UNIVERSIDADE FEEVALE

JONATHA MARTINS CARDOSO

MINERAÇÃO DE TEXTO EM FONTES DA INTERNET, COM ENFOQUE NA ADMINISTRAÇÃO PÚBLICA E ATIVIDADE POLÍTICA

Novo Hamburgo

2019

JONATHA MARTINS CARDOSO

MINERAÇÃO DE TEXTO EM FONTES DA INTERNET, COM ENFOQUE NA ADMINISTRAÇÃO PÚBLICA E ATIVIDADE POLÍTICA

Trabalho de Conclusão de Curso, apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação, pela Universidade Feevale.

Orientadora: Prof.ª. Dr.ª Marta Rosecler Bez

Novo Hamburgo

2019

AGRADECIMENTOS

Em primeiro lugar, eu agradeço a Deus, por vários motivos. Não apenas por quem Ele é e por aquilo que ele fez e tem feito na minha vida. Por tantas bênçãos e conquistas, principalmente nesta década, desde que vim para Novo Hamburgo, no ano de 2011. Se estou concluindo este bacharelado, após seis anos de estudo, tendo estudado sempre com dedicação e excelência, Ele é o principal responsável. Toda a sabedoria que conquistei ao longo do tempo vem dEle. Toda honra e glória sejam à Ele - somente à Ele. Aliás, se você está lendo esse texto, saiba que Jesus te ama, quer te salvar e tem um plano especial para a sua vida. Entregue-se à Ele!

Agradeço muito à minha esposa, Ana Daniela. Embora ela esteja ao meu lado desde 2017, já na fase final do curso, seu apoio e ajuda foram muito importantes, durante a redação e estudo para este trabalho. Aliás, várias das melhorias que implementei em minha ferramenta partiram de sugestões dela. Sem dúvida, não sei como seria fazer esse Trabalho de Conclusão sem que tivesse uma auxiliadora ao meu lado – e é exatamente o que Deus me deu com Ana, alguém que me ajuda em tudo. Obrigado por todas as vezes que me ajudastes, principalmente na fase final. Te amo para sempre!

Não poderia deixar de agradecer à minha mãe (Nidia), bem como meu irmão Jefferson. Mesmo triste por não ter mais meu pai (Helder) para me aconselhar e me acompanhar nesta fase, tenho certeza que estou honrando a ele com meu esforço – aliás, se estou na área de TI, muito foi por incentivo e ajuda dele. Tenho muita alegria em dizer que meus pais sempre me apoiaram e me incentivaram, e com certeza, chegar ao final deste curso é uma demonstração de orgulho deles para comigo. E, da mesma forma, que possa continuar servindo de inspiração para meu querido irmão, em sua vida acadêmica e profissional.

Quero agradecer também a Câmara Municipal de Novo Hamburgo, onde sou servidor público com muito orgulho, e pude, não apenas ter o apoio de vários colegas, mas a oportunidade de validar este trabalho, e usá-la como minha inspiração para cria-lo. Também quero fazer uma menção honrosa ao Instituto de Previdência dos Servidores de Novo Hamburgo – onde comecei minha carreira como servidor público e, em 2013, pude iniciar este bacharelado.

Também quero agradecer à minha amada orientadora, Profª. Drª. Marta Bez. Não me esquecerei quando em 2017, sem ideia do que fazer no meu TCC, tive uma

boa conversa contigo, e começou a surgir a temática deste trabalho. Foi o suficiente para decidir que serias minha orientadora. E a cada semana que passou de conversa, desde 2018, cada vez mais tive certeza de que fiz a melhor escolha. Terei muito orgulho, para o resto da vida, de dizer que meu trabalho foi orientado pela Profª. Marta. Serei eternamente grato, e posso me considerar, humildemente, um discípulo seu.

Também sou grato por todos os professores que tive até o momento na FEEVALE, por todo o conhecimento adquirido e todas as experiências vividas. Sei que um pouquinho de cada um está neste trabalho, pois tive de usar de quase todas as áreas estudadas para desenvolvê-lo. Mas não posso deixar de mencionar alguns professores em especial: Prof. Roberto Scheid, Prof. Juliano e Prof. Ricardo, que, em várias vezes, me inspiraram e me ajudaram a desenvolver algo importante neste trabalho. Também agradeço aos meus avaliadores, Prof. Daniel Dalalana e, principalmente, Prof. Adriana, pois suas ajudas, críticas e conselhos foram fundamentais para que esse trabalho chegasse ao nível que pude alcançar.

Por fim, agradeço a todos que me auxiliaram, em algum momento, e em maior ou menor nível, nesta minha caminhada.

Obrigado a todos!

“Pois o Senhor é quem dá sabedoria; de sua boca procedem o conhecimento e o discernimento.” ***(Provérbios 2:6)***

RESUMO

Nos dias atuais, é importante que tecnologias como as mídias sociais e as mídias tradicionais sejam utilizadas como fonte de informação. Duas das áreas que necessitam disto são a Administração Pública e a Atividade Política. Embora existam meios computacionais de obter tais informações, como a mineração de texto, eles são desconhecidos ou pouco explorados pelo usuário comum. Assim, é necessário criar uma forma acessível para que se possa usar tais técnicas. Desta forma, o presente trabalho tem como objetivo criar um modelo de extração de informação, a partir de fontes da Internet, baseado em *text mining* e utilizando ferramentas gratuitas existentes, que possibilite gerar conhecimento relacionado a assuntos que sejam de interesse da administração pública e da atividade política. Com relação à metodologia, o trabalho pautou-se pela pesquisa bibliográfica e experimental. Baseando-se na área de *text mining*, foi realizada a pesquisa de ferramentas para todas as fases do processo, culminando em uma rigorosa seleção, bem como no desenvolvimento de um software de análise de dados. O modelo criado tem caráter genérico e é independente de ferramentas, sendo que a partir deste modelo uma representação prática foi criada, utilizando o conjunto de ferramentas selecionadas. Após validação na Câmara de Vereadores de Novo Hamburgo/RS, verificou-se que ele é capaz de extrair informação e conhecimento, tanto para Gabinetes Parlamentares, como para Assessorias de Comunicação. Além disto, por ser genérico, o mesmo também pode ser aplicado em outras áreas do conhecimento.

Palavras-chave: Mídias sociais e tradicionais. Administração Pública. Atividade Política. Mineração de texto. Modelo de extração de informação.

ABSTRACT

Nowadays, it is important that technologies such as social and traditional media be used as source of information. Two of the areas that need this are Public Administration and Political Activity. Although, there are technologies to obtain that kind of information, such as text mining, they are either unknown or little explored by the average user. Thus, it is necessary to create an accessible way in order to use such techniques. Therefore, the present work aims to create a model of information extraction from internet sources, based on text mining and making use of existing free tools, which allows to generate knowledge related to subjects that are of interest to the political administration and political activity. Regarding the methodology, the work was guided by bibliographic and experimental research. Based on the text mining area, tools were researched for all phases of the process, culminating in a rigorous selection, as well as the development of a data analysis software. The model created has a generic character and it is independent of tools, and from this model, a practical representation was created using the selected toolset. After validation at City Council of Novo Hamburgo/RS, it was found that the model can extract information and knowledge, both for Parliamentary Offices, as for Press Advisories. Moreover, because it is generic, it also can be applied in other areas of knowledge.

Keywords: Social and traditional media. Public Administration. Political Activity. Text mining. Information extraction model.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

3D	*Three dimensions*
API	*Application Programming Interface*
CMC	*Computer-mediated communication*
CSE	*Custom Search Engine*
CSV	*Comma Separated Values*
cURL	*Command Line Granddaddy*
GREL	*Google Refine Expression Language*
HTML	*Hypertext Markup Language*
HTTP	Hypertext Transfer Protocol
ID	Identificador
JSON	*JavaScript Object Notation*
NH	Novo Hamburgo
ODS	*Open Document Sheet*
RPG	*Role-playing game*
RS	Rio Grande do Sul
RSS	*Really Simple Syndication*
SQL	*Structured Query Language*
TI	Tecnologia da Informação
TSV	*Tab Separated Values*
UGC	*User-generated content*
UFRGS	Universidade Federal do Rio Grande do Sul
URL	*Uniform Resource Locator*
XLS	*Microsoft Excel Spreadsheet*
XLSX	*Microsoft Excel Spreadsheet XML Format*
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Não há como negar que a Tecnologia da Informação tem influenciado o mundo atual. Uma dessas fortes influências, graças à Internet, são as mídias sociais, cujo crescimento tem se intensificado nos últimos anos, como alternativa às mídias tradicionais – inclusive aquelas que divulgam suas notícias na Internet. Entretanto, ainda existem áreas da sociedade que dão o mesmo nível de importância para ambas as mídias, e até mesmo uma importância maior para as tradicionais. Entre elas, estão a Administração Pública e a Atividade Política – incluindo a atuação de parlamentares.

Ambas as áreas são afetadas por esta evolução. Nota-se que elas valorizam a força das mídias tradicionais, embora estejam cientes da capacidade de alcance das sociais. É fundamental para estas duas áreas utilizarem ambas, não apenas como meio de comunicação e divulgação, mas também como fornecedoras de informações, que podem ajudar na tomada de decisão. A Câmara Municipal de Novo Hamburgo, no Rio Grande do Sul, não está afastada desta situação. Os quatorze vereadores, com o auxílio dos seus gabinetes parlamentares, conhecem o potencial de ambas as mídias.

Um dos problemas para que o uso das mídias – principalmente as sociais – como fonte de informação se intensifique, está exatamente na dificuldade em obter tais informações a partir de conteúdos publicados. Nota-se como motivo principal o desconhecimento da existência de ferramentas, bem como de técnicas computacionais que realizam tal tarefa.

A área da mineração de texto é capaz de auxiliar neste problema, visto que ela possui como foco a busca por informações em textos, que antes eram desconhecidas. Entretanto, existe um desconhecimento por parte do usuário comum, não apenas a respeito desta área, mas também da existência de ferramentas que permitem buscar tais informações – inclusive gratuitas, de uso livre. Constata-se o uso de métodos manuais complexos e morosos para obter informações simples dessas fontes, que, por meio da mineração de texto, por exemplo, seriam facilmente obtidas.

O modelo tradicional de mineração de texto geralmente é representado por meio de um fluxograma, aonde suas etapas são distribuídas. Embora tal modelo possa ser compreendido e aplicado por profissionais da área, é pouco provável que

um usuário comum possa compreendê-lo e, consequentemente, aplicá-lo, na prática, a uma necessidade de busca por informações em textos.

Por este motivo, entendeu-se que era necessário criar um modelo, baseado em *text mining*, mas que apresentasse, de forma mais detalhada, o processo para extrair informação a partir de uma determinada fonte. Embora o foco deste trabalho esteja nas fontes da Internet, principalmente mídias sociais e tradicionais, criou-se um modelo genérico, e que pode ser usado para qualquer fonte.

Desta forma, o objetivo geral do presente trabalho consiste em criar um modelo de extração de informação, a partir de fontes da Internet, baseado em *text mining* e utilizando ferramentas gratuitas existentes, que possibilite gerar conhecimento relacionado a assuntos que sejam de interesse da administração pública e da atividade política[1].

Desde já, cabe aqui salientar que o modelo criado, buscando atender ao seu caráter genérico, não está alicerçado em um conjunto finito de ferramentas. Antes, possibilita que qualquer ferramenta, capaz de executar ao menos uma das etapas da mineração de texto, possa ser utilizada. Para tanto, criou-se uma representação prática do modelo genérico, com base no conjunto de ferramentas que foram testadas e selecionadas neste trabalho. O uso de ferramentas diferentes exige a alteração/recriação desta representação prática, porém não afeta o modelo criado.

Para atingir o objetivo mencionado acima, foram estabelecidos cinco objetivos específicos[2]: 1) identificar se e como são utilizados, pelos governos e políticos, dados obtidos de redes sociais e sites de notícias; 2) investigar de que forma é possível obter dados de mídias sociais e tradicionais, com base nos conceitos de mineração de texto; 3) analisar as ferramentas existentes, que permitem a obtenção dos dados, bem como a sua posterior mineração e extração de informação; 4) criar o modelo de extração de informação, baseado nos conceitos de *text mining*, aplicando

[1] Embora o objetivo geral apresentado no anteprojeto do presente trabalho apresente a expressão “modelo de obtenção de dados”, entendeu-se que o modelo criado não teria apenas o caráter de coletar/extrair dados, mas sim – até por ser baseado em text mining – de realizar todo o processo, cujo foco é o de obter informação e conhecimento. Logo, este é um modelo que extrai informação para obter e gerar conhecimento.

[2] Tendo em vista que o objetivo geral foi reescrito, conforme explicado na nota acima, os objetivos específicos sofreram algumas correções, a fim de atender ao objetivo geral e ao propósito do trabalho.

as ferramentas selecionadas e testadas; e 5) validar o modelo proposto na Câmara Municipal de Novo Hamburgo/RS[3].

Com relação aos objetivos mencionados anteriormente, a pesquisa é exploratória, tendo em vista que foi baseada em pesquisa bibliográfica nas áreas de *text mining*, mídias sociais e tradicionais, bem como o uso destas tecnologias na Administração pública e atividade política. Desta forma, foi possível criar o modelo de obtenção de dados, bem como estudar os resultados obtidos.

Tal modelo foi validado na Câmara de Vereadores de Novo Hamburgo, tanto por Gabinetes Parlamentares, quanto pela Assessoria de Comunicação. O mesmo mostrou-se eficaz para atender ao seu objetivo: extrair informação e gerar conhecimento sobre assuntos de seu interesse.

Com relação à metodologia, o trabalho pautou-se pela pesquisa bibliográfica e experimental. A pesquisa bibliográfica mostra-se essencial para que se tenha contato direto com os textos publicados no meio acadêmico, que tenham relação com os assuntos abordados. Para elaborar o referencial teórico, fez-se uso de artigos científicos, dissertações, teses, bem como documentos técnicos que tenham relação com as áreas abordadas neste trabalho.

Por sua vez, a pesquisa experimental proporcionará o estudo de ferramentas que visam auxiliar na coleta de dados nas fontes da Internet abordadas no presente trabalho, e a posterior análise, demonstrando como a tecnologia pode facilitar a extração de informação e geração de conhecimento, dando suporte à área pública e política. Durante a fase de experimentação, as ferramentas pesquisadas passaram por estudo, testes e análise, e, desta forma, foram incluídas e classificadas de acordo com o modelo criado.

Este trabalho encontra-se dividido em sete capítulos. Além desta introdução, que representa o primeiro capítulo, tem-se o segundo, onde são tratados os temas das mídias sociais e mídias tradicionais, incluindo conceitos, história, evolução e situação atual, bem como a relação entre ambas. No terceiro, está a Administração Pública e a Atividade Política, trazendo a relação da tecnologia e das mídias sociais com elas, a situação atual e perspectivas. O referencial teórico é concluído no quarto

[3] Embora no anteprojeto conste que a validação seria com os parlamentares, observou-se que seria mais vantajoso fazê-la com a equipe do gabinete, que geralmente possui maior envolvimento com tais atividades, bem como com a Assessoria de Comunicação, que interage diariamente com tais mídias.

capítulo, ao falar da mineração de texto, abordando as técnicas existentes, bem como as etapas necessárias do processo.

O quinto capítulo trata as ferramentas de coleta de dados pesquisadas e testadas, tanto para as mídias sociais como para as tradicionais. Já o sexto aborda o desenvolvimento da ferramenta RAnalyzer, expondo os motivos que levaram à sua criação, bem como as suas funcionalidades e forma de disponibilização.

Por sua vez, no sétimo capítulo é apresentado o modelo de extração de informação, incluindo sua representação gráfica, as diferenças com relação ao modelo tradicional, e a representação prática com as ferramentas. Em seguida, o oitavo capítulo expõe a validação do modelo na Câmara de Novo Hamburgo, explicando como ocorreu o planejamento e o transcurso desta validação, bem como a análise dos questionários respondidos pelos participantes.

Por fim, são apresentadas as conclusões do presente trabalho, bem como são abordadas as perspectivas para o futuro, incluindo trabalhos e projetos que podem ser desenvolvidos a partir deste.

2 MÍDIAS SOCIAIS E TRADICIONAIS

O conceito de “mídia” está ligado ao conceito de comunicação, tendo relação direta com a linguagem, a cultura e a tecnologia. Aliás, Perles (2007) salienta que, embora o termo tenha o significado de “informação”, sua origem latina traz a ideia de “comunhão” e “comunidade”. Ou seja, está ligado ao processo de transmitir e recuperar informações em uma comunidade. Com o passar dos anos - e principalmente nos últimos séculos -, tecnologia e comunicação passaram a se relacionar diretamente.

Por sua vez, a evolução das tecnologias tem ocupado um papel fundamental na sociedade. Uma das principais demonstrações disto ocorreu com o desenvolvimento da Tecnologia da Informação, como lembram Nie e Erbring (2002). Principalmente no final do século passado, segundo Hahl *et al.* (2013), o surgimento da Internet é considerado o ápice da evolução tecnológica. E, conforme Giglietto, Rossi e Bennato (2012), as mídias sociais compõem um dos principais agentes desta evolução.

Desta forma, o presente capítulo busca conceituar as mídias sociais e tradicionais, de forma a entender sua origem, evolução e importância, bem como a relação entre elas, e as perspectivas para o futuro.

2.1 MÍDIAS SOCIAIS

O termo “mídia social” surgiu a partir da necessidade de descrever ferramentas que surgiam com a evolução tecnológica, como *blogs*, leitores de notícias, colaboração social e *sites* de perfis. Porém, segundo Kaplan e Haenlein (2010), o conceito de “mídia social” possui uma compreensão limitada do seu significado.

Por exemplo, existem descrições superficiais que apontam que uma mídia é social quando as pessoas podem conversar na Internet, ou que todas as mídias são sociais, incluindo rádio, televisão e jornal. Há, ainda, muitos que entendem que o termo “rotula um conjunto”, composto por termos como “Facebook”, “*status*”, “Twitter”, “*tweet*”, “comentários”, entre outros. Assim, é necessário abordar alguns conceitos e definições, para melhor compreensão do que é “mídia social”.

2.1.1 Conceitos e definições

Para definir as mídias sociais, Kaplan e Haenlein (2010) traçaram uma linha aos conceitos de Web 2.0 e UGC. Segundo eles, as mídias sociais são o "[...] grupo de aplicações para Internet construídas com base nos fundamentos ideológicos e tecnológicos da Web 2.0, e que permitem a criação e troca de Conteúdo Gerado pelo Usuário" (p. 61, tradução nossa). Conforme Kente (2017), elas são as diversas ferramentas eletrônicas existentes, que permitem aos seus usuários a publicação e acesso a informações, bem como a colaboração e construção de relacionamentos.

É necessário entender os três conceitos que possuem relação com estas. Kaplan e Haenlein (2010, p. 60, tradução nossa) afirmam que "[...] uma definição formal do termo requer, primeiro, traçar uma linha a dois conceitos relacionados [...]: Web 2.0 e Conteúdo Gerado pelo Usuário." Outro conceito básico, mas fundamental para a compreensão é "rede social".

Christakis e Fowler (2009, p. 13, tradução nossa) definem que uma rede social é um "[...] conjunto organizado de pessoas, que consistem em dois tipos de elementos: os seres humanos e as conexões entre eles". Embora essa seja uma ferramenta antiga, de acordo com Huberman, Romero e Wu (2008), as redes sociais evoluíram a partir do surgimento das ferramentas de CMC, ou *Computer-Mediated Communication*, que permitem esta interação e comunicação, dando origem às redes sociais digitais. Com isso, passaram a existir também na Internet (RECUERO, 2006).

Cabe salientar que alguns autores tratam o termo "rede social" como um assunto ligado exclusivamente à tecnologia. Por exemplo, Hahl *et al.* (2013, p. 1) afirmam que "Redes sociais são plataformas sociais virtuais [...]". Porém, Lourenço (2011) salienta que o conceito de redes sociais não surgiu com a Web 2.0, nem com as ferramentas de CMC:

> Esta é uma estrutura social existente desde sempre. Todos nós nos movemos em redes sociais – a nossa família, os nossos amigos, os nossos colegas de trabalho, os nossos vizinhos. [...] Transpô-las para um formato digital [...] não implica que, de repente, se fale em redes sociais. Elas sempre fizeram parte da nossa vida. (LOURENÇO, 2011, p. 20).

Murugesan (2007) ressalta o grande passo dado com o surgimento da segunda geração da Internet, chamada de Web 2.0. Surgida no início do atual século, tem como pilares fundamentais a interatividade, a participação e a

colaboração entre os usuários, trazendo não apenas um conjunto de novas ferramentas e tecnologias, mas também novas estratégias de negócios e tendências sociais.

Os usuários, que antes só acessavam conteúdo, tornaram-se parte vital desta nova era. Eles passaram a ter o poder de produzir e receber mais e melhor conteúdo, informação e conhecimento (CHANG; KANNAN, 2008). Isso foi possível graças as características dessa "nova Internet", tais como a criação e atualização de páginas Web com facilidade, interfaces ricas e responsivas, a colaboração entre usuários na criação e modificação de conteúdo, entre outras (MURUGESAN, 2007).

Kaplan e Haenlein (2010), por sua vez, apontam que o conceito de "conteúdo gerado pelo usuário" (UGC, ou *User Generated Content*) descreve as várias formas de conteúdo que estão disponíveis publicamente e que são criadas por usuários – em suma, é o conjunto de todas as formas nas quais as pessoas usam as mídias sociais. Segundo os autores, o conceito deve estar firmado nas seguintes características:

> [...] primeiro, precisa ser publicado ou em um *site* acessível publicamente ou em um *site* de rede social acessível a um grupo selecionado de pessoas; segundo, precisa mostrar uma certa quantidade de esforço criativo; e finalmente, precisa ter sido criado fora de rotinas ou práticas profissionais (KAPLAN; HAENLEIN, 2010, p. 61, tradução nossa),

Os autores salientam que, com a evolução tecnológica, representada pela evolução do *hardware*, pela popularização da Internet de banda larga, pela disponibilização de mais ferramentas para a criação e, principalmente, pelo surgimento de uma geração de jovens "nativos digitais" com conhecimento técnico, a UGC evoluiu ao longo dos anos. Isto exposto, Kaplan e Haenlein (2010) entendem que é possível definir seis tipos de mídias sociais, conforme a Figura 1:

Figura 1 - Classificação das Mídias Sociais

		Presença social / Riqueza de mídia		
		Baixa	**Média**	**Alta**
Auto apresentação/ Auto revelação	**Alta**	Blogs	Sites de redes sociais (ex.: Facebook)	Mundos sociais virtuais (ex.: Second Life)
	Baixa	Projetos colaborativos (ex.: Wikipédia)	Comunidades de conteúdo (ex.: YouTube)	Jogos de mundos virtuais (ex.: World of Warcraft)

Fonte: Traduzido de Kaplan e Haenlein (2010).

- Projetos colaborativos: visam a criação conjunta e simultânea de conteúdo, como as *wikis* e os aplicativos de *bookmark* social.
- *Blogs:* são páginas *web* com foco pessoal. Possuem variações, como diário pessoal ou conteúdo relacionado a uma área específica.
- Comunidades de conteúdo: possuem como foco o compartilhamento de mídia entre usuários, como fotos, vídeos e slides. Nessa classificação encontram-se o YouTube e o Instagram.
- *Sites* de redes sociais: são aplicativos que permitem a conexão de usuários por meio de perfis de informações pessoais. Nessa classificação encontram-se o Facebook e o Twitter – embora eles também trabalhem com conteúdo.
- Jogos de mundos virtuais: mundos virtuais são plataformas que simulam um ambiente em 3D, onde usuários se apresentam com avatares personalizados. Nos jogos, os usuários se comportam conforme regras estabelecidas em um contexto de RPG (*Role-Playing Game*) multijogador.
- Mundos sociais virtuais: os usuários são livres para escolher seu comportamento, de forma semelhante à vida real, sem restrições.

Tendo em vista que os autores consideram que é necessário que a "mídia social" contenha o conceito de *User Generated Content* (UGC), eles desconsideram o *e-mail*, os mensageiros instantâneos, artigos ou textos jornalísticos compartilhados, *blogs* sem conteúdo e comentários, e conteúdos produzidos com foco comercial.

2.1.2 Histórico e evolução

Kaplan e Haenlein (2010) trazem o que pode ser considerado o embrião da mídia social: a *Usenet*, criada no final dos anos 1970 na *Duke University,* que era um sistema de discussão mundial, permitindo aos usuários a publicação de mensagens. No final dos anos 1980, surgiu o *site* de rede social *Open Diary*, oferecendo a oportunidade de reunir escritores de diários *on-line*. Estas ferramentas evoluíram para o que se conhece como *blog*, amplamente utilizado em todo o mundo.

Ao longo da década de 2000 surgiram vários *sites* de redes sociais. O Orkut, criado em 2001 e que se desenvolveu a partir de 2004, após ser adquirido pela Google (RECUERO, 2009). Segundo Hahl *et al.* (2013), tornou-se a primeira bem-

sucedida mídia social, com inúmeras ferramentas de interação, como a criação de perfis, comunidades (grupos), entre outros. Na mesma época surgiu o MySpace, com grande sucesso, devido a maior personalização em comparação ao Orkut. Era possível ter e criar comunidades, além de interagir por meio de perfis, grupos, fotos, músicas etc.

Em fevereiro de 2004 surge o Facebook (KAPLAN; HAENLEIN, 2010). Criado por Mark Zuckerberg, tinha foco em alunos que saiam do ensino secundário que ingressavam na universidade, e que necessitavam de uma rede de contatos. Ele permite aos seus participantes publicar sua rotina, por meio de atualizações de *status*, fotos, vídeos, jogos e *links*, podendo reagir (“curtidas”, por exemplo), comentar ou compartilhar publicações (HAHL *et al.*, 2013). Kente (2017) complementa, afirmando que a rede oferece inúmeras possibilidades, por causa dos seus incontáveis dados armazenados, relacionados a pessoas, grupos, produtos, entre outros.

Surgiu na mesma época um novo serviço que se mostrou, desde o início, um sucesso mundial: o microblog, cujo precursor foi o Twitter, criado por Jack Dorsey, Biz Stone e Evan Williams. O termo “microblog” vem do tamanho curto das postagens, que no início era de 140 caracteres. Sua estrutura é baseada em seguidores, onde cada pessoa escolhe quem deseja seguir, bem como ser seguida por outras. Ele também fornece outros recursos, como o envio de mensagens em modo privado, mensagens direcionadas, construção de um perfil simples etc.

Outra mídia social importante é o YouTube, que surgiu em fevereiro de 2005, como uma ferramenta de compartilhamento de vídeo. Seu sucesso está calcado na sua simplicidade, pois não é necessário possuir conhecimentos de edição e filmagem, mas apenas ter uma câmera qualquer, filmar e publicar no *site*. Sua estrutura permite que qualquer vídeo seja encontrado por meio de uma ferramenta de pesquisa. Além disso, um vídeo pode ser compartilhado quando, onde e para quem quiser, e um usuário pode ainda ter uma página própria de vídeos.

Além destas, outras ferramentas que surgiram como mídias sociais, mencionadas por Recuero (2009), foram o Flickr em 2004 (fotos), o Fotolog em 2002 (fotos), o Plurk (*microblog*), o Foursquare (rede social baseada em localização), o *Second Life* (mundo virtual) e o Google Wave. Isto atesta o que Telles (2011, p. 4) afirmou: “As mídias sociais fazem parte de uma revolução poderosa, influenciam decisões, perpetuam ou destroem marcas e elegem presidentes.”

2.1.3 Uso das mídias sociais

As mídias sociais permitiram que, em menor tempo ainda, um número elevado de pessoas tenha acesso a alguma informação ou notícia, segundo Zhang *et al.* (2017). Como afirmam Wu *et al.* (2017, p. 3062, tradução nossa), “A mídia social é hoje globalmente onipresente e predominante.” Christakis e Fowler (2009) afirmam que as pessoas passaram a ter uma noção de interconectividade. Por isto, diversas áreas utilizam-nas como fonte de informação.

Batrinca e Treleaven (2014) salientam que o mundo dos negócios foi o primeiro a não apenas usar, mas também a analisar os dados das mídias sociais, onde as empresas as usam para divulgação de suas marcas, melhoria de produtos e do atendimento ao cliente, criação estratégias de marketing e publicidade, entre outros. Kaplan e Haenlein (2010) salientam que muitas decisões que as empresas tomam, tem como foco o uso rentável de aplicações como Facebook, Twitter e YouTube.

Zago e Bastos (2013) também trazem a importância das mídias para a publicação de notícias. Embora existam sites de notícias em meios digitais, a divulgação em mídias sociais é uma importante estratégia para a difusão. Embora seja considerado um “canal acessório”, a atividade dos usuários nas redes sociais “[...] tem impacto significativo na difusão das notícias de cada jornal [...]” (ZAGO; BASTOS, 2013, p. 118).

O uso das redes também permite que os usuários possam contribuir diretamente na repercussão das notícias, pois elas permitem replicações (compartilhamentos) e comentários e, desta forma, ampliando a visibilidade das notícias, recebendo mais atenção e audiência e, consequentemente, mais leitura e acesso. Isso porque, como dizem os autores, “[...] o conteúdo, ao ser replicado, pode vir a ser acessado por uma gama maior de usuários e potencialmente encadear novas redes de difusão do mesmo material.” (p. 119)

Outro exemplo de área que tem reconhecido as redes sociais como úteis é a das ciências sociais, segundo Wilson, Gosling e Graham (2012). Mais do que o seu impacto na vida social das pessoas, pesquisadores de diversas disciplinas têm visto nas mídias uma ferramenta para observar comportamento, testar hipóteses e recrutar participantes de diferentes países e grupos demográficos, cujos resultados têm sido cada vez mais publicados em periódicos e anais.

A relação da Administração Pública e Atividade Política com as mídias sociais será abordada de forma detalhada no subcapítulo 313.1.2.

2.2 MÍDIAS TRADICIONAIS

A partir da relação direta que surgiu entre a tecnologia e a comunicação, deu-se origem ao que se chama de “mídia tradicional”, envolvendo principalmente o jornal, a revista, o rádio e a televisão. Embora já se possa ver uma interação entre as mídias tradicionais e as emergentes – principalmente as sociais -, Kavanaugh *et al.* (2012) apontam que profissionais de relações públicas que trabalham nos governos, geralmente se concentram nos meios tradicionais de comunicação, como boletins informativos, *releases*, jornais, televisão e rádio.

Devido a esta preferência que profissionais da área de Relações Públicas - e até mesmo governos - dão às mídias tradicionais, faz-se necessário abordá-las, visto que serão uma das fontes utilizadas no presente trabalho.

2.2.1 Histórico e evolução

De acordo com Perles (2007), não existe uma definição precisa de quando o ser humano começou a se comunicar com os outros. Inicialmente, por meio de sinais e instrumentos. A escrita, por sua vez, surgiu há cerca de 6 mil anos e desenvolveu-se no último milênio. Por sua vez, Zago e Bastos (2013) lembram que a circulação de notícias remonta à Roma Antiga, no final do século anterior à Cristo, quando as primeiras notícias sobre o governo eram divulgadas.

A grande evolução da mídia tradicional ocorreu com o sistema de prensa tipográfica, inventado por volta de 1440. Desta forma, foi possível transformar um processo, antes manual e demorado, em mecânico e ágil, permitindo criar publicações em escala antes inimaginável, possibilitando a difusão do conhecimento e da informação (PERLES, 2007). Os primeiros jornais surgiram no século XVII, na Europa, sendo que no Brasil os primeiros começaram a circular no início do século XIX.

Duas tecnologias que ajudaram na difusão da informação, no século XIX e XX, de acordo com Zago e Bastos (2013), foi o telégrafo, que permitia a transmissão de mensagens por meio de cabos eletromagnéticos, e o rádio, que surgiu no final do

século XIX. Outra mídia tradicional de grande importância foi a televisão, que surgiu no final da década de 1920, desenvolvendo-se na década seguinte (PERLES, 2007).

Segundo o autor, a comunicação por rádio e, principalmente, pela televisão teve mais uma evolução na década de 1960, com os satélites, que permitiam a transmissão entre diversas localidades do mundo, em um curto espaço de tempo. Desta forma, nos anos seguintes, desenvolve-se uma integração entre os meios de comunicação.

Perles (2007) ressalta que o mundo tem passado por várias mudanças, rápidas e intensas, nas últimas décadas – principalmente nos últimos 25 anos do século XX, considerado sem precedentes. Sabe-se que o futuro será ainda mais diferente, graças a revolução tecnológica. Neste sentido, Jambeiro (1998) salienta que, desde a década de 1970, o mundo direciona-se para uma convergência das comunicações com a eletrônica e a informática.

Entretanto, segundo Gomes Jr. (2014), até hoje o uso da tecnologia é visto às vezes com maus olhos na área do jornalismo. Isso impediu, segundo Machado (2003), uma modificação na forma de produção de conteúdo nas organizações jornalísticas. É importante entender que a tecnologia permite um novo formato de jornalismo, atuante em todas as etapas de produção do conteúdo, desde a apuração até a circulação. E é possível compreender como a tecnologia está afetando a mídia tradicional, graças ao surgimento de inúmeras plataformas digitais, que permitem novas formas de produção e consumo de conteúdo jornalístico.

Machado (2003) menciona exemplos de projetos onde movimentos sociais produzem seus próprios conteúdos, sem depender da mídia tradicional para terem exposição; do mesmo modo, veículos cada vez mais utilizam o conceito de “leitor-repórter”, onde o próprio leitor ou usuário pode produzir conteúdo.

Uma vantagem do crescimento dos *sites* de notícias, conforme Fernandes (2018), é o desenvolvimento de um imenso banco de dados de notícias - alimentado por estes *sites* - que está à disposição das pessoas. Zago e Bastos (2013) corroboram e complementam, afirmando que as pessoas cada vez mais utilizam a tecnologia para terem acesso às notícias, por meio de *feeds RSS*, assinaturas *on-line*, ou mesmo no próprio *site* do veículo, acessando quantas notícias desejar.

Graças a essas mudanças, segundo Fernandes (2018), as pessoas consumidoras de notícias têm mudado a maneira como se informam. Nonato, Pimenta e Pereira (2012) ressaltam que a expansão da tecnologia às classes de

baixa renda no Brasil tem permitido o acesso a informações, até então desconhecidas ou inacessíveis. E a tendência, segundo eles, é de que “Quanto mais cedo têm o acesso a esta forma de consumir informação, provavelmente maior será o seu distanciamento das mídias tradicionais [...] (NONATO; PIMENTA; PEREIRA, 2012, p. 2)”

Sem dúvida, as mídias tradicionais têm sido influenciadas pela mídia *on-line*, principalmente relacionadas a cobertura jornalística. Os tradicionais meios de comunicação possuem alcance representativo; porém, o alcance amplificado e os consequentes efeitos da mídia pela Internet demonstram-se mais profundos, graças a quantidade e dinâmica que carrega consigo (HAßLER, MAURER; HOLBACH, 2013).

2.2.2 Relação com as mídias sociais

Fernandes (2018) ressalta a evolução que as tecnologias contemporâneas de informação e comunicação – principalmente a Web 2.0 e o acesso à Internet banda larga – trouxeram às pessoas, principalmente para as plataformas de informação e acesso a notícias. Porém, existe uma diferença notória entre as diferentes mídias:

> O rádio configura-se eminentemente como local, embora possa percorrer longas distâncias. E a televisão, apesar de ter suas produções realizadas nos grandes centros urbanos pelas emissoras matrizes, reserva espaços para produção de programas locais, principalmente aos programas jornalísticos. Os sites de informação e as redes sociais, até pela sua característica de velocidade da informação, tem um alcance do público receptor muito além destes meios tradicionais. (NONATO; PIMENTA; PEREIRA, 2012, p. 5)

Segundo Kavanaugh *et al.* (2012), as mídias tradicionais apresentam uma carência em relação às sociais, pois estas permitem obter informações importantes, como opiniões e pontos de vista diversos, bem como alcançar públicos tradicionalmente menos representados, como os jovens e a população de baixa renda. Esse é um dos motivos, segundo Turcotte *et al.* (2015), para que a mídia tradicional passe por um processo de declínio, visto que há uma abundância de mídias disponíveis – tendo as mídias sociais como grande expoente.

Há pesquisas apontando que cada vez mais as pessoas consomem notícias no Facebook ou Twitter. Outras, por sua vez, mostram que a quantidade de notícias lidas por uma pessoa nas mídias sociais, que foram disponibilizadas e/ou

compartilhadas por seus amigos, correspondem ao dobro das publicadas por veículos tradicionais ou jornalistas, em suas mídias sociais oficiais (GOMES JR., 2014) – ou seja, há mais chance de uma determinada notícia ser acessada por um perfil pessoal em uma mídia social do que por uma página oficial, na mesma rede social.

Um ponto a ser destacado é que a mídia tradicional, conforme Kavanaugh *et al.* (2012), não é capaz de sanar dúvidas, como a localização e frequência da ocorrência de eventos importantes, as diferentes visões de um evento e os usuários influentes em uma comunidade, isto devido ao estilo jornalístico baseado em relatos envolvendo fontes tradicionais. Eles salientam que "A análise profunda de fluxos de mídia social também pode fornecer acesso a segmentos da comunidade que não tem participado nas formas tradicionais" (KAVANAUGH *et al.*, 2012, p. 481, tradução nossa).

Os autores, então, analisam o uso das mídias sociais e tradicionais em um cenário de governança pública – tema que possui relação com o presente trabalho. Eles entendem que:

> Governos, organizações locais e cidadãos continuarão a usar uma combinação de métodos tradicionais de comunicação (por exemplo, jornais, jornais, televisão, revistas, telefones) e ferramentas emergentes, telefones inteligentes e mídias sociais. Os governos sabem que possuem diversas audiências com diferentes necessidades e preferências. As mídias sociais são apenas mais um conjunto de canais de comunicação para divulgar e servir os interesses de cidadãos diferentes (principalmente os mais jovens). Os cidadãos continuarão a usar diferentes mídias para obter e compartilhar informações, não apenas uns com os outros, mas com o governo (KAVANAUGH *et al.*, 2012, p. 489, tradução nossa).

Por isto, de acordo com os autores, acredita-se que tanto os governos quanto os cidadãos continuarão a usar as mídias tradicionais, em combinação com as mídias sociais, bem como *smartphones* e outras tecnologias.

2.3 CONSIDERAÇÕES SOBRE O CAPÍTULO

Este capítulo apresentou as mídias sociais e tradicionais. Com relação às sociais, foram apresentados alguns conceitos e definições, bem como seu histórico e evolução até os dias atuais, e, por fim, seu uso. Sobre as mídias tradicionais, além do histórico e evolução, tratou-se da relação entre mídias sociais e tradicionais.

Pode-se entender que o capítulo buscou dar os subsídios necessários para que seja possível atingir o primeiro objetivo específico, que é o de identificar se e como são utilizados, pelos governos e políticos, dados obtidos de redes sociais e sites de notícias.

Assim, o próximo capítulo tratará da Administração Pública e da Atividade Política, cujo principal foco está em analisar como elas observam e utilizam as Tecnologias de Informação e Comunicação - com destaque para as mídias sociais.

3 ADMINISTRAÇÃO PÚBLICA E ATIVIDADE POLÍTICA

O foco do presente trabalho está no uso de fontes da Internet pela administração pública, e na atividade política – ambos no Brasil. Porém, antes de compreender como ocorre o uso (ou não) destas fontes, pelos governos e políticos, faz-se necessário entender alguns conceitos básicos.

3.1 ADMINISTRAÇÃO PÚBLICA

O foco do presente trabalho, com relação à Administração Pública, está nos Poderes Legislativo e Executivo, tendo em vista que seus integrantes são escolhidos por meio de eleições quadrienais, possuem poderes para trabalhar em prol do bem da população e, principalmente, são submetidos a constantes fiscalizações por parte da imprensa e da população.

Coelho (2012) descreve a relação que deve haver entre os Poderes Legislativo e Executivo, bem como de cada um deles para com a sociedade:

> Em um regime democrático – em que os governantes são eleitos e têm seus atos constantemente submetidos ao escrutínio da opinião pública e dos formadores de opinião – a força de um governo depende, em grande parte, do apoio que suas propostas políticas e proposições legislativas encontrarem no parlamento; da sintonia entre suas ações e as expectativas dos eleitores; e da relação mantida com os diferentes grupos organizados da sociedade – meios de comunicação, sindicatos e associações, empresas e ONGs etc. (COELHO, 2012, p. 19)

A Constituição de 1988 corrobora isto, ao separar os Títulos III, que trata da Organização do Estado, e IV, que trata da Organização dos Poderes, para descrever os Poderes Executivo e Legislativo dos entes federados. Por exemplo, a função de controle e fiscalização por parte do Poder Legislativo, e/ou promoção do bem-geral do povo brasileiro, bem como a observância das leis, por parte do Poder Executivo.

Por isto, faz-se necessário entender de que forma ambos os poderes utilizam as novas tecnologias - com destaque para as mídias sociais – no desenvolvimento dos seus governos e mandatos, bem como na fiscalização dos agentes políticos. Isso permitirá identificar o *status quo,* bem como se e de que maneira a adoção das mídias sociais como fonte de informação poderia beneficiar a vida da população de uma cidade, estado, ou até mesmo do Brasil. Cabe salientar que, conforme dito no

capítulo anterior, existe uma preferência maior dos governos para as mídias tradicionais.

3.1.1 Relação entre Administração Pública e tecnologia

O uso da tecnologia da informação no governo, segundo Picazo-Vela, Gutiérrez-Martínez e Luna-Reyes (2012), vem das décadas de 1950 e 1960, com as primeiras compras de *mainframes* com processamento em lote. Porém, sua implantação foi lenta e gradual. Por exemplo, Diniz *et al.* (2008) lembram que, com a crise fiscal que atingiu o Brasil na década de 1980, tornou-se necessária uma reforma e modernização no setor público brasileiro, visando excelência na administração e na prestação dos serviços.

Para que isso fosse possível, ações de implantação da tecnologia nos governos eram necessárias e passaram a ser mais frequentes nos governos a partir da década de 1990. Porém, foi com o desenvolvimento da Internet que esse uso começa a crescer nos governos ao redor do mundo. Nesta época, surgiram os primeiros *sites* de governos federais e programas governamentais, com a finalidade de usar a TI como uma ferramenta estratégica.

Assim, surge o conceito de *e-government* (e-governo ou governo eletrônico), sendo definido como o uso das Tecnologias de Informação e Comunicação, e principalmente da Internet, com o alvo de buscar um melhor governo (BONSÓN *et al.*, 2012). Barbosa (2017) salienta que, desta forma, a administração pública buscou melhorar sua capacidade de governança, ou seja, implantar suas políticas públicas.

Acompanhando a tendência mundial, na segunda metade dos anos 1990, o conceito de governo eletrônico começou a ser posto em prática no Brasil, principalmente com o uso de bancos de dados (BARBOSA, 2017). Em 2000 foi implantado o Programa de Governo Eletrônico no Brasil.

De acordo com Barbosa (2017), o Comitê Executivo de Governo Eletrônico, surgido pouco tempo depois, tinha como intuito estabelecer diretrizes e ações de implementação da área de TI, dentro da Administração Pública Federal. Diniz *et al.* (2008, p. 37) apontam algumas áreas envolvidas: “[...] inclusão digital; [...] implementação do software livre; [...] infraestrutura de redes; [...] gestão do conhecimento e informação estratégica [...].”

Na década seguinte, conforme Bonsón *et al.* (2012), novos programas foram lançados, com foco na melhoria de processos, infraestrutura de TI e políticas de informação, buscando um governo mais eficaz, eficiente e confiável. No Brasil, por exemplo, foram desenvolvidos projetos como o Plano Nacional de Banda Larga e o Plano de Telecentros Comunitários.

Embora Bonsón *et al.* (2012) apontem que as iniciativas para usar o e-governo sejam encontradas em quase todos os programas de modernização dos governos, Picazo-Vela, Gutiérrez-Martínez e Luna-Reyes (2012) lembram que muitos lugares não possuem estratégias integradas, que observem as vantagens do investimento em TI como benéficas para os governos, principalmente por causa de seus governantes, que, por suas visões político-partidárias e ideológicas, deixam de investir, ou reduzem os investimentos no e-governo.

O surgimento do chamado “Governo 2.0”, embora inspirado pela Web 2.0, não está focado apenas em redes sociais ou tecnologia, mas evolui-se para um arranjo focado em cooperação e colaboração, com transparência e respeito. Na verdade, a relação entre uma entidade, seja ela pública ou privada, e os seus *stakeholders*, deve se basear na colaboração, por meio de interações ativas, e no engajamento, por meio de resultados benéficos (BONSÓN *et al.*, 2012).

Eles também ressaltam que a tecnologia tem proporcionado novas formas de participação no governo, melhorando a consciência social dos cidadãos, bem como seu engajamento, valendo-se inclusive da criatividade. O crescente interesse na política, bem como no desenvolvimento e implementação de políticas e programas públicos, é potencializado pela Web 2.0.

Em suma, pode-se apontar que o principal motivo pela qual está sendo adotado o governo eletrônico ao redor do mundo é “[...] fortalecer a transparência e a prestação de contas e mudar o papel passivo que cidadãos como clientes tinham (BONSÓN *et al.*, 2012, p. 123).”

Concorda Barbosa (2017), ao afirmar que o uso da TI, acompanhado de novas políticas públicas visando uma reforma administrativa, são fundamentais para que o governo melhore a interatividade entre cidadãos, empresas e órgãos do governo.

3.1.2 Mídias sociais nos serviços governamentais

As pessoas usam cada vez mais as mídias sociais para se comunicarem, não apenas com pessoas próximas, mas também com o poder público, criando formas de comunicação e interação, seja com o governo, com autoridades ou mesmo uns com os outros. Isso tem auxiliado líderes comunitários, autoridades e servidores públicos a, não apenas informarem, mas também serem informados pelos cidadãos (KAVANAUGH *et al.*, 2012).

Picazo-Vela, Gutiérrez-Martínez e Luna-Reyes (2012) apontam que, antigamente os governos usavam tecnologias como o e-mail para se comunicar com os cidadãos. Porém, tal evolução não foi acompanhada de uma mudança organizacional, o que fez com que, muitas vezes, respostas a contatos por e-mail levassem dias ou semanas para serem enviadas – quando não eram perdidas.

Kavanaugh *et al.* (2012) ressaltam o desenvolvimento na comunicação, por meio da Web 2.0 e das mídias sociais. Um dos grandes contribuintes para esse novo momento é o uso de *smartphones*, pois ele vai além da posse de computadores – um celular, atualmente, é um computador de bolso e, para muitas pessoas, é o único computador ao qual tem acesso.

Inclusive, os autores lembram que muitas dessas pessoas fazem parte de segmentos da população que, antes dessas tecnologias, eram difíceis de serem alcançadas pelas formas tradicionais e, até mesmo, eram subestimadas. Atualmente, todos os grupos socioeconômicos da sociedade têm a possibilidade de acompanhar os governos, tornando-os mais acessíveis às pessoas.

Bonsón *et al.* (2012) corroboram essa posição, e entendem que essas ferramentas, sustentadas pela Web 2.0, podem ser usadas no engajamento de cidadãos, na troca de opiniões, no debate e compartilhamento de informações relacionadas a problemas políticos e sociais. Kavanaugh *et al.* (2012) apontam que várias agências governamentais têm reconhecido a mídia social como uma importante fonte de informações.

Por isto, de acordo com os autores, o uso das mídias sociais pelos governos, bem como de serviços *on-line* com conteúdo gerado pelos usuários, pode ajudar a melhorar a comunicação entre governos e cidadãos. Picazo-Vela, Gutiérrez-Martínez e Luna-Reyes (2012) apontam que muitas organizações governamentais têm usado

as mídias sociais como um poderoso conjunto de ferramentas para aprimorar e reinventar a comunicação entre o governo e o cidadão.

Isso traz vários benefícios ao governo, como a eficiência, a responsabilidade, a melhoria da confiança e da democracia – inclusive, Bonsón *et al.* (2012) apontam que a disponibilidade de informação é fundamental para a democracia. Outro benefício é que o uso da mídia social permite aos governos uma economia de dinheiro e de recursos maior do que usando as mídias tradicionais.

Aliás, o uso das mídias sociais não só melhora a comunicação e a participação dos cidadãos, como também dá mais transparência aos atos da administração. Bonsón *et al.* (2012) apontam que os governos têm enfrentado exigências de transparência muito grandes, principalmente por causa das tecnologias de mobilização. Eles entendem que "Como uma das dimensões do governo eletrônico é a disponibilidade de certas informações aos cidadãos, a transparência é uma base importante do governo eletrônico" (BONSÓN *et al.*, 2012, p. 126, tradução nossa).

Segundo Marques, Aquino e Miola (2014), tanto agentes quanto instituições do Estado são continuamente provocados a refletirem sobre a adoção de mídias digitais, que são práticas e cômodas, para reforçar a legitimidade das suas práticas. Ou seja, o uso da tecnologia – principalmente das mídias sociais - aumenta o envolvimento dos cidadãos, gerando mais oportunidades de fazê-los participarem e colaborarem com o governo, principalmente quando relacionado aos serviços públicos.

De acordo com Picazo-Vela, Gutiérrez-Martínez e Luna-Reyes (2012), as mídias sociais têm sido usadas para melhorar os serviços governamentais, como no compartilhamento de informações dentro do governo e entre agências, bem como a disseminação de informações para o público. Kavanaugh *et al.* (2012) lembram que é possível monitorar, identificar e responder questões e problemas que ocorrem nas cidades, em diversas áreas.

Cada vez mais, segundo os autores, é feito o monitoramento de mídias sociais para detectar, obter informações e monitorar discussões a respeito de grandes eventos públicos, como manifestações políticas. Isso pode, inclusive, ajudar a resolver eventuais problemas, como a violência e o vandalismo. Outro exemplo é o monitoramento de doenças e epidemias, sendo possível reconhecer, de forma precoce, informações sobre quarentenas, campanhas de vacinação, entre outros.

Kavanaugh *et al.* (2012) lembram que os serviços *on-line* e as mídias sociais disponibilizam uma imensa quantidade de informações, e que elas estão à disposição do governo para serem utilizadas na melhoria dos serviços públicos. Por exemplo, em um monitoramento contínuo de uma situação, é possível analisar o retorno da população a campanhas, ou mesmo obter informações sobre deficiências da administração e/ou de serviços prestados por ela.

3.1.3 Situação atual e perspectivas

De acordo com Kavanaugh *et al.* (2012, p. 483), corroborado por pesquisas realizadas, "[...] a pessoa de relações públicas de várias agências governamentais, normalmente não estava familiarizada nem se sentia à vontade com as mídias sociais." Isso prejudica os governos, tornando difícil o gerenciamento desses canais de comunicação. Em síntese, eles entendem que:

> É possível que um gerente de relações públicas possa se concentrar nos meios tradicionais de comunicação, tais como, cartas de imprensa, comunicados à imprensa e entrevistas por telefone com a TV e o rádio locais. No entanto, para gerenciar a comunicação com um público mais diversificado, eles precisam de alguém que use mídia social. Então, ou eles (os governos) têm que reciclar os atuais gerentes de relações públicas, ou eles têm o custo adicional de inserir outra pessoa para gerenciar as atividades de relações públicas que envolvem a mídia social. (KAVANAUGH *et al.*, 2012, p. 489, tradução nossa)

Os estudos preliminares dos autores apontaram nessa mesma direção. Na maioria das vezes, quem postava no Twitter ou gerenciava a página do Facebook não era da liderança da organização, mas um estudante universitário, ou mesmo um jovem, responsável por publicar anúncios, atualizações e outras informações. Picazo-Vela, Gutiérrez-Martínez e Luna-Reyes (2012) reforçam esse entendimento.

Os autores também apontam que cada organização governamental tem utilizado as mídias sociais de uma forma própria, obtendo resultados diferentes, e nem sempre as vendo como uma ferramenta estratégica a longo prazo. Mais do que isto: muitas adotam como uma opção na abordagem de tentativa e erro, gerando custos desnecessários.

Em síntese, Kavanaugh *et al.* (2012) ressaltam que os governos usam as mídias sociais sem saber seus custos ou benefícios, sem conhecer seu público, sem analisar a pessoa responsável por monitorar as comunicações, e sem compreender o efeito das suas comunicações ao público.

Porém, os autores entendem que o uso dessas novas mídias pode fornecer ao governo informações sobre a comunidade, das quais métodos tradicionais não tem condições de obter, devido ao custo alto e alcance limitado. Bonsón *et al.* (2012) apontam que a figura do gerente de mídia social, que existe há vários anos no setor privado, está aos poucos tornando-se popular no setor público.

Porém, sabe-se que a comunicação tradicional, como jornal, rádio, televisão, revista e telefone, tem se tornado digital e acessível na Internet – inclusive, com a tendência de convergência. Desta forma, é possível que, com o passar dos anos e com o desenvolvimento tecnológico, seja possível uma transição dos métodos tradicionais para métodos cada vez mais emergentes.

Por esses motivos, os autores ressaltam a necessidade de abordar alguns questionamentos e entendimentos, para dar eficácia ao uso dessas tecnologias:

> Quais mídias sociais o governo deve usar para se comunicar de maneira mais eficiente com um público diversificado? Como as mensagens devem ser formadas e enquadradas nas mídias sociais para serem eficazes? [...] Que papel as mídias sociais desempenham na mistura geral de fontes de informação para os cidadãos se comunicarem sobre a vida cívica, uns com os outros e com o governo? As mídias sociais afetam a participação cívica e, em caso afirmativo, para quem e que tipos de participação cívica? (KAVANAUGH *et al.*, 2012, p. 481)

Embora, como dito anteriormente, o uso das mídias sociais possa exigir um investimento – por exemplo, em treinamento de funcionários, contratação de pessoal para cuidar especificamente delas, campanhas de educação e orientação para seu uso -, e os orçamentos governamentais têm sido cada vez mais reduzidos, devido a questões econômicas, é notório, segundo os autores, que seu uso possui grande potencial, e é capaz de trazer um retorno importante à atuação do governo.

Picazo-Vela, Gutiérrez-Martínez e Luna-Reyes (2012, p. 510, tradução nossa) concluem, que “[...] as redes sociais eletrônicas e outras mídias sociais vieram para ficar, e os governos precisam começar a usá-las para ter uma participação mais ativa na formação das novas maneiras de interação entre indivíduos e organizações [...]”, e que “[...] as novas mídias podem até ter o potencial de transformar o atual sistema de governança nos níveis de cidade, estado ou país.”

3.2 ATIVIDADE POLÍTICA E PARLAMENTAR

Silva e Ferreira Jr. (2013) expõem que o uso da tecnologia e da Internet para campanhas eleitorais existe desde o final dos anos 1990, porém, foi no final da década de 2000 que seu uso superou qualquer expectativa, demonstrando claramente a possibilidade de alcance e interação entre políticos e eleitores com o mundo digital.

O grande impacto disto está no marketing político, visto que a meta deste é dar a possibilidade de vitória a um candidato, e isso passa pelo uso das diversas tecnologias existentes. E uma das principais são as mídias sociais, devido a capacidade que elas têm de atingir crenças e sentimentos de pessoas, o que pode se tornar um voto na eleição.

Há algum tempo, durante os períodos eleitorais, são feitas análises em mídias, devido ao alto engajamento de cidadãos e candidatos. Segundo Marques, Aquino e Miola (2014), muitas vezes as campanhas políticas possuem poucos recursos financeiros, e o uso das mídias permite diminuir essa dificuldade – embora existam formas pagas de impulsionar o uso dessas tecnologias. Mas, mais do que isto, o uso delas atualmente fornece interatividade, discussão e *feedback* de atitudes e opiniões como jamais visto no século passado.

Os autores mostram a necessidade desta interatividade permanecer após a eleição, lembrando que o uso de redes sociais, por parte dos agentes políticos, como deputados, deve ser analisado com base na estratégia de comunicação da política, na disputa por visibilidade junto ao público eleitor – mesmo passada a eleição -, e também na imagem que o político tem de si mesmo.

E é exatamente por este motivo que os autores salientam a evolução trazida pelas mídias sociais. Os antes “políticos tradicionais”, que sempre tiveram uma base eleitoral consolidada, proporcionando sucessivas eleições em cargos públicos, estão sendo desbancados por novos atores, que, muitas vezes, não são políticos “profissionais”, mas que se valem das mídias sociais para adquirirem o capital social necessário para serem eleitos – e as últimas eleições, como dito anteriormente, apontam exatamente neste sentido. Esse entendimento é confirmado pelo estudo de Fenoll e Cano-Orón (2017), apontando que há um interesse maior por partidos mais novos do que pelos partidos e políticos tradicionais.

Dados obtidos por Marques, Aquino e Miola (2014, p. 182), em pesquisa realizada com congressistas brasileiros no Twitter, demonstram que parlamentares mais jovens utilizam o *microblog* com maior frequência, bem como os deputados mais bem votados e os que ocupam cargos importantes, como líder de partido ou integrante da mesa diretora. Porém muitos deputados as utilizam apenas durante a campanha, deixando-as de lado após o pleito.

O risco dessa atitude é alertado por Silva e Ferreira Jr. (2013, p. 11), ao salientarem que o político deve ir além das promessas de campanhas, mas sim conquistar seu eleitor e abrir caminho, de forma a registrar sua posição, criar um rumo, bem como alcançar seu intuito na campanha política.

Assim, pode-se entender que o uso da tecnologia – principalmente as mídias sociais – é útil não apenas para os políticos em campanha, mas também para um político eleito em mandato, seja para cargo executivo ou legislativo, pois poderá não apenas manter o vínculo com os seus eleitores, mas também exercer seu trabalho enquanto agente político dentro do governo.

Desta forma, faz-se necessário o uso de técnicas e tecnologias que permitem ao político e parlamentar a possibilidade de obter informações do seu eleitorado, a partir de suas mídias sociais, bem como de outras mídias. Isto pode mostrar-se como um diferencial, não apenas para sua atuação política ou parlamentar, mas também para eleições futuras. E é neste sentido que o presente trabalho se coloca.

3.3 CONSIDERAÇÕES SOBRE O CAPÍTULO

Em síntese, o presente capítulo tratou dos conceitos relacionados à Administração Pública e a Atividade Política, envolvendo questões sobre o uso da tecnologia, o uso de meios de comunicação e mídias sociais, a relação entre as mídias sociais e os serviços governamentais, e ainda a situação atual, com algumas perspectivas futuras.

Assim, o presente capítulo atingiu o objetivo específico de identificar se e como são utilizadas, pelos governos e políticos, dados obtidos de redes sociais e *sites* de notícias.

Prosseguindo, o capítulo a seguir trata do processo de mineração de texto, as técnicas utilizadas, bem como as etapas necessárias para a sua execução.

4 MINERAÇÃO DE TEXTOS

Conforme exposto no subcapítulo 3.1.2, faz-se necessário que tanto a Administração Pública quanto a Atividade Política ou Parlamentar lancem mão de ferramentas e técnicas que permitam obter informações de mídias sociais e tradicionais. Com o crescimento destas, é cada vez maior o volume de informações que circula, por exemplo, em uma página de um vereador no Facebook, ou um *blog* de notícias de uma cidade.

Tal volume, por necessidade do trabalho ou da atividade, exige análise e extração de informações. Por exemplo, saber quem é a pessoa que mais comenta em uma página, ou quais os termos mais falados, ou ainda qual o sentimento em geral dos textos publicados (positivo ou negativo).

Sendo assim, encontra-se na mineração de textos uma ferramenta capaz de auxiliar nessa extração. O grande motivo para tal entendimento está na sua finalidade, conforme aponta Pezzini (2016), ao afirmar que ela é um processo que extrai informações de textos em linguagem natural, que são desconhecidas e que tem potencial utilidade. Além disto, tendo em vista que as informações coletadas no trabalho basear-se-ão exclusivamente em texto, entende o autor que há um alto valor no resultado obtido ao aplicar as técnicas de mineração de texto.

4.1 CONCEITO

A mineração de textos, segundo Pezzini (2016), é considerada um paradigma de programação, cujo foco é resolver e entender o relacionamento existente entre textos. Muitas vezes, tal relação mostra-se confusa e ambígua, sendo que eventuais dificuldades não conseguem ser resolvidas por formas tradicionais.

Por sua vez, Aranha e Passos (2006, p. 1) trazem que a mineração de texto é um campo “[...] multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Linguística e Ciência Cognitiva.”

Isto tornou-se necessário, principalmente, quando do desenvolvimento da Internet e das redes computacionais, onde cada vez mais os documentos e textos virtuais passaram a ser o principal meio de armazenar dados e informações. Segundo os autores supracitados, a maioria do conteúdo disponível na Internet está em formato de texto, e geralmente não estão estruturados.

Em síntese, pode-se entender, de acordo com Serapião, Suzuki e Marques (2010), que o propósito da mineração de texto é a busca por termos e expressões relevantes, muitas vezes em documentos de grande volume de dados. Isto ocorre por meio de ferramentas, que, ao analisar um texto ou documento, podem identificar informações relevantes, de acordo com a temática desejada. Além disto, é possível criar agrupamentos por assuntos, analisando também a frequência de ocorrência. Isso, segundo Corrêa, Marcacini e Rezende (2012, p. 6), "[...] viabiliza sobremaneira a análise exploratória de documentos [...]"

A partir das temáticas e frequências, a mineração de texto permite identificar o teor do assunto de um texto ou conjunto destes, e sua relevância para o pesquisador. Para isto ser possível, várias áreas da Tecnologia da Informação são envolvidas, como a mineração de dados, o aprendizado de máquina, a recuperação de informação, a estatística e a linguagem computacional (PEZZINI, 2016).

A mineração de textos possui grande potencial de uso, tendo em vista que, com a crescente quantidade de dados e informações que são obtidos ou armazenados por meio de texto, faz-se necessário usar a tecnologia para substituir tarefas antes feitas pelas pessoas, trazendo mais rapidez e menor custo. Por exemplo, conforme salienta Pezzini (2016), a análise de sentimentos através da expressão na forma escrita, pode ser feita de modo mais rápido por meio do *text mining*, do que se feito por uma ou várias pessoas, de modo manual.

Esta área pode atender a outras demandas, que tradicionalmente não seriam realizadas. Carrilho Junior (2007) sugere que pesquisas de opinião podem deixar de ser objetivas e passar a serem discursivas, ou seja, ao invés de escolherem respostas pré-determinadas, eles escrevem o que desejarem, e a mineração de texto entra como analisador dos termos e respostas escritos. Outros exemplos de áreas que podem ser abrangidas, como o atendimento ao cliente feito pelas empresas e a análise de textos e documentos da área da medicina, como pesquisado por Serapião, Suzuki e Marques (2010).

Cabe aqui salientar, de acordo com Aranha e Passos (2006), que a mineração de texto não é um mecanismo de busca. Enquanto aquela está relacionada a descoberta de informações antes desconhecidas, esta relaciona-se com informações cujo usuário já sabe o que e aonde procurar. Além disto, a mineração de texto não tenta simular qualquer comportamento humano. Por outro lado, conforme aponta Pezzini (2016, p. 61), "[...] mesmo com toda a automação

fornecida pela mineração de dados, ainda é necessário alguém ao final do processo para avaliar os resultados obtidos na mineração.”

4.2 TÉCNICAS DE MINERAÇÃO DE TEXTO

Pezzini (2016) apresentam três técnicas utilizadas na área de mineração de dados, e que podem ser aplicadas na mineração de texto: o Processamento de Linguagem Natural, a Recuperação de Informação e a Extração de Informação.

O Processamento de Linguagem Natural, de acordo com o autor, utiliza técnicas que permitem melhorar a compreensão da linguagem natural utilizando computadores. Nela, são usados recursos como a manipulação de texto (*strings)*, de forma que seja possível processar textos com rapidez.

A Recuperação de Informação baseia-se em métodos e medidas, tanto de semântica quanto de estatística, que permitem processar textos automaticamente, de forma a encontrar os documentos ou textos que possuem uma “[...] resposta para a questão [...]” – embora não se encontre a resposta em si (MACHADO *et al.*, 2010 apud PEZZINI, 2016, p. 59). Ou seja, ela aponta aonde é possível encontrar algum termo, mas não diretamente onde está este termo.

A Extração de Informação tem como finalidade buscar partes relevantes de um texto ou de um documento, permitindo extrair informações específicas. Neste caso, a compreensão da linguagem natural é inferior à que ocorre nas outras técnicas, não por ser defasada ou simples, mas porque não demanda tal compreensão.

Desta forma, o trabalho optou pela técnica da Extração de Informação, pois esta visa a busca de partes relevantes em textos, bem como extrair informações destas partes, sem a necessidade de uma compreensão da linguagem natural. Não que a primeira técnica não pudesse ser aplicada ao presente trabalho, porém, ela demandaria mais estudo e testes, algo que não seria possível contemplar neste trabalho.

Cabe colocar que Aranha e Passos (2006) também trazem duas técnicas adicionais. A primeira é a Indexação, que permite a busca em texto com base em palavras-chave. É dependente de uma estrutura de dados organizada, sendo possível, a partir dela, a recuperação de texto, bem como a realização de cálculos usando palavras-chave como busca.

Outra técnica abordada pelos autores é a Mineração de Dados, que também permite identificar conhecimentos relevantes a partir de uma base textual, embora também dependam de um banco de dados organizado e previamente processado. Entretanto, como os próprios ressaltam, “Mineração de textos é um conjunto de métodos usados para navegar, organizar, achar e descobrir informação em bases textuais. Pode ser vista como uma extensão da área de *Data Mining*, focada na análise de textos.” (ARANHA; PASSOS, 2006, p. 2)

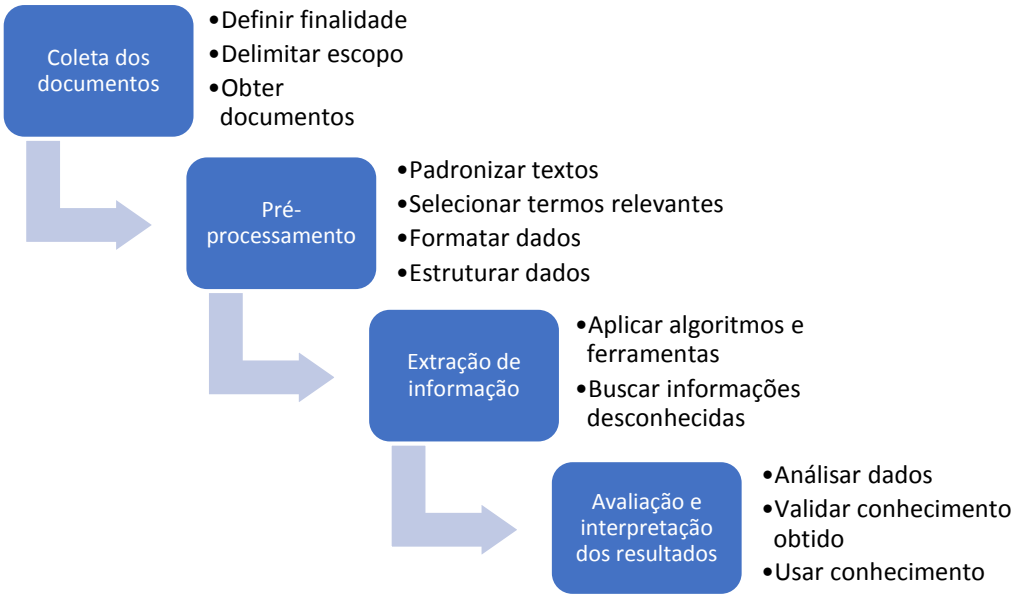
Ainda são mencionadas pelo autor as técnicas de Classificação, a Clusterização e a Otimização.

4.3 ETAPAS DO PROCESSO DE MINERAÇÃO DE TEXTO

De acordo com Martins *et al.* (2003), existem quatro etapas básicas no processo de mineração de texto: a coleta de documentos, o pré-processamento, a extração de informação e conhecimento, e a avaliação e interpretação dos resultados. O modelo a ser apresentado no presente trabalho será criado a partir destas quatro etapas, conforme será explicado no capítulo 7.

O processo de mineração de texto pode ser resumido conforme na Figura 2:

Figura 2 - Processo de mineração de texto e ações principais



Fonte: Do autor (2019)

O que é importante mencionar é que não existe uma esquematização única do processo de *text mining*, sendo que alguns podem ter alguma modificação,

conforme exemplificado no subcapítulo 7.2, sendo incluída alguma etapa, ou dividindo uma etapa em duas. O que será apresentado neste capítulo trata-se de uma abordagem comum ao processo de mineração de texto.

4.3.1 Coleta dos documentos

A primeira etapa consiste em buscar todo o tipo de documento que seja relacionado ao que se deseja obter. Geralmente, recebe a denominação de “corpus”. Segundo Corrêa, Marcacini e Rezende (2012), é nessa etapa – chamada por eles de “Identificação do Problema” – que o escopo do problema é delimitado, com o desígnio de definir aonde será aplicada a mineração, como serão usados os dados, as eventuais restrições, entre outros.

No presente trabalho, cujo foco está nas mídias sociais e tradicionais, os documentos relacionados são, principalmente, as páginas das redes sociais e os *sites* de notícias, onde o escopo está em obter informações desconhecidas de ambas as mídias. É possível, de acordo com Martins *et al*. (2003), obter dados também em livros, e-mails, entre outras fontes. Por este motivo, o caráter genérico do modelo que será criado permitirá o uso de qualquer fonte.

Cabe salientar que, conforme Corrêa, Marcacini e Rezende (2012), não há uso de informações externas, ou seja, que se baseiem em conhecimento prévio das pessoas ou do ambiente relacionado. Tal coleta segue procedimentos e utiliza a tecnologia sem a necessidade de supervisão ou seleção. Para exemplificar, na coleta de postagens no Facebook, não se está preocupado em analisar o conteúdo, se possui relação com o que se deseja, ou mesmo se é uma propaganda. Seu objeto é, basicamente, extrair os dados desses documentos, sendo até possível uma categorização ou tematização primária.

4.3.2 Pré-processamento

É importante ressaltar que essa etapa, segundo Corrêa, Marcacini e Rezende (2012), é fundamental, pois é ela que permitirá transformar texto em linguagem natural para representações manipuláveis por algoritmos. Segundo Aranha e Passos (2006), por serem dados considerados como não-estruturados, é necessário previamente estruturar os textos.

Assim, é necessário formatar os dados obtidos dos documentos, de forma a dar a eles uma organização padronizada – de acordo com Martins *et al*. (2003), uma estrutura atributo-valor, como em um vetor computacional, mas de uma forma que suas características não sejam prejudicadas. Isso permitirá que os dados possam ser usados ao longo dos próximos passos.

Há alguns passos importantes para a continuação do processo, de acordo com os autores. O primeiro é a padronização dos textos, onde geralmente são feitas conversões, conforme o tipo de dado desejado. Por exemplo, conversão de texto para data ou número.

Outro passo é a seleção dos termos representativos presentes. Tendo em vista que o texto coletado está em linguagem natural, é necessário remover as chamadas *stopwords*[4]. Desta forma, há uma quantidade menor de termos a serem analisados, auxiliando no processamento. Também existe a possibilidade de identificar e remover variações morfológicas, como, por exemplo, conjugações verbais e sinônimos. Termos compostos que possuem um único significado semântico também podem ser removidos.

Existem técnicas que permitem isto, como o *stemming* e o *thesaurus*, bem como o uso de dicionários léxicos. Aliás, Aranha e Passos (2006) ressaltam a importância da linguística, principalmente com relação à compreensão da sintaxe e da semântica.

Outro passo necessário é a compactação da informação, ou agrupamento conceitual. Mais do que simplesmente sinônimos, os autores trazem a necessidade de unificar conceitos, que muitas vezes aparecem representadas de formas diferentes, mas que se referem a um único termo. Por exemplo, os termos “NH”, “Novo Hamburgo”, “N. Hamburgo” e “N.H” se referem a um mesmo conceito. Tal processo pode ser feito de forma manual ou automática, de acordo com a ferramenta usada.

Corrêa, Marcacini e Rezende (2012) também mencionam a seleção por medidas estatísticas, como a frequência do termo (frequência de aparição de um termo em uma coleção textual) e a frequência de documentos (frequência de aparição de um termo em vários documentos).

[4] As *stopwords* são palavras que são consideradas irrelevantes para um conjunto de resultados a ser exibido ou utilizado, como, por exemplo, artigos e pronomes.

Existem alguns problemas trazidos por Aranha e Passos (2006) que podem dificultar esta fase. Por exemplo, a parametrização, que é a definição de regras que atendam a todas as possíveis construções linguísticas do idioma; o dinamismo, que é a evolução da língua ao longo do tempo, fazendo surgirem novas regras; a existência de ruído, que são erros ortográficos, palavras sem acento, ausência de espaço entre as palavras, entre outros; e a ambiguidade, que é a existência de mais de um significado para um mesmo termo, conforme o contexto.

4.3.3 Extração de informação e conhecimento

A etapa seguinte – também chamada de “extração de padrões” por Corrêa, Marcacini e Rezende (2012) ou de análise exploratória dos dados por Martins *et al*. (2003) – tem relação com várias técnicas, como a aplicação de algoritmos e ferramentas, e busca informações desconhecidas, mas que, ao serem descobertas, tem o potencial de serem úteis.

Após o texto passar pelo pré-processamento, é possível extrair informações relacionadas a frequência de ocorrência das palavras. Isto permite, por exemplo, analisar as palavras que mais ocorreram, dando uma noção da temática abordada no conjunto textual.

Outro ponto analisado é a correlação entre palavras, que consiste em verificar qual o grau (em porcentagem) de ligação entre duas palavras ou dois termos/expressões. É possível, a partir destes dados, identificar a relação entre expressões e o grau de intensidade desta. Se, por exemplo, há uma alta correlação entre os termos “vereador” e “saúde”, isto pode indicar a atuação de um vereador na saúde do município, ou apontar que algum vereador está com problema de saúde.

Geralmente, utiliza-se tabelas, gráficos e outras ferramentas visuais, que auxiliam na compreensão das informações extraídas. Com o desenvolvimento da Internet, surgiram as nuvens de palavras, que permitem a visualização da frequência de palavras. Outra ferramenta muito utilizada é a de grafos de palavras, que utilizando as estruturas de nós e grafos permite, por exemplo, exibir a correlação existente entre palavras de um documento ou texto.

Aranha e Passos (2006) trazem algumas outras informações que podem ser extraídas, como as entidades. Por meio de correferências, acrônimos e anáforas, é

possível extrair, de um texto, a resposta para perguntas como “Quem”, “O quê”, “Quando”, “Como” e “Onde”.

Associações lógicas também podem ser feitas, de forma a obterem conhecimento. Por exemplo, ao analisar a expressão “o vereador hamburguense João da Silva”, é possível identificar, por meio de algoritmos inteligentes, que “João da Silva” é vereador de Novo Hamburgo.

Há outra ferramenta interessante que é a sumarização de conteúdo, que permite, analisando frequências e correlações, utilizar técnicas para resumir um texto, com base em seu conteúdo relevante (ARANHA; PASSOS, 2006). De um texto de 10 parágrafos, é possível resumi-lo em quatro linhas, sem prejudicar a compreensão e o entendimento do assunto.

4.3.4 Avaliação e interpretação dos resultados

A etapa final necessita apenas que um usuário interessado faça sua análise sobre o que foi extraído e gerado - se é satisfatório e se atende o desejado. Caso não contemple isto, ele pode apontar as deficiências e o que poderia ser feito para melhorar. Corrêa, Marcacini e Rezende (2012) dividem essa etapa em duas.

A primeira é o chamado “pós-processamento”, pois ela valida o conhecimento obtido. Isto pode ser feito de forma subjetiva, conforme o conhecimento de uma pessoa, ou objetiva, utilizando estatística para definir a qualidade dos resultados obtidos. No presente trabalho, será feita uma validação por meio de análise subjetiva,. Não que o uso de tais métodos não seja positivo, mas acabaria por tornar a presente etapa mais complexa.

A segunda, de acordo com os autores, é o “uso do conhecimento”, onde, após a validação dos resultados, eles estão aptos a serem usados, de acordo com os propósitos desejados, as fontes pesquisadas e as informações extraídas.

Por exemplo, ao final da mineração de texto aplicada por Serapião, Suzuki e Marques (2010) em laudos eletrônicos de mamografia, os autores identificaram a quantidade de termos utilizados, bem como aqueles escritos com grafia incorreta. Foi possível concluir que havia boa aderência a uma taxonomia relacionada à mamografia, bem como a necessidade de corrigir a escrita para evitar termos com erro. São pontos que, sem essa tecnologia, seriam difíceis de serem identificados.

Ressalta-se que, para o modelo que será criado, a atuação do usuário não será limitada ao passo final, quando da análise, mas o mesmo poderá atuar em todo o processo, o que pode não apenas facilitar a identificação de falhas ou deficiências, bem como determinar, de forma mais clara e transparente, o que e como são obtidos os dados coletados em um determinado documento ou fonte.

4.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

O capítulo que se encerra teve como alvo apresentar o conceito de mineração de texto e o seu uso. Também foram apresentadas técnicas aplicáveis, e, por fim, o processo de mineração de texto e suas respectivas etapas.

Ele é fundamental para que seja possível atingir o segundo objetivo específico, que é o de investigar de que forma é possível obter dados de mídias sociais e tradicionais, com base nos conceitos de mineração de texto.

Prosseguindo, o capítulo seguinte abordará as ferramentas que permitem obter, extrair e analisar dados de mídias sociais e tradicionais, de acordo com os conceitos e etapas de mineração de texto abordados neste capítulo.

5 FERRAMENTAS

O objetivo do presente trabalho é “criar um modelo de extração de informação, a partir de fontes da Internet, baseado em *text mining* e utilizando ferramentas gratuitas existentes [...]”. Porém, conforme dito na introdução, tal modelo possui caráter genérico, e não se restringe a um conjunto finito e definitivo de ferramentas. Para isto, criar-se-á uma representação prática deste modelo, conforme será exposto no subcapítulo 7.3, baseando-se em uma seleção de ferramentas.

O foco desse capítulo é abordar esse processo, onde, após diversas pesquisas e estudos, bem como testes e validações, foi possível selecionar algumas ferramentas, dentre as várias pesquisadas.

5.1 SELEÇÃO DE FERRAMENTAS

Após pesquisa bibliográfica, bem como buscas na Internet, foram encontradas cerca de trinta ferramentas, capazes de realizar ao menos uma das etapas abordadas no subcapítulo 4.3 – algumas até com a possibilidade de realizar todo o processo.

Entretanto, para a seleção foram definidos dois critérios básicos:

a) os *softwares* deveriam ter boa usabilidade (*user-friendly*), ou seja, deveriam ser de fácil uso para o usuário comum, não sendo necessário um conhecimento técnico de difícil compreensão, como ocorreria com uma linguagem de programação ou algum *software* que exige complexa configuração.
b) os *softwares* deveriam ser livres e gratuitos. Este ponto é importante, tendo em vista que o uso de ferramentas com essa característica universaliza o modelo proposto, sem a necessidade de adquirir *softwares*, onerando os cofres públicos.

Todos os softwares testados que foram pesquisados deveriam atender a estes dois critérios para serem selecionados. Com o objetivo de não tornar o processo de seleção demorado e complexo, não foi adotada uma análise com critérios objetivos de usabilidade, como uma validação com usuários ou outros métodos de aferição – a escolha foi estritamente subjetiva, após testar as

ferramentas. Algumas até foram aprovadas nos critérios, mas apresentaram algum problema, como na exportação.

As ferramentas que não foram selecionadas foram:

a) **Clementine** (ARANHA; PASSOS, 2006), que utiliza as ferramentas do software SPSS da IBM para *text mining*. Possui inúmeras funcionalidades, mas seu uso exige um certo conhecimento técnico.
b) **Cortex Competitiva** (ARANHA; PASSOS, 2006), ferramenta de *text mining* aplicada à Inteligência Competitiva. Possui diversas fontes pesquisadas, e pode gerar inúmeros gráficos e informações.
c) **dtSearch** (KLEMANN; REATEGUI; RAPKIEWICZ, 2011) é uma ferramenta de pesquisa em documentos, com filtros e buscas. Também possibilita extrair e analisar o texto. Destaque para a nuvem de palavras.
d) **GATE** (KLEMANN; REATEGUI; RAPKIEWICZ, 2011) é um *framework* de mineração de texto, que trabalha com processamento de linguagem, extração de informação e algoritmos de aprendizado de máquina.
e) **GAWK** (BRUNS; LIANG, 2012), consiste em uma versão GNU = da linguagem AWK, designada para processamento de texto e usada para extração de dados.
f) **Gephi** (SILVA; STABILE, 2016), que permite a visualização de redes (grafos) de palavras.
g) **Google Alerts** (FERNANDES, 2018) é um serviço que permite monitorar resultados de uma determinada busca de um termo. Não permite limitar a busca em certos *sites*, embora seja possível definir a fonte de origem.
h) **iScience Maps** (REIPS; GARAIZAR, 2011), de pré-processamento, mas não foi encontrada.
i) **KNIME Analytics Plataform** (RANJAN; AGARWAL; VENKATESAN, 2017), que é um software de análise de dados, além de gerar relatórios. Integra vários componentes do *machine learning*, além de *data mining*.
j) **Media Style** (ARANHA; PASSOS, 2006), que apresenta soluções de extração de informação baseada em palavras-chave. Não foi possível visualizá-lo ou testá-lo.

k) **Netvizz** (RIEDER, 2013), que possui várias ferramentas interessantes, como grafos de palavras conforme o conteúdo postado em uma página, a listagem de postagens de uma página, entre outros.
l) **NodeXL** (BRUNS; LIANG, 2012), que permite a visualização de redes (grafos) de palavras. Existem versões que permitem a análise de redes sociais, após a coleta de dados.
m) **Orange Data Mining** (RANJAN; AGARWAL; VENKATESAN, 2017), um software baseado em componentes, com foco na visualização de dados, aprendizado de máquina, mineração de dados e análise de dados.
n) **RapidMiner** (KLEMANN; REATEGUI; RAPKIEWICZ, 2011), focada em Ciência de Dados, permite descobrir conhecimento, além de mineração de dados. Possui versão gratuita, mas a versão completa é paga.
o) **SAS Text Miner** (ARANHA; PASSOS, 2006), que utiliza um poderoso conjunto de ferramentas, podendo gerar gráficos e informações. Possui inúmeras funcionalidades, mas seu uso exige conhecimento técnico.
p) **Text Analyst** (ARANHA; PASSOS, 2006), que, segundo os autores, gera uma rede de semântica do texto baseada em um algoritmo. Não foi possível visualizá-lo ou testá-lo.
q) **Text Mining Suite** (ARANHA; PASSOS, 2006), *software* produzido pela empresa Intext Mining. Permite analisar frequências e associações (correlações) entre conceitos. Gera apenas tabelas.
r) **TextAlyser** (KLEMANN; REATEGUI; RAPKIEWICZ, 2011), disponível *online*, que analisa texto ou arquivo, trazendo informações sobre frequência de palavras. Interface simples em inglês, com tabelas.
s) **Weka** (RANJAN; AGARWAL; VENKATESAN, 2017), é um *software* de mineração, que permite executar os passos necessários, incluindo o pré-processamento. Possui algoritmos e ferramentas para análise, classificação e visualização de informações.
t) **Wordcounter** (ARANHA; PASSOS, 2006), que está disponível na Internet. Ele se restringe a contar as palavras e os caracteres.
u) **WordStat** (MEYER; HORNIK; FEINERER, 2008), onde os textos são categorizados automaticamente usando um dicionário de palavras.

Excelente ferramenta, exibindo gráficos e extraindo ótimas informações, mas é paga.

v) **Yahoo! Pipes** (LOMBORG; BECHMANN, 2014), para pré-processamento, porém foi descontinuada.

Salienta-se, por fim, que a representação será criada com base nessa seleção, e é plenamente possível elaborar uma nova seleção de ferramentas, ou mesmo substituir uma por outra, de forma a alterar ou recriar uma representação do modelo – sem que este seja alterado, visto que é um modelo genérico.

5.2 FERRAMENTAS DE COLETA DE DADOS

Com relação à coleta de dados, a pesquisa foi auxiliada pela tecnologia API (*Application Programming Interface*). Lomborg e Bechmann (2014) explicam que ela é uma estrutura de *backend*, onde desenvolvedores podem conectar-se a um determinado serviço para executar tarefas, como a obtenção de dados, com alta capacidade de coleta e mineração, e permitem fazer uma análise qualitativa e quantitativa destes.

Todas as APIs utilizadas são baseadas em HTTP, ou seja, são baseadas em requisições GET em HTTP, da mesma forma que se obtém uma página acessada pelo navegador. A diferença está no retorno, que ao invés de uma página *web*, obtém-se uma estrutura em JSON (*JavaScript Object Notation)*[5]. A partir desta estrutura, é possível manipular os dados e utilizá-los para qualquer finalidade.

5.2.1 Mídias sociais

O Facebook utiliza uma API chamada *Graph*, citada por Arun e Nayagam (2014), que é considerada a forma básica de manipular dados na rede social, sendo possível, por exemplo, publicar conteúdo ou coletar dados. A rede social é composta de três objetos básicos: os nós (*nodes*), que são os objetos individuais; as bordas (*edges*), que são as conexões entre um objeto e um conjunto de objetos, e os campos (*fields*), que são dados de um objeto.

[5] Trata-se de um formato de troca de dados entre sistemas, simples e prático, utilizando uma estrutura mais legível aos humanos do que o XML.

O foco do trabalho, com relação ao Facebook, está em obter dados de três nós: usuários (perfis), páginas e grupos. Percebe-se que estes são os principais da rede, e é onde a Administração Pública e os políticos procuram ou podem procurar dados e informações úteis para suas atividades. O modelo proposto no trabalho pode, por exemplo, fornecer informações ricas e profundas sobre assuntos e personagens discutidas neste e em outros grupos.

Todas as APIs testadas no trabalho permitem o acesso aos dados mediante autorização. Isso ocorre por meio de um *token*, que em algumas APIs deve ser gerado pelo usuário e informado na URL, e em outras é possível fazer o *login*, sendo automaticamente gerado e colocado na URL. Para isto, é necessário um vínculo com a rede – por exemplo, para utilizar a Graph, é necessário ser usuário do Facebook.

Ao longo das versões lançadas, várias modificações foram realizadas, visando a segurança e o sigilo. Desde a versão 3.0, não é mais possível obter dados de grupos. Além disto, após alguns testes, também não foi possível obter dados de perfis pessoais. Isto poderia prejudicar o trabalho, e fez com que se buscasse uma alternativa para obter tais dados, o qual será explicado dentro do subcapítulo 5.2.4.

Por meio da montagem de URLs personalizadas, é possível buscar vários dados de páginas da rede social, como postagens e/ou comentários de uma postagem, incluindo, entre outros campos, a data de criação e a quantidade de comentários ou reações. É possível, também, filtrar por data. Para a busca, é necessário informar o código do objeto a ser buscado, que pode ser o nome de usuário[6] ou o ID[7] do objeto.

Por configuração da própria tecnologia, existe um limite máximo de registros obtidos, que é de 100 nós por solicitação. Para facilitar o processo de busca em mais de 100 registros, existe um sistema de paginação, que por meio do último registro, chamado de *offset*, é possível obter um código que direciona para a próxima página – ou até a página anterior. Em testes, foi possível obter as postagens de diversas páginas, tanto pequenas quanto grandes. As APIs do

[6] Praticamente todas as páginas utilizam nome de usuário. Por exemplo, na página da Feevale (www.facebook.com/feevale), o usuário é “feevale”.
[7] Todos os nós, sejam páginas, perfis ou grupos, possuem um ID. Aqueles que não possuem um usuário, podem obter o ID na URL. Aqueles que possuem, podem obter por meio de serviços disponibilizados na Internet que permitem a conversão de usuário para ID.

YouTube e Twitter também possuem paginação, sendo de 100 registros neste e 50 registros naquele.

De acordo com Courtois, Mechant e Marez (2011), a API do YouTube permite a interação com a plataforma, bastando ter uma conta no Google. A partir disto, é possível fazer diversas consultas para os recursos (*resources*), que são entidades com dados individuais, como um canal, um vídeo ou uma *playlist*. Também é possível obter e atualizar dados disponibilizados.

No que diz respeito ao presente trabalho, focou-se no recurso “pesquisa” (*search*). Existem duas possibilidades: uma pesquisa simples e uma temática. Em ambas se utiliza como filtro um termo desejado, de igual forma que se faz na pesquisa do *site*. É possível filtrar o tipo de recurso desejado (vídeo, canal ou *playlist*), definir região ou idioma de pesquisa, entre outros. Embora a documentação apresente que é possível filtrar por data, os testes demonstraram o contrário. É possível ainda uma pesquisa temática, informando o código do tópico – existem vários tópicos, como “Negócios” e “Saúde”, tendo cada um o seu respectivo código.

A API do Twitter, mencionada por Lomborg e Bechmann (2014), permite o acesso a informações publicadas ou compartilhadas de forma pública, e para isto também exige um *token* gerado por uma conta no Twitter. É possível obter ou publicar dados relacionados a contas, usuários, *tweets*, mídia, entre outros.

Com relação a este trabalho, é possível obter as publicações relacionadas a uma palavra ou expressão pesquisada, ou mesmo a uma *hashtag*. É possível também pesquisar a *timeline* de um usuário, seja *tweets* ou menções. Em ambas as opções, ela permite filtrar resultados recentes, resultados populares ou uma mistura destes. Também há a possibilidade de buscar os *tweets* até uma determinada data.

5.2.2 Mídias tradicionais

Conforme mostrado no subcapítulo 3.1.1, é necessário buscar dados em mídias tradicionais, tendo em vista a importância que ainda é dada pela Administração Pública e pela Atividade Política. Essa é uma parte complexa pois, devido a diversidade de *sites* de notícias, era necessário encontrar um “denominador comum” para esta coleta.

Com base em Papaioannou *et al*. (2017) e Fernandes (2018), foi analisado o *site* News API para pesquisa e busca de artigos e notícias em tempo real na Internet,

podendo buscar notícias mais acessadas, novas notícias, menções ou resenhas, compartilhamentos, entre outros. Ele permite filtros por palavra-chave ou expressão, por data, fonte da notícia, domínio do site ou idioma.

Vários testes foram realizados, buscando notícias e artigos com relação aos temas propostos no presente trabalho – no caso, pesquisou-se o termo “Novo Hamburgo”, para tentar obter notícias com relação à respectiva cidade. Porém, o resultado não se mostrou satisfatório, tendo em vista que, por muitos *sites* de notícias não estarem disponíveis para busca, apenas foram obtidas notícias relacionadas ao time de futebol Novo Hamburgo. Também se testou a pesquisa em domínio, porém nenhum dos domínios de notícias da região de Novo Hamburgo/RS pode ser alcançado. Com isto, chegou-se à conclusão de que ele é uma ferramenta ótima para pesquisa em assuntos que tenham foco em temas genéricos, como saúde ou educação, mas não em temas específicos de uma cidade.

Desta forma, para atingir um maior número de *sites* de notícias, optou-se por coletar os dados por meio de *feeds* RSS, conforme diversos autores, como Haßler, Holbach e Maurer (2014), apontam. A tecnologia consiste em um sistema de sumarização de itens de um *blog* ou página Web, que permite a um usuário inscrever-se para receber alertas de atualizações.

Tendo em vista que o trabalho se baseia no uso de API, para que não fosse necessário buscar uma técnica diferente na coleta, baseou-se em Hurtado (2018), que utilizou a API *rss2json* para transformar um *feed* RSS em uma estrutura JSON possível de ser obtida por meio de uma chamada API usando uma URL. Foram feitos testes com vários *sites* de notícias com RSS e em todos a API se mostrou positiva. A desvantagem é a obtenção de um número pequeno de notícias por requisição – geralmente de 10 a 30, conforme a disponibilidade do RSS. Além disto, não há paginação, que permitiria obter mais notícias.

Entretanto, existe um problema: nem todos os *sites* possuem RSS. Por exemplo, os principais jornais e meios de comunicação em Novo Hamburgo, que possuem páginas na Internet, não utilizam *feeds*. Para solucionar tal situação, o subcapítulo 5.2.5 trata de uma ferramenta do Google que permite tal pesquisa.

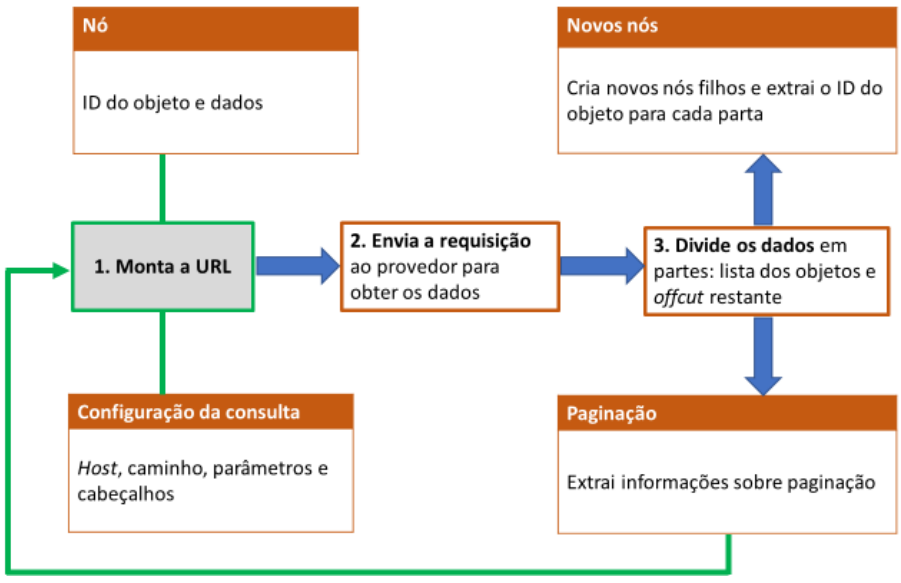
5.2.3 Facepager

Para auxiliar a construção das URLs de chamada às APIs, baseou-se em Salloum, Al-Emran e Shaalan (2017), Haßler, Holbach e Maurer (2014) e Salman (2017) para pesquisar a ferramenta Facepager. Desenvolvida por Jünger e Keyling (2018), é uma aplicação para recuperar dados genéricos por meio de APIs.

Cabe salientar que existem outras ferramentas para requisições API, como o cURL (*Command Line Granddaddy*), o HTTPie e o Hurl.it. Entretanto, as mesmas funcionam em modo texto, ou, no caso do último, exige o preenchimento manual dos parâmetros desejados.

Conforme demonstrado na Figura 3, a ferramenta monta uma URL a partir de dados inseridos sobre nós (ID do objeto e dados) e das configurações da pesquisa (caminho, parâmetros, cabeçalhos e host). Em seguida, ela busca os dados na Internet, para, então, serem divididos em nós filhos, que permitem a consulta ou o reuso para novas buscas ou exportação, e em paginação, extraindo a informação para percorrer as demais páginas – como dito anteriormente, quando a busca supera o número máximo de linhas retornadas.

Figura 3 - Fluxo de trabalho do Facepager



Fonte: Traduzido de Jünger e Keyling (2018)

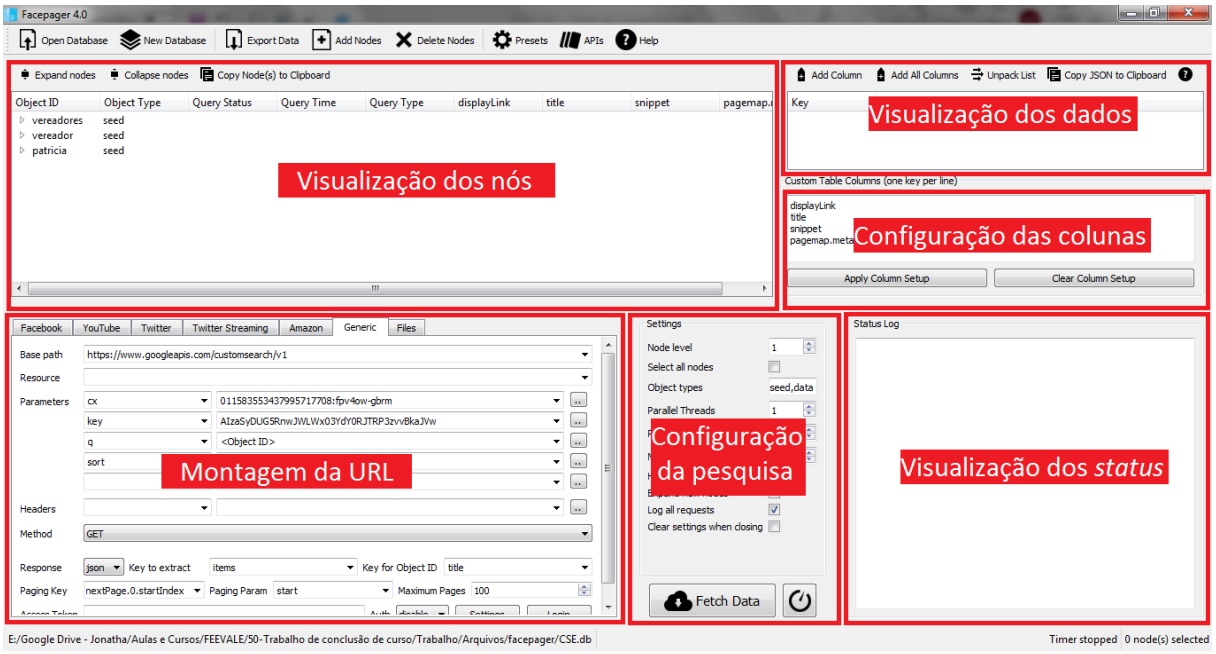
Por meio de dispositivos, como os campos substitutos (*placeholders*) que auxiliam na montagem das URLs e no reuso das pesquisas anteriormente

realizadas, e as chaves (*keys*), permitindo a exibição ou uso de valores encadeados retornados de uma pesquisa, é possível facilmente obter dados de APIs.

Algumas vantagens que ele apresenta são, por exemplo, ter opções nativas de busca no Facebook, Twitter e YouTube, bem como em outros endereços que usam API, e até mesmo dentro de arquivos. Não é necessário informar *token* de acesso nas mídias sociais, sendo apenas necessário fazer o *login*. Também há a definição livre das colunas a serem exibidas, o uso de bancos de dados para armazenar os dados coletados e o armazenamento de configurações predefinidas (*presets*) para uso posterior. Tais funcionalidades podem ser vistas na Figura 4.

Cabe salientar que, embora a janela possua várias opções, devido às predefinições previamente configuradas, o uso do *software* é facilitado.

Figura 4 - Layout da tela do Facepager



Fonte: Adaptado de Jünger e Keyling (2018)

As coletas de dados realizadas na Graph API (Facebook), Twitter API, YouTube API, News API e rss2json foram preparadas e efetuadas com o auxílio do Facepager. Ele se mostrou ágil e prático na pesquisa e obtenção dos dados, bem como na montagem da URL, com baixa complexidade. Outra vantagem é a reutilização dos dados, que foram obtidos previamente, para realizar novas pesquisas, ou ainda para consultas posteriores, pois ele os armazena em arquivos de bancos de dados.

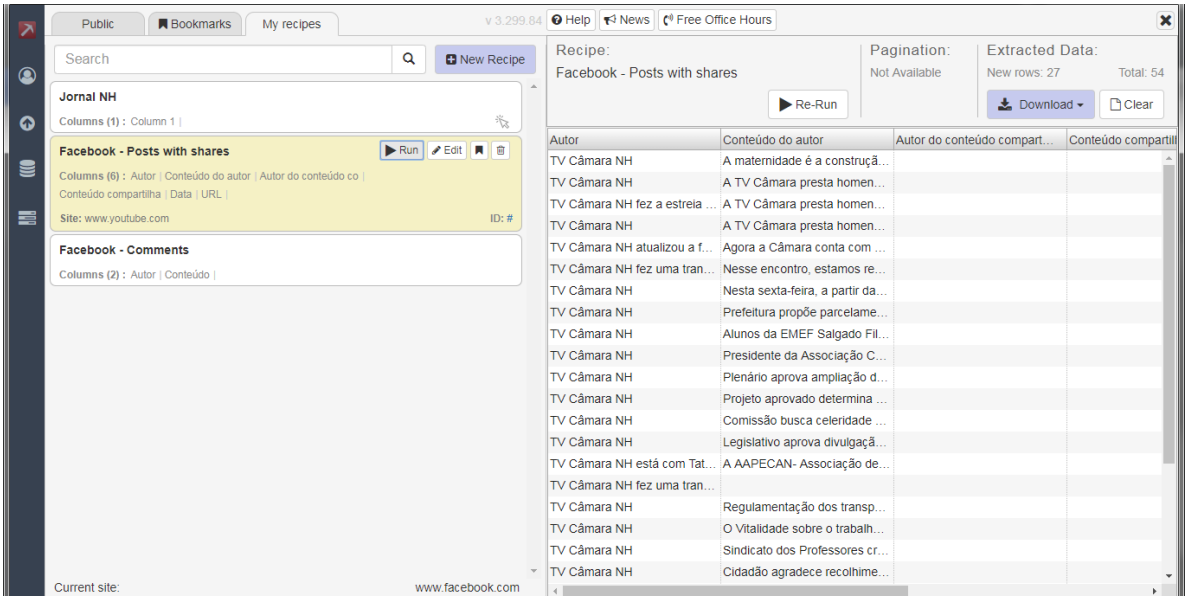
5.2.4 Data Miner

Como dito anteriormente, ao ser estudada a documentação das APIs das mídias sociais, obteve-se a informação de que em fevereiro de 2018 o Facebook removeu funcionalidades da sua API, como a pesquisa em grupos e perfis. Tendo em vista que isto prejudicaria o trabalho, visto que grupos e perfis são amplamente usados, entre outras coisas, para discussões políticas, debates sobre questões relacionadas as cidades, fazia-se necessário buscar uma alternativa viável para obter tais dados, tendo em vista que por meio da API não seria possível.

Após exaustiva pesquisa, encontrou-se o aplicativo Data Miner. Desenvolvido pela Software Innovation Lab, ele consiste em uma extensão do navegador Google Chrome, que permite extrair dados de qualquer página da Internet, por meio de receitas (*recipes*), que são pré-configurações estruturadas de acordo com a disposição dos dados exibidos. Isto é possível graças à estrutura HTML exibida em uma página, visto que ele utiliza *tags*, classes, *ids* e outros componentes de um código HTML para obter determinado tipo de dado. Ele consegue trabalhar inclusive com navegação por páginas de resultados, escolha de colunas exibidas e exportação dos dados coletados em formato de planilha ou de texto CSV.

Após estudar a ferramenta, foi possível criar *recipes* específicas para postagens e comentários, seja para grupo, perfil ou página, visto que a disposição das informações é o mesmo nestas três funcionalidades do Facebook. Cabe aqui salientar que tal coleta depende do acesso do usuário ao conteúdo, ou seja, ele só pode coletar as postagens ou comentários em grupos públicos ou que ele participa. Isso vale para perfis, buscando apenas postagens públicas de uma pessoa ou todas as postagens dos seus amigos. Ou seja, o Data Miner facilita a coleta de dados que já são acessíveis pelo usuário. A janela que exibe as *recipes*, bem como os dados obtidos, é apresentada na Figura 5..

Figura 5 - Exemplos de *recipes* e dados obtidos no Data Miner



Fonte: Do autor (2019)

Havia duas questões que necessitavam serem resolvidas. A primeira era a rolagem infinita, visto que com essa tecnologia, o conteúdo é carregado conforme é feita a rolagem manual do usuário. Ou seja, era necessária alguma forma de rolar automaticamente a página para carregar mais conteúdo a ser importado, sem necessitar uma rolagem manual. Assim, foi encontrada a extensão Scroll it![8] para Google Chrome, desenvolvida por Bolotniuk (2017), que permite a rolagem automática – geralmente 30 rolagens, visto que ele não consegue executar infinitas rolagens. Assim, seria possível obter mais postagens na coleta pelo Data Miner.

Da mesma forma, o outro problema se relacionava aos comentários ocultos, e que só aparecem quando o próprio usuário clica nos *links* de expansão. Para resolver isto, encontrou-se uma ferramenta[9] escrita em JavaScript por Farley (2015), que permite esta expansão – ela funciona como um "favorito" adicionado ao navegador, sendo que, após o clique, ele expande comentários e respostas a comentários nas postagens exibidas na página.

Devido à excelente experiência do Data Miner, foi testada uma *recipe* em um *site* de notícias, o que demonstrou a possibilidade de utilizá-lo tanto para coletar dados em mídias sociais quanto em mídias tradicionais, quando não for possível realizá-lo pelos meios tradicionais, como ocorre com o uso de APIs.

[8] Disponível em: <https://github.com/always-oles/ScrollIt>. Acesso em: 12 mai. 2019
[9] Disponível em: <http://com.hemiola.com/215/08/29/expand-all/>. Acesso em: 12 mai. 2019

5.2.5 Google Search Engines

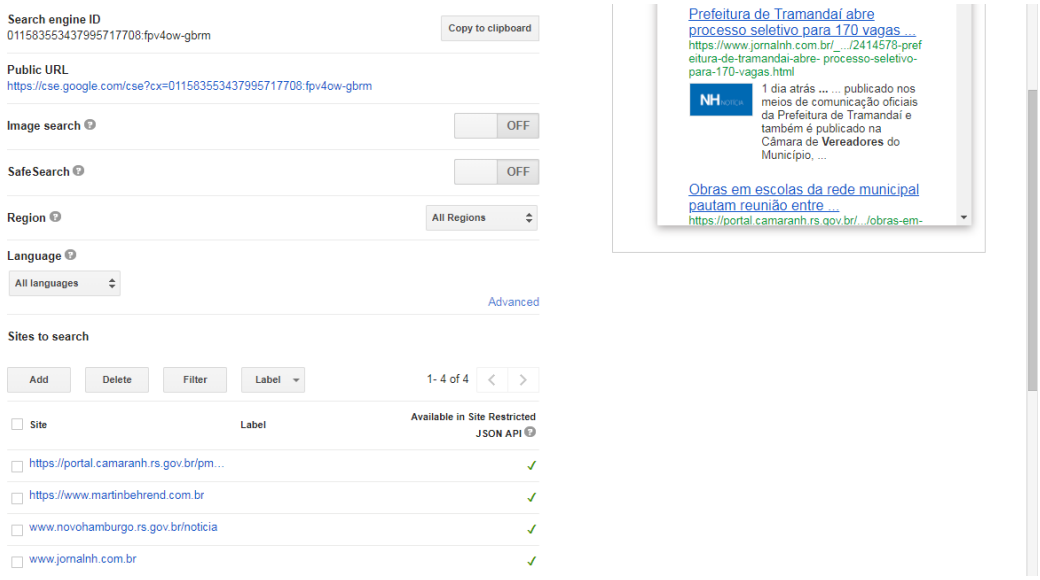
Após os testes realizados no News API, bem como na pesquisa em *feeds* RSS, percebeu-se que muitos *sites* de notícias não permitem sua obtenção, ou por não serem rastreadas pela API, ou por não disponibilizarem *feed* RSS – inclusive há vários portais de notícias que abandonaram o uso, por disponibilizarem meios mais práticos, como extensões de navegadores ou aplicativos para *smartphones* – a ascensão das mídias sociais como agregadores de notícias também motivou isto.

Por exemplo, pouquíssimas notícias foram obtidas dos quatro principais portais de notícias de Novo Hamburgo[10]: o Jornal NH[11], o blog do jornalista Martin Behrend[12], a Prefeitura de Novo Hamburgo[13] e a Câmara de Vereadores de Novo Hamburgo[14]. Desta forma, seria fundamental encontrar uma ferramenta que permitisse tal busca.

A solução encontrada foi apontada por Dhara (2016), bem como por Abdulhayoglu e Thijs (2017): o uso do Google Custom Search Engines (CSE), que consiste em uma API que permite criar ferramentas de pesquisas personalizadas, para busca em um site ou *blog* específico, ou um conjunto destes. Por exemplo, é possível criar uma CSE para pesquisar dentro de determinados *sites* ou *blogs* de notícias, ligados a um determinado município. Um exemplo está na Figura 6:

[10] Existem outros portais de notícias em Novo Hamburgo, como o Jornal Canudos (www.jornalcanudos.com.br) e o Portal Novo Hamburgo (http://novohamburgo.org), porém eles são pouco atualizados.
[11] Disponível em: <https://www.jornalnh.com.br>
[12] Disponível em: <https://www.martinbehrend.com.br>
[13] Disponível em: <https://novohamburgo.rs.gov.br/noticia>
[14] Disponível em: <https://portal.camaranh.rs.gov.br/pm3/informacao_e_conhecimento/noticias>

Figura 6 – Tela de configurações de uma CSE



Fonte: Do autor (2019)

Ela também possui, além da funcionalidade de escolher *sites* para buscar o conteúdo, a possibilidade de escolher região ou idioma, a busca em imagens, opções de filtro por data e ordenamento por data. A ferramenta pode ser usada, por padrão, com uma caixa de busca personalizada, semelhante à do Google, mas com a possibilidade de ser implantada dentro de um *site*. Seguindo o proposto no trabalho, obteve-se a possibilidade de retornar a busca via API, usando o Facepager.

5.3 FERRAMENTAS DE PRÉ-PROCESSAMENTO

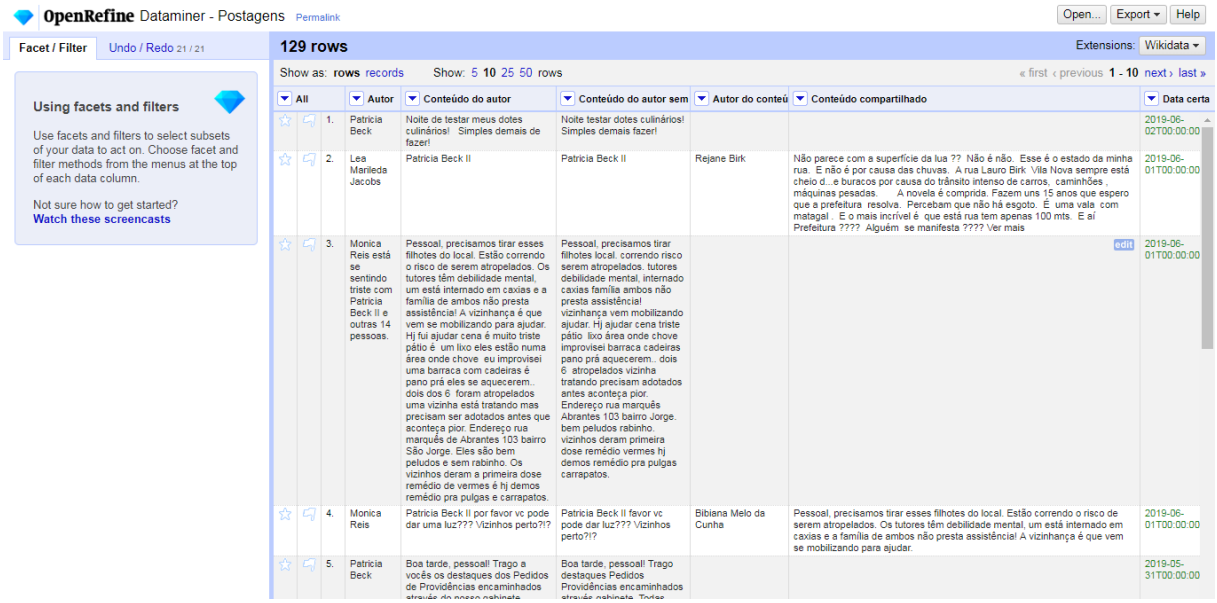
Para esta fase, foram testadas e validadas duas ferramentas: o *software* OpenRefine, e a linguagem de programação R.

5.3.1 OpenRefine

Lomborg e Bechmann (2014), Silva e Stabile (2016) e Batrinca e Treleaven (2014) fazem menção à esta ferramenta, anteriormente chamada de Google Refine. Ela é usada para limpeza e transformação de dados para vários formatos, e trabalha com a ideia de tabelas no estilo de planilhas eletrônicas. Em síntese, permite a análise e organização dos dados, antes de que elas ingressem em ferramentas de extração de informação.

A ferramenta permite a importação do texto por diversos formatos, como TSV, CSV, planilhas, JSON e XML. É possível, inclusive, selecionar bancos de dados, conteúdo copiado da área de transferência e planilhas on-line no Google. Após alguns testes, verificou-se que a importação de planilhas deve ocorrer no formato XLS, visto que não foi possível fazer no formato XLSX. Além disto, durante a importação de arquivos CSV exportados no Facepager, não houve organização correta das colunas. Portanto, entendeu-se que a forma mais precisa de importação é a conversão para o formato XLS. Após isto, os dados ficam à disposição para manipulação, conforme Figura 7:

Figura 7 - Manipulação de dados no OpenRefine



Fonte: Do autor (2019)

Após a importação, inúmeras funcionalidades de limpeza e preparação são fornecidas pela ferramenta, que permitem uma completa limpeza e organização dos dados. É possível excluir colunas e linhas, sendo que a exclusão destas ocorre com ou sem a necessidade de filtros aplicados. Também há a possibilidade de marcar/sinalizar as linhas ou registros com uma estrela ou com uma bandeira.

Duas funções importantes são as de “filtros” e “facetas”. A primeira permite filtrar em uma coluna; a segunda, mais complexa, é uma mescla de filtros e agrupamentos conforme o tipo de dado, como texto, número e data – há várias outras opções, inclusive personalizáveis. Por exemplo, é possível visualizar uma linha do tempo para colunas de data, um histograma para colunas de número, ou uma lista de categorias agrupadas para colunas de texto. Isso permite filtrar a busca

por categorias, períodos, valores, entre outros. É possível ainda criar uma categoria em uma coluna, renomeando determinados agrupamentos já existentes.

Também é possível converter colunas ou registros para determinados tipos de texto. Por exemplo, para criar uma linha do tempo, é necessário que as colunas sejam do tipo “data”. Para isto, basta fazer a conversão para esse tipo. Também é possível converter para outros formados, como número, texto em minúsculo/maiúsculo, entre outros. Porém, o mais interessante é a possibilidade de transformar dados de uma coluna por meio de comandos de programação, usando as linguagens e Jython (implementação da linguagem Pyhton escrita em Java).

Por meio de códigos nestas linguagens foi possível converter uma coluna com data em formato de texto – às vezes, até por extenso – em uma de formato de data. Esta é uma das principais vantagens da ferramenta, que inclusive permite que essas alterações impactem na coluna ou que gerem uma nova. É ainda possível renomear ou mover colunas, bem como, dividi-las por meio de caracteres ou fazer união de colunas.

Cabe salientar que todas as mudanças podem ser desfeitas ou refeitas – inclusive, tal funcionalidade é atrelada ao projeto, ou seja, é possível, após abrir um projeto anteriormente fechado, refazer ou desfazer ações. Este é um dos motivos pela qual a ferramenta possui avançada posição em relação às planilhas eletrônicas convencionais. Por fim, é possível exportar os dados por diversas maneiras, como TSV, CSV, planilha XLS, XLSX ou ODS, código SQL ou ainda uma exportação customizada, onde são selecionadas as colunas a serem exportadas, o formado em que o arquivo será salvo e outras configurações.

Todas essas funcionalidades permitem que o OpenRefine, embora seja uma ferramenta de pré-processamento, exiba informações extraídas, antes mesmo de chegar na respectiva fase. Agrupamentos de coluna permitem obter a distribuição de conteúdo conforme cada grupo. A linha do tempo também permite isto, além de delimitar um tempo desejado. Em síntese, além de limpar e preparar, o OpenRefine permite filtrar os dados.

5.3.2 Linguagem de programação R

Inúmeros autores, como Williams (2009), Meyer, Hornik e Feinerer (2008) e Fellows (2012) tratam a respeito da linguagem R. Entretanto, embora possua os

conceitos e paradigmas de linguagens de programação tradicionais, ela é uma poderosa ferramenta estatística que permite, por meio de vários pacotes disponíveis, o uso de técnicas computacionais para solução de diversos problemas e necessidades de grande complexidade (FELLOWS, 2012).

Meyer, Hornik e Feinerer (2008) expõem de forma clara e completa quão poderosa é essa linguagem quando aplicada à mineração de texto, principalmente no tocante ao pré-processamento e à extração de informação..

É importante salientar que, como afirma Fellows (2012), a ferramenta precisa ser chamada e usada por meio de comandos de texto e, por esse motivo, é necessário que o usuário possua conhecimento de nível básico sobre algoritmos para entender o funcionamento das chamadas de funções ou atribuições de variáveis. Esta questão mostrou-se vital para o andamento do trabalho, e que culminou no Capítulo 6.

Para o uso da linguagem, existe o software RStudio, que permite sua configuração, a escrita de códigos de programação, *console* para execução de comandos, tela de visualização de saídas, entre outros. Entretanto, é importante colocar desde já que, diferente de ferramentas abordadas no trabalho, o RStudio não será selecionado, pois ele é apenas um meio para que se use a linguagem R, por meio de códigos de programação, e não um programa de *text mining* em si.

A mineração de texto no R se dá pelo pacote *tm* (FEINERER; HORNIK, 2018), que possui um conjunto de funções que envolvem tanto o pré-processamento como a extração de informação[15]. Existem outros pacotes que são importantes para a análise de textos e pré-processamento, como o *tidyverse* (WICKHAM, 2017) e o *tidytext* (QUEIROZ *et al.,* 2019), que são focados na manipulação de texto, e o *quanteda* (BENOIT *et al,,* 2019), que permite a divisão de texto por frase – por exemplo, ao invés de se analisar um parágrafo, analisa-se cada uma das frases contidas nele, individualmente.

O texto deve ser carregado por meio de um arquivo. Embora tenham sido feitos testes em arquivos TSV e CSV, aconselha-se o uso de arquivos TXT. Outro ponto a ser considerado é a acentuação e a presença de caracteres especiais, visto que, após alguns testes, percebeu-se que o pré-processamento falhava em palavras

[15] Existem pacotes que permitiriam a obtenção dos dados por meio de API, porém tal questão não será analisada, visto que já foi obtido com sucesso uma ferramenta para tal busca – o Facepager.

com mais de um caractere especial ou com acento. Desta forma, retirou-se toda a acentuação do texto carregado.

O texto é transformado em um objeto do tipo “Corpus” que, segundo a documentação do pacote, baseia-se em coleções de documentos[16] que contém linguagem natural. A partir desta conversão, é possível manipular os textos carregados, principalmente fazendo-se uso da função *tm_map*. Por meio dela, ocorre, por exemplo, a remoção de caracteres especiais, a remoção de pontuação, números, espaços em branco excedentes, termos relacionados a URLs ou endereços de e-mail, bem como a remoção de *stopwords* – inclusive é possível fazê-la por meio da indicação de um idioma e/ou de uma lista de *stopwords*[17].

Para que possa ser manipulado por diversas funções da etapa de extração de informação, é gerada uma matriz de palavras – a chamada *TermDocumentMatrix* -, por meio da qual pode-se fazer diversas operações, como a filtragem e a ordenação dos registros. Também é possível converter tais dados em um *Data Frame*, que é um objeto usado para armazenar tabelas de dados.

5.4 FERRAMENTAS DE EXTRAÇÃO DE INFORMAÇÃO

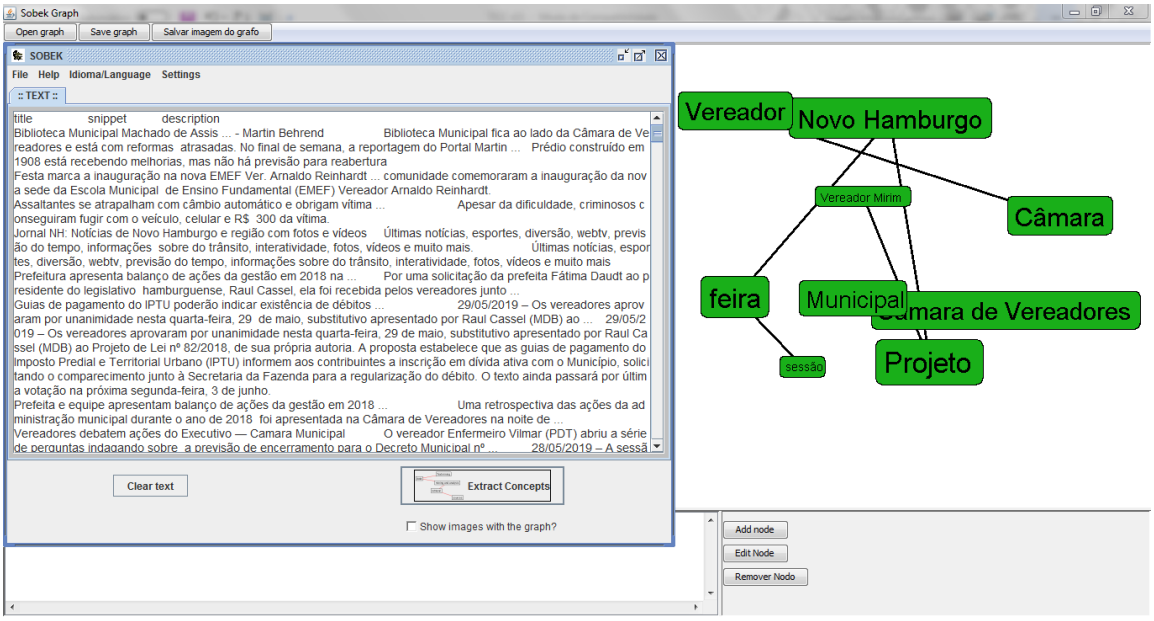
A fase final da mineração de texto se dá com a extração de informação. Foram analisadas e validadas duas ferramentas: o Sobek e a própria linguagem R.

5.4.1 Sobek

Mencionada por Klemann, Reategui e Rapkiewicz (2011), Macedo *et al*. (2016), bem como Oliveira, Azevedo e Gomes (2019), a ferramenta foi desenvolvida em 2007 pelo Grupo de Pesquisa GTech.Edu, da Universidade Federal do Rio Grande do Sul (UFRGS). Busca ser uma ferramenta para ajudar professores a revisarem trabalhos de alunos. Posteriormente, começou a ser usado na compreensão de leitura e na sumarização de texto, bem como em outras questões relacionadas à mineração de textos.

[16] Cada linha carregada de um arquivo é caracterizada como “documento” pelo Corpus.
[17] O OpenRefine não possui de forma nativa a remoção de *stopwords*, mas demanda o uso de códigos em Jython ou em GREL.

Figura 8 - Interface do Sobek



Fonte: Do autor (2019)

Possui uma interface muito simples, conforme Figura 8, sendo possível colar um texto ou carregar de um arquivo – em alguns testes, arquivos muito grandes apresentaram problemas e não puderam ser carregados. Após o processo, basta clicar no botão para extrair os conceitos. Após o processamento, é gerada uma rede de grafos, que mostra a relação entre palavras, bem como a importância e relevância de cada uma delas dentro do texto. Os grafos são manipuláveis, os nós podem ser removidos, e, ao clicar em um nó, é possível ver em quais frases a palavra é usada – aliás, não apenas palavras, mas também expressões.

No software ainda é possível definir o número médio de conceitos (palavras ou expressões) a serem exibidos, definir as *stopwords* que serão utilizadas – o Sobek realiza este pré-processamento de forma nativa – e definir a frequência mínima, ou seja, o número mínimo de vezes que determinada palavra ou expressão aparece.

5.4.2 Linguagem de programação R

Assim como exposto anteriormente, a linguagem R possui várias ferramentas que permitem a extração de informação a partir de um texto. Novamente, os testes foram realizados usando o RStudio, e este assunto será abordado em detalhes no capítulo 6.

A partir do *DocumentTextMatrix* e, após, do *DataFrame* gerado na fase de pré-processamento, é possível gerar inúmeras saídas, tanto gráficas quanto textuais. Como saídas textuais, mas que exibem ótimas informações, há dois comandos: o *findFreqTerms*, que exibe os termos mais frequentes de uma *DocumentTextMatrix*, conforme a correlação mínima; e o *findAssocs*, que analisa associações/correlações entre palavras com base em uma palavra.

A nuvem de palavras pode ser gerada a partir do pacote *wordcloud2* (LANG; CHIEN; 2018), onde basta informar o *DataFrame* e configurar algumas outras opções, como a frequência mínima a ser exibida na nuvem e questões estéticas, como rotação, fonte e cores. Inclusive, a nuvem é interativa, pois ao posicionar o mouse em cima de uma palavra, aparece a quantidade de ocorrências dela no texto.

Outra ferramenta gráfica é o histograma, que, a partir de dados inseridos, permite criar gráficos estatísticos sobre as palavras. Por meio do pacote *ggplot2* (WICKHAM *et al.*, 2019), é possível gerar várias opções de gráficos, como o de frequências e de correlações. A geração de gráficos é riquíssima, e possui inúmeras funcionalidades, como definições sobre títulos, eixos, fontes, entre outros.

Com boas referências ao Sobek, é possível mencionar a construção de grafos de palavras, por meio do pacote *visNetwork* (THIEURMEL; ROBERT, 2019). Ele funciona a partir de um *DataFrame*, onde são analisadas as correlações entre as palavras. Após definida a quantidade de palavras e a correlação mínima entre elas, são exibidas as ligações entre as palavras (representadas por círculos) por meio de setas.

Ainda é possível, por meio do pacote *lexiconPT* (GONZAGA, 2017), realizar a análise de sentimentos das frases ou parágrafos, a partir de dois dicionários léxicos: o OpLexicon (SOUZA *et al*., 2011) e o SentiLex (SILVA; CARVALHO; SARMENTO, 2012). O processo consiste em analisar cada palavra de uma frase ou parágrafo/linha e indicar se a mesma é positiva, neutra ou negativa. Usando o pacote *ggplot2*, podem ser gerados gráficos de dispersão e histogramas.

Embora haja outros pacotes que permitem a obtenção de mais informações, verifica-se que tais funcionalidades são suficientes para realizar a extração de informação a partir do R, devido à quantidade de informações que permitem gerar e, a partir delas, conhecimento que pode trazer novos entendimentos a respeito do uso das mídias sociais, bem como da repercussão em mídias tradicionais.

5.5 CONSIDERAÇÕES SOBRE O CAPÍTULO

O capítulo que se encerra abordou sobre as ferramentas que serão utilizadas no presente trabalho, destacando os pilares da seleção. Em seguida foram abordadas as ferramentas de coleta, tanto para mídias sociais quanto para as tradicionais - incluindo as dificuldades encontradas e as soluções aplicadas. Também foram abordadas as ferramentas da fase de pré-processamento, bem como as de extração de informação. Ao longo do capítulo, foram expostas as ferramentas selecionadas, bem como as que foram testadas mas que não foram aprovadas.

Foram pesquisadas e testadas 28 (vinte e oito) ferramentas, sendo que, por não atenderem ao dois pilares apontados, foram excluídas 22 (vinte e duas), listadas no subcapítulo 5.1, e foram selecionadas cinco: Facepager, Data Miner, Google Custom Search Engines, OpenRefine, Sobek.

Assim, considera-se que o capítulo atingiu os dois objetivos específicos de investigar de que forma é possível obter dados de mídias sociais e tradicionais, com base nos conceitos de mineração de texto, bem como, analisar as ferramentas existentes, que permitem a obtenção dos dados e a sua posterior mineração e extração de informação.

Entretanto, durante o estudo da linguagem de programação R, entendeu-se que não seria possível selecioná-la, visto que exige conhecimento de programação do usuário, fugindo totalmente dos pilares determinados. Mesmo o uso do RStudio – o que poderia melhorar a tarefa de programação – não seria capaz de reduzir uma possível dificuldade no uso.

Sendo assim, percebeu-se que seria necessário reunir todas as funcionalidades de mineração de texto em uma ferramenta gráfica simples, abstraindo do usuário final qualquer código, mas apenas usando *widgets* simples. Por isto, o próximo capítulo tratará do desenvolvimento desta ferramenta, que contém o pré-processamento e a extração de informação. Desta forma, adiciona-se ao conjunto de cinco ferramentas selecionadas a linguagem R.

6 DESENVOLVIMENTO DE FERRAMENTA

Conforme relatado anteriormente, verificou-se a alta capacidade da linguagem R em atuar em todo o processo da mineração de texto, sendo ela capaz de dar origem a funcionalidades e ferramentas. Porém, devido ao foco da simplicidade, conforme exposto no subcapítulo 5.1, entendeu-se que uma linguagem de programação fugiria disto, pois exige do usuário conhecimentos de algoritmos e linguagem de programação. Na verdade, a linguagem enquadra-se no grupo de ferramentas de maior complexidade, que demandam conhecimento técnico.

Desta forma, devido ao conhecimento proporcionado pelo Curso de Sistemas de Informação, e após pesquisa, verificou-se que seria possível agrupar as funcionalidades apresentadas nos subcapítulos 5.3.2 e 5.4.2, que tratavam da linguagem, e utilizá-las por meio de uma interface gráfica, simples e de fácil uso a qualquer pessoa.

Para isto, encontrou-se o pacote *shiny*, que permite a construção de aplicativos interativos para a web, com uma facilidade considerável. Desta forma, ao invés de utilizar a linguagem R por meio de código haveria um *software* específico, utilizando toda a tecnologia proporcionada pela linguagem R, abordando o pré-processamento e, principalmente, a extração de informação.

Cabe salientar que, ao mencionar que a linguagem R é capaz de atuar em toda a mineração de texto, inclui-se aqui a coleta de dados. Entretanto, entendeu-se não colocar tal etapa no desenvolvimento da ferramenta por dois motivos:

O primeiro é que a decisão de criar a ferramenta ocorreu apenas quando do andamento das etapas finais de pesquisa das ferramentas, e a seleção das relacionadas à coleta de dados e ao pré-processamento foi plenamente satisfatória – inclusive resolvendo o problema da coleta em páginas e grupos do Facebook. O segundo é que a ferramenta é um meio para alcançar o objetivo, e não o objetivo em si, ou seja, desenvolver a ferramenta não era a intenção do presente trabalho, mas mostrou-se um passo fundamental e necessário para atingir o objetivo.

6.1 RANALYZER

A ferramenta desenvolvida consiste em uma interface web, baseando-se em todas as funcionalidades explicadas nos subcapítulos 5.3.2 e 5.4.2, onde as

operações de seleção do conteúdo a ser analisado e a posterior análise e extração de informação podem ser realizadas, conforme a Figura 9.

Figura 9 - Tela do RAnalyzer

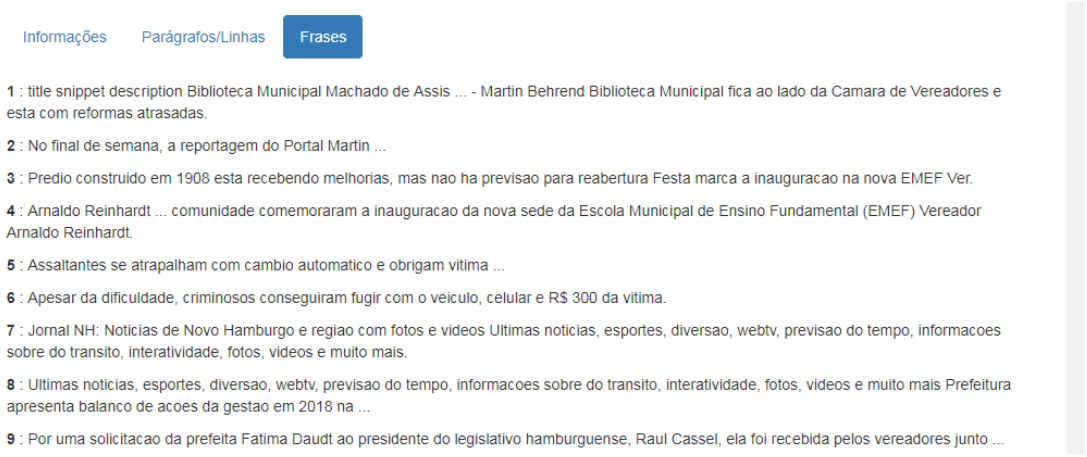
Fonte: Do autor (2019)

Basicamente, a ferramenta possui uma barra de opções (abas), onde cada uma dá acesso a um recurso gráfico, apresentando a análise e extração, bem como opções de filtro, que alteram o recurso exibido – seja um gráfico, uma tabela ou outro.

A primeira aba dedica-se a carregar o arquivo a ser analisado, que deve estar salvo no formato TXT. O procedimento de envio e processamento é controlado pela aba Arquivo. Após o arquivo ser carregado – sendo possível escolher o número máximo de linhas a ser carregado, com a finalidade de não prejudicar o processamento –. o *software* executa o procedimento de pré-processamento, exatamente como descrito no subcapítulo 5.3.2.

Além disto, são exibidas as informações básicas, como nome, quantidade de linhas (parágrafos), quantidade de frases e quantidade de palavras. Também é possível ver o texto na íntegra, de duas formas: dividido por linha/parágrafo ou divido por frase (Figura 11). Ao lado de cada linha ou frase, é exibido o respectivo número da linha ou frase.

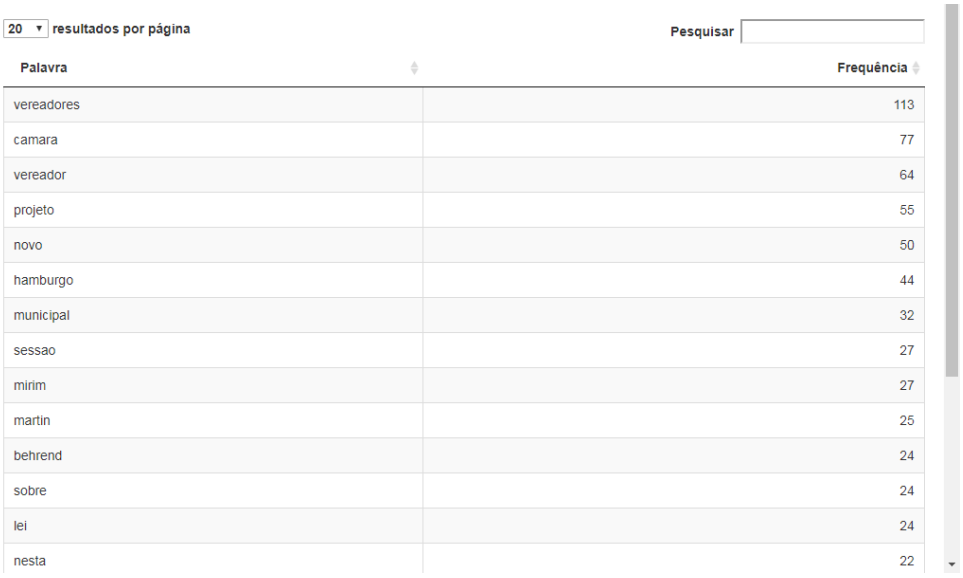
Figura 10 - Texto dividido por frases



Fonte: Do autor (2019)

Após isto, é possível escolher qualquer uma das outras abas. A próxima aba é a Lista de Palavras. A mesma é responsável, como o nome diz, por exibir a lista completa de palavras (Figura 11), com a sua respectiva frequência de ocorrência. Tem como diferencial a possibilidade de filtrar as palavras, ordenar as colunas por palavra ou frequência em ordem crescente ou decrescente, e ainda estabelecer a frequência mínima de ocorrência – de 1 a 200 ocorrências, sendo que selecionando o valor mínimo de 1, todas as palavras são exibidas.

Figura 11 - Lista de palavras



Palavra	Frequência
vereadores	113
camara	77
vereador	64
projeto	55
novo	50
hamburgo	44
municipal	32
sessao	27
mirim	27
martin	25
behrend	24
sobre	24
lei	24
nesta	22

Fonte: Do autor (2019)

Em seguida, está a aba Nuvem, que exibe uma nuvem de palavras (Figura 12), onde é possível visualizar o grau de frequência das palavras no texto de forma

proporcional – quanto maior a palavra, maior sua frequência. Ao colocar o *mouse* sobre uma palavra, é possível saber quantas vezes a mesma apareceu. A tela possui um ajuste de frequência, variando entre 0.5 e 10, sendo que quanto maior o valor, menor é a frequência máxima considerada e, consequentemente, menos palavras são exibidas.

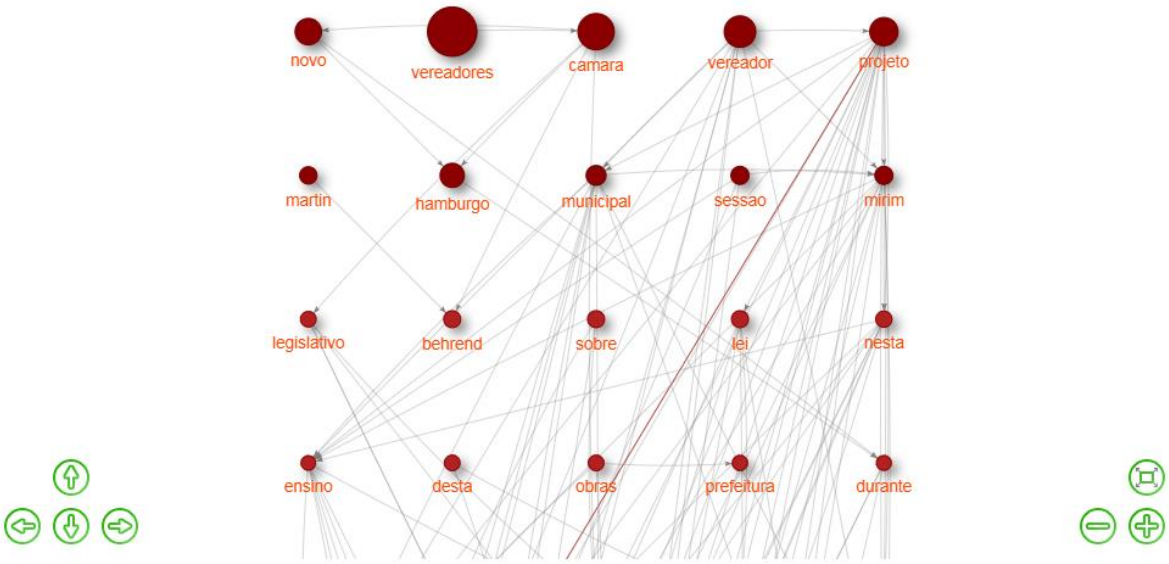
Figura 12 - Nuvem de palavras

Fonte: Do autor (2019)

Na aba Grafo, é possível visualizar a relação entre as palavras em formato de grafo (Figura 13), de forma semelhante a ferramenta Sobek. É possível escolher a quantidade de palavras exibida no grafo (exibir de 10 a 100 palavras, dentre as mais frequentes). Quanto maior o círculo, e quanto mais escuro o tom de vermelho, maior é a quantidade de ocorrências da palavra no texto. Ao colocar o *mouse* em cima da palavra, é exibida tal quantidade.

As ligações entre as palavras são exibidas por meio de setas, sendo que, ao clicar em uma palavra, ele exibe todas as setas que chegam nela e/ou que partem dela. É possível escolher o grau mínimo de correlação entre estas palavras (entre 5 e 100%), sendo que quanto menor o grau, mais setas aparecerão.

Figura 13 - Grafo de palavras



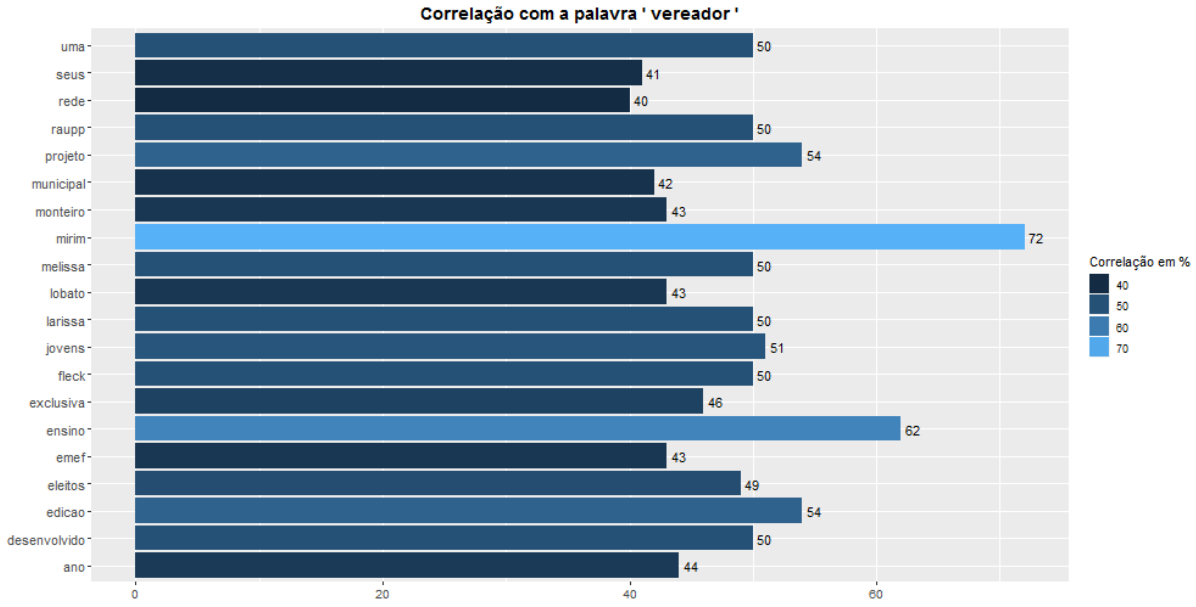
Fonte: Do autor (2019)

Cabe salientar que a versão atual foi desenvolvida ao longo do mês de outubro, após a validação descrita no subcapítulo 8.2. O grafo antigo era estático e menos interativo. Por exemplo, não havia diferenciação no tamanho das palavras conforme a frequência. Além disto, os ajustes de tamanho da imagem eram manuais, e demoravam. Isto ocorreu porque, à época, não se encontrou um mecanismo prático para visualizar o grafo, tal qual no Sobek – o que corroborará, na conclusão, o contínuo aperfeiçoamento desta ferramenta.

A aba Frequências dedica-se a expor de forma simples a frequência das palavras por meio de um gráfico de barras (histograma) - inclusive utilizando a cor azul em diferentes tons, sendo que quanto mais claro, maior é a frequência. É possível selecionar o número máximo de palavras a serem exibidas – entre 5 e 30 – e a frequência de ocorrência das palavras, da mesma forma que na aba Grafo.

Outro histograma é disponibilizado na aba Correlações, mas com uma funcionalidade interessante: detalhar correlações entre palavras. A partir de uma palavra selecionada, dentre as que constam no texto, são exibidas as palavras que mais se relacionam com esta – relação expressa em porcentagem, conforme Figura 14. É possível exibir de 5 a 30 palavras, bem como ajustar a correlação mínima de exibição.

Figura 14 - Histograma de correlação



Fonte: Do autor (2019)

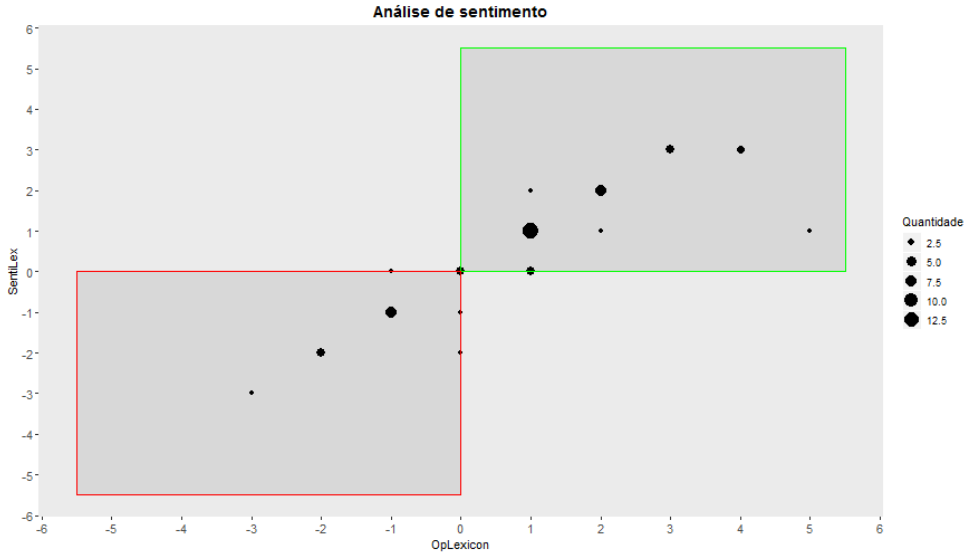
Cabe aqui ressaltar que a quantidade de palavras exibida nestas duas abas é influenciada por duas variáveis: frequência de ocorrência e correlação. Ou seja, mesmo que na aba Frequências esteja selecionada a quantidade de 30 palavras, se a correlação escolhida for alta, existe a possibilidade de serem exibidas menos de 30 palavras no gráfico.

Outro ponto importante de salientar é que existe uma diferença entre este histograma e o grafo de correlações. Enquanto naquele são exibidas as palavras que mais se correlacionam com a selecionada, neste são exibidas as correlações dentro do grupo de palavras mais frequentes. Ou seja, uma palavra pode aparecer no histograma mas não aparecer no grafo, visto que não está entre as mais frequentes.

Por fim, para enriquecer a ferramenta, foi adicionada uma aba de Análise de Sentimentos. Após a análise feita pelos dois dicionários léxicos, mencionados no subcapítulo 5.4.2, é feito um somatório dos pontos atribuídos, para, então, chegar-se à conclusão se o conjunto de frases ou parágrafos é negativo ou positivo. Isto ocorre por meio de um gráfico de dispersão, onde o eixo das abcissas representa o dicionário SentiLex – quanto mais à direita, mais positivo o texto é – e o eixo das ordenadas representa o dicionário OpLexicon – quanto mais acima, mais positivo o texto é.

O gráfico (Figura 15) vai de -5 a +5, sendo que são tirados os *outliers*, considerados, segundo pesquisa, fora destes limites. Cada ponto representa a pontuação daquele conjunto de frases ou parágrafos, tanto para um quanto para o outro dicionário, e quanto maior o tamanho do ponto, mais frases/parágrafos receberam aquela pontuação. Para facilitar o entendimento, destacou-se em verde (ou azul) os pontos positivos para ambos os dicionários, e em vermelho os pontos negativos para ambos.

Figura 15 - Gráfico de análise de sentimentos

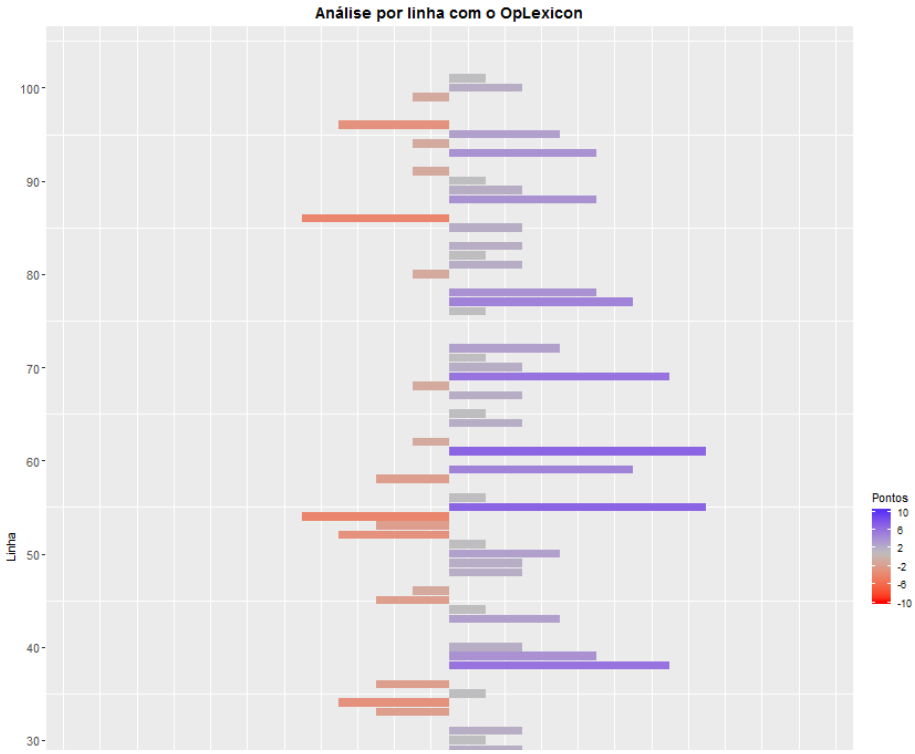


Fonte: Do autor (2019)

Entretanto, ao analisar a ferramenta, verificaram-se duas questões. A primeira é que ele analisava apenas parágrafos. Ou seja, um texto completo, sem parágrafos, teria apenas um ponto no gráfico. Por esse motivo, adicionou-se uma opção à esquerda, que consiste em escolher se a análise se dará por parágrafo ou frase. Por parágrafo, entende-se cada linha do arquivo de texto. Por frase, tem-se a divisão, por meio de um pacote do R, dos textos e parágrafos em frases.

Outra questão é que a quantidade de pontos divergia do total de linhas, parágrafos ou frases. Após uma análise, percebeu-se que isto acontecia porque é feita uma junção (*join*) entre os dois dicionários. Porém, nem sempre os dois são iguais, o que faz com que, muitas vezes, de um número determinado de linhas, apenas a uma parte delas fosse analisada. Assim, foram criadas duas visualizações adicionais, que exibem um histograma (Figura 16) para cada dicionário, com o detalhamento do sentimento por linha/parágrafo ou frase.

Figura 16 - Análise de sentimento por linha/parágrafo ou por frase



Fonte: Do autor (2019)

A numeração das linhas/parágrafos ou frases pode ser ajustada, para aumentar ou diminuir o detalhamento das barras, por meio de um controle deslizante à esquerda. Além disto, para analisar cada linha ou frase, é possível ir na aba Arquivo, onde é exibido o detalhamento destes, incluindo a numeração da linha ou frase.

Cabe salientar que outra ferramenta que auxilia nesta análise foi adicionada à esquerda dos gráficos de cada dicionário: são exibidas a frase ou parágrafo com pontuação mais positiva e mais negativa, bem como sua respectiva pontuação. Tais escolhas levam em consideração cada um dos dois dicionários analisados.

Para auxiliar no uso do RAnalyzer, foi adicionado um botão de ajuda em cada aba, denominado “Leia mais”, explicando o funcionamento da tela, os mecanismos ela possui e que tipo de informação ela exibe. Além disto, cada botão/controle possui uma explicação sucinta, acima da mesma.

6.2 DISPONIBILIZAÇÃO E PUBLICAÇÃO

Conforme dito no capítulo 5, buscou-se, ao desenvolver o RAnalyzer, proporcionar facilidade no uso de ferramentas ao usuário comum. Por este motivo,

buscou-se, após a conclusão da ferramenta, uma forma de disponibilizá-la para uso.. Foram analisadas duas opções: a primeira era criar um programa para instalação local, e a segunda era publicar a ferramenta na Internet.

Para a primeira opção, entendeu-se que seria positivo, visto que as demais ferramentas utilizadas possuem instalação local. Havia a possibilidade de criar um instalador, utilizando o pacote RInno[18]. Entretanto, mesmo seguindo os passos pesquisados e indicados, não foi possível criar tal instalador.

Com relação à disponibilização na Internet, o pacote *shiny* possui uma plataforma *online*, chamada “shinyapps.io”, que permite, ao desenvolvedor de R usando o pacote, publicar seu programa na Internet, sem custo. De certa forma, esta era a opção mais promissora a ser trabalhada.

Durante algum tempo, não foi possível disponibilizar o RAnalyzer nesta plataforma, devido a problemas em pacotes utilizados – principalmente relacionados à versão antiga dos grafos. Porém, com a mudança nos grafos implantada em outubro, tal disponibilização foi possível. Após inúmeros testes, o *software* funcionou perfeitamente e passou a funcionar, sem necessidade de instalação, na Internet[19].

Por outro lado, devido aos problemas relatados, partiu-se para uma opção alternativa: um arquivo executável BAT, contendo os comandos de instalação do R e seus pacotes, bem como do próprio Ranalyzer, a fim de garantir que ele será executado e funcionará de acordo com o que foi desenvolvido no presente trabalho.

Os arquivos para instalação do *software*, bem como as instruções de instalação, estão disponíveis no GitHub[20]. Isto é explicado em detalhes no manual escrito para o presente trabalho (APÊNDICE A - Manual de instalação e uso das ferramentas selecionadas). Desta forma, tornou-se possível instalar e usar o *software* tanto pela Internet (*online*), quanto localmente (*offline*).

6.3 CONSIDERAÇÕES SOBRE O CAPÍTULO

Este capítulo teve como foco abordar a ferramenta RAnalyzer, que foi desenvolvida para permitir o uso da linguagem R e de todo o seu leque de ferramentas pelo usuário comum. Basicamente, além de explicar o contexto de

[18] Disponível em: <https://github.com/ficonsulting/RInno>. Acesso em: 08 jul. 2019.
[19] Disponível em: <https://ojonathacardoso.shinyapps.io/ranalyzer/>. Acesso em: 27 out. 2019.
[20] Disponível em: <https://github.com/ojonathacardoso/ranalyzer>. Acesso em: 05 ago. 2019.

criação, foram apresentadas todas as suas funcionalidades, bem como foi explicado de que forma a mesma foi disponibilizada para uso.

Sendo assim, o capítulo mostrou-se complementar ao objetivo específico de analisar as ferramentas existentes, que permitem a obtenção dos dados e a sua posterior mineração e extração de informação. Isto será fundamental para atingir o objetivo específico de criar o modelo de extração de informação, baseado nos conceitos de *text mining* e utilizando as ferramentas selecionadas e testadas – o foco do próximo capítulo.

7 MODELO DE EXTRAÇÃO DE INFORMAÇÃO

Todo o trabalho escrito até o presente momento – e principalmente a partir do capítulo 4, sobre mineração de texto – teve como foco construir a base necessária para que se chegasse neste ponto: a criação do modelo de extração de informação , que é o objetivo geral estabelecido.

Entretanto, conforme deu-se tal construção, observou-se que seria, não apenas possível, mas também importante que a criação do modelo fosse pautada pela sua genericidade, ou seja, que seu foco não estivesse exclusivamente nas mídias sociais e/ou tradicionais. Utilizar qualquer tipo de fonte para realizar a mineração de texto, seria um diferencial estratégico para ele: possibilitar que outras fontes da Internet, arquivos, bancos de dados, entre outros, pudessem ser atendidos por este modelo.

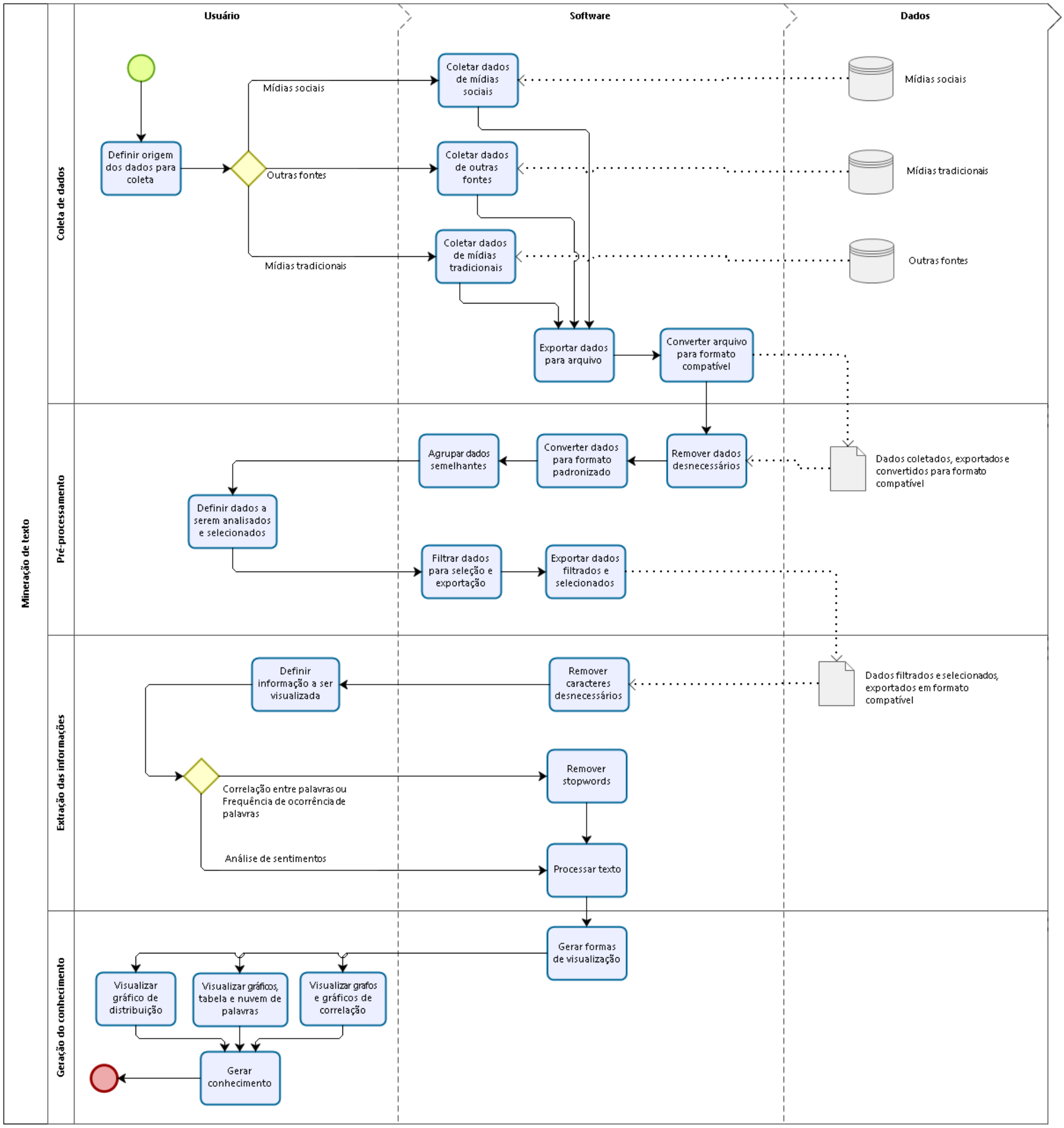
Além disto, devido à evolução tecnológica, como já mencionado várias vezes no trabalho, embora as ferramentas pesquisadas e selecionadas sejam *software* livre, não é plausível que o modelo baseie-se nelas para que o processo de mineração ocorra – basta lembrar que o Facebook, por exemplo, removeu de sua API, em 2018, o acesso à grupos e perfis por meio de requisições via URL. Assim, confirmando a sua genericidade, o modelo não está baseado em ferramentas. Na prática, o conjunto de ferramentas de obtenção dos dados pode mudar conforme a fonte desejada, conforme as funcionalidades disponibilizadas e a necessidade do usuário.

Ou seja, para atingir o objetivo do trabalho, criou-se um **modelo universal**, **aplicável a qualquer fonte de texto**, o qual permite transformar dados/texto bruto em **informação que gere conhecimento**, **sem que haja vínculo a qualquer conjunto de ferramentas específico**.

Tendo em vista que o trabalho possui como importante pilar a mineração de texto, entendeu-se como fundamental criar o modelo a partir do processo de mineração de texto. Desta forma, o modelo será uma representação teórico-prática detalhada das técnicas expostas no subcapítulo 4.3, tendo em vista a alta capacidade delas em extrair informação e conhecimento, independente do assunto ou da fonte.

7.1 REPRESENTAÇÃO GRÁFICA E DESCRIÇÃO DO MODELO

Figura 17 - Modelo de extração de informação



Fonte: Do autor (2019)

Com base no exposto anteriormente, foi criado e diagramado um modelo genérico (Figura 17), demonstrando todo o processo.

O modelo é composto de três atores principais:

- o usuário, que define o objeto de sua pesquisa; que, ao longo do processo, apresenta-se como operador dos softwares – embora, em alguns momentos, defina alguns aspectos que influenciarão no fluxo e no *software*; e que, após o processo, tem acesso às informações exibidas, capazes de gerar conhecimento sobre os dados inseridos;
- o *software*, que consiste no conjunto de ferramentas que é utilizado durante o processo – conforme dito antes, operadas pelo usuário -, sendo que, conforme suas definições, usar-se-á um ou outro *software*, ou uma ou outra funcionalidade;
- e a fonte de dados, de onde os softwares importam ou exportam os dados/textos, conforme o processo avança.

A primeira fase do processo refere-se à Coleta de dados. Neste momento, o usuário definirá a origem dos dados para a coleta. Estão estabelecidas no modelo três fontes: mídias sociais, mídias tradicionais e outras fontes. Considera-se como mídia social, como já exposto no trabalho, qualquer um dos *sites* que sejam atendidos pelo conceito abordado por Kaplan e Haenlein (2010), conforme o subcapítulo 2.1.1 deste.

Para mídias tradicionais, considera-se qualquer *site* de notícias, seja referente a um jornal, revista, *blog* ou semelhante, cujo propósito primeiro seja a divulgação de notícias. “Outras fontes” referem-se a qualquer outro tipo de fonte, seja um banco de dados, um conjunto de textos, entre outros – dos quais seja possível extrair texto.

Conforme a fonte escolhida, será realizada, por meio de *software* específico, a respectiva coleta. O que é importante entender é que, em cada caso, é possível usar um mesmo *software*, ou até mesmo necessitar mais de um. Além disto, o usuário pode fazer a coleta em mais de uma fonte ao mesmo tempo. É possível, no mesmo *software*, que se colete os dados de mídias sociais diferentes.

Para trazer um exemplo prático: neste trabalho, o mesmo *software* usado para mídias sociais pode ser usado para mídias tradicionais – ao mesmo tempo, inclusive; porém, em uma das mídias sociais, é necessário utilizar outro *software*. Ou seja, para tornar o modelo genérico, para cada fonte deve ser analisado o meio eletrônico (*software*) mais adequado para obter os dados.

Mesmo que os *softwares* sejam diferentes, há dois passos iguais para todos: a exportação dos dados obtidos para um arquivo e a conversão deste para um formato compatível. Nem sempre o formato que o *software* gera é o ideal para que seja possível avançar à próxima fase. Sendo assim, se prevê no modelo, não apenas a exportação, mas também a conversão, visando compatibilidade. Assim, conforme exposto na respectiva raia, surgem os “Dados coletados, exportados e convertidos para formato compatível”.

A fase seguinte refere-se ao pré-processamento, onde inicialmente os dados antes exportados serão importados. O primeiro passo consiste em remover dados desnecessários, isto porque, muitas vezes, os dados gerados anteriormente podem vir com registros ou conteúdo inútil e desnecessário, que apenas atrapalham o processo e que podem ser removidos sem prejuízo às etapas posteriores. Após, os dados precisam ser convertidos para um formato personalizado. Dependendo da fonte de origem, dados importantes, como data e número, por exemplo, aparecem identificados como “texto” e, por isto, não são manipuláveis na filtragem. Assim, essa conversão é necessária, para que seja possível criar filtros.

Para completar a limpeza, há o passo de agrupamento de dados semelhantes. Muitas vezes, há dados que pertencem a uma mesma categoria, porém estão desagrupados. A fim de tornar o filtro preciso, em algumas situações, convém agrupar os dados. Então, o usuário pode definir quais dados ele deseja analisar e selecionar – por exemplo, selecionar os dados de uma pessoa ou de um período de datas. Após isto, ele já pode filtrá-los e exportá-los.

A próxima fase consiste na extração das informações. Os dados anteriormente exportados são importados em *software*. Eles passam por duas fases de limpeza adicional. A primeira é a remoção de caracteres desnecessários, como pontuação e, dependendo do *software*, acentuação. A segunda é a remoção de *stopwords*, como já explicado no trabalho. Cabe salientar que estes processos, embora de pré-processamento, são comuns em ferramentas de extração das informações, o que fez serem colocados aqui.

A questão é que estas *stopwords* são fundamentais para o processo de análise de sentimentos. Desta forma, como pode-se observar no modelo, a remoção não ocorre quando dessa análise, mas apenas para as visualizações relacionadas a frequência de palavras e correlação entre palavras – esta decisão de qual das

formas de visualização será acessada e utilizada é do usuário. Após isto, o texto é processado e as informações são geradas.

Ficam à disposição do usuário a visualização de histogramas, gráficos de dispersão, tabelas, nuvem de palavras e grafos de correlação. Claro que, conforme o *software*, mais formas podem ser disponibilizadas. Porém, estas visualizações são consideradas, com base em pesquisa acadêmica sobre mineração de texto, como métodos gráficos comuns para exibição das informações, sendo o mínimo que um *software* deva ter para ser considerado como partícipe desta fase.

O usuário, neste caso, poderá consultar as várias formas de visualização ao mesmo tempo, comparando diferentes informações de uma mesma fonte. As visualizações não são simultâneas – como em um *dashboard*. Além disto, para que seja possível comparar diferentes fontes, é necessário acessar diferentes instâncias dos *softwares* – cada uma com as informações de uma fonte específica. Não está previsto, no modelo, o uso de um software que permita, não apenas a visualização de informações em tela única, como em um *dashboard*, como também a comparação entre informações de fontes diferentes – mas não há impedimento de se escolher uma ferramenta que faça isto.

Para concluir, o usuário, a partir das informações geradas, terá a capacidade de, ao visualizá-las, gerar conhecimento a partir dos dados inseridos. Por exemplo, se os textos são positivos ou negativos, quais as palavras mais usadas, qual a correlação entre as principais palavras, quais as palavras que mais se relacionam com uma determinada, entre outras. Com isto, o fluxo Dado -> Informação -> Conhecimento se encerra, cumprindo o propósito do modelo de transformar dados/texto bruto em conhecimento útil ao usuário.

7.2 DIFERENÇAS EM RELAÇÃO AO MODELO TRADICIONAL

Embora inspirado no modelo tradicional de mineração de texto (Figura 18), algumas diferenças são importantes. A primeira é o nível de detalhamento, tendo em vista que o modelo tradicional é superficial, enquanto o proposto neste trabalho é detalhado.

Figura 18 - Etapas do processo de mineração de textos

Fonte: Machado (2003)

Agrega-se a isto o nível de entendimento. Os modelos e processos existentes exigem conhecimento técnico do usuário. O proposto é de melhor compreensão ao usuário comum – ou seja, é mais fácil de entender e colocar em prática. Basta analisar o processo de mineração de textos na Figura 19.

Figura 19 - Processo de mineração de textos

Fonte: Adaptado de Aranha e Passos (2006, p. 4)

Os exemplos acima apresentados são semelhantes a outros apresentados e descritos em textos acadêmicos – são modelos técnicos. Importante salientar que, muitas vezes, textos que descrevem tais imagens, como os de Aranha e Passos (2006) e de Correa, Marcacini e Rezende (2012), detalham um pouco o processo. O que se buscou no modelo proposto é aprofundar o processo, sem deixar de lado o seu caráter genérico.

Agrega-se a esta vantagem o fato de que o modelo apresentado é simples e independente de ferramentas o suficiente para que um usuário, sem grande conhecimento de tecnologia, possa realizar mineração de texto. O usuário tem a

capacidade de extrair conhecimento com mais facilidade, com base nos formatos gráficos gerados a partir das informações apresentadas

Outro diferencial é que é possível, no modelo apresentado, iniciar o processo a partir de uma fase ou etapa avançada. Por exemplo, se um usuário já possuir arquivos de texto prontos, sem necessidade de coleta e pré-processamento, pode ir diretamente para as fases finais, para extrair o conhecimento. Ou seja, além das vantagens mencionadas, este é um modelo dotado também de **flexibilidade**, sendo utilizado conforme a necessidade do usuário.

7.3 REPRESENTAÇÃO PRÁTICA DO MODELO COM FERRAMENTAS

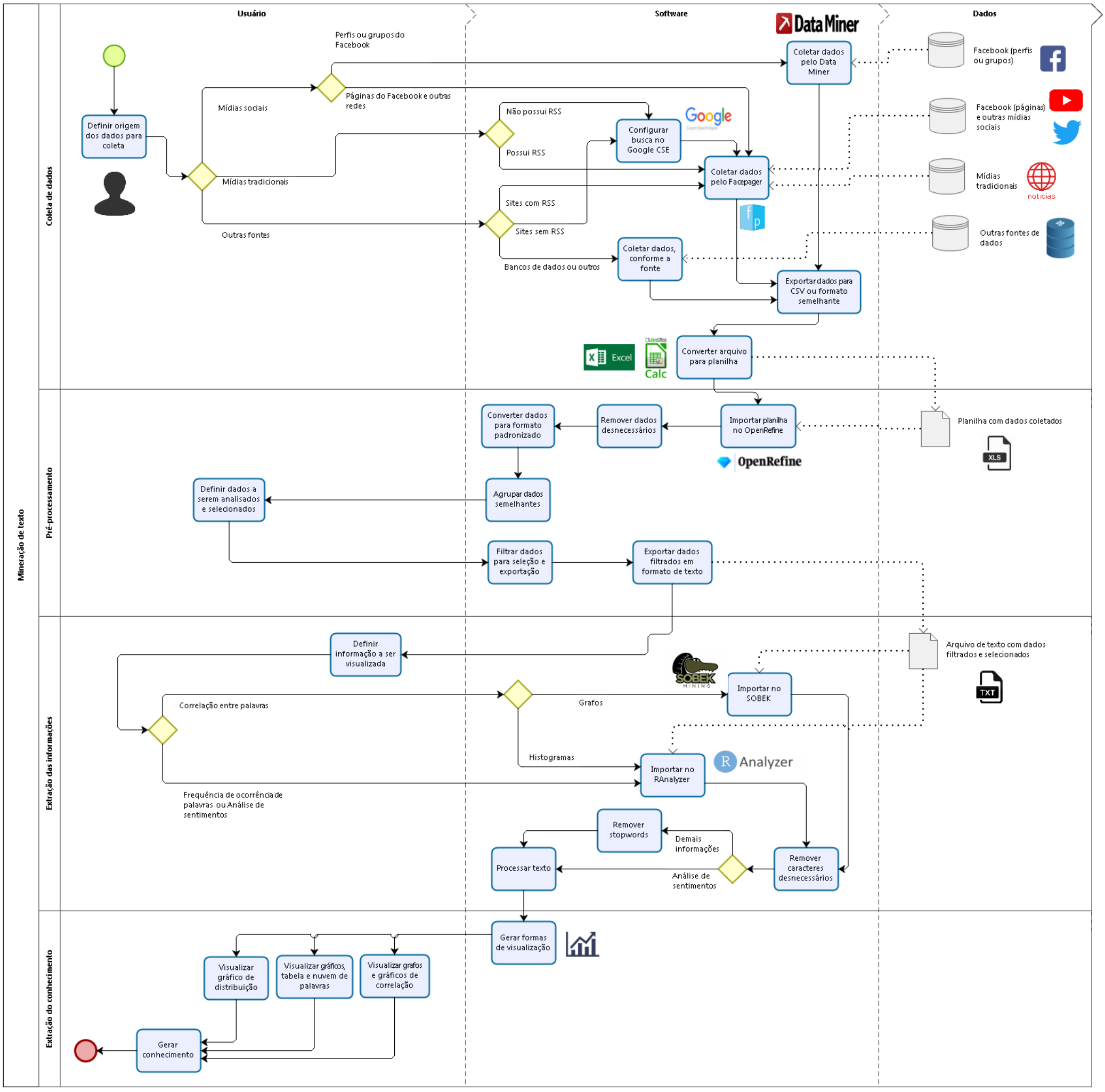
De forma a demonstrar a capacidade adaptativa deste modelo conforme o conjunto de ferramentas, foi criada uma representação do modelo proposto anteriormente (Figura 20), porém usando ferramentas que atendem às fases do modelo.

A partir das pesquisas e testes feitos nos capítulos 5 e 6, definiu-se que serão utilizadas as seguintes ferramentas ou *softwares*:

- Facepager: para coleta de dados de mídias sociais e mídias tradicionais que utilizam API – no caso das mídias tradicionais que usam RSS, a coleta será auxiliada por uma plataforma *online*, chamada “rss2json”; no caso das que não usam, será auxiliada pela plataforma *Google Custom Search Engines*;
- Data Miner, para coleta de dados de mídias sociais que não utilizam ou não permitem a coleta por meio de API. No Facebook, haverá o auxílio das ferramentas *Scroll it! e Expand*;
- OpenRefine, para o pré-processamento;
- Sobek, para a extração de informação em formato de grafos de palavras;
- RAnalyzer, para a extração de informações diversas. Junto dela, estão a própria linguagem R e o complemento RTools.

Obviamente, as ferramentas, bem como a representação prática são desenvolvidas a partir do *status quo* da pesquisa. Por exemplo, caso o Facebook proíba o acesso aos dados de páginas por meio da API, esta representação obriga-se a ser alterada. Porém, salienta-se: o modelo genérico permanece imutável.

Figura 20 - Representação prática do modelo



Fonte: Do autor (2019)

Inicialmente, ocorre a definição da origem dos dados que serão coletados – mídias sociais, mídias tradicionais ou outras fontes. Para mídias sociais, é analisado se a coleta ocorre em perfis ou grupos do Facebook – cuja coleta deve ser feita usando o Data Miner – ou se ocorre em páginas do Facebook ou outras redes – cuja coleta deve ser feita usando o Facepager.

No caso de mídias tradicionais, ambas passam pelo Facepager. A diferença está em ter ou não o RSS. As mídias que tiverem, podem ter sua coleta feita diretamente no Facepager. As que não tiverem, obrigam-se a ter uma busca criada e configurada no Google CSE. O mesmo vale para outras fontes: *sites* com RSS no Facepager, *sites* sem RSS no Google CSE e depois no Facepager, e bancos de dados ou outros tem seus dados coletados de maneira específica, conforme a fonte.

Independente, os dados coletados são exportados para formato CSV ou semelhante – no caso do Data Miner, em formato de planilha XLSX. Porém, pelo uso do OpenRefine, faz-se necessário converter estes arquivos para planilha em formato XLS. Após isto, tem-se a planilha com dados coletados, para ser importada no OpenRefine.

Todos os passos de remoção de dados desnecessários, conversão de dados para formato padronizado e agrupamento de dados semelhantes são feitos por meio de procedimentos e códigos explicados no manual (APÊNDICE A). Após isto, o usuário terá a capacidade de definir de que forma filtrará e selecionará os dados. Por fim, eles são exportados em formato TXT.

Após a exportação, o usuário decide o que visualizará, para, a partir disto, saber qual ferramenta usar. Caso ele deseje ver a correlação entre palavras por meio de grafos, deve importar no Sobek. Caso deseje ver por meio de histogramas, ou se quiser ver questões relacionadas a frequência de ocorrência ou análise de sentimentos, a importação deve ocorrer no RAnalyzer.

Ambos os *softwares* removem caracteres desnecessários e – a não ser que se queira ver a análise de sentimentos – são removidas as *stopwords*. Por fim, o texto é processado e é possível visualizar os gráficos de dispersão, os histogramas, a tabela de palavras, os grafos de palavras e as nuvens de palavras.

7.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

O capítulo que se encerra apresentou o modelo genérico proposto, com o foco de realizar todo o processo de mineração de texto, desde a obtenção dos dados até a transformação destes em informação, que gera conhecimento. Da mesma forma, foi apresentada a representação prática do modelo, utilizando as ferramentas testadas e selecionadas nos capítulos 5 e 6.

Desta forma, o capítulo atingiu o objetivo específico de criar o modelo de extração de informação, baseado nos conceitos de *text mining*, aplicando as ferramentas selecionadas e testadas. Ao atingi-lo, também se deu o principal passo para atingir por completo o objetivo geral do presente trabalho.

O próximo capítulo apresenta a validação do modelo junto a Câmara de Vereadores de Novo Hamburgo, utilizando as ferramentas estudadas nos capítulos anteriores e visando atingir o último objetivo específico. Desta forma, será possível concluir com sucesso a proposta deste trabalho.

8 VALIDAÇÃO DO MODELO

No presente capítulo será apresentado o processo de validação do modelo genérico e de sua representação prática, conforme apresentados no capítulo 7. Como já mencionado, a validação ocorreu na Câmara Municipal de Novo Hamburgo/RS, que é a sede do Poder Legislativo da cidade. Porém, antes de expor esse momento, o trabalho abordará os preparativos que foram feitos para que a validação fosse possível.

8.1 ELABORAÇÃO DO MANUAL

Durante a criação do modelo prático, entendeu-se que era necessário desenvolver um manual de instalação, configuração e uso das ferramentas selecionadas. Devido à quantidade de passos que são necessários, tal documento seria de grande utilidade e importância.

Basicamente, o mesmo possuía três seções, explicando o processo de instalação, de configuração e de uso das ferramentas. Ele foi desenvolvido de forma a ser o mais explicativo possível, usando inúmeras imagens ilustrativas – um dos motivos que o tornou extenso, com dezenas de páginas.

Todo o manual foi desenvolvido com base no Windows, nas versões 7 e 10. Embora algumas ferramentas sejam disponíveis em Linux e/ou Mac OS, entendeu-se necessário baseá-lo no sistema operacional mais utilizado nos computadores pessoais.

Após várias revisões, realizou-se uma validação de teste com dois estudantes da FEEVALE, da área de Tecnologia da Informação. O primeiro tem 28 anos e está no 3º semestre do curso de Ciência da Computação, e o segundo tem 20 anos e está no 5º semestre do curso Tecnólogo em Jogos Digitais.

Ao longo da validação, perceberam-se três problemas: o primeiro foi que algumas funções ocorreram de modo diferente do que estava no manual – inclusive, de forma diferente entre os dois computadores. Isto permitiu entender que, por mais que os testes sejam feitos em um único computador, os procedimentos podem apresentar mudanças ou problemas diferentes, conforme cada computador em que for executado. Assim, além das divergências serem corrigidas e inseridas, seriam

feitos novos testes para tentar abranger a maior quantidade possível de divergências entre os ambientes.

O segundo problema foi a demora em fazer o processo. Para se ter uma ideia, apenas o processo de instalação demorou uma hora. O grande responsável por isto – que é o terceiro problema – era a baixa automatização na instalação. Muitos eram os passos necessários para instalação e configuração. Estava claro de que era necessário automatizar o processo de instalação e configuração – inclusive, uma validação com estudantes de enfermagem estava marcada para o dia seguinte, mas teve de ser desmarcada.

Trabalhou-se, então, em um processo de automatização. Após inúmeros testes e estudos, foi criado um arquivo BAT executável, que fazia o processo de instalação e configuração de quase todas as ferramentas. As únicas que tal processo se mostrou inviável foram o Data Miner e o Scroll It!, não apenas por serem extensões do Google Chrome, mas também porque o primeiro exige conta do Google e criação das *recipes* – como explicado no subcapítulo 5.2.45.2.5.

Entretanto, tal automatização foi um avanço, visto que, ao final dela, as demais ferramentas estão prontas e configuradas. Claro que essa automatização passou por uma evolução ao longo das semanas, onde várias melhorias foram implantadas. Por exemplo, inicialmente, entendia-se necessário usar o RStudio para executar o RAnalyzer, porém, após pesquisa e testes, foi possível eliminá-lo da instalação. Outro exemplo foram os pacotes adicionais da linguagem R, pois foi possível instalá-los durante o processo completo.

Tal processo de instalação foi disponibilizado no GitHub, por meio do endereço github.com/ojonathacardoso/ranalyzer-adds, onde, ao baixar o pacote e descompactá-lo, basta ao usuário executar o arquivo instalador completo, dentro da pasta “Install”.

Após este trabalho – que inclusive impactou na reescrita do Manual, unificando os capítulos de instalação e configuração -, foi feita uma validação com uma usuária, que, embora não fosse utilizar as ferramentas em seu trabalho, tem conhecimento didático avançado – a mesma tem 36 anos e é Professora de Inglês – e um conhecimento básico de Informática, podendo assim tornar o Manual de mais fácil compreensão.

Após três dias de uso, inúmeras melhorias na escrita e na organização do documento foram implantadas. Por exemplo, as imagens receberam destaques em

vermelho, os processos de uso foram explicados em passo a passo, inúmeras referências a subcapítulos anteriores foram implantadas, entre outras modificações.

8.2 VALIDAÇÃO NA CÂMARA DE VEREADORES

Escrito o manual (APÊNDICE A) e realizados os testes, começou-se a planejar a validação junto a Câmara de Vereadores de Novo Hamburgo. O primeiro passo foi definir com quem, o que seria validado e por qual período. A definição de quem participaria da validação foi que ela seria aplicada em dois ambientes distintos: a Assessoria de Comunicação e os Gabinetes Parlamentares.

A Assessoria de Comunicação da Câmara conta, atualmente, com cinco jornalistas e três estagiários. Destes, há três jornalistas que são responsáveis pelas mídias sociais da Câmara, e que fazem o acompanhamento de notícias nas mídias sociais e tradicionais. Assim, a validação com a área de comunicação seria aplicada com eles.

Com relação aos Gabinetes Parlamentares, a Câmara possui 14 (quatorze) vereadores, sendo que cada um possui um Assessor Parlamentar, um Coordenador de Gabinete e dois estagiários – geralmente, um da área de Comunicação e outro da área de Direito. Todos os vereadores possuem mídias sociais – ao menos, uma página no Facebook – e ao menos um dos funcionários do gabinete faz a gestão das mídias, além de acompanhar assuntos da atividade parlamentar e do poder público municipal nas mídias sociais e/ou tradicionais.

Entretanto, embora haja quatorze gabinetes, o autor entendeu que seria não apenas desnecessário, mas principalmente inviável aplicar a validação em todos os gabinetes, por alguns motivos. Primeiro, a impossibilidade de conciliar o trabalho como servidor na Casa Legislativa de Novo Hamburgo com o acompanhamento e auxílio nesta validação. Segundo, a falta de conhecimento básico em Informática por parte de alguns usuários, o que tornaria difícil a aplicação visando um prazo curto. Terceiro, a pouca movimentação nas mídias sociais por parte de alguns vereadores. E quarto, a indisponibilidade de computador compatível, visto que em vários gabinetes há o uso de Linux para os estagiários de comunicação.

Assim, foram selecionados três gabinetes que possuem páginas no Facebook, sendo que dois destes possuem páginas com mais de 10 mil curtidas cada, e o terceiro possui cerca de 2 mil curtidas. Cabe salientar que não houve

nenhuma preferência política na seleção das páginas – inclusive, os nomes dos vereadores não serão mencionados no presente trabalho.

Com relação ao período de validação, entendeu-se que duas semanas seria um período suficiente para o processo, pois não é tão curto que impossibilite o aprendizado e o uso das ferramentas, e nem é longo para não prejudicar o cronograma. Assim, ficou estabelecido que o processo iniciaria na última semana de setembro de 2019, indo até a segunda semana de outubro.

A definição das ferramentas a serem utilizadas na validação necessitava passar por uma aprovação da Gerência de TI da Câmara Municipal de Novo Hamburgo. Mesmo que seja por um prazo curto, os *softwares* necessitavam ser analisados e aprovados ou não para serem instalados nos computadores institucionais. Após a análise, o Facepager e o Sobek foram reprovados, devido a questões de direitos de uso. O Facepager, além disto, tem o problema de impossibilitar a configuração de *proxy* de rede, o que prejudicaria o seu uso na Câmara.

Mesmo com essa reprovação, é possível fazer a validação utilizando o Data Miner – visto que utiliza o Google Chrome -, o OpenRefine e o RAnalyzer. Elas permitiram que fosse possível coletar dados do Facebook, fazer o pré-processamento e a extração de informação.

Paralelo a isto, foi desenvolvido um questionário a ser aplicado aos jornalistas da Assessoria de Comunicação (APÊNDICE B - Questionário aplicado na Assessoria de Comunicação) e dos Gabinetes Parlamentares (APÊNDICE C - Questionário aplicado em Gabinetes Parlamentares), que, além de obter dados de perfil, teria como foco mensurar a validação, contendo perguntas a respeito do modelo, das ferramentas, da relação atual com as mídias sociais e tradicionais e as perspectivas futuras de uso delas.

Com isto, no dia 24 de setembro foi feita a apresentação do modelo e das ferramentas à Assessoria de Comunicação, tendo sido feita a instalação entre os dias 24 e 27. Obviamente que a instalação não poderia levar tanto tempo, e isto fez com que várias modificações fossem implantadas no instalador, a fim de torná-lo mais ágil e prático.

Já de início percebeu-se que os jornalistas gostaram, principalmente o Gerente de Comunicação, que vislumbrava várias pesquisas que poderiam ser feitas, e que sem este modelo não seriam. Assim, entendeu-se que havia uma

perspectiva positiva na validação com a Assessoria de Comunicação. Com relação aos gabinetes, foi organizado para fazer a apresentação no dia 30.

Neste dia, foi feita a apresentação a dois gabinetes escolhidos – por indisponibilidade do gabinete, a apresentação ao terceiro gabinete ocorreu no dia seguinte. Conforme o modelo ia sendo apresentado e, principalmente, quando as ferramentas selecionadas foram apresentadas, era notória a admiração por parte dos gabinetes.

Em suma, o principal motivo para esta admiração era a necessidade atual de fazer essas tarefas de modo manual. Antes, por exemplo, a exportação de comentários no Facebook era totalmente manual, sendo necessário expandi-los um a um, copiar seu texto e seu autor, para, depois, colocar em algum documento.

Se o modelo se resumisse apenas a coleta dos dados, por si só, ele já traria uma grande contribuição ao trabalho deles. Porém, ao visualizar as ferramentas seguintes, maior era a admiração pelas informações disponibilizadas. Por exemplo, a possibilidade de saber as pessoas que mais comentaram em um bloco de postagens, as palavras que apareciam (ou não) nos textos, a relação entre elas, a análise dos sentimentos.

Pode-se concluir, ao final das explicações e da disponibilização das ferramentas, que elas – bem como o modelo proposto – teriam grande chance de serem úteis ao trabalho tanto de gabinetes quanto da área de Imprensa. Mas para tirar as conclusões com precisão, seria necessário aguardar o final do prazo para testes e preenchimento dos respectivos questionários. A data para conclusão da validação, com o consequente recolhimento dos questionários, foi 11 de outubro de 2019.

Menciona-se, para concluir, que o autor do presente trabalho se colocou à disposição para, se necessário, auxiliar no uso das ferramentas – o que ocorreu apenas no último dia, mas que será comentado melhor no próximo subcapítulo. Além disto, no dia 7 foi enviado um e-mail a todos, colocando-se novamente à disposição e salientando a data final de validação.

Cabe apenas salientar que as ferramentas poderiam ser utilizadas sem problemas após o período de validação, tanto nos computadores da Câmara quanto em computadores pessoais – com o acréscimo de que eles poderiam utilizar o Facepager e o Sobek.

8.3 ANÁLISE DOS QUESTIONÁRIOS

Os questionários aplicados, conforme mencionado anteriormente, estão nos apêndices B e C. Aqui, elas serão mencionadas e suas respostas analisadas conforme a numeração de cada pergunta que está em cada questionário.

As questões de 1 a 3 referem-se ao perfil dos usuários – idade, sexo e escolaridade. A idade de todos varia entre 23 e 47 anos, notando-se que, nos gabinetes, os que responderam são mais jovens – geralmente estagiários da área de Comunicação. A distribuição por sexo é de 2 para 1 – para cada dois usuários do sexo masculino, uma do sexo feminino. Já a escolaridade é majoritariamente de Ensino Superior completo ou incompleto – novamente, quando incompleto, refere-se aos estagiários.

As questões de 4 a 11 referem-se à divulgação de informações nas fontes da Internet. Com relação às mídias sociais usadas para divulgação (questão 4), todos utilizam o Facebook. O Twitter e o YouTube são mais utilizados pela Câmara, não tanto pelos gabinetes. Com relação a divulgação em mídias tradicionais (questão 5), usa-se jornal local, embora há também o uso de *sites* (principalmente o da Câmara), *blog* e até mesmo informativo impresso.

Duas perguntas importantes e diretamente ligadas à pesquisa – a 6 e a 7 – tratam da busca de informações em mídias sociais (q. 6) e tradicionais (q. 7). Todos utilizam o Facebook para obter informações, seja do parlamentar, da Câmara, da Prefeitura ou de outros assuntos. Com relação às mídias tradicionais, diversas fontes são usadas, sendo que todos usam jornal local e *sites* de notícias. Também há registros de jornais de alcance regional ou nacional, revistas e *blogs*.

Cabe aqui colocar alguns comentários. Primeiro, de que a busca de alternativas para obter dados de perfis e grupos do Facebook foi um grande acerto, e que sem isso, provavelmente o trabalho não seria tão bem sucedido. Outro fato é a busca em *sites* e jornais locais, que, muitas vezes, não possuem formas de obtenção dos dados a não ser visualizando os mesmos. Neste ponto, o Data Miner e o Google CSE – mais usados no trabalho para este fim – também são importantes para a aplicação prática do modelo.

Por fim, cabe mencionar o Instagram, que não foi pesquisado no trabalho, mas que é utilizado por quase todos os pesquisados, seja para divulgar ou para

obter informações. Isto já aponta para uma necessidade futura: estudar o uso de API, Data Miner ou outro meio para coleta de dados no Instagram.

As questões 8 e 9 são muito interessantes. Para a Assessoria de Comunicação, o uso do Google é importante. Aliás, em conversa informal, foi colocada a necessidade de obter informações de sites, os quais suas ferramentas nativas de pesquisa não são bem sucedidas. O uso do Google CSE, mesmo em teste, foi plenamente satisfatório, e é um avanço em relação ao uso do Google tradicional.

Por outro lado, os gabinetes nunca utilizaram esse tipo de ferramenta, não apenas por não haver o interesse, mas principalmente pela falta de conhecimento deste tipo de ferramenta. Por exemplo, um dos gabinetes falou que fazia essa busca de forma manual, com o auxílio de planilhas no Excel.

As questões 10 e 11 apenas confirmaram o que já se imaginava. É importante o acesso a informações sobre mídias sociais e tradicionais, tanto para o gabinete de um vereador como para a Assessoria de Comunicação de uma Casa Legislativa.

Após esta parte, o questionário aborda o uso do modelo e das ferramentas. A diferença entre os dois questionários é que para a Assessoria de Comunicação, foram adicionadas duas questões a mais – 14 e 15. A diferença estava no detalhamento do modelo, visto que para os jornalistas perguntou-se sobre as quatro camadas, enquanto para os gabinetes perguntou-se apenas da primeira e última. Tal escolha se fez para focar a pesquisa na coleta e na análise final, sem deixar o questionário mais complexo para compreensão e resposta.

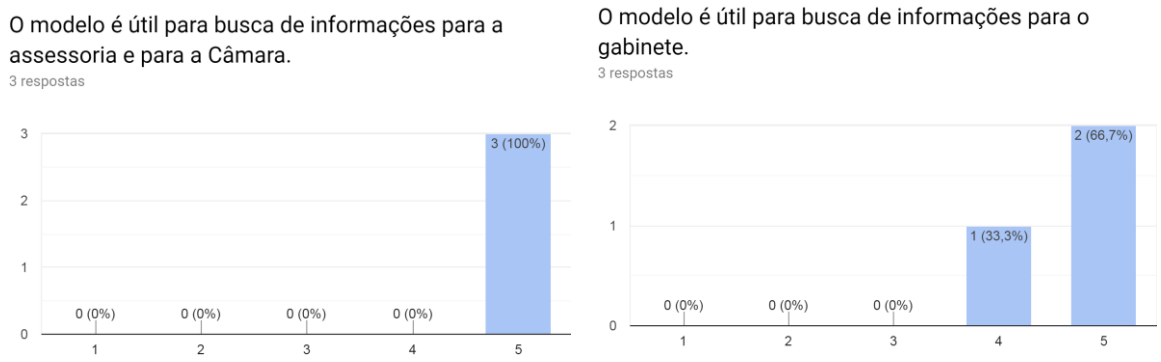
A questão 12 aborda o uso do modelo. Cabe salientar que quase todos marcaram que usaram o modelo com pouca frequência – apenas um deles usou com mais frequência. Isso pode ser entendido ou explicado por duas questões: a primeira é que, como dito antes, nenhum gabinete ou jornalista solicitou ajuda para o uso delas, seja por desinteresse ou – mais provável – por pouco tempo para conciliar trabalho e validação.

O único gabinete que reportou algo foi exatamente o que assinalou que usou bastante as ferramentas, sendo que responderam ao e-mail mencionado anteriormente, informando que estavam usando as ferramentas, sem dúvidas. A segunda questão é que um gabinete e uma jornalista solicitaram ajuda para usar as ferramentas, exatamente no último dia da validação.

As questões 13 a 16 (imprensa) e 13 a 14 (gabinete) tratavam do modelo. Todos os gabinetes e jornalistas entenderam que as camadas atenderam as necessidades para coleta, bem como possuíam as etapas necessárias ou importantes. Em síntese, segundo todos, o modelo permite obter dados de uma fonte, e, após passar por uma fase de pré-processamento, extrair informação que possa gerar conhecimento. Isso é corroborado pelas questões 17 a 19 (imprensa) e 15 e 16 (gabinete), onde todos marcaram que foi possível extrair conhecimento sobre a Câmara, sobre os vereadores e sobre outros assuntos.

A conclusão desta seção se dá ao questionar, aos gabinetes, se as informações apresentadas condiziam com a realidade do vereador, ao passo que todos responderam que sim. Além disto, demonstrando a eficácia do modelo, todos marcaram, nas questões 20 (imprensa) e 18 (gabinete) que ele é útil para buscar informações, seja para o gabinete, seja para a Câmara (Figura 21).

Figura 21 - Resposta das questões sobre a utilidade do modelo



Fonte: Do autor (2019)

Tais respostas corroboram a percepção dos usuários sobre a relevância dos resultados obtidos com o uso do modelo, desde o início da validação. Por exemplo, a possibilidade de saber quem é a pessoa que mais comenta nas postagens de uma página ou perfil do Facebook, ou saber quais as palavras mais faladas, ou ainda entender o sentimento dos seguidores ao se manifestarem nos comentários, são informações que não eram fáceis de serem obtidas. Porém, este modelo, auxiliado pelo conjunto de ferramentas, possibilitou tal busca.

As questões 21 a 26 (imprensa) e 19 a 24 (gabinete) se referem a cada uma das formas de visualização da última camada, providenciadas pelo RAnalyzer. O que se percebeu é todas elas foram aprovadas, consideradas interessantes ou muito

interessantes – claro, cada um com uma preferência específica. Algo que se pode perceber é que o grafo teve um interesse um pouco menor.

O grande motivo para isto – e confirmado pela resposta sobre se as formas de visualização eram claras e compreensíveis (questão 26 da imprensa e 24 de gabinetes) – é que a versão do *software* que foi instalada ainda não tinha passado pelas melhorias apresentadas no subcapítulo 6.1. O grafo de palavras, por exemplo, era da versão estática, menos interativa. A análise de sentimentos também era da versão mais simples, que agora é mais clara e detalhada.

As questões 27 a 29 (imprensa) e 25 a 27 (gabinetes) tratavam do manual. É importante analisar que as respostas foram variadas, predominantemente positivas, mas também houve respostas negativas, de que o manual não foi útil, não estava tão claro e que precisaria de melhorias. Cabe salientar que o manual foi escrito apenas para auxiliar a instalação e o uso das ferramentas, e ao longo da sua escrita, passou por várias melhorias.

Além disto, nem todas as ferramentas descritas no manual foram disponibilizadas – o que se pode discutir a necessidade de ter sido disponibilizada uma versão específica. Por fim, não foi feita nenhuma explicação detalhada do manual, nem usando-o como base. Todas as ferramentas foram mostradas na prática, mas o manual foi disponibilizado para consultas e dúvidas, bem como para, se desejado, implantar as ferramentas em computador pessoal do usuário.

A antepenúltima pergunta tratava do trabalho simplificado em relação ao que havia antes. Interessantemente, metade das respostas foi que o modelo não simplificou. Fica difícil analisar tal comportamento, visto que as respostas em geral mostraram o contrário. Inclusive a penúltima pergunta mostrou que praticamente todos pretendem utilizar as ferramentas.

Entre os gabinetes, entende-se que as ferramentas são interessantes, facilitam e agilizam o trabalho antes manual. Porém, é importante uma dedicação maior para sua compreensão – algo que é compreensível, mas discutível, visto que as respostas foram majoritariamente de que o modelo foi pouco usado. Por parte da Assessoria de Comunicação, a importância das ferramentas foi até mais destacada, vislumbrando a possibilidade de usá-las com frequência, para melhorias na produção de publicações e ações institucionais, bem como, na análise das atividades da Câmara e vereadores,

Ao concluir a análise das respostas dadas, tanto por gabinetes quanto pela Assessoria de Comunicação, fica claro que o uso das mídias sociais e tradicionais para divulgação e pesquisa não é apenas uma opção, mas uma realidade e uma necessidade para o trabalho, tanto de um quanto de outro. Também se mostrou inegável que o uso de ferramentas que auxiliem todo o processo apontado pelo modelo – desde a coleta até a análise final – leva o trabalho a um outro patamar de obtenção de informação e conhecimento.

Aliás, isto deve ser destacado: ficou claro, não apenas na apresentação inicial, como também na análise dos questionários, que, assim como mencionado ao longo do presente trabalho, a inexistência ou desconhecimento desse tipo de ferramenta é um fato que esse tipo de usuário enfrenta. Por fim, cabe destacar que, embora a validação tenha sido de apenas duas semanas aproximadamente, as ferramentas não exigiam uma grande complexidade de aprendizado, de forma que o modelo fosse inviável ao usuário comum. Nesse sentido, o RAnalyzer, desenvolvido neste trabalho, foi muito importante para essa análise – mesmo que tenha sido usada a versão com menos recursos.

8.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

O capítulo que se encerra teve por finalidade realizar a validação do modelo e das ferramentas, na Câmara Municipal de Novo Hamburgo. Desta forma, foi explicado o desenvolvimento do manual de instalação e uso das ferramentas selecionadas. Foi apresentado como ocorreu o processo de validação, e, por fim, realizada uma análise das respostas apresentadas nos questionários, tanto por jornalistas da Câmara quanto por gabinetes parlamentares. Sendo assim, conclui-se que o capítulo atingiu o último objetivo específico, que é o de validar o modelo proposto na Câmara Municipal de Novo Hamburgo/RS.

Além disto, ao atingir tal objetivo específico, é possível concluir que o objetivo geral – qual seja, o de criar um modelo de extração de informação, a partir de fontes da Internet, baseado em *text mining* e utilizando ferramentas gratuitas existentes, que possibilite gerar conhecimento relacionado a assuntos que sejam de interesse da administração pública e da atividade política - foi plenamente atingido, e o trabalho está concluído.

CONCLUSÃO

O presente trabalho teve como objetivo criar um modelo de extração de informação, a partir de fontes da Internet, baseado em *text mining* e utilizando ferramentas gratuitas existentes, que possibilite gerar conhecimento relacionado a assuntos que sejam de interesse da administração pública e da atividade política. No caso do presente trabalho, tal modelo passou por validação na Câmara de Vereadores de Novo Hamburgo, no Rio Grande do Sul. Ao longo dos capítulos, foi apresentado o referencial teórico necessário para atingir tal objetivo, bem como de que forma o mesmo seria alcançado.

O segundo capítulo tratou das mídias sociais e tradicionais. Foi apresentada a origem, conceitos e evolução das mídias sociais, bem como o seu uso na sociedade. Também se tratou da origem das mídias tradicionais, abordando seu histórico, evolução e relação com as mídias sociais. Tal capítulo deu os subsídios necessários para que se atingisse o objetivo específico de identificar se e como são utilizados, pelos governos e políticos, dados obtidos de redes sociais e sites de notícias.

O terceiro capítulo tinha como tema a Administração Pública e a Atividade Política, abordando alguns conceitos básicos, o envolvimento com a tecnologia, o uso das duas mídias, e a relação entre as mídias sociais e os serviços governamentais. Por fim, tratou da situação atual, bem como perspectivas futuras. Com este capítulo, foi possível atingir o objetivo específico de identificar se e como são utilizados, pelos governos e políticos, dados obtidos de redes sociais e sites de notícias.

Para concluir o referencial teórico, o quarto capítulo teve como finalidade apresentar o conceito de mineração de texto e a sua aplicabilidade. Também foram apresentadas técnicas aplicáveis e, por fim, o processo em si, que envolve quatro etapas. Ele foi fundamental para que o segundo objetivo específico, que é o de investigar de que forma é possível obter dados de mídias sociais e tradicionais, com base nos conceitos de mineração de texto fosse atingido.

Dando sequência, o capítulo quinto tratou das ferramentas que seriam selecionadas, de acordo com os pilares da seleção previamente definidos. Foram abordadas as ferramentas de coleta, tanto para mídias sociais quanto para as tradicionais - incluindo as dificuldades encontradas e as soluções aplicadas.

Também foram abordadas as ferramentas da fase de pré-processamento, bem como as de extração de informação.

Com este capítulo, foram atingidos dois objetivos específicos: investigar de que forma é possível obter dados de mídias sociais e tradicionais, com base nos conceitos de mineração de texto; e analisar as ferramentas existentes, que permitem a obtenção dos dados, bem como a sua posterior mineração e extração de informação.

Durante o estudo da fase final do processo de mineração de texto, percebeu-se que seria necessário reunir estas funcionalidades em uma ferramenta gráfica de uso simples. Sendo assim, o sexto capítulo teve como foco abordar a ferramenta RAnalyzer, apresentando todas as suas funcionalidades, bem como a forma de disponibilização para uso. Desta forma, o capítulo mostrou-se complementar ao objetivo específico de analisar as ferramentas existentes, bem como o de criar o modelo de extração de informação, baseado nos conceitos de *text mining*, aplicando as ferramentas selecionadas e testadas.

O penúltimo capítulo abordou o modelo genérico proposto, que demonstra o processo de mineração de texto, indo desde a obtenção dos dados até a transformação destes em informação e conhecimento. Também foi apresentada a representação prática deste modelo, utilizando as ferramentas testadas e selecionadas. Assim, o capítulo atingiu o objetivo específico de criar o modelo de extração de informação, baseado nos conceitos de *text mining*, aplicando as ferramentas selecionadas e testadas.

O último capítulo teve por objetivo realizar a validação do modelo e das ferramentas, sendo apresentado como ocorreu a redação do manual de instalação e uso das ferramentas selecionadas, bem como explicado de que forma ocorreu o processo de validação. Por fim, foi realizada uma análise das respostas apresentadas nos questionários, tanto pelos jornalistas da Câmara quanto pelos gabinetes parlamentares selecionados.

Com este último capítulo, foi possível não apenas atingir o último objetivo específico - validar o modelo proposto na Câmara Municipal de Novo Hamburgo/RS - como o objetivo geral. Assim, entende-se que o trabalho foi concluído com sucesso, e contribuiu para incentivar o uso de uma poderosa e importante área da tecnologia, principalmente por usuários que possuem pouco conhecimento técnico.

Vislumbram-se vários trabalhos futuros a partir do trabalho realizado. Em primeiro lugar, embora com resultado positivo, faz-se necessário realizar uma validação em larga escala, para que seja possível receber um *feedback* mais amplo e, quiçá, mais preciso. Tal validação deve ocorrer por um período maior, pois notou-se que o período de duas semanas é insuficiente para entender o funcionamento e testar todas as ferramentas mencionadas na representação prática – salientando que algumas ferramentas não puderam ser validadas.

O trabalho teve como enfoque as mídias sociais e tradicionais, porém, é importante colocar que uma validação com outras fontes, da Internet ou não, é necessária. Por exemplo, algumas modificações foram realizadas no RAnalyzer, de forma que ele pudesse analisar textos, como dissertações e discursos. Nesse sentido, devido ao seu caráter genérico, uma validação em outras áreas do conhecimento também é importante.

Com base nos questionários respondidos, dois pontos precisam ser analisados. O primeiro é que as críticas dos usuários devem ser resgatadas, principalmente com relação ao manual de uso das ferramentas. Por mais revisões que sejam feitas para sua melhoria, dificilmente um texto escrito conseguirá, em pouco tempo, explicar algo novo a um usuário comum. O segundo é que se faz necessária uma pesquisa e uma validação do Instagram, de forma a analisar a necessidade (ou não) de alteração/inclusão da ferramenta para sua coleta, tendo em vista que vários pesquisados responderam que a utilizam.

Como já foi exposto no trabalho, o desenvolvimento do RAnalyzer teve como alvo atender a camada de extração de informação. Entretanto, durante a pesquisa, notou-se que é possível, sim, criar uma ferramenta que execute quase todos os passos que as demais ferramentas selecionadas executam – ou seja, da coleta à análise. Isto não foi feito por dois motivos: 1) é um desenvolvimento que demanda um grande tempo – provavelmente, vários meses de trabalho; 2) uma ferramenta mais completa não deixaria de demandar o uso de outras ferramentas, como, por exemplo, para coletar dados de perfis e grupos no Facebook.

Por último, todo este trabalho permite o surgimento de dois projetos: o primeiro é um curso ou treinamento sobre o uso de ferramentas para mineração de texto, mostrando na prática, desde a coleta até a análise; o segundo é o uso de vídeos explicativos, apresentando não apenas as ferramentas, mas explicando a própria área de *text mining*.

Para concluir, é importante que essa área – e isto inclui as ferramentas - seja mais divulgada entre as pessoas, tendo em vista que vivemos em um mundo cuja informação possui grande valor, não apenas para empresas, mas também para as pessoas. Valer-se da mineração de texto para extrair informações, pode ser a diferença – tomando a temática do presente trabalho - entre um político ser eleito e outro não, ou de um gabinete parlamentar ser mais atuante do que outro. Tudo isso graças à um dos principais bens deste século: a informação.

REFERÊNCIAS

ABDULHAYOGLU, Mehmet; THIJS, Bart. Use of ResearchGate and Google CSE for author name disambiguation. **Scientometrics**, Basileia, v. 111, p. 1965-1985, jun. 2017. Disponível em: <https://doi.org/10.1007/s11192-017-2341-y>. Acesso em: 02 mar. 2019.

AMAZON AWS. **Text Mining example codes (tweets)**. Disponível em: <https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html>. Acesso em: 11 jun. 2019.

ARANHA, Christian; PASSOS, Emmanuel. A Tecnologia de Mineração de Textos. **Revista Eletrônica de Sistemas de Informação**, Rio de Janeiro, n. 2, 2006. Disponível em: <http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171>. Acesso em 19 abr. 2019.

ARUN, K.; NAYAGAM, M. Building applications with social networking API's. **International Journal of Advanced Networking and Applications**, Tamil Nadu, v. 5, p. 2070-2075, 2014. Disponível em: <http://www.ijana.in/papers/V5I5-7.pdf>. Acesso em: 02 mar. 2019.

BARBOSA, Albenir Rêgo. Perfil da produção científica brasileira sobre governo eletrônico. **Revista Eletrônica Gestão e Serviços**, São Paulo, v. 8, n. 1, p. 1785-1810, jan./jul. 2017. Disponível em: <https://www.metodista.br/revistas/revistas-metodista/index.php/REGS/article/view/6681/5841>. Acesso em: 24 mar. 2019.

BATRINCA, Bogdan; TRELEAVEN, Philip. Social media analytics: a survey of techniques, tools and platforms. **AI & Society**, Basileia, v. 30, p. 89-116, fev. 2015. Disponível em: <https://doi.org/10.1007/s00146-014-0549-4>. Acesso em: 02 mar. 2019.

BENOIT, KENEETH et al. Package "quanteda". **The Comprehensive R Archive Network**. 2019. Disponível em: <https://cran.r-project.org/web/packages/quanteda/quanteda.pdf>. Acesso em: 24 jun. 2019.

BONSÓN, Enrique et al. Local e-government 2.0: Social media and corporate transparency in municipalities. **Government Information Quarterly**, Amsterdam, v. 29, p. 123-132, out. 2012. Disponível em: <https://doi.org/10.1016/j.giq.2011.10.001>. Acesso em: 14 mar. 2019.

BRASIL. **Constituição da República Federativa do Brasil (1988)**. Brasília, DF: Presidência da República. Disponível em: <http://www.planalto.gov.br/ccivil_03/Constituicao/Constituiçao.htm>. Acesso em: 30 mar. 2019.

BRUNS, Axel; LIANG, Yuxian E. Tools and methods for capturing Twitter data during natural disasters. **First Monday**, Bridgman, v. 17, n. 4, abr. 2012. Disponível em: <https://ojphi.org/ojs/index.php/fm/article/view/3937/3193>. Acesso em: 02 nov. 2018

CARDOSO, Jonatha Martins. Desenvolvimento de software RAnalyzer para mineração de textos. In: FEIRA DE INICIAÇÃO CIENTÍFICA, 11, 2019, p. 209, **Anais...**, Novo Hamburgo, 2019. Disponível em: <https://www.feevale.br/Comum/midias/e8786f81-38c3-446e-87c9-d60d84f64f69/Anais%20FIC%202019.pdf>. Acesso em: 30 out. 2019.

CARRILHO JUNIOR, João Ribeiro. **Desenvolvimento de uma metodologia para mineração de textos**. 2007. 96 f. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: <http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1>. Acesso em: 10 mai. 2019.

CHANG, Ai-Mei; KANNAN, P. K. **Leveraging Web 2.0 in Government**. IBM Center for the Business of Government: Washington, 2008. Disponível em: <http://www.businessofgovernment.org/sites/default/files/LeveragingWeb.pdf>. Acesso em: 02 mar. 2019.

CHANG, Winston et al. Package “shiny”. **The Comprehensive R Archive Network**. 2019. Disponível em: <https://cran.r-project.org/web/packages/shiny/shiny.pdf>. Acesso em: 28 jul. 2019.

CHRISTAKIS, Nicholas; FOWLER, James. **Connected:** The surprising power of our social networks and how they shape our lives. New York: Little, Brown and Company, 2009. Disponível em: <https://www.researchgate.net/publication/258568206_Nickolas_A_Christakis_and_James_H_Fowler_2009_Connected_The_Surprising_Power_of_our_Social_Networks_and_How_they_Shape_our_Lives_Little_Brown_New_York_NY_353_pages>. Acesso em: 24 fev. 2019.

COELHO, Ricardo Corrêa. **Estado, governo e mercado**. UFSC: Florianópolis, 2012. Disponível em: <http://www.transparencia.fai.ufscar.br/Transparencia/Documentos/2611R02.pdf>. Acesso em: 24 mar. 2019.

CORRÊA, Geraldo Nunes; MARCACINI, Ricardo Marcondes; REZENDE, Solange Oliveira. **Uso da mineração de textos na análise exploratória de artigos científicos**. Instituto de Ciências Matemáticas e de Computação. USP: São Carlos, set. 2012. Disponível em: <http://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_383.pdf>. Acesso em: 17 mai. 2019.

COURTOIS, Cédric; MECHANT, Peter; DE MAREZ, Lieven. Teenage uploaders on YouTube: Networked public expectancies, online feedback preference, and received on-platform feedback. **Cyberpsychology, Behavior, and Social Networking**, New

Rochelle, v. 14, n. 5, p. 315-322, 2011. Disponível em: <http://dx.doi.org/10.1089/cyber.2010.0225>. Acesso em: 02 mar. 2019.

DHARA, Arup. A personalized discovery service using Google custom search engine. **Annals of Library and Information Studies**, New Delhi, v. 63, p. 298-305, dez. 2016. Disponível em: <http://op.niscair.res.in/index.php/ALIS/article/viewFile/13880/1106>. Acesso em: 02 mar. 2019.

DINIZ, Eduardo Henrique et al. O governo eletrônico no Brasil: perspectiva histórica a partir de um modelo estruturado de análise. **Revista de Administração Pública**, São Paulo, v. 43, n. 1, p. 23-48, jan./fev. 2008. Disponível em: <http://bibliotecadigital.fgv.br/ojs/index.php/rap/article/view/6678/5261>. Acesso em: 24 mar. 2019.

FACEBOOK FOR DEVELOPERS. **Graph API**. Disponível em: <https://developers.facebook.com/docs/graph-api/>. Acesso em: 14 abr. 2019.

FEINERER, Ingo; HORNIK, Kurt. Package “tm”. **The Comprehensive R Archive Network**. 2018. Disponível em: <https://cran.r-project.org/web/packages/tm/tm.pdf>. Acesso em: 24 jun. 2019.

FELLOWS, Ian. Deducer: A Data Analysis GUI for R. **Journal of Statistical Software**, Innsbruck, v. 49, n. 8, jun. 2012. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v049i08/v49i08.pdf>. Acesso em: 26 mai. 2019.

FENOLL, Vicente; CANO-ORÓN, Lorena. Participación ciudadana en los perfiles de Facebook de los partidos españoles. Análisis de comentarios en la campaña electoral de 2015. **Communication & Society**, Pamplona, v. 30, n. 4, p. 131–148, 2017. Disponível em: <http://roderic.uv.es/handle/10550/63163>. Acesso em: 07 abr. 2019.

FERNANDES, Pedro André F. C. **Online News Recommendation System**. 2018. 74 f. Dissertação (Mestrado Integrado em Engenharia Informática e Computação) – Faculdade de Engenharia, Universidade do Porto: Porto, 2018. Disponível em: <https://repositorio-aberto.up.pt/bitstream/10216/114179/2/278383.pdf>. Acesso em: 02 mar. 2019.

GGPLOT2: Axis manipulation and themes. **RStudio Pubs Static**. Disponível em: <https://rstudio-pubs-static.s3.amazonaws.com/3364_d1a578f521174152b46b19d0c83cbe7e.html>. Acesso em: 14 jun. 2019.

GIGLIETTO, Fabio; ROSSI, Luca; BENNATO, Davide. The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source. **Journal of Technology in Human Services**, v. 30, p. 145-159, 2012. Disponível em: <http://dx.doi.org/10.1080/15228835.2012.743797>. Acesso em: 13 abr. 2019

GITHUB. **OpenRefine wiki**. Disponível em: <https://github.com/OpenRefine/OpenRefine/wiki>. Acesso em: 12 mai. 2019.

GOMES JR., Paulo Pinheiro. Big data e o consumo de notícias nas redes sociais. **Revista Gestão e Desenvolvimento**, Novo Hamburgo, v. 11, n. 1, p. 46-57, jan. 2014. Disponível em: <https://periodicos.feevale.br/seer/index.php/revistagestaoedesenvolvimento/article/view/71>. Acesso em: 16 abr. 2019.

GONZAGA, Silas. O Sensacionalista e Text Mining: Análise de sentimento usando o lexiconPT. **Paixão por dados**. Disponível em: <https://sillasgonzaga.github.io/2017-09-23-sensacionalista-pt01/>. Acesso em: 14 jul. 2019.

_____. Package “lexiconPT”. **The Comprehensive R Archive Network**. 2017. Disponível em: <https://cran.r-project.org/web/packages/lexiconPT/lexiconPT.pdf>. Acesso em: 24 jun. 2019.

GOOGLE DEVELOPERS. **Custom Search Engine**: API Reference. Disponível em: <https://developers.google.com/custom-search/v1/>. Acesso em: 14 abr. 2019.

GOOGLE DEVELOPERS. **YouTube Developer Documentation**. Disponível em: <https://developers.google.com/youtube/documentation/>. Acesso em: 14 abr. 2019.

GTECH.EDU RESEARCH GROUP. **Educational Text Mining**: Mining with a single click. Universidade Federal do Rio Grande do Sul. Disponível em: <https://biblioteca.furg.br/images/documentos/Formularios/sobek_quick_reference_guide.pdf>. Acesso em: 24 jun. 2019.

HAHL, Bruno et al. A influência das redes sociais nas relações interpessoais. **Revista Eletrônica do Colégio Mãe de Deus**, Porto Alegre, v. 4, p. 11, 2013. Disponível em: <http://www.colegiomaededeus.com.br/revistacmd/revistacmd_v42013/artigos/a2_redes_sociais_cmdset2013.pdf>. Acesso em: 29 mar. 2019.

HAßLER, Jörg; MAURER, Marcus; HOLBACH, Thomas. Advancement through technology? The analysis of journalistic online content by using automated tools. **Studies in Communication and Media**, Baden-Baden, v. 3, n. 2, 2014. Disponível em: <https://www.scm.nomos.de/fileadmin/scm/doc/SCM_14_02_02.pdf>. Acesso em: 02 mar. 2019.

HUBERMAN, Bernardo A.; ROMERO, Daniel M.; WU, Fang. Social networks that matter: Twitter under the microscope. **First Monday**, Bridgman, v. 14, n. 1, jan. 2009. Disponível em: <https://firstmonday.org/article/view/2317/2063>. Acesso em: 02 mar. 2019.

HURTADO, Luis Cobo. **Estudio y aplicación del robot Pepper para la interacción con personas mayores**. 2018. 192 f. Dissertação (Mestrado em Eletrônica Industrial e Automática) – Universidad de Valladolid, Valladolid, 2018. Disponível em: <http://uvadoc.uva.es/bitstream/10324/29358/1/TFM-I-808.pdf>. Acesso em: 12 mai. 2019.

JAMBEIRO, Othon. Gestão e tratamento da informação na sociedade tecnológica. **São Paulo em Perspectiva**, São Paulo, v. 12, n. 4, p. 3-10, 1998. Disponível em: <http://produtos.seade.gov.br/produtos/spp/v12n04/v12n04_01.pdf>. Acesso em: 3 mar. 2019.

JÜNGER, Jakob; KEYLING, Till. **Facepager:** An application for generic data retrieval through APIs. 2018. Código fonte e *releases*. Disponível em: <https://github.com/strohne/Facepager>. Acesso em: 30 abr. 2019.

KAPLAN, Andreas M.; HAENLEIN, Michael. Users of the world, unite! The challenges and opportunities of social media. **Business Horizons**, Bloomington, v. 53, p. 59-68, 2010. Disponível em: <www.elsevier.com/locate/bushor>. Acesso em: 25 fev. 2019.

KAVANAUGH, Andrea L. et al. Social media use by government: From the routine to the critical. **Government Information Quarterly**, Amsterdam, v. 29, p. 480-491, out. 2012. Disponível em: <https://doi.org/10.1016/j.giq.2012.06.002>. Acesso em: 02 mar. 2019.

KENTE, Malcolm. **Social network analysis**. 2017. 29 f. Tese (Software Engineering and Management), Department of Computer Science and Engineering, University of Gothenburg, Gothenburg, 2017. Disponível em: <https://gupea.ub.gu.se/bitstream/2077/54662/1/gupea_2077_54662_1.pdf>. Acesso em: 23 mar. 2019.

KLEMANN, Miriam; REATEGUI, Eliseo; RAPKIEWICZ, Clevi. Análise de ferramentas de mineração de textos para apoio à produção textual. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 22, 2011, **Anais**... Aracaju, 2011. Disponível em: <https://www.br-ie.org/pub/index.php/sbie/article/download/1866/1632>. Acesso em: 20 mai. 2019.

LANG, Dawei; CHIEN, Guan-tin. Package “wordcloud2”. **The Comprehensive R Archive Network**. 2018. Disponível em: <https://cran.r-project.org/web/packages/wordcloud2/wordcloud2.pdf>. Acesso em: 24 jun. 2019.

LOMBORG, Stine; BECHMANN, Anja. Using APIs for data collection on social media. **The Information Society: An International Journal**, Indiana, v. 30, p. 256-265, 2014. Disponível em: <https://doi.org/10.1080/01972243.2014.915276>. Acesso em: 02 mar. 2019.

LOURENÇO, Patrícia. **Comunicação integrada e redes sociais: uma questão de influência**. 2011. 97 f. Dissertação (Mestrado em Comunicação, Cultura e Tecnologias da Informação) – Escola Superior de Comunicação Social, Instituto Universitário de Lisboa, Lisboa, 2011. Disponível em: <https://repositorio.iscte-iul.pt/bitstream/10071/4554/1/TESE de Patrícia Vale Lourenço.pdf>. Acesso em: 29 mar. 2019.

MACEDO, Alexandra Lorandi et al. Uma ferramenta de mineração de texto para apoio à leitura e escrita autoral. **Informática na Educação: teoria & prática**, Porto

Alegre, v. 19, n. 2, p. 123-139, jun./set. 2016. Disponível em: <https://seer.ufrgs.br/InfEducTeoriaPratica/article/viewFile/55613/39083>. Acesso em: 21 mai. 2019.

MACHADO, Elias. **O ciberespaço como fonte para os jornalistas**. Universidade Federal da Bahia: Salvador, 2003. Disponível em: <http://www.bocc.ubi.pt/pag/machado-elias-ciberespaco-jornalistas.pdf>. Acesso em: 25 fev. 2019.

MARQUES, Francisco Paulo Jamil Almeida; AQUINO, Jakson Alves De; MIOLA, Edna. Parlamentares, representação política e redes sociais digitais: perfis de uso do Twitter na Câmara dos Deputados. **Opinião Pública**, Campinas, v. 20, n. 2, p. 178-203, ago. 2014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-62762014000200178&lng=pt&tlng=pt>. Acesso em: 23 mar. 2019.

MARTINS, Claudia Aparecida et al. **Uma Experiência em Mineração de Textos Utilizando Clustering Probabilístico e Clustering Hierárquico**. Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo: São Carlos, 2003. Disponível em: <http://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_205.pdf>. Acesso em: 18 abr. 2019.

MEYER, David; HORNIK, Kurt; FEINERER, Ingo. Text Mining Infrastructure in R. **Journal of Statistical Software**, Innsbruck, v. 25, n. 8, mar. 2008. Disponível em: <http://epub.wu.ac.at/3978/1/textmining.pdf>. Acesso em: 23 jun. 2019.

MURPHY, Philip. **Basic Text Mining in R**. Disponível em: <https://rstudio-pubs-static.s3.amazonaws.com/265713_cbef910aee7642dc8b62996e38d2825d.html>. Acesso em: 11 jun. 2019.

MURUGESAN, San. Understanding Web 2.0. **IT Professional**, Piscataway, v. 9, jul.-ago. 2007. Disponível em: <https://doi.org/10.1109/MITP.2007.78>. Acesso em: 02 mar. 2019.

NEWS API. **Documentation**. Disponível em: <https://newsapi.org/docs>. Acesso em: 28 abr. 2019.

NIE, Norman H.; ERBRING, Lutz. Internet and society: a preliminary report. **IT & Society**, v. 1, n. 1 p. 275-283, verão 2002. Disponível em: <http://www.itandsociety.org>. Acesso em: 23 jan. 2019.

NONATO, Murillo Nascimento; PIMENTA, Thaís Ariane Ferreira; PEREIRA, Francis José. Geração Z: Os desafios da mídia tradicional. In: CONGRESSO DE CIÊNCIAS DA COMUNICAÇÃO NA REGIÃO NORDESTE, 14., 2012, **Anais**... Recife, 2012. Disponível em: <http://exame.abril.com.br/marketing/noticias/geracao-z-e-mais-conectada-fuma->. Acesso em: 27 fev. 2019.

OLIVEIRA, Luciana Gonçalves de; AZEVEDO, Breno Fabrício Terra; GOMES, Cleidiane Basílio Almeira. **Softwares de mineração de texto na análise de**

produções textuais de estudantes do ensino fundamental: possibilidades interdisciplinares. In: SEMINÁRIO INTERNACIONAL DE EDUCAÇÃO, TECNOLOGIA E SOCIEDADE, 23, 2018, Taquara, 2018.

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT. **Participative web:** user-created content. Paris, abr. 2007. Disponível em: <https://www.oecd.org/sti/38393115.pdf>. Acesso em: 28 fev. 2019.

PAPAIOANNOU, Ioannis et al. Alana: Social Dialogue using an Ensemble Model and a Ranker trained on User Feedback. In: ALEXA PRIZE, 2017, Seattle. **1st Proceedings**…, 2017. Disponível em: <http://alexaprize.s3.amazonaws.com/2017/technical-article/alana.pdf>. Acesso em: 02 mar. 2019.

PERLES, João Batista. **Comunicação:** conceitos, fundamentos e história. 2007. Disponível em: <http://www.bocc.ubi.pt/pag/perles-joao-comunicacao-conceitos-fundamentos-historia.pdf>. Acesso em: 27 fev. 2019.

PEZZINI, Anderson. Mineração de textos: conceito, processo e aplicações. **Revista Eletrônica do Alto Vale do Itajaí**, Ibirama, v. 5, n. 8, p. 1-13, 2016. Disponível em: <http://www.revistas.udesc.br/index.php/reavi/article/viewFile/6750/6415>. Acesso em: 15 abr. 2019.

PICAZO-VELA, Sergio; GUTIÉRREZ-MARTÍNEZ, Isis; LUNA-REYES, Luis Felipe. Understanding risks, benefits, and strategic alternatives of social media applications in the public sector. **Government Information Quarterly**, Amsterdam, v. 39, p. 504-511, out. 2012. Disponível em: <https://doi.org/10.1016/j.giq.2012.07.002>. Acesso em: 02 mar. 2019.

PRABHAKARAN, Selva. How to make any plot in ggplot2? **r-statistics.co**. Disponível em: <http://r-statistics.co/ggplot2-Tutorial-With-R.html>. Acesso em: 19 jun. 2019.

PREDICTIVE Analytics Today. **Top 27 free software for text analysis, text mining, text analytics.** Disponível em: <https://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/>. Acesso em: 11 mai. 2019.

QUEIROZ, Gabriela de. et al. Package “tidytext”. **The Comprehensive R Archive Network**. 2019. Disponível em: <https://cran.r-project.org/web/packages/tidytext/tidytext.pdf>. Acesso em: 24 jun. 2019.

RANJAN, R.; AGARWAL, S.; VENKATESAN, S. Detailed analysis of data mining tools. **International Journal of Engineering Research & Technology (IJERT)**. v. 6, n. 5, mai. 2017. Disponível em: <https://www.ijert.org/research/detailed-analysis-of-data-mining-tools-IJERTV6IS050459.pdf >. Acesso em: 20 mai. 2019.

REATEGUI, E. et al. Sobek: A Text Mining Tool for Educational Applications. In: International Conference on Data Mining, 2011, **Anais…**, Las Vegas, 2011, p. 59-64. Disponível em:

<https://pdfs.semanticscholar.org/119b/9fdda9a9e3bfc5764ad07cfbe655231c3443.pdf>. Acesso em: 23 jun. 2019.

RECUERO, Raquel. **Comunidades em redes sociais na internet:** proposta de tipologia baseada no Fotolog.com. 2006. Tese (Doutorado em Comunicação e Informação) – Programa de Pós-Graduação em Comunicação e Informação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006. Disponível em: <http://www.lume.ufrgs.br/handle/10183/8614>. Acesso em: 22 fev. 2019.

_____. **Redes Sociais na Internet**. 1. ed. Porto Alegre: Meridional, 2009. Disponível em: <http://www.ichca.ufal.br/graduacao/biblioteconomia/v1/wp-content/uploads/redessociaisnainternetrecuero.pdf>. Acesso em: 22 fev. 2019

RIEDER, Bernhard. **Facebook's app review and how independent research just got a lot harder**. Disponível em: <http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/>. Acesso em: 16 abr. 2019.

rss2json.com. **API Documentation**. Disponível em: <https://rss2json.com/docs>. Acesso em: 28 abr. 2019.

SALLOUM, Said.; AL-EMRAN, Mostafa; SHAALAN, Khaled. Mining social media text: Extracting knowledge from Facebook. **International Journal of Computing and Digital Systems**, v. 6, n. 2, p. 73-81, mar. 2017. Disponível em: <https://www.researchgate.net/publication/314095118_Mining_Social_Media_Text_Extracting_Knowledge_from_Facebook>. Acesso em: 23 mar. 2019.

SALMAN, Nurul. **Achieving trusted data in big data by using network platform**. 2017. 21 f. Curso de Ciência da Computação, University Sultan Zainal Abidin, Terengganu, 2017. Disponível em: <http://greenskill.net/suhailan/fyp/report/037673.pdf>. Acesso em: 02 mar. 2019.

SERAPIÃO, Paulo Roberto Barbosa; SUZUKI, Kátia Mitiko Firmino; MARQUES, Paulo Mazzoncini de Azevedo. Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia. **Revista Radiologia Brasileira**. São Paulo, v. 43, n. 2, p. 103-107, mar.-abr. 2010. Disponível em: <http://www.scielo.br/pdf/rb/v43n2/a10v43n2.pdf>. Acesso em: 19 abr. 2019.

SILVA, Mário; CARVALHO, Paula; SARMENTO, Luís. Building a sentiment lexicon for social judgement mining. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 10, 2012, **Proceedings…** Coimbra, p. 218-228, https://dl.acm.org/citation.cfm?id=2261082. Acesso em: 17 jul. 2019.

SILVA, Luciano Timoteo da; FERREIRA JUNIOR, Achiles Batista. Marketing político e sua importância através das mídias sociais. **Revista Temática**, João Pessoa, n. 8, p. 12, ago. 2013. Disponível em: <http://www.insite.pro.br/2013/agosto/marketing_politico_midiasdigitais.pdf>. Acesso em: 23 mar. 2019.

SILVA, Tarcízio; STABILE, Max (Org.). **Monitoramento e pesquisa em mídias sociais**: Metodologias, aplicações e inovações. São Paulo: Instituto Brasileiro de Pesquisa e Análise de Dados, 2016.

SOUZA, Marlo et al. Construction of a Portuguese Opinion Lexicon from multiple resources. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 8, 2011, **Proceedings**... Cuiabá, 2011, p. 59-66. Disponível em: <https://www.aclweb.org/anthology/W11-4507.pdf>. Acesso em: 17 jul. 2019.

Statistical tools for high-throughput data analysis (STHDA). **Text mining and word cloud fundamentals in R: 5 simple steps you should know**. Disponível em: <http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know>. Acesso em: 11 jun. 2019.

Statistical tools for high-throughput data analysis (STHDA). **Text mining and word cloud fundamentals in R**: 5 simple steps you should know. Disponível em: <http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know>. Acesso em: 11 jun. 2019.

TELLES, André. **A revolução das mídias sociais:** Cases, conceitos, dicas e ferramentas. São Paulo: M. Books, 2011. Disponível em: <https://www.andretelles.net.br/downloads/a-revolucao-das-midias-sociais-andre-telles.pdf>. Acesso em: 27 abr. 2019.

TEXT Mining example codes (tweets). **RStudio Pubs Static**. Disponível em: <https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html>. Acesso em: 12 jun. 2019.

THIEURMEL, Benoit; TITOUAN, Robert. Package “visNetwork”. **The Comprehensive R Archive Network**. 2019. Disponível em: <https://cran.r-project.org/web/packages/visNetwork/visNetwork.pdf>. Acesso em: 24 jun. 2019.

TURCOTTE, Jason et al. News recommendations from social media opinion leaders: Effects on media trust and information seeking. **Journal of Computer-Mediated Communication**, v. 20, p. 520–535, 2015. Disponível em: <https://academic.oup.com/jcmc/article/20/5/520-535/4067592>. Acesso em: 23 mar. 2019.

TWITTER DEVELOPERS. **API reference index**. Disponível em: <https://developer.twitter.com/en/docs/api-reference-index>. Acesso em: 14 abr. 2019.

WELDON, Ashley. Top 30 Big Data Tools for Data Analysis (Updated in 2019). **Octoparse**. Disponível em: <https://www.octoparse.com/blog/top-30-big-data-tools-for-data-analysis>. Acesso em: 10 mai. 2019.

WICKHAM, Hadley. Package “tidyverse”. **The Comprehensive R Archive Network**. 2017. Disponível em: <https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf>. Acesso em: 24 jun. 2019.

WICKHAM, Hadley et al. Package “ggplot2”. **The Comprehensive R Archive Network**. 2019. Disponível em: <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>. Acesso em: 24 jun. 2019.

WILLIAMS, Graham. Rattle: A Data Mining GUI for R. **The R Journal**, v. 1, n. 2, dez. 2009. Disponível em: <https://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf>. Acesso em: 23 jun. 2019.

WILSON, Robert; GOSLING, Samuel; GRAHAM, Lindsay. A Review of Facebook Research in the Social Sciences. **Perspectives on Psychological Science**, v. 7, n. 3, p. 203-220, 2012. Disponível em: <http://journals.sagepub.com/doi/10.1177/1745691612442904>. Acesso em: 05 mar. 2019.

WU, Bo et al. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In: IJCAI, 26., 2017, **Proceedings**... Melbourne, 2017. Disponível em: <https://www.ijcai.org/proceedings/2017/0427.pdf>. Acesso em: 01 mar. 2019.

ZAGO, Gabriela; BASTOS, Marco. Visibilidade de notícias no Twitter e no Facebook: Análise comparativa das notícias mais repercutidas na Europa e nas Américas. **Brazilian Journalism Research**, Brasília, v. 9, n. 1, p. 116-133, jun. 2013. Disponível em: <https://bjr.sbpjor.org.br/bjr/article/view/510/445>. Acesso em: 02 mar. 2019.

ZHANG, Daokun et al. User Profile Preserving Social Network Embedding. In: IJCAI, 26., 2017, **Proceedings**... Melbourne, 2017. Disponível em: <http://static.ijcai.org/proceedings-2017/0472.pdf>. Acesso em: 22 fev. 2019.

APÊNDICE

APÊNDICE A - Manual de instalação e uso das ferramentas selecionadas

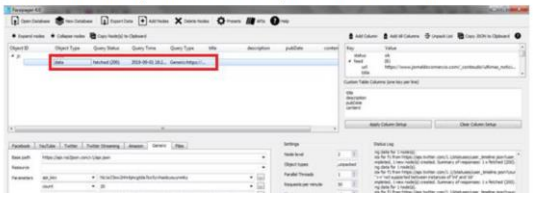
Devido à sua extensão – mais de 80 páginas -, o mesmo não foi colocado nesta seção. Entretanto, o mesmo pode ser acessado, em sua íntegra, no endereço github.com/ojonathacardoso/ranalyzer-adds/blob/master/Manual.pdf.

SUMÁRIO

Etapa 5: Obter os dados, conforme explicado no subcapitulo 2.1.1.4. O limite de dados obtidos varia conforme o RSS.

Etapa 6: Expandir os dados obtidos.

Após a busca, diferentemente das demais buscas, aparece apenas uma linha, cujo valor na coluna "Object Type" é "data". Isso não quer dizer que apenas um registro retornou, mas que deve ser feito um procedimento adicional.



etapa e, então, copie, cole e altere os códigos.

Primeiro, vamos ajustar as linhas que contém a expressão "Ontem". Para isto, converta a coluna "Data" usando o código abaixo.

```
replace(value,'Ontem','DATAONTEM')
```

A expressão DATAONTEM deve ser substituída pela data na qual você obteve os dados. Por exemplo, se os dados foram obtidos no Data Miner em 20 de agosto, a expressão "Ontem" se relaciona a "19 de agosto". Lembre-se que não se pode usar o "º" em 1º.

```
replace(value,'Ontem','19 de agosto')
```

Segundo, vamos ajustar as demais linhas que contém quantidade de horas, como "2 h" ou "12 h" – ou seja, que se referem a "Hoje". Elas devem ser substituídas pelo dia da coleta, assim como acima mostrado.

```
if(indexOf(value,' de ') == -1, 'DATAHOJE', value)
```

APÊNDICE B - Questionário aplicado na Assessoria de Comunicação

Questionário

Olá! Meu nome é Jonatha Cardoso, sou acadêmico do Curso de Sistemas de Informação, pela Universidade FEEVALE. Estou desenvolvendo meu Trabalho de Conclusão de Curso, denominado "Mineração de Texto em Fontes da Internet, com enfoque na Administração Pública e Atividade Política", orientado pela Profª Dra. Marta Rosecler Bez.

Após utilizar o Modelo de Extração de Conhecimento, bem como testar as ferramentas apresentadas, gostaria de pedir a colaboração desta Assessoria de Comunicação para responder algumas questões.

Antes disto, informe seus dados. Os mesmos serão utilizados para fins estatísticos.

*Obrigatório

1. **Idade:** *

2. **Sexo:** *
   Marcar apenas uma oval.
   - ◯ Masculino
   - ◯ Feminino
   - ◯ Prefiro não responder
   - ◯ Outro: _____

3. **Escolaridade:** *
   Marcar apenas uma oval.
   - ◯ Ensino Fundamental Incompleto
   - ◯ Ensino Fundamental Completo
   - ◯ Ensino Médio Incompleto
   - ◯ Ensino Médio Completo
   - ◯ Ensino Superior Incompleto
   - ◯ Ensino Superior Completo

Relação da Assessoria de Comunicação com fontes da Internet

4. **Indique quais mídias sociais a assessoria utiliza para divulgar informações sobre a Câmara.** *
   Marque todas que se aplicam.
   - ☐ Facebook
   - ☐ Twitter
   - ☐ YouTube
   - ☐ Instagram
   - ☐ Outro: _____

5. **Indique quais mídias tradicionais a assessoria utiliza para divulgar informações sobre a Câmara.** *

Marque todas que se aplicam.

- [ ] Jornal local
- [ ] Jornal estadual ou nacional
- [ ] Revista
- [ ] Site de notícias
- [ ] Blog
- [ ] Outro: _____

6. **Indique quais mídias sociais a assessoria utiliza para pesquisar informações sobre a Câmara ou sobre seus membros.** *

Marque todas que se aplicam.

- [ ] Facebook
- [ ] Twitter
- [ ] YouTube
- [ ] Instagram
- [ ] Outro: _____

7. **Indique quais mídias tradicionais a assessoria utiliza para busca de informações sobre a Câmara, seus membros, a Prefeitura, ou outros assuntos.** *

Marque todas que se aplicam.

- [ ] Jornal local
- [ ] Jornal estadual ou nacional
- [ ] Revista
- [ ] Site de notícias
- [ ] Blog
- [ ] Outro: _____

8. **A assessoria já utilizou ou utiliza alguma ferramenta para buscar informações de alguma das fontes anteriormente listadas?** *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

9. **Se a resposta na pergunta anterior foi "Sim", indique qual ou quais ferramentas já utilizou. Se a resposta foi "Não", explique por que nunca utilizou-se nenhuma ferramenta.** *

_____

_____

_____

_____

_____

10. **É importante para a Câmara ter acesso a informações sobre suas mídias sociais.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

11. **É importante para o Câmara ter acesso a informações sobre mídias tradicionais, como jornais locais e sites de notícias.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

Modelo e ferramentas utilizadas

12. **A assessoria utilizou o modelo com que frequência?**
Marcar apenas uma oval.

	1	2	3	4	5	
Nenhuma	◯	◯	◯	◯	◯	Muita

13. **A primeira camada do modelo (Coleta de dados) atende a necessidade de busca de dados nas fontes consideradas como importantes.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

14. **A segunda camada do modelo (Pré-processamento) engloba todas as etapas relacionadas a limpeza e filtragem dos dados coletados na primeira camada. Desta forma, a camada possui todas as etapas necessárias ou consideradas como importantes.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

15. **A terceira camada do modelo (Extração das informações) engloba todas as etapas de processamento do texto escolhido, gerando informações a respeito de correlação entre palavras, frequêcia de palavras e análise de sentimentos. Desta forma, a camada possui todas as etapas necessárias ou consideradas como importantes.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

16. **A última camada do modelo (Análise de conhecimento) permite a visualização de gráficos, tabelas, grafos e outras representações visuais das informações geradas. Desta forma, a camada atende a necessidade de extração de informação e conhecimento a partir dos dados inseridos na primeira camada.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

17. **A partir das ferramentas utilizadas, foi possível extrair conhecimento sobre a Câmara.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

18. **A partir das ferramentas utilizadas, foi possível extrair conhecimento sobre a atuação dos vereadores.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

19. **A partir das ferramentas utilizadas, foi possível extrair conhecimento sobre outros assuntos.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

20. **O modelo é útil para busca de informações para a assessoria e para a Câmara.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

Formas de visualização

Das formas de visualização gráfica que são geradas, marque a seguir as preferências da assessoria.

21. **Nuvem de palavras** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

22. **Grafo de palavras** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

23. **Gráfico de frequências** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

24. **Gráfico de correlação entre palavras** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

25. **Análise de sentimentos** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

26. **As formas de visualização gráfica, em geral, são claras e compreensíveis.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

Manual

27. **O manual de instalação, configuração e uso das ferramentas validadas neste experimento estava claro e de fácil compreensão.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

28. **O manual foi útil e importante para compreender o processo e aprender como devem ser usadas as ferramentas** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

29. **Quais melhorias devem ser implementadas no manual?**

Perspectivas para o futuro

30. **O fato de ter utilizado o modelo e as ferramentas durante duas semanas, simplificou o trabalho da assessoria?** *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

31. **A assessoria pretende utilizar as ferramentas validadas neste experimento?** *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não
- ◯ Talvez

32. **Justifique a resposta anterior:** *

APÊNDICE C - Questionário aplicado em Gabinetes Parlamentares

Questionário

Olá! Meu nome é Jonatha Cardoso, sou acadêmico do Curso de Sistemas de Informação, pela Universidade FEEVALE. Estou desenvolvendo meu Trabalho de Conclusão de Curso, denominado "Mineração de Texto em Fontes da Internet, com enfoque na Administração Pública e Atividade Política", orientado pela Profª Dra. Marta Rosecler Bez.

Após utilizar o Modelo de Extração de Conhecimento, bem como testar as ferramentas apresentadas, gostaria de pedir a colaboração deste gabinete para responder algumas questões.

Antes disto, informe seus dados. Os mesmos serão utilizados para fins estatísticos.

*Obrigatório

1. **Idade:** *

2. **Sexo:** *
   Marcar apenas uma oval.
   - ◯ Masculino
   - ◯ Feminino
   - ◯ Prefiro não dizer
   - ◯ Outro: _____

3. **Escolaridade:** *
   Marcar apenas uma oval.
   - ◯ Ensino Fundamental Incompleto
   - ◯ Ensino Fundamental Completo
   - ◯ Ensino Médio Incompleto
   - ◯ Ensino Médio Completo
   - ◯ Ensino Superior Incompleto
   - ◯ Ensino Superior Completo

Relação do gabinete com fontes da Internet

4. **Indique quais mídias sociais o gabinete utiliza para divulgar a atuação do vereador.** *
   Marque todas que se aplicam.
   - ☐ Facebook (Página)
   - ☐ Facebook (Perfil pessoal)
   - ☐ Facebook (Grupo)
   - ☐ Twitter
   - ☐ YouTube
   - ☐ Instagram
   - ☐ Outro: _____

5. **Indique quais mídias tradicionais o gabinete utiliza para divulgar informações sobre a atuação do vereador.** *

Marque todas que se aplicam.

- [ ] Jornal local
- [ ] Jornal estadual ou nacional
- [ ] Revista
- [ ] Site de notícias
- [ ] Blog
- [ ] Outro: _____

6. **Indique quais mídias sociais o gabinete utiliza para buscar informações sobre a atuação do vereador ou sobre a Câmara.** *

Marque todas que se aplicam.

- [ ] Facebook (Página)
- [ ] Facebook (Perfil pessoal)
- [ ] Facebook (Grupo)
- [ ] Twitter
- [ ] YouTube
- [ ] Instagram
- [ ] Outro: _____

7. **Indique quais mídias tradicionais o gabinete utiliza para buscar informações sobre o vereador, a Câmara, a Prefeitura ou outros assuntos.** *

Marque todas que se aplicam.

- [ ] Jornal local
- [ ] Jornal estadual ou nacional
- [ ] Revista
- [ ] Site de notícias
- [ ] Blog
- [ ] Outro: _____

8. **O gabinete já utilizou ou utiliza alguma ferramenta para buscar informações de alguma das fontes anteriormente listadas?** *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

9. **Se a resposta na pergunta anterior foi "Sim", indique qual ou quais ferramentas já utilizou. Se a resposta foi "Não", explique por que nunca utilizou-se nenhuma ferramenta.** *

_____

_____

_____

_____

_____

10. **É importante para o gabinete ter acesso a informações sobre suas mídias sociais.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

11. **É importante para o gabinete ter acesso a informações sobre mídias tradicionais, como jornais locais e sites de notícias.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

Modelo e ferramentas utilizadas

12. **O gabinete utilizou o modelo com que frequência?** *
Marcar apenas uma oval.

	1	2	3	4	5	
Nenhuma	◯	◯	◯	◯	◯	Muita

13. **A primeira camada do modelo (Coleta de dados) atende a necessidade de busca de dados nas fontes consideradas como importantes para o gabinete.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

14. **A última camada do modelo (Análise de conhecimento) atende a necessidade de extração de informação e conhecimento a partir dos dados inseridos na primeira camada.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

15. **A partir das ferramentas utilizadas, foi possível extrair conhecimento sobre a atuação do vereador.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

16. **A partir das ferramentas utilizadas, foi possível extrair conhecimento sobre outros assuntos.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

17. **As informações apresentadas pela ferramenta condizem com a realidade do vereador e de sua atuação como parlamentar.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

18. **O modelo é útil para busca de informações para o gabinete.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

Formas de visualização

Das formas de visualização gráfica que são geradas, marque a seguir as preferências do gabinete.

19. **Nuvem de palavras** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

20. **Grafo de palavras** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

21. **Gráfico de frequências** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

22. **Gráfico de correlação entre palavras** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

23. **Análise de sentimentos** *
Marcar apenas uma oval.

	1	2	3	4	5	
Pouco interessante	◯	◯	◯	◯	◯	Muito interessante

24. **As formas de visualização gráfica, em geral, são claras e compreensíveis.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

Manual

25. **O manual de instalação, configuração e uso das ferramentas validadas neste experimento estava claro e de fácil compreensão.** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

26. **O manual foi útil e importante para compreender o processo e aprender como devem ser usadas as ferramentas** *
Marcar apenas uma oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

27. **Quais melhorias devem ser implementadas no manual?**

_____

_____

_____

_____

_____

Perspectivas para o futuro

28. **O fato de ter utilizado o modelo e as ferramentas durante duas semanas, simplificou o trabalho do gabinete?** *
Marcar apenas uma oval.

◯ Sim
◯ Não

29. **O gabinete pretende utilizar as ferramentas validadas neste experimento?** *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não
- ◯ Talvez

30. **Justifique a resposta anterior:** *

_____

_____

_____

_____

_____