de distribuição dos documentos em cada classe das bases. Note que a classe Conservador possui apenas 4,3% de todos os documentos pertencentes à base Marco Civil I. Com relação a Marco Civil II, as classes Observação e Informação, juntas, possuem aproximadamente um total de 60% dos documentos da base, sendo 9 o número de classes

que a compõe.

Observe agora a coluna denominada Erro na Tabela 5.5, note que as classes com maior número de documentos obtiveram menor taxa de erro no processo, comprovando que o modelo de aprendizado foi polarizado pelas classes com maior número de documentos.

Classe	Proporção %	Erro %
Marco Civil I		
Conservador	4,3	98,85
Neutro	42,3	57,29
Progressista	53,4	20,21
Marco Civil II		
Alerta	5	44,33
Antagonismo	1,9	40
Apoio	5,9	30,64
Complacência	1,5	48,38
Explicação	10,3	43,57
Indignação	7,4	58,33
Informação	40,6	21,47
Mobilização	9,4	40,47
Observação	18	41,20

Tabela 5.5: Proporção de documentos e erros de classificação para as classes pertencentes as bases de Marco Civil I e II.

5.7.2 Evolução do Modelo

Neste trabalho as três versões do modelo propostas foram colocadas a prova e avaliadas em três quesitos: diminuição de esforço humano, qualidade e estabilidade. Um interessante passo para evolução do modelo é o trabalho em conjunto de suas versões, essa interação pode proporcionar a evolução do modelo em relação a:

- confiabilidade dos resultados: observando documentos que no processo de agrupamento, em todas as versões do modelo, permanecessem juntos, pode-se afirmar que tais documentos tem maior probabilidade de pertencerem a mesma classe;

- qualidade de classificação: com o aumento da probabilidade de um agrupamento ser homogêneo, salientado no item anterior, a qualidade de classificação também tende a evoluir, pois as classes estarão espacialmente mais distantes umas das outras;

- descoberta de documentos com maior probabilidade de confundir o aprendizado: documentos em que as versões do modelo sempre divergem no processo de agrupamento, ou seja, cada modelo o agrupa com documentos diferentes, tem grande probabilidade de confundir o processo de aprendizado, visto que nossos modelos não focam a semântica latente existente no texto dos documentos.

Além da interação entre as versões do modelo, a análise de semântica latente pode proporcionar resultados melhores aos obtidos neste trabalho. Nas versões propostas do modelo, exceto aquela que faz uso da Indexação por Semântica Latente, não são usadas ferramentas e metodologias que objetivem extrair e fazer uso da semântica latente existente no texto, tais técnicas podem proporcionar resultados mais semelhantes ao que o especialista obteria se realizasse a classificação manual das bases de dados.

Representar coocorrências de palavras em grafos é uma das estruturas fundamentais para aplicação de mineração semântica (RACHAKONDA et al., 2014). Muitos pesquisadores têm empregado seus esforços na exploração dessa ferramenta para alcançar seus objetivos em relação a extração de conhecimento semântico em bases de dados de interesse como pode ser visto em (XU et al., 2015), (VARGA et al., 2014), (RACHAKONDA et al., 2014) e (HSU et al., 0).

Muitas vezes, baseados em estudos científicos com respeito a capacidade cognitiva humana, metodologias de mineração semântica usando grafos, têm sido, com sucesso, empregadas em tarefas de busca e indicação de documentos para leitura de notícias, classificação

de micro postagens provenientes de redes sociais, organização de documentos armazenados de modo não estruturado, dentre outros. Essa ferramenta pode proporcionar a evolução da acurácia do modelo proposto neste trabalho e, possivelmente, a diminuição do esforço humano na classificação.

Capítulo 6

Conclusões e Trabalhos Futuros

A análise de grandes conjuntos de dados provenientes de redes sociais gera alto custo, quanto ao esforço humano, se realizado manualmente. Ela exige a leitura e classificação manual de cada documento pertencente ao conjunto. Outros métodos de menor custo como a classificação baseada em *hashtags* ou palavras-chave não analisam a semântica de todo o corpo de texto do documento, tornando tais métodos mais suscetíveis a erros caso um documento possua uma *hashtags* mas trate de um assunto divergente ao representado pela mesma.

Este trabalho propôs um modelo iterativo semiautomático para classificação de dados. O modelo busca diminuir o esforço humano no processo de classificação utilizando técnicas de aprendizado de máquina e estatísticas, em que, baseado no conhecimento extraído de uma pequena porção de dados manualmente classificados pelo especialista, todos os outros documentos são automaticamente classificados.

Após o processo de revisão bibliográfica, levantamento de hipóteses e testes, três versões do modelo combinando técnicas de aprendizado de máquina foram apresentados. Para avaliação de cada versão, submetemos dois conjuntos de dados previamente classificados por especialistas: Marco Civil I e II. Ambas as bases são formadas por 2044 documentos.

As bases citadas sofreram análise prévia sendo submetidas à técnicas de processamento de texto como análise léxica, remoção de *stopwords* e lematização com objetivo de remover ruídos. Em alguns casos, tais bases também foram submetidas ao LSI (*Latent Semantic Index*) buscando redução de dimensionalidade e melhor exploração da semântica do texto.

A versão do modelo que combina agrupamento hierárquico e classificação retornou melhores resultados, sendo assim, foi foco da discussão no trabalho. Os experimentos realizados mostraram que o modelo é solução viável para a problemática em questão nesta pesquisa. Para Marco Civil I foi obtido valor mediano de F_1 igual a 0,61, para Marco Civil II, a mediana de F_1 foi igual a 0,56. Quanto ao esforço humano, para Marco Civil I o especialista precisaria rotular apenas 73 documentos, cerca de 3,5% do total de documentos da base, para Marco Civil II apenas 880 documentos, correspondendo a cerca de 42,3% da base.

O resultado da execução do modelo é a construção, também de modo semiautomático, de uma tabela que sumarize as categorias emergentes da análise de conteúdos realizada com auxílio do modelo proposto. Com exceção do campo conteúdo, em que o especialista informa manualmente um pequeno resumo a respeito das classes emergentes, todos os outros campos, incluindo documentos exemplos para representar cada classe, são encontrados automaticamente.

A fim de validar a escolha dos documentos exemplos, um documento de cada classe mais similar ao texto informado pelo especialista no campo conteúdo foi selecionado, além deste, outros dois mais distantes entre todos os documentos de uma mesma classe são também escolhidos.

Visando a melhora do modelo proposto, outras técnicas podem ser adicionadas ao mesmo. No processamento de documentos, um corretor de texto poderia proporcionar alguma melhora, assim como um código capaz de identificar e tratar abreviações. Por serem coletadas em redes sociais, as bases em estudo contém muitos erros gramaticais e abreviações.

O desenvolvimento de um método capaz de encontrar a quantidade de fatores ideal para

indexar um conjunto de documentos qualquer, a fim de unir as vantagens trazidas pelo LSI as obtidas pelo modelo com uso do agrupamento hierárquico, pode também proporcionar uma melhora relevante no resultado do modelo quanto o esforço do especialista e a qualidade de classificação da base.

Apêndice A

Análise de Conteúdos

Tabela A.1: Categorias Emergentes da Análise de Conteúdo - MCI.

Categorias	Conteúdo	Palavras frequentes	Exemplos
Conservador	Tuítes que mostram o temor de alguns usuários em relação à aprovação do Marco Civil da Internet, devido a um suposto fim da liberdade de expressão na rede.	parcial, piratasparana, povo, setores, unicamente	RT @PiratasParana: Matria da Gazeta do Povo extremamente parcial ao escutar unicamente os setores contrrios ao Marco Civil. http://t.co/lIeZ6KWc
Neutro	Tuítes puramente informativos, a respeito do Marco Civil da Internet, que esclarecem o projeto de lei, avisam com respeito às datas de votação e etc.	votado, ser, votao, cmara, internet	RT @atarde: Polmicas adiam de novo votao do Marco Civil da Internet http://t.co/nd4xRD3W internet marcocivil RT @marcosjornpercu: @pauloteixeira13 @CarlosZarattini Segundo quem conhece, dep., est saindo do marco civil da internet A NeutraLIDADE DA REDE. Essas teles.... RT @reporterbrasil: Marco Civil entra na pauta prioritria da Cmara e pode ser votado nesta semana. http://t.co/0tzx6lhk
Progressista	Tuítes de pessoas que apoiam a aprovação do Marco Civil da Internet, no Brasil, e influenciam outros a fazer o mesmo.	google, livre, neutralidade, votao, internet	RT @Tecnocrata: Cmara adia votao do marco civil da internet http://t.co/YWehFyHr MarcoCivil Internet Cibercultura RT @EditorPatentEd: As Feared, Brazil's 'Anti-ACTA' Marco Civil Killed Off By Lobbyists http://t.co/o76qGnvh RT @AVGBrasil: Google, Facebook e Mercado Livre divulgam carta em apoio ao Marco Civil http://t.co/CeD5wabT via @IDGNow

Tabela A.2: Categorias Emergentes da Análise de Conteúdo - MCII.

Categorias	Conteúdo	Palavras frequentes	Exemplos
Alerta	Os tuítes desta classe tinham o objetivo de informar aos interessados os fatos que poderiam influenciar no processo de votação do Marco Civil da internet como conflitos de interesse, bem como as consequências, caso a lei seja ou não aprovada.	rede, querem, neutralidade, teles, internet	RT @Advox: Brazilian Congress and lobbyists kill world's first 'Internet bill of rights', the Marco Civil http://t.co/2mM9cn0H RT @OccupiedMuslim: Brazilian Congress and lobbyists kill world first internet Bill of Rights http://t.co/ATEPJHnz RT @cavallini: hoje hein. deputados por favor votem o marco civil com neutralidade de verdade! http://t.co/jkCxmfbo
Antagonismo	Tuítes que se opõem ao Marco Civil da Internet ou apenas alguns pontos dele.	web, opiceblum, mpf, projeto, internet	RT @idgnow: Convergncia Digital - Internet - Marco Civil: Para Minicom, Internet grande ambiente de negcios http://t.co/zSvMmTPM RT @OpiceBlum: RT @OpiceBlum: MPF aponta falhas no PL que cria marco civil da internet marcocivil direitoeletronico crimeseletronicos http://t.co/3i ... RT @EstadaoLink: Ainda h resistncia ao projeto, diz o relator, Alessandro Molon marcocivil http://t.co/4tjZvHBQ / Qual a novidade?
Apoio	Tuítes de apoio à aprovação do Marco Civil da Internet no Brasil.	mercadolivre, apoio, internet, facebook, google	RT @hashonomyrafa: Google, Facebook e MercadoLivre declaram apoio ao Marco Civi... http://t.co/ioNP3aD6 google facebook internet (vi ... RT @meioemensagem: Facebook, Google e MercadoLivre manifestam apoio ao Marco Civil em carta aberta http://t.co/v7mLpZ0s RT @cbaraodeitarare: Provedores de Internet saem em defesa do Marco Civil e da neutralidade da rede Baro de Itarar http://t.co/Hc7yFOdz
Complacência	Tuítes que reforçam a necessidade da aprovação do Marco Civil da Internet.	incansvel, luta, molon, alessandromolon, internet	RT @Getschko: Todo apoio ao Dep. Molon @alessandromolon em sua incansvel luta pela aprovao do Marco Civil! No esmorecer! RT @Getschko: (reenviando) Todo apoio ao Dep. Molon @alessandromolon em sua incansvel luta pela aprovao do Marco Civil! No esmorecer! RT @rodrigoantao: Emendas so ameaa ao Marco Civil da Internet - http://t.co/ZE4yf49G via @JornalOGlobo
Explicação	Tuítes que ajudam a entender melhor o objetivo do Marco Civil da Internet no Brasil, seja para o bem ou para o mal.	of, votao, lobby, liberdade, internet	RT @Roanna: Marco Civil da Internet: entre o lobby e a liberdade http://t.co/ZQYRkDQ4 Por uma Internet com menos lobby e mais liberdade. RT @RadarParlamento: New Version of Marco Civil Threatens Freedom of Expression in Brazil:... http://t.co/xfd1O8es RT @pauloteixeira13: Cmara pode votar Marco Civil da Internet nesta semana http://t.co/8KWnoxRu MarcoCivil
Indignação	Tuítes que evidenciaram a ocorrência de fatos que desagradaram ao autor.	by, killed, lobbyists, off, internet	RT @coolvibe: Anti Internet: As Feared, Brazil's 'Anti-ACTA' Marco Civil Killed Off By Lobbyists Techdirt http://t.co/WqOmvVZW RT @EditorPatentEd: As Feared, Brazil's 'Anti-ACTA' Marco Civil Killed Off By Lobbyists http://t.co/o76qGnvh RT @portaldailha: Plenrio pode votar marco civil da internet e fator previdencirio http://t.co/xohza6aV floripa news
Informação	Tuítes com objetivo de informar, principalmente, os fatos ocorridos nos órgãos governamentais em relação ao Marco Civil da Internet.	votado, ser, cmara, votao, internet	RT @atarde: Polmicas adiam de novo votao do Marco Civil da Internet http://t.co/nd4xRD3W internet marcocivil RT @intervozes: O Marco Civil da Intervet voltou a pauta na Comisso Especial de deputados! A presso surtiu efeito: a votao em 19/09 ... RT @CaroliCamargos: Matria sobre marcocivil na Agncia Cmara: http://t.co/SKjg9Esm
Mobilização	Tuítes que buscavam motivar os ideais daqueles que apoiaram ou não o Marco Civil da Internet.	depmarcomaia, morrer, caro, livre, internet	RT @biagranja: Caro @DepMarcoMaia, eu quero minha internet livre. No deixe o Marco Civil morrer! RT @stelles13: Marco Civil da Internet: Molon pede Mobilizacao da sociedade para votao http://t.co/5Ru96Xqy PTnaCamara RT @cbaraodeitarare: Entidades publicam carta defendendo Marco Civil com neutralidade e sem direito autoral Baro de Itarar http://t.co/ALCCcICQ
Observação	Tuítes que estimulam o leitor a refletir a respeito da ideia nele apresentada.	rights, ser, cmara, votao, internet	RT @observatorio: MARCO CIVIL DA INTERNET - Esquenta disputa pelo controle da internet http://t.co/odEB0ACV RT @uoltecnologia: Votao do marco civil volta pauta da Cmara aps ser adiada por cinco vezes http://t.co/XvQAErLl RT @jcanistrum: Sou favorvel neutralidade da rede. Isso para proteger ningum menos do que o consumidor, aquele que tem que http://t.co/fnqE8v4U

Referências Bibliográficas

AGARWAL, A. et al. Sentiment analysis of twitter data. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Languages in Social Media*. Portland, Oregon, 2011. p. 30–38.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. *Modern Information Retrieval*. 2. ed. Boston, MA, USA: ACM press New York, 2011.

BETTIO, R. W. d. et al. Inter-relaão das técnicas term extration e query expansion aplicadas na recuperação de documentos textuais. Florianópolis, SC, 2007.

BORTOLON, B.; REGATTIERI, L. L.; MALINI, F. L. de L. Avenida brasil: Eu assisti, você assistiu ea rede estava lá. In: *Texto apresentado no XVIII Congresso de Ciências da Comunicação na Região Sudeste. Bauru*. Bauru, SP, Brasil: Intercom, 2013.

BRUNS, A.; LIANG, Y. E. Tools and methods for capturing twitter data during natural disasters. *First Monday*, v. 17, n. 4, 2012.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 6, n. 1, p. 1–6, jun. 2004. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1007730.1007733>.

DEERWESTER, S. C. et al. Indexing by latent semantic analysis. *JASIS*, v. 41, n. 6, p. 391–407, 1990.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, v. 55, n. 10, p. 78–87, 2012.

DUARTE, J. M.; FRED, A. L.; DUARTE, F. J. F. A constraint acquisition method for data clustering. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Havana, Cuba: Springer, 2013. p. 108–116.

ELLISON, N. B. et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, Wiley Online Library, v. 13, n. 1, p. 210–230, 2007.

FISHER, D. H. Knowledge acquisition via incremental conceptual clustering. *Machine learning*, Springer, v. 2, n. 2, p. 139–172, 1987.

FRAKES, W. B.; BAEZA-YATES, R. *Information Retrieval: Data Structures and Algorithms*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1992.

G1. *Eleições Brasileiras Geraram Quase 40 Milhões de Tuítes, diz Twitter*. 2014. Http://g1.globo.com/politica/eleicoes/2014/noticia/2014/10/eleicoes-brasileiras-geraram-quase-40-milhoes-de-tuites-diz-twitter.html.

HADGU, A. T.; GARIMELLA, K.; WEBER, I. Political hashtag hijacking in the u.s. In: INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE. *Proceedings of the 22nd international conference on World Wide Web companion*. Republic and Canton of Geneva, Switzerland, 2013. p. 55–56.

HAN, E.-H. S.; KARYPIS, G. *Centroid-based Document Classification: Analysis and Experimental Results*. London, UK, UK: Springer, 2000.

HARTUV, E.; SHAMIR, R. A clustering algorithm based on graph connectivity. *Information processing letters*, Elsevier, Salt Lake City, USA, v. 76, n. 4, p. 175–181, 2000.

HAYKIN, S. Neural networks: A compreensive foundation. Prentice-Hall, New Jersey, 1999.

HSU, P.-L. et al. Mining various semantic relationships from unstructured user-generated web data. *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, v. 0, n. 0, 0. ISSN 1570-8268. Disponível em: <http://www.websemanticsjournal.org/index.php/ps/article/view/389>.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. *ACM computing surveys (CSUR)*, Acm, v. 31, n. 3, p. 264–323, 1999.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, MCB UP Ltd, v. 28, n. 1, p. 11–21, 1972.

JUNIOR, J. F. H. et al. *Análise Multivariada de Dados*. Upper Saddle River, NJ, USA: Prentice Hall, 2005.

KARYPIS, G. *CLUTO-a Clustering Toolkit*. Minneapolis, MN, USA, 2002.

LEE, K. et al. Twitter trending topic classification. In: *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*. Washington, DC, USA: IEEE Computer Society, 2011. (ICDMW '11), p. 251–258. ISBN 978-0-7695-4409-0. Disponível em: <http://dx.doi.org/10.1109/ICDMW.2011.171>.

LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, IBM, v. 1, n. 4, p. 309–317, 1957.

MITCHELL, T. Machine learnin g. *McCraw Hill*, 1996.

OLIVEIRA, E. et al. Combining clustering and classification approaches for reducing the effort of automatic tweets classification. In: . Roma, Italy: 6th International Conference on Knowledge Discovery and Information Retrieval, 2014.

ORENGO, V.; HUYCK, C. A stemming algorithmm for the portuguese language. In: IEEE COMPUTER SOCIETY. *String Processing and Information Retrieval, International Symposium on*. The Burroughs, London, 2001. p. 0186–0186.

RACHAKONDA, A. R. et al. Editorial: A generic framework and methodology for extracting semantics from co-occurrences. *Data Knowl. Eng.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 92, p. 39–59, jul. 2014. ISSN 0169-023X. Disponível em: <http://dx.doi.org/10.1016/j.datak.2014.06.002>.

RECUERO, R.; BASTOS, M. T.; ZAGO, G. Narrative and violence: The brazilian autumn coverage on twitter. *Revista Matrizes*, v. 191, p. 191–217, 2014.

REGATTIERI, L. L. et al. Marcocivil: Visualizing the civil rights framework for the internet in brazil. 2014.

REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Barueri, SP, Brasil: Editora Manole Ltda, 2003. 89-114 p.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, Emerald Group Publishing Limited, v. 60, n. 5, p. 503–520, 2004.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995. ISBN 0-13-103805-2.

SALTON, G.; WONG, A.; YANG, C.-S. A vector space model for automatic indexing. *Communications of the ACM*, ACM, v. 18, n. 11, p. 613–620, 1975.

SCOTT, S.; MATWIN, S. Feature engineering for text classification. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. (ICML '99), p. 379–388. ISBN 1-55860-612-2. Disponível em: <http://dl.acm.org/citation.cfm?id=645528.657484>.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, ACM, v. 34, n. 1, p. 1–47, 2002.

SOUCY, P.; MINEAU, G. W. A simple knn algorithm for text categorization. In: IEEE. *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. San Jose, CA, USA, 2001. p. 647–648.

SOUZA, F. P.; CIARELLI, P. M.; OLIVEIRA, E. de. Combinando fatores de ponderação para melhorar a classificação de textos. *Anais do Computer on the Beach*, p. p–32, 2014.

VARGA, A. et al. Linked knowledge sources for topic classification of microposts: A semantic graph-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, v. 26, n. 0, 2014. ISSN 1570-8268. Disponível em: <http://www.websemanticsjournal.org/index.php/ps/article/view/360>.

VENS, C.; VERSTRYNGE, B.; BLOCKEEL, H. Semi-supervised clustering with examples cluster. 5th International Conference on Knowledge Discovery and Information Retrieval, 2013.

WAGSTAFF, K.; CARDIE, C. Clustering with instance-level constraints. *AAAI/IAAI*, v. 1097, 2000.

WALL, M. *Big Data: Are You Ready for Blast-off?* 2014. Http://www.bbc.com/news/business-26383058.

XU, Z. et al. Knowle: a semantic link network based system for organizing large scale online news events. *Future Generation Computer Systems*, Elsevier, v. 43, p. 40–50, 2015.

YANG, Y.; LIU, X. A re-examination of text categorization methods. In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1999. (SIGIR '99), p. 42–49. ISBN 1-58113-096-1. Disponível em: <http://doi.acm.org/10.1145/312624.312647>.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. (ICML '97), p. 412–420. ISBN 1-55860-486-3. Disponível em: <http://dl.acm.org/citation.cfm?id=645526.657137>.

ZELIKOVITZ, S.; HIRSH, H. Using lsi for text classification in the presence of background text. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2001. (CIKM '01), p. 113–118. ISBN 1-58113-436-3. Disponível em: <http://doi.acm.org/10.1145/502585.502605>.

ZENG, H.-J. et al. Cbc: Clustering based text classification requiring minimal labeled data. In: *ICDM*. IEEE Computer Society, 2003. p. 443–450. ISBN 0-7695-1978-4. Disponível em: <http://dblp.uni-trier.de/db/conf/icdm/icdm2003.html\#ZengWCLM03>.