

UNIVERSIDADE ESTADUAL PAULISTA

Instituto de Geociências e Ciências Exatas - IGCE

Curso de Bacharelado em Ciências da Computação

SAMUEL CAETANO DA SILVA

**Mineração de texto aplicado a um estudo de caso no
Twitter**

Orientadora: Prof^a Dr^a Adriane Beatriz de Souza Serapião

Coorientadora: Prof^a Dr^a Verônica Oliveira de Carvalho

Rio Claro - SP

2017

Mineração de texto aplicado a um estudo de caso no Twitter

Trabalho de Conclusão de Curso, na modalidade Trabalho de Graduação, realizado no período de agosto à dezembro de 2017, apresentado no 2º semestre de 2017 à disciplina ES/TG do Curso de Bacharelado em Ciências da Computação do Instituto de Geociências e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de Rio Claro, segundo a Portaria IGCE/DTA nº 043/2012.

Aluno: Samuel Caetano da Silva

Orientadora: Prof^a Dr^a Adriane Beatriz de Souza Serapião
Coorientadora: Prof^a Dr^a Verônica Oliveira de Carvalho

Rio Claro - SP

2017

Este documento foi confeccionado em *Latex*.

Resumo

Diante do novo contexto de interações sociais baseadas nas redes sociais virtuais na chamada Era da Informação, a possibilidade de se obter conhecimento a partir dessas novas plataformas de convívio se mostra potencial, uma vez que a quantidade de dados disponíveis nessas redes é grande e a extração de conhecimento desses dados proporciona vantagem competitiva para qualquer pessoa ou grupo capaz de obtê-lo. A plataforma social virtual escolhida foi o Twitter. Para tanto, esse projeto propõe utilizar esses dados para descobrir quais são as palavras mais frequentemente reproduzidas pelos seguidores dos perfis de empresas do mercado segurador no Twitter. Entretanto, a extração de conhecimento desses dados não é um processo trivial, uma vez que esses estão desestruturados e dispersos pelas plataformas sociais virtuais, levando à problemática do modo como esses dados podem ser coletados e processados. Como proposta de solução, este projeto aplica técnicas de Mineração de Dados e de Textos para, respectivamente, solucionar esse problema inicial de coleta e tratamento dos dados. Para a extração de conhecimento foi aplicado técnicas de clusterização com o intuito de realizar agrupamentos sobre as postagens dos seguidores dos perfis das empresas seguradoras. Por fim, esse projeto se mostrou relevante pois a quantidade de informação obtida com apenas o agrupamento das postagens dos seguidores mencionados foi grande, tornando possível que diversas análises sejam tomadas sobre essas informações, mostrando que a proposta de se utilizar as redes sociais virtuais para obtenção de conhecimento estratégico é algo relevante e poderoso.

Abstract

Considering the new context of social interactions based on virtual social networks in the called Information Age, the possibility of obtaining knowledge from those platforms shows potential, since the available amount of data in those networks is very high and the knowledge extraction from these data provides a competitive advantage for those capable to obtaining it. The virtual social network chosen was Twitter. This project proposes to make use of these data to discover which words are most frequently reproduced by the followers of the insurance company's profiles on Twitter. However, the knowledge extraction from these data is not a trivial process, because these are unstructured and dispersed over the virtual social networks, bringing to the problematic of how these data can be collected and processed. As a solution, this project applies Data and Texts Mining techniques to, respectively, solve this initial problem of collecting and treating data. To the knowledge extraction was applied clustering techniques with the purpose of grouping the posts of the followers of the profiles of the insurance companies. As conclusion, this project showed to be relevant because of the amount of information obtained by grouping of the posts of the followers, which was great, making possible that other analysis to be taken on this information, revealing that the proposal of using virtual social networks to obtain strategic knowledge is relevant and powerful.

Listas de Figuras

Figura 1 – Grafo simétrico e bidirecional.	21
Figura 2 – Grafo assimétrico e unidirecional.	22
Figura 3 – <i>Overview</i> do processo de análise de dados.	25
Figura 4 – Etapas de processamento de dados textuais.	26
Figura 5 – Esquema do procedimento metodológico.	31
Figura 6 – Diagrama de Venn para as três maiores contas da base.	43
Figura 7 – Figura comparativa de dimensionalidade para cada conta.	45
Figura 7 – Figura comparativa de dimensionalidade para cada conta - <i>continuação</i>	46
Figura 7 – Figura comparativa de dimensionalidade para cada conta - <i>continuação</i>	47
Figura 7 – Figura comparativa de dimensionalidade para cada conta - <i>continuação</i>	48
Figura 8 – Figura com os <i>boxplots</i> para cada conta após a redução de dimensão.	49
Figura 8 – Figura com os <i>boxplots</i> para cada conta após a redução de dimensão - <i>continuação</i>	50
Figura 9 – Figura com a distribuição dos termos de conta após a redução de dimensão.	52
Figura 9 – Figura com a distribuição dos termos de conta após a redução de dimensão - <i>continuação</i>	53
Figura 10 – Figura de clusterização particional para cada conta.	55
Figura 10 – Figura de clusterização particional para cada conta - <i>continuação</i>	56
Figura 11 – Figura de clusterização particional para cada conta.	58
Figura 11 – Figura de clusterização particional para cada conta - <i>continuação</i>	59
Figura 12 – Figura de clusterização hierárquica para cada conta.	61
Figura 12 – Figura de clusterização hierárquica para cada conta - <i>continuação</i>	62
Figura 13 – Figura de gráfico de bolha para cada conta.	64
Figura 13 – Figura de gráfico de bolha para cada conta - <i>continuação</i>	65
Figura 14 – Figura de nuvem de palavras para cada conta.	67
Figura 14 – Figura de nuvem de palavras para cada conta - <i>continuação</i>	68
Figura 15 – Mapa de localização dos seguidores das contas de estudo.	71
Figura 15 – Mapa de localização dos seguidores das contas de estudo - <i>continuação</i>	72
Figura 16 – Mapa de calor da localização dos seguidores das contas de estudo.	73
Figura 16 – Mapa de calor da localização dos seguidores das contas de estudo - <i>continuação</i>	74

Lista de Tabelas

Tabela 1 – Tabela de termos usados no Twitter.	23
Tabela 2 – Porcentagem de distribuição dos seguidores cadastrados	42
Tabela 3 – Porcentagem dos seguidores compartilhados entre seguradoras	44
Tabela 4 – Comparaçāo da quantidade de <i>clusters</i> gerados por diferentes métodos	60

List of Algorithms

1	Algorithm for collecting tweets from the Twitter platform.	32
2	Pre-processing algorithm	34
3	Algorithm for k -means.	37
4	Algorithm for calculating dissimilarity.	39
5	Algorithm for generating matrix X.	40

Sumário

1	INTRODUÇÃO	10
2	REVISÃO BIBLIOGRÁFICA	13
2.1	O Mundo pós-Internet	13
2.1.1	A revolução da informação	13
2.1.2	A Era da Informação	15
2.1.3	O mercado na era digital	17
2.2	As redes sociais e o Twitter	19
2.2.1	Redes e grafos sociais	19
2.2.2	A plataforma do Twitter	22
2.3	Mineração de textos	24
2.4	Trabalhos Relacionados	27
3	METODOLOGIA	29
3.1	Estudo de caso: mercado segurador como objeto de aplicação . . .	29
3.2	Procedimento metodológico	31
3.3	A coleta de dados e o pré-processamento	31
3.3.1	O tratamento dos documentos	33
3.3.2	A pesagem dos termos	34
3.4	A tarefa de pós-processamento	35
3.4.1	O agrupamento dos documentos pré-processados	35
3.4.2	A clusterização esférica	38
3.4.3	O problema da complexidade espacial e temporal no estudo de caso	39
4	RESULTADOS	42
4.1	Análise para geração dos resultados	42
4.2	Análise dos resultados obtidos	54
4.2.1	A análise extra-cluster	54
4.2.2	A análise intra-cluster	66
4.2.3	Geovisualização dos dados	69
5	CONCLUSÕES	75
	REFERÊNCIAS	76

"O homem é um animal social"

Aristóteles

"Conhecimento é poder"

Bacon, Francis

1 Introdução

Atualmente, as redes sociais virtuais já fazem parte do dia a dia das pessoas; seja em instituições, empresas ou na vida pessoal, a troca de dados realizada por meio dessas torna a informação compartilhada muito valiosa. Empresas, por exemplo, acabam utilizando os dados disponíveis nas redes sociais para descobrir informações que as ajudem a divulgar um produto com maior precisão, no sentido de direcionar o produto "certo" para um (possível) cliente "certo", de modo a encontrar o perfil de consumo de seus clientes, a fim de tornar seus planos de negócio mais assertivos e menos imprecisos.

A Internet das Coisas tem se tornado um termo cada vez mais importante desde que foi concebida e permite visualizar esse nova iteração entre sociedade e Internet. Esse conceito permite uma visão onde todas as coisas *online* estão conectadas e podem ser facilmente controladas e monitoradas, podem ser identificadas automaticamente e/ou podem se comunicar entre si (TSAI et al., 2014).

As coisas que compõe a Internet das Coisas podem ser visualizadas como depositantes de dados *online* na rede. Nesses dados existem padrões e esses padrões podem ser extraídos por meio de técnicas específicas que ajudam na obtenção de conhecimento. Recipientes desses dados podem ser as redes sociais virtuais.

Dentre as diversas redes sociais virtuais existentes como o Facebook, LinkedIn, Twitter, Google+ e WhatsApp, o Twitter se mostra tão relevante quanto o Facebook, apesar do número de usuários ativos ser inferior (PEWINTERNET, 2016). Essa é uma rede social onde um usuário *segue* outros usuários (*follower*) e tais usuários podem ser seguidos por outros (*followed*). Um *followed* publica mensagens com até 140 caracteres cada e esse tuíte (*tweet*) é visualizado por todos seus *followers*. Por meio dos tuítes se tem, então, uma troca intensiva de dados entre usuários.

É possível observar, diante do exposto, que a quantidade de informação possível de ser extraída do Twitter é imensa. Todo o conteúdo textual disponível nessa plataforma possui informações que são relevantes, as quais permitem n análises, desde o termo mais frequente entre os *followers* de uma dada empresa até a análise de sentimento¹ dos *tuítes* desses *followers*. Esse *background* permite vantagens competitivas para aqueles que são capazes de entender essas informações.

Se uma empresa, com acesso aos dados de seus *followers*, decide analisar as informações obtidas, essa poderia descobrir, por exemplo, se determinado produto está sendo lucrativo ou não, se determinado produto tem apresentado falhas ou afins, se seus

¹ Análise de sentimento é o campo que analisa as opiniões de pessoas sobre determinadas entidades ou objetos, classificando essa opinião como positiva, negativa ou neutra (HADDI; LIU; SHI, 2013).

clientes mudaram de gosto sobre determinado produto ou sobre a empresa, etc.

A mineração de texto (MT), que é uma subárea da Inteligência Artificial, surge como um meio prático de se analisar os dados discutidos de modo a encontrar informações. A MT é um processo o qual se busca extraír conhecimentos relevantes ao usuário a partir de dados não estruturados. O objetivo é revelar padrões ao usuário de modo a permiti-lo obter conhecimento. O processo de MT é composto pelas seguintes etapas: (i) coleta dos documentos (textos) a serem analisados, (ii) tratamento dos documentos selecionados, (iii) extração de padrões a partir dos documentos tratados e, por fim, (iv) análise e exploração dos padrões obtidos. A etapa (ii) pode ser chamada de pré-processamento.

Diversas tarefas podem ser aplicadas durante a etapa (iii) para permitir que as análises sejam realizadas posteriormente. Uma dessas tarefas pode ser o agrupamento (*clustering*). É por meio da tarefa de *clusterização* que se agrupa dados semelhantes, de acordo com alguma métrica. No contexto desse trabalho, há o agrupamento dos *followers* por similaridade² dos *tuítes* (documentos).

Logo, diante do exposto, esse trabalho tem por objetivo aplicar a MT nos *tuítes* recuperados do Twitter. Com os *tuítes* recuperados se realiza o agrupamento para descobrir qual a palavra mais utilizada pelos *followers* de um dado *followed*. O que permite conhecer (ainda que superficialmente) qual o perfil desses *followers* e, consequentemente, qual o perfil das pessoas que curtem uma determinada conta no Twitter. Através de um estudo de caso aplicado, nesse trabalho, se procurará compreender todo o processo necessário para a operacionalização de uma aplicação real.

Para tanto, é necessário entender todo o processo, implementá-lo e testá-lo com exemplos reais. Define-se, portanto, o mercado segurador como objeto de exemplo. A escolha por tal segmento do mercado se justifica pela força econômica representada pelas seguradoras. Contudo, os conceitos e análises a serem apresentadas são adequáveis à quaisquer contextos.

Esse projeto se mostra bastante atual e relevante, considerando o atual estágio da Internet e como a iteração entre essa e a sociedade propiciou o acúmulo de dados *online*, fazendo surgir o *Big Data* e a crescente necessidade de ferramentas computacionais para a análise de dados (TSAI et al., 2014). O atual cenário é esse e as técnicas para a análise desses inúmeros dados dispostos na Internet, como as que serão expostas nesse trabalho, são demasiadamente importantes devido à quantidade ilimitada de conhecimento e de aplicações que podem ser extraídas das informações que esses dados contêm.

Para o desenvolvimento da aplicação proposta neste trabalho foram utilizadas, em maior grau, os pacotes SciKitLearn, Matplotlib e SciPy, os códigos (*scripts*) foram

² Entende-se por similaridade o grau de semelhança que um objeto possui com outro, neste contexto os objetos são documentos textuais e o grau de semelhança é medido pelas aparições de palavras que estão contidas em cada documento

implementados através do KDevelop 5.1.0 usando a linguagem Python e utilizou-se a linguagem Sqlite para implementação do banco de dados.

A estrutura do trabalho é composta assim: o Capítulo 2 contextualiza o mundo atual com as redes sociais e sua importância de modo mais abrangente, a fim de elucidar conceitos e motivações; o Capítulo 3 apresenta a metodologia utilizada e o estudo de caso, aplicando à teoria os exemplos reais e o Capítulo 4 apresenta os resultados obtidos da aplicação no mercado segurador e o Capítulo 5 conclui esse trabalho com as considerações finais.

2 Revisão bibliográfica

A chegada da Era Digital, por meio da Revolução da Informação, tornou-se um marco na história da humanidade e pode ser facilmente entendida como motor de um novo mundo, o mundo virtual, e é nesse mundo que a sociedade da informação existe. Neste capítulo vários conceitos sobre o contexto presente na Revolução da Informação e em técnicas de tratamento da informação são introduzidos, relativos ao escopo da aplicação proposta.

2.1 O Mundo pós-Internet

O mundo pós-Internet, que é o atual, permitiu um rápido avanço no compartilhamento de informação e isso trouxe novas formas de interação social e econômica. Esta influência digital e "*online*" está configurando as relações em todos os aspectos da vida humana, como será abordado a seguir.

2.1.1 A revolução da informação

Indubitavelmente, as redes sociais fazem parte das interações humanas. O que quer que as pessoas comam, ouçam, falem ou onde quer que vão, são coisas que estão todas disponíveis nas dezenas de redes sociais virtuais que existem na rede mundial. A vida social *offline* foi transformada em vida social *online* e isso tem transformado as relações humanas como nada antes.

A interação **humano-humano**, que era a principal forma de conexão entre pessoas na realidade *pré-Internet*, está mais para uma relação do tipo **humano-computador-humano** na realidade *pós-Internet*, que é a qual vivemos. Essa relação é concluída a partir das seguintes proposições (LÉVY, 1997):

1. As pessoas se conectam à Internet através de um computador;
2. A Internet é a via pela qual os dados *online* transitam;
3. Os dados são disponibilizados na Internet pelas pessoas através de um computador;
4. Um dado deixa de ser *offline* quando passa a estar na Internet.

Das observações acima, se um dispositivo eletrônico de processamento de dados é um computador, então *smartphones* e *tablets* são computadores. Portanto, cada dado que está *online* é acessível, só e somente, via computadores. Desse modo, entre a pessoa

P_a e a pessoa P_b deve haver um dispositivo computacional envolvido que as permitam se comunicar na Internet.

Estudos da Fundação Getúlio Vargas (FGV) tornados públicos pelo jornal Folha de São Paulo, indicam que existem quase 1,6 computadores por habitantes no Brasil (FGV, 2016). Outro fato de relevância é o grau de dependência que a sociedade tem dos computadores e da Internet.

Esse fato leva à onipresença de computadores no dia-a-dia dos indivíduos do século XXI. Isso fica evidente quando o despertador, o rádio, a câmera, a lanterna, a televisão se tornam parte do *smartphone* (computador), ou seja, quando se combina encontros com amigos através de aplicativos de comunicação no *smartphone*, ou quando se entra numa discussão por meio de informações dispostas na Internet que são acessadas via computadores, ou quando adquirimos conhecimento ou o compartilhamos na nossa plataforma social *online*.

As pessoas gostam de compartilhar suas experiências com aquelas que compõem seu círculo de convívio, logo, essa relação não poderia ocorrer de maneira diferente no mundo virtual. Com a popularização da Internet, por volta dos anos 90, as pessoas começaram a descobrir as facilidades que a Internet lhes proporcionava, principalmente no sentido da comunicação, com a popularização dos *emails*.

A partir de então, uma revolução na comunicação e na interação entre indivíduos aconteceu: as pessoas, a quilômetros de distância uma da outra, poderiam trocar informações a custo de alguns cliques. O que antes necessitava papel, tinta, endereço, código postal, burocracia e muito tempo agora se fazia com um navegador de *Internet*, um endereço de *email* e alguns segundos. Essa revolução aproximou países e continentes, trazendo o mundo para a tela do computador (LÉVY, 1997).

Assim, atualmente, seria muito difícil se imaginar um mundo menos conectado. *Emails* ainda não supriam uma necessidade básica das pessoas: permitir que P_a expressasse suas experiências, opiniões e imagens importantes com todos seus conhecidos de modo simultâneo. De início, o surgimento de foruns *online* supriram a necessidade por essa exposição, entretanto, foi somente com as primeiras redes sociais *online* que as interações humanas começaram a ser propriamente virtualizadas, no sentido da quantidade de conexões feitas por uma pessoa P_a e da rápida interação entre elas (instantaneidade).

O surgimento do *email*, dos fóruns *online*, dos *chats online* acompanhavam a popularização dos computadores nos lares domésticos e quanto mais popular ficavam os computadores tanto mais sofisticadas ficavam os meios de interação social virtual. Estritamente ligado à massificação dos computadores, a informação na Internet também foi massificada e com isso um tipo de comércio virtual surgiu. O uso da informação disponível na rede tornou possível que um novo jeito de se entender as relações de mercado

surgisse. Se no mundo pré-Internet as relações econômicas se davam por poder e capital, no mundo pós-Internet as relações se dão por informação e computadores (FLORIDI, 1999).

Computadores são o único modo pelo qual a humanidade pode manipular as grandes quantidades de dados a fim de obter informação e conhecimento. Possivelmente, a curto prazo, não haverá modo mais eficaz de se processar dados, a partir das quantias disponíveis hoje na Internet, se não por avanços nos próprios computadores e na computação (FLORIDI, 1999). Portanto, a humanidade se encontra num estágio irreversível, onde não há espaço para se imaginar um futuro no qual tenha menos computadores, menos informação e menos interação virtual. Pelo contrário, o futuro será aquele no qual os computadores e a troca de informações serão tão naturais e implícitos no cotidiano que até as tarefas mais simples (por exemplo, escolher filmes, músicas ou lugares para lazer) serão computadorizadas e serão impensadas fora dos computadores pela geração contemplada.

A Segunda Revolução Industrial foi o domínio da capacidade humana sobre a eletricidade e isso tornou possível o surgimento dos circuitos eletrônicos, que levou os homens ao cenário de uma Terceira Revolução Industrial, que se deu pela aplicação de tecnologia de ponta nos processos industriais. Disso, não seria impossível dizer que vivencia-se uma outra revolução, essa menos industrial, no sentido dos processos e das relações, e mais social, no sentido das relações entre pessoas e das consequências dessa nas relações de mercado. Logo, neste contexto, vivencia-se uma Revolução da Informação (FLORIDI, 1999).

Segundo Floridi (1999), essas mudanças levaram a sociedade contemporânea a ser aquela que vivencia o maior fluxo de avanços tecnológicos e mudanças sociais correspondentes a esses avanços: por exemplo, a massificação do acesso ao conhecimento. A informação é, conforme escrito pelo autor, o novo *digital gold* e é o recurso mais valioso que a sociedade da informação possui e o único modo que essa conhece para interpretar sua realidade.

2.1.2 A Era da Informação

De acordo com Floridi (1999), a popularização dos computadores domésticos, assim como sua sofisticação, ocorreu juntamente ao avanço e popularização da Internet. O termo "Internet" se refere a *international network*, que nada mais é do que a rede mundial de comunicação, e pode ser entendida, sob uma visão técnica, em três grandes dimensões: física, digital e ciberspatial, que são respectivamente: (i) infraestrutura, (ii) capacidade de memória¹ e (iii) espaço semântico.

Entende-se dimensão física (infraestrutura) aquela organização de computadores e de conectividade que permite a transferência de dados entre esse computadores conectados.

¹ Memória como a capacidade que essa rede possui de reter dados e não no sentido de HD, *hard disk*, memória RAM ou memória física

A transferência de dados ocorre através de protocolos, por exemplo TCP/IP². Essa combinação, permite uma plataforma de memória global, que se resulta da coesão de todas as memórias de todos os computadores que compõem a rede, sendo extensível ilimitadamente, mas que é sempre finita. Por fim, dada essa capacidade não-física de reter dados, esses formam um conjunto que possui algum significado semântico ou conceitual para a rede, que é o ciberspatial.

Como apresentado em Floridi (1999), dada uma rede ρ e dado um documento x , para cada x , $x \in \rho$, provido por uma URL³ há um documento y , $y \in \rho$, que é acessado diretamente, sem a passagem por um documento z secundário. De modo que essa relação $\text{URL}(x) \rightarrow y$ descreve o ciberespaço da rede ρ .

Logo, o tráfego de dados de uma rede pode ser entendido como a relação anteriormente descrita. Então torna-se evidente que um documento x é um dado, assim como um documento y , e que a interação que leva de x à y constitui a comunicação de uma rede qualquer.

Seja a Internet um exemplo real de rede, esta possui uma quantidade finitamente grande de dados (documentos); essa vastidão de dados permite a obtenção de conhecimento para se entender quase qualquer coisa do aspecto econômico e social que se deseja. Ainda mais na sociedade da informação, onde a rede é quase que total e plenamente acessível (no sentido de que quase todos acessam a Internet) e desse modo, a disposição de dados *online* se torna totalmente descentralizada, ou seja, a produção de conteúdo *online* não é comandada apenas por grandes corporações da área da computação e informática, mas por pessoas e principalmente por elas.

Devido à crescente inclusão digital e, consequentemente, à expansão das redes sociais, os dados se tornaram massivos na rede e sua devida análise permite obter conhecimento sobre esses usuários que tornam seus dados *offline* em dados *online*. Comumente, as análises desses dados e o uso dessas informações se dá no mercado.

Os dados *online* são *commodities*⁴ e representam vantagens estratégicas de redirecionamento para venda e divulgação de produtos (FLORIDI, 1999). Mas, embora a visão mercadológica sobre o conhecimento extraído dos dados *online* seja majoritária, é possível fugir dessa abordagem, pois, num contexto de redes sociais virtuais, onde seus usuários publicam dezenas de informações, há dados que descrevem fenômenos sociais como sensação de segurança, visão política, aprovação ou reprovação de instituições públicas, por exemplo.

² *Transmission Control Protocol/Internet Protocol* é um conjunto de protocolos que permitem que dois ou mais computadores se comuniquem, essa comunicação é ponto-a-ponto, ou seja, um computador *host* envia dados para outro.

³ *Uniform Resource Locator* é uma referência a um documento na Internet

⁴ *Commodities podem ser entendidos como bens de input para a produção de outros bens e serviços* (INVESTOPEDIA, 2016a).

A sociedade da informação existe numa era onde, diferentemente de eras anteriores, a informação está em todos os lugares e é totalmente acessível a partir de quase todos os lugares. Durante séculos o conhecimento esteve restrito a determinadas pessoas e a determinadas instituições e seu acesso era apenas para camadas sociais específicas, enquanto que hoje todas as grandes encyclopédias estão disponíveis na palma da mão. Em tempos anteriores, um indivíduo somente poderia ter acesso a algum livro por meio de bibliotecas. A revolução da informação tirou o conhecimento das mãos de poucos e colocou-o nas mãos da maioria.

2.1.3 O mercado na era digital

A Era da Informação estabeleceu, além de novas relações sociais, novas relações econômicas. Por causa dos inúmeros avanços tecnológicos das últimas décadas e da massificação ao acesso à informação, as relações de negócios foram globalizadas e isso acarretou na virtualização das relações entre entidades que compõem o jogo econômico: fornecedores, produtores, varejistas, consumidores e fiscalizadores, por exemplo (TOMAÉL; ALCARÁ; CHIARA, 2005).

À medida que as interações sociais se faziam *online*, uma moderada virtualização das interações econômicas também acontecia. Relações de comércio *offline* agora poderiam se dizer *online*, de modo que se encenavam no palco da grande rede mundial de computadores, sem a presença física dos atores, e era propiciada pela crescente influência que os novos meios de comunicação exercia na sociedade (WIGAND, 1997).

O dinamismo proporcionado pela Revolução da Informação, no que se refere à mudança "daquilo que é", reflete nas interações econômicas entre indivíduos de um modo que se expande a definição usual, ou *mainstream*, do que é **comércio**. Se na Era Pré-Digital (pré-Internet) o comércio se dava numa relação canalizada por atores definidos em seus papéis, na Era Digital (pós-Internet) essa definição é mais confusa e imprecisa, uma vez que os atores não são definidos, já que podem ser qualquer indivíduo *online* (WIGAND, 1997). Ou seja, o comércio pode ser entendido como linear (ou centralizado) no pré-Internet, e não-linear (ou descentralizado) na Era Digital (pós-Internet).

De acordo com Wigand (1997), o comércio descentralizado se caracteriza pela falta da noção de fronteira das organizações, já que virtualizadas podem estar em quaisquer lugares; valores agregados⁵ mudam dinamicamente e atividades são novamente distribuídas, de acordo com as relações anteriormente conhecidas na era pré-Internet; e indivíduos se tornam empreendedores por si próprios, uma vez que as relações **humano-computador-humano** possuem dois indivíduo, um em cada ponta, e isso torna qualquer um protagonista e espectador diante desse modo virtual de interagir com outros.

⁵ Valor agregado descreve um melhoramento que uma empresa dá ao seu produto ou serviço antes de repassá-lo para os seus consumidores (INVESTOPEDIA, 2016b)

O próprio desenvolvimento e massificação do acesso à Internet e aos computadores demonstra o interesse e a importância credida pela iniciativa empresarial ao fomento dessa revolução e pela construção dessa era. Esses avanços fomentados por esses atores econômicos propiciaram, de acordo com Wigand (1997), os seguintes items:

1. avanços na quantidade de informação a ser comunicada;
2. uma relação mais próxima entre vendedor e comprador;
3. a realidade de um grande pátio onde vendedores e compradores existem juntos e compararam ofertas;
4. uso estratégico dos dados *online* disponíveis como modo de adquirir vantagem competitiva.

A partir disso, empresas que usufruem dessas consequências da Era Digital são aquelas que terão mais vantagem competitiva no campo mercadológico e essas (empresas) são as mais propensas a se sobressaírem em detrimento daquelas que insistem em negar a influência e o poder contido nos dados *online*. Um análise de caso, aplicado ao abordado anteriormente, é o de uma grande empresa nos Estados Unidos que, em 1995, iniciou suas atividades *online* no comércio de livros, e com o passar do tempo foi adquirindo muito capital e destacando-se nesse segmento. Tamanho foi o sucesso dessa empreitada que outros segmentos do mercado passaram a utilizar essa mesma abordagem, o que abriu um novo caminho, provavelmente sem volta (SILVA et al., 2016).

Num espaço amostral de tempo de apenas 12 anos, entre 2000 e 2012, houve um aumento de 1.500% no número de usuários da Internet brasileira (SILVA et al., 2016). Isso significa um aumento de potenciais consumidores que, agora, navegam na rede mundial de computadores.

O comércio descentralizado permite uma relação mais próxima entre vendedores e consumidores, de modo que essa relação tende a permitir preços mais acessíveis aos consumidores, se comparado ao comércio centralizado, que ainda existe e é o *mainstream* nas relações econômicas *offline*. Embora o primeiro seja aquele que se torna mais relevante e o segundo aquele que se torna mais obsoleto.

Durante o primeiro semestre de 2016 foi descoberto que, apesar da grande crise econômica vivenciada pelos brasileiros, houve um aumento significativo de 31% no número de consumidores ativos do mercado virtual (*e-commerce*). Isso mostra que as relações construídas no comércio virtual são aquelas que menos são afetadas pelas flutuações políticas e econômicas do mundo real (SILVA et al., 2016). O mercado virtual brasileiro é composto em 19% pelo setor de moda e acessórios, em 18% pelo segmento de cosméticos e

perfumaria, em 10% pelo de eletrodomésticos, em 9% pelo de livros e revistas e em 7% pelo setor de informática (SILVA et al., 2016).

Adquirir conhecimento a partir dos dados *online* é importante para obter vantagem competitiva no mercado. Uma possível aplicação utilizando essa abordagem mercadológica sobre o conhecimento extraído da Internet é a análise de sentimento realizada sobre dados *online*, de uma plataforma social virtual, durante um período determinado de tempo (SANTOS, 2016). Essa abordagem permitiu analisar as opiniões de clientes de uma dada marca (empresa) e polarizar essas opiniões como positiva ou negativa, de modo a classificar as opiniões do público sobre os produtos ou serviços fornecidos por uma empresa.

A análise de sentimentos realizada por (SANTOS, 2016) se mostra demasiadamente eficaz, no sentido de que uma empresa poderá saber com mais precisão onde realizar possíveis ajustes de modo a maximizar a satisfação do público, seja no fator preço, seja no fator qualidade. Essa precisão nos ajustes sobre seus produtos, provavelmente, não seria utilizada por empresas, ou instituições, que não tenham esse conhecimento, o que implicaria no uso do acaso e da sorte para que tais ajustes fossem bem-sucedidos, causando, possivelmente, prejuízo temporal e financeiro, o que levaria essa empresa a uma considerável desvantagem na competição por clientes, comparando-a a uma outra empresa que possui esse conhecimento e que atua na mesma área.

Para tanto, torna-se muito difícil obter vantagem competitiva no mercado na Era Digital ignorando a existência das interações virtuais que ocorrem a todo instante e que fazem circular uma infinidade de dados que contêm em si, preciosa informação que descreve⁶ as relações reais e consequentemente o mundo real.

2.2 As redes sociais e o Twitter

A interação social virtual se dá em plataformas. Existem várias e uma delas é o Twitter. Essa seção aborda as particularidades de algumas plataformas sob uma visão computacional.

2.2.1 Redes e grafos sociais

Dada a relação **humano-computador-humano**, temos P_a e P_b como entidades que se relacionam por meio de um computador α . Essa abordagem nos permite visualizar uma relação entre dois terminais que são conectados, fisicamente ou não. Sob tal abordagem, é possível encarar tal relação sob a forma de um grafo.

⁶ Essa descrição do mundo real não deve ser tomada como verdade absoluta, mas como um norteador para a tomada de decisões

Um grafo (G) é uma estrutura de dados composta de vértices (V) e arestas, (A), onde um vértice V_0 se conecta a um vértice V_1 através de uma aresta A . V_0 pode se conectar com n outros vértices somente por meio de arestas.

G pode ser do tipo dirigido (dígrafo), onde cada vértice possui um sentido de início (parte de) e fim (chega em), e nesse caso $A_{0,1} \neq A_{1,0}$. O que não ocorre em grafos do tipo não-dirigido, onde cada aresta não tem qualquer orientação e, portanto, $A_{0,1} = A_{1,0}$, onde os subíndices da aresta indicam início e fim, respectivamente. Dada essa abordagem, é possível visualizar a relação entre P_a e P_b , proposta acima, como um grafo; essa abordagem permite organizar o mundo virtual e por meio de plataformas,

A partir dessa abordagem, é possível organizar as relações sociais *online* por meio de grafos, que são os grafos sociais. Um grafo social representa uma rede, que também é social. Assim, no mundo virtual existem alguns grafos sociais e, consequentemente, algumas redes sociais, pois para cada organização dessas relações *online*, tem-se uma diferente abordagem sobre essas relações. Cada abordagem sobre as relações entre P_a e P_n deriva uma rede social.

Uma rede social pode ser entendida como um modo de se abordar as relações virtuais entre as pessoas que compõem grupos *online*. Essa abordagem remete às diferentes formas que essas interações acontecem nessa plataforma, ou rede, e essa "abordagem" determina como se dará, por exemplo, a comunicação e o armazenamento desses dados *online* dentro de uma plataforma. Ou seja, considere duas plataformas sociais (ou redes sociais) N_0 e N_1 , ambas possuem os mesmos usuários, mas cada uma possui uma abordagem diferente sobre o modo com o qual essas pessoas irão interagir nessas plataformas.

Em N_0 , os usuários podem se comunicar de modo que essa comunicação é armazenada num robusto banco de dados, onde em qualquer momento no futuro essas mensagens podem ser visualizadas pelas partes que se comunicavam. Porém, esses mesmos usuários não dispõem dessa funcionalidade na plataforma N_1 , pois todas as mensagens trocadas são deletadas da base de dados algum tempo depois de terem sido lidas pelo recebedor.

Apesar de ser um exemplo trivial e simples, é suficiente para ilustrar como o modo que se vê e organiza-se as relações sociais virtuais geram diferentes plataformas com diferentes características entre si, embora com algum grau de semelhança: N_0 e N_1 permitem a comunicação entre seus usuários. Todas as redes sociais têm um proposto em comum: a comunicação. O que torna uma rede social diferente de outra é o modo como essa decide permitir e conduzir isso.

Apesar do exemplo acima expor uma abordagem técnica, isso somente se torna viável a partir de uma abordagem teórica e abstrata, a qual ocorre por meio dos grafos e o modo como cada plataforma possui seu grafo social.

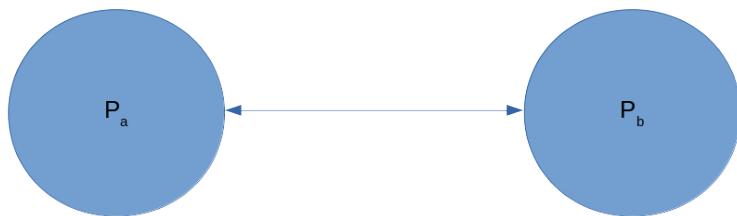


Figura 1 – Grafo simétrico e bidirecional.

Plataformas como o Facebook, o LinkedIn e o MySpace, por exemplo, possuem um grafo social que pode ser entendido como uma espécie de dígrafo, ou seja, cada usuário somente pode ser contato (ou amigo, no *lexicon* do Facebook) de um outro, se e somente se, esse outro permitir. De modo que se tem um grafo como o da Figura 1, onde P_a e P_b são usuários e a aresta que os conectam representa a "amizade mútua" entre esses usuários. Esse vértice é simétrico, uma vez que uma amizade nunca é unidirecional ou assimétrica, ou seja, para um ser amigo de outro, o outro deve ser amigo de um (RUSSEL, 2014). Essa abordagem tenta virtualizar com mais precisão as relações humanas e são as que podem representar melhor o mundo real no mundo virtual, no sentido de realismo.

Uma abordagem diferente tem o WhatsApp, por exemplo, onde um usuário não necessariamente precisa ser contato de outro para se comunicar com esse, o que gera um grafo onde as arestas estão conectadas por vértices não-direcionais até que se tenha uma primeira comunicação, que faz com que esse vértice adquira sentido e direção. Pode ser entendido como bidirecional ou simétrico, pois um canal de comunicação é aberto entre os dois nós (arestas).

Essa abordagem é semelhante às redes telefônicas, no sentido de que cada telefone representa uma aresta e que a rede telefônica representa os inúmeros vértices. Cada telefone, quando realiza chamadas para outro, muda o *status* de seu vértice de não-direcional para bidirecional e assim que a chamada é finalizada, esse vértice perde seu *status* de bidirecional e volta a ser não-direcional, no sentido de que não há mais fluxo de informação saindo de quaisquer pontas desse vértice. Essa análise é análoga para o WhatsApp, de modo que não há o *background* telefônico, mas a Internet e os agentes dessa plataforma não são telefones, mas computadores (*smartphones*).

Diferentemente das anteriores, o Twitter se destaca pela abordagem de seu grafo. Na plataforma do Twitter, um pode ser contato de outro mesmo se o outro não for contato de um, ou no *lexicon* do Twitter: um usuário pode ser seguidor de outro, independente



Figura 2 – Grafo assimétrico e unidirecional.

do outro. Essa abordagem do Twitter se diferencia da abordagem do Facebook pelo fato de que não há a tentativa se representar com fidelidade as relações humanas *offline* no mundo virtual. O Twitter não virtualiza as amizades, mas resgata uma relação de seguidor (**follower**) e seguido (**followed**).

A Figura 2 indica como pode ser representado o grafo social do Twitter. Isso simboliza que P_a ou P_b pode ser tanto *follower* um do outro, como também pode ser *followed* um do outro. O mútuo pode acontecer. O vértice indica o sentido de *follower*; a Figura 2 indica que P_a é *follower* de P_b e, consequentemente, que P_b é *followed* por P_a .

2.2.2 A plataforma do Twitter

O Twitter é uma rede social que é um serviço de *microblogging*⁷, e que permite que seus usuários se comuniquem por meio de curtas mensagens de no máximo 140 caracteres, que no *lexicon* do Twitter é denominado *tuíte* (tuíte). Os tuítes transmitem ideias e pensamentos. Essa característica do Twitter permite que seus usuários se comuniquem sobre aquilo que eles querem se expressar, sobre algo que querem compartilhar e tudo isso de um jeito simples e rápido.

A Era Digital e a sociedade da informação se caracterizam pela velocidade que suas interações sociais ou não se dão. Tudo se torna obsoleto rapidamente e logo já existe algo novo para substituir o antigo; o mesmo ocorre com o compartilhamento de informações e conhecimentos na Internet, tudo precisa ser rápido, a informação se torna velha e antiquada na mesma velocidade. Não obstante a isso, a plataforma do Twitter virtualiza, e identifica essas relações e intensifica a percepção sobre como novas informações são compartilhadas.

O grafo social do Twitter é um grafo composto de arestas e vértices direcionais e assimétricos, diferentemente do grafo social do Facebook, LinkedIn ou até mesmo do WhatsApp. A assimetria surge quando um usuário é *seguidor* de outro, mas o outro não

⁷ *Microblogging* é um serviço que possibilita que uma pessoa realize publicações e expresse-se acerca daquilo que considera relevante.

precisa saber nem que está sendo *seguido* pelo um. E é essa assimetria que permite suprir, ao virtualizar, a curiosidade humana no mundo *online*.

Segundo Russel (2014), essa abordagem permite entender o grafo social do Twitter como um grafo de interesse, que é a relação entre usuários e seus mais arbitrários interesses na plataforma, sendo esse um dos motivos pelos quais o Twitter se mostra uma plataforma rica para a exploração de dados.

A Tabela 1 apresenta termos da plataforma do Twitter e sua conotação.

Objeto	Conotação
<i>Tweets</i>	Bloco atômico básico na construção da plataforma do Twitter. É uma mensagem de no máximo 140 caracteres que pode ser textual ou multimídia.
<i>Users</i>	São todos os usuários que compõem essa plataforma. Podem ser seguidores e seguidos.
Entidades	Disponibilizam metadata e informações adicionais sobre conteúdo postado no Twitter, normalmente atrelado ao tuíte.
<i>Timeline</i>	Pilha de eventos de um usuário ou dos eventos compartilhados pelos seguidos por esse usuário.
<i>Hashtag</i>	Símbolo '#' que marca tópicos nos tuítes. Marca o assunto que o tuíte se refere

Tabela 1 – Tabela de termos usados no Twitter.

Os tuítes são os átomos que compõe a plataforma e são sempre publicados por um usuário. Esse usuário pode ser tanto um *follower* quanto um *followed*. Seja f_r um usuário seguidor (*follower*) de outros três usuários (*followed*) (f_{d_0} , f_{d_1} , f_{d_2}), sendo que f_{d_0} segue mutualmente f_r . A ideia de *timeline* (linha do tempo) consiste em fazer com que f_r receba uma lista em ordem cronológica dos tuítes publicados por todos os usuários do quais ele é *follower*, f_{d_0} recebe somente os tuítes publicados por f_r ; a *timeline* é como uma pilha, na qual o primeiro tuíte que entrou é o mais antigo e o último a entrar é o mais recente, temporalmente, e essa pilha é preenchida somente com os tuítes daqueles que são seguidos por um usuário qualquer.

O Twitter se torna relevante não só pelas possibilidades de abordagens mercadológicas sobre os dados *online*, mas por todo o arcabouço humano que compõe essa rede e pelo modo no qual sua dinâmica permite que as interações virtuais aconteçam. Um dos fatores de popularização do Twitter está além da possibilidade da comunicação na "velocidade do pensamento", mas no modelo assimétrico de seu grafo social. Como já mencionado, esse grafo social assimétrico permite ao Twitter suprir uma necessidade essencial do ser-humano: a curiosidade (RUSSEL, 2014).

2.3 Mineração de textos

A popularização da Internet, e consequentemente das redes sociais virtuais, elevou o número de indivíduos que interagem na rede e com isso o aumento na quantidade dos dados *online* disponíveis. Segundo Mattos (1982), dados, sejam *onlines* ou *offline*s, são referências não interpretadas e não estruturadas a algum objeto (físico ou não) e enquanto não estruturados e não classificados é impossível se extrair quaisquer informações deles.

Por exemplo, seja x um número qualquer. x é um dado; mas um dado que não traz qualquer informação consigo, de modo que somente é possível obter informação desse número lhe dando algum significado. Logo, dependendo do significado que é dado a x será possível, para quem o lê, interpretá-lo de algum modo. Ainda segundo o autor, a informação obtida a partir dos dados é um acréscimo de conhecimento.

O acesso à informação e o aumento do conhecimento proporcionam o entendimento do mundo real, de modo que permite, a quem tem o conhecimento, uma melhor abordagem sobre determinado problema. Assim, o melhor modo de se obter conhecimento, de se tentar entender o mundo virtual e como as relações sociais se dão na sociedade da informação na Era Digital é através das informações que estão contidas nos dados *online* espalhados pela Internet. As redes (plataformas) sociais comportam os dados que melhor descrevem as relações sociais virtuais.

Um dos modos de se realizar a interpretação dos dados da Internet é através do processo de *Data Mining* (Mineração de Dados), o qual consiste em descobrir padrões e correlações entre os dados analisados por meio de técnicas estatísticas e matemáticas (LAROSE, 2005). A Figura 3 ilustra uma visão geral do processo executado para obtenção e análise dos dados disponíveis. A base de dados, no caso desse trabalho, é a Internet, mais especificamente a plataforma social do Twitter; o processo de recuperação de dados é a coleta dos dados disponíveis na plataforma, e consiste em, partindo de algum critério, coletá-los e armazená-los para futuros processamentos. As técnicas de Mineração de Dados consistem no processo de dar sentido aos dados, estruturando e identificando-os. A análise dos padrões obtidos pelo processo anterior consiste nas n possíveis visualizações que os padrões encontrados permitem, e essa etapa acaba sob a subjetividade e o olhar do observador desses padrões. Por fim, a informação e o conhecimento são o resultado obtido por meio da visualização escolhida sobre os padrões dos dados da base.

A Mineração de Dados se mostra importante quando se parte da premissa de que é através do conhecimento que se aprimora técnicas ou costumes até então vigentes. Além disso, alguns outros fatores tornam propício o crescimento do uso das técnicas de mineração e, de acordo com Larose (2005), são os seguintes:

1. o crescimento da coleta de dados: a popularização da Internet e o crescimento da



Figura 3 – *Overview* do processo de análise de dados.

virtualização das relações entre pessoas possibilitaram um cenário propício para que os dados fossem postos *online* por essas pessoas;

2. a crescente disponibilidade de acesso aos dados da rede: a organização das plataformas virtuais permitiram o acesso dos dados disponíveis em suas plataformas através de determinadas políticas de privacidade;
3. o crescimento da capacidade de computação: o aumento da capacidade de armazenamento de dados em *datacenters*.

Somente a partir dessas técnicas é que se torna possível identificar, por exemplo, quais as novas tendências tanto consumidoras quanto sociais das pessoas. É possível determinar quando as pessoas estão mais propensas a comprarem um produto de uma determinada categoria ou não, por exemplo. Tal mecanismo permite o redirecionamento e a publicidade eficiente de produtos para os consumidores.

Considerando a rede social do Twitter, com toda sua complexidade e estrutura, é possível identificar que análises desse tipo são interessantes de se aplicar nela. O Twitter é baseado no tuíte e isso constitui o átomo de toda a plataforma, portanto, análises sobre os tuítes se mostram relevantes e essa abordagem é justificada por toda explicação acima realizada.

O tuíte é um dado textual, pois dele a informação mais relevante vem de um conteúdo escrito em linguagem natural que expressa algo sobre algum objeto. Para tanto, a mineração executada sobre dados desse tipo passa a ser mais específica, nesse caso, pode-se considerar a Mineração de Textos como um conjunto de técnicas e processos relevantes e úteis de se aplicar (MATHINA; SHANHI; NANDHINI, 2015).

A Mineração de Textos (MT) é um processo de análise de dados textuais que visa buscar e encontrar padrões entre esses dados (GODFREY et al., 2014). No *lexicon* da MT, um dado é um **documento**. Há um conjunto de passos a serem realizados e esses consistem em três grandes etapas, como mostrado na Figura 4. O pré-processamento consiste na estruturação do dado, de modo que se realiza: (i) a limpeza do documento retirando lixo textual⁸; (ii) a tokenização, que é a identificação das palavras (*tokens*), ou termos, que compõem o documento; (iii) remoção de *stop-words*⁹ e, por fim, ocorre (iv) a stemização¹⁰ dos *tokens* (termos). Ao final dessas etapas de pré-processamento obtém-se um documento *limpo*, *tokenizado* e *stemizado*.

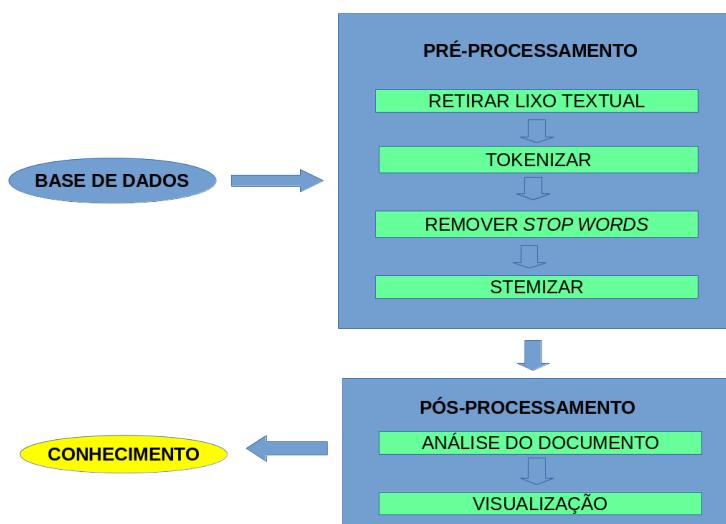


Figura 4 – Etapas de processamento de dados textuais.

Combinado a isso, o pós-processamento vem como uma análise feita sobre o documento tratado. Diversas abordagens podem ser tomadas durante o pós-processamento, uma das abordagens comuns para quando há a necessidade de se analizar dados textuais é a realização da "pesagem" dos termos que consiste em, por exemplo, verificar a frequência de aparição desse termo sobre o documento ou sobre um conjunto de documentos.

Por exemplo, dado um documento $\delta = [\tau_0, \tau_1, \dots, \tau_n]$, temos que $\forall \tau_n \in \delta, \exists \varphi = \frac{\text{número de ocorrências de } \tau_n \text{ em } \delta}{n}$, onde τ é um termo e φ é o peso desse termo no documento. Essa é apenas uma das possíveis abordagens. A etapa de pré-processamento é importante pois trata o documento, a fim de deixá-lo processável, e a etapa de pós-processamento permite todos os seguintes passos de análise e visualização dos padrões que esses documentos têm em si.

⁸ Quaisquer caracteres não pertencentes ao alfabeto latino

⁹ Palavras que são irrelevantes no contexto de análise textual, como artigos e preposições, por exemplo

¹⁰ Stemizar consiste em reduzir uma palavra à sua raiz. Exemplo:

trabalhando ⇒ *trabalh*

trabalhei ⇒ *trabalh*

cantar ⇒ *cant*

Para cada um dos processos superficialmente discutidos acima, há um algoritmo possível de ser aplicado e tal discussão será abordada no Capítulo 3.

2.4 Trabalhos Relacionados

Trabalhos, cuja abordagem utilizaram-se da plataforma virtual do Twitter são alguns, entre eles está o trabalho desenvolvido por Ribeiro, Tavares e Cohen (2014), que procura identificar usuários da plataforma que são influentes no tópico "cerveja", de modo a aplicar uma análise sobre grafos relacionais, além de técnicas de MT como ferramenta auxiliar de compreensão dos dados obtidos. O autor, neste trabalho, visa fornecer um modo capaz de identificar usuários influentes sob um "tópico" específico, de modo a partir esse conhecimento por empresas que se interessam.

Já no trabalho realizado por Franca e Oliveira (2014), há a proposta de análise de sentimento sobre os tuítes referentes aos protestos que ocorreram no Brasil no meio do ano de 2013. A abordagem proposta pelo autor consiste em coleta, pré-processamento e pós-processamento dos dados coletados da plataforma virtual do Twitter, de modo a ser possível descobrir quais foram os sentimentos dos usuários do Twitter com relação aos protestos que ocorreram.

Petró (2010) realizou uma abordagem voltada para o ramo de Comunicação Social, com o intuito de mostrar que as relações sociais virtuais são aquelas que melhor constróem um relacionamento entre consumidores e empresas, pois se trata de uma relação horizontal. As relações construídas no mundo virtual das redes sociais são aquelas que melhor permitem às empresas atingir grupos cada vez maiores de usuários, que podem ser, também, possíveis consumidores de seus produtos; tal possibilidade aproxima consumidores das empresas.

A autora identifica que para a inserção de uma empresa no mundo social virtual deve haver muito mais mudança de comportamento em relação à comunicação do que mudança técnica. Esse *upload* para as relações sociais virtuais permite a empresa adquirir um capital social até então desconhecido e é por esse capital que a empresa pode buscar melhorar suas relações com seus clientes.

Melo, Franklin e Vianna (2015) propõem o uso dessa abordagem como uma ferramenta de monitoramento de mensagens relacionadas ao mercado segurador, de modo a permitir uma regulação por parte da SUSEP através da análise de sentimento realizada sobre os tuítes. Assim, um documento marcado como "reclamação" ou "insatisfação" através da análise de sentimento permite que o órgão regulador estatal, a SUSEP, fiscalize melhor as organizações reguladas, as empresas seguradoras. Essa abordagem permite políticas de regulação econômica por parte de órgãos estatais de um modo mais eficientes, de acordo com o autor, por se basearem em reclamações "em tempo real".

Por fim, no trabalho proposto na presente monografia, no campo das funções de Mineração de Dados, a clusterização, com o objetivo de identificar os padrões presentes nos documentos (tuítes) dos seguidores de contas objeto de estudo, com a finalidade de permitir identificar qual o possível perfil dos seguidores de cada conta, além de permitir a comparação entre os seguidores dessas. No campo de técnicas, uso de algoritmos de aprendizado de máquina e no campo de aplicação, busca fornecer conhecimento de vantajoso uso, principalmente para *marketing* e comunicação.

3 Metodologia

A aplicação do mercado segurador como objeto de estudo no presente trabalho tem como finalidade analisar o mercado segurador em uma rede social virtual e determinar qual o possível perfil dos usuários que seguem cada empresa seguradora, se há ou não semelhanças entre esse público e, se houver, tornar possível identificar essa variação de modo a permitir uma vantagem estratégica para a abordagem das empresas interessadas em seu público-alvo.

3.1 Estudo de caso: mercado segurador como objeto de aplicação

O mercado segurador sempre esteve presente nas atividades econômicas. A forma mais primitiva de seguro remonta à antiga Babilônia onde os mercadores, nas caravanas, precisavam proteger seus bens de ladrões, do mal tempo e de ataques de animais. As caravanas transitavam em lugares inóspitos e perigosos, e isso representava uma tomada de risco, sem pessoas dispostas a tomar riscos as caravanas teriam uma baixa probabilidade de existir e, consequentemente, os avanços culturais e econômicos que elas propiciaram (CHARTERED INSURANCE INSTITUTE, 2017).

O seguro existe para garantir que os riscos de uma determinada atividade sejam menos danosos àquele que se arriscou a iniciá-la; para além disso, serve como garantia a preservação de propriedade: permite que alguém que perde sua casa ou automóvel seja restituído em valor do bem perdido. Conforme exposto por Chartered Insurance Institute (2017), os seguros são aquilo que permite as pessoas a seguirem inovando, preocupando-se menos com os riscos e mais com os sucessos; ou seja, é uma das ferramentas que torna possível o mundo receber inovação.

Segundo dados da *Tudo sobre seguros* (2017), uma iniciativa da Escola Nacional de Seguros, o mercado de seguros brasileiro se concentra em três grandes áreas: (i) saúde, (ii) pessoas, que compreende previdência e proteção contra acidentes, e (iii) automóveis. O seguro saúde compreende os planos de saúde e é o tipo de seguro que, no Brasil, possui maior receita: 67,8%. O seguro de previdência corresponde aos mecanismos de proteção contra riscos de aposentadoria e velhice; esse tipo de seguro teve um considerável aumento de participação de receita nos últimos anos, uma vez que nas últimas décadas houve no Brasil aumento tanto na expectativa de vida quanto na renda *per capita*. O seguro de automóveis já representou a maior fatia da receita capitada por seguradoras, mas nos últimos anos houve uma queda na participação para outros ramos, muito em decorrência do aumento pela procura de seguros de previdência e pela competição entre as seguradoras de automóveis.

As três áreas juntas somaram cerca de 86,6% da receita do mercado de seguros, em 2015. Além disso, ramos não tradicionais como seguro habitacional e rural foram os ramos do mercado segurador que mais cresceram entre 2011 e 2015, ainda segundo *Tudo sobre seguros (2017)*.

Em âmbito nacional, o Conselho Nacional de Seguros Privados (CNSP) é o orgão máximo regulador do mercado de seguros. O CNSP fixa diretrizes e normas regulatórias sobre as políticas gerais envolvendo seguros e tem por função regular a existência de seguradoras e de corretores de seguros. Os corretores de seguros são empresas ou profissionais liberais sem vínculo com as seguradoras e têm um nível de atuação mais próximo ao cliente.

A Superintendência de Seguros Privados (SUSEP) é o orgão responsável pela regulação e fiscalização do mercado segurador, com exceção dos seguros saúde. Essa autarquia tem por funções: implementar as políticas determinadas pelo CNSP, supervisionar a indústria seguradora, criar regulamentos relativos a operações envolvendo seguros, aprovar limitações às ações das seguradoras, entre outras.

Para a escolha do setor como objeto de estudo não houve discriminação entre ramos seguradores. Foram selecionadas 11 seguradoras com negócios no Brasil; essas são aquelas que obtiveram maior receita em 2016 e, portanto, foram as consideradas relevantes para objeto de estudo. A lista das seguradoras selecionadas seguiu a publicação do site *Valor econômico (2016)*, que é a seguinte:

1. Porto Seguro;
2. Bradesco Seguros;
3. Itaú Corretora;
4. SulAmérica Seguros;
5. Caixa Seguradora;
6. Seguros Unimed;
7. Allianz Brasil;
8. Liberty Seguros;
9. HDI Seguros Brasil;
10. Mapfre Brasil;
11. Tokio Marine Seguros;

3.2 Procedimento metodológico

Esse trabalho tem como objetivo realizar a identificação dos termos mais frequentemente usados por seguidores específicos de um dado seguidor, que no contexto desse trabalho serão as empresas seguradoras, de modo a permitir tanto uma possível análise de perfil desses seguidores quanto para apresentar a abordagem realizada como um método para adquirir vantagem estratégica de publicidade e comunicação por parte daqueles interessados em expandir seu público-alvo.

As análises foram realizadas através da tarefa de clusterização particional (esférica) e hierárquica, que é uma tarefa de agrupamento. Além disso, foi usado *tag cloud*¹, gráficos e geolocalização como ferramentas auxiliares à análise. O presente trabalho foi realizado em três divisões: (i) coleta de dados da plataforma do Twitter, (ii) pré-processamento dos dados e (iii) pós-processamento. Como pode ser visto na Figura 5:

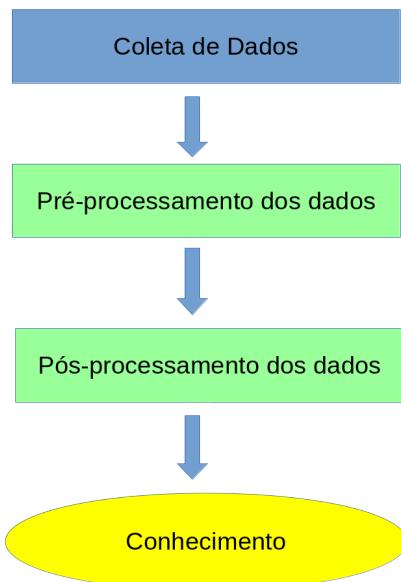


Figura 5 – Esquema do procedimento metodológico.

Nas próximas seções, cada etapa metodológica será descrita detalhadamente.

3.3 A coleta de dados e o pré-processamento

A coleta de dados, nesse trabalho, é a tarefa de recuperar e inserir na base de dados as seguradoras listadas como objetos de estudo, todos seus seguidores disponíveis juntamente com um máximo de 200 tuítes para cada seguidor. De modo a se obter o

¹ Tag Cloud é uma representação visual de dados textuais, onde cada tag é uma palavra de um documento e é exibida com tamanhos proporcionais à sua importância no documento ou na coleção de documentos

universo formado pelas entidades seguidos, seguidor e os tuítes desses. O algoritmo para a coleta de dados realizada está no Algoritmo 1.

Algoritmo 1 – Algoritmo para coleta de tuítes da plataforma do Twitter.

```

1  for Seguradora in Lista de Seguradoras:
2      insere Seguradora no banco de dados
3      recupera lista de seguidores de Seguradora # realiza chamada GET na
4          plataforma do Twitter
5      for seguidor in lista de seguidores:
6          insere seguidor no banco de dados
7          sleep(1) # força uma pausa de 1 segundo
8          recupera lista de tuítes de seguidor # realiza chamada GET na
9              plataforma do Twitter
10         for tuite in lista de tuítes:
11             insere tuite no banco de dados

```

A plataforma do Twitter possui limitações e restrições tanto à privacidade de seus usuários quanto ao tempo de chamadas a serem realizadas para a plataforma. Há usuários que tornam suas contas no Twitter privadas, portanto, não públicas e não acessíveis através de chamadas pela API². Assim, nem todos os seguidores das seguradoras foram considerados, uma vez que ou não possuíam quaisquer postagens ou tinham suas contas bloqueadas.

Outro fator relevante na tarefa de coleta foi a limitação no número de tuítes recuperados da plataforma para a base de dados, que no caso foi estabelecido em 200 tuítes. A escolha desse valor limite máximo seguiu o exposto na documentação³ para a função *GET statuses / user_timeline* da API, cujo parâmetro *count* é fixado num máximo de 200 tuítes por chamada.

Caso haja a necessidade de se recuperar uma quantia acima de 200 tuítes, torna-se necessário o uso de *GET statuses / user_timeline* em repetição; o uso da repetição dessa chamada faz surgir a problemática do tempo de acesso para requisições à API, que é limitado em 15 minutos. Se houver estouro nesse tempo de acesso, deve-se esperar 15 minutos, que é o tempo da API permitir novas chamadas. Portanto, para evitar complexidade temporal muito alta, devido ao grande número de seguidores por parte de algumas seguradoras e pelo número possivelmente alto de tuítes disponíveis de cada seguidor, resolveu-se limitar as chamadas de tuítes para cada seguidor em apenas uma.

Outra abordagem realizada nesse trabalho foi forçar uma espera de 1 segundo antes de recuperar a lista de tuítes de um seguidor, com o objetivo de não alcançar o limite de chamadas da API.

² API, Application-Program Interface é um comando de instruções e padrões de acesso que ditam como programas individuais devem acessar alguma plataforma

³ https://dev.twitter.com/rest/reference/get/statuses/user_timeline

A coleta dos dados se iniciou em 7 de março de 2017 às 10h10 e terminou em 10 de maio de 2017 às 4h30. Observa-se um total de 101.084 seguidores inseridos, 8.716 seguidores sem quaisquer tuítes e 12.939 seguidores protegidos (contas privadas).

A etapa (ii) vem no sentido de dar início à limpeza dos dados obtidos da plataforma virtual para que seja possível trabalhá-los posteriormente. Um tuíte é um dado textual, um documento e, como qualquer outro dado com conteúdo textual escrito em linguagem natural, há de haver lixos e/ou palavras que são irrelevantes para a análise de dados. O pré-processamento ocorre em quatro fases: (i) remoção de lixos, (ii) separação das palavras do documento, (iii) eliminação de artigos e/ou preposições e (iv) redução de uma palavra à sua raiz morfológica.

3.3.1 O tratamento dos documentos

A remoção dos lixos se deu de modo a substituir letras do alfabeto latino por seus "equivalentes" no alfabeto inglês, onde ç ⇒ c, (á|à|â|ã) ⇒ a, (é|è|ê) ⇒ e, (í|í|î|ĩ) ⇒ i, (ò|ó|ô|õ) ⇒ o, (ú|ù|û|ũ) ⇒ u e '‑' ⇒ ε, considerando ε como palavra vazia (nada).

Aplicou-se as seguintes expressões regulares em Python para realizar a remoção de marcações HTML, menções (*mentions*) a outros usuários da plataforma, *hashtags*, números, URLs e *smiles*, respectivamente.

```

HTML : ( <[^>]+> )
Mentions : ( '@[\w]+ ')
Hashtags : ( '\#+[\w_]+[\w\_\-\_]*[\w_]+ ')
Numeros : ( '\d+' )
URL : ( http[s]?://(?:[a-z] | [0-9] | [$_@.&+]) | [!*\\(\(),] | 
(?:%[0-9a-f][0-9a-f]) )+
Smiles : ( [:=;][oO\‐]?[D\)\]\(\]/\0pP] )

```

A expressão regular de remoção das marcações HTML consiste em reconhecer palavras que se iniciem com o símbolo <, seguido de qualquer símbolo exceto >, onde esse é o símbolo terminal da palavra. A remoção de menções a outros usuários é feita reconhecendo palavras que se iniciem com @ seguido de qualquer símbolo alfanumérico. A remoção de *hashtags* consiste em reconhecer palavras que se iniciem com # seguido de qualquer símbolo alfanumérico, ou _ ou -. A remoção de URLs consiste em reconhecer palavras que se iniciem por *http* ou *https*, seguido por ':', seguido por qualquer símbolo alfanumérico ou -, ou __, ou @, ou (, ou).

O passo seguinte consistia em rearranjar o documento de modo a dá-lo uma estrutura que tornasse possível sua manipulação posteriormente. A ideia dessa fase consiste

em converter o documento de uma cadeia de caracteres em um vetor de *tokens* (palavras), ou seja, $\tau_0\tau_1\tau_2\dots\tau_n \Rightarrow [\tau'_0, \tau'_1, \tau'_2, \dots, \tau'_n]$.

O dado estruturado obtido ainda possui falhas, para tanto, ainda é necessária a aplicação de outros filtros sobre o documento; aplicou-se, então, a remoção de *stop words*. A ferramenta em Python utilizada para aplicação dessa fase e da seguinte foi o Natural Language ToolKit, NLTK⁴. A lista de *stop words* em português consiste em artigos, preposições, pronomes, entre outros, por exemplo⁵: 'a', 'ao', 'aos', 'aquele', 'aqueles', 'aquele', 'aqueles', 'aquilo', 'as', 'até'.

A última fase é a redução de palavras para o seu radical morfológico, a *stemização*. A aplicação dessa fase consiste em valorizar o sentido léxico das palavras em detrimento do semântico. Por exemplo, em "você é um marinheiro", "cruzeiro marítimo" e "nós somos marujos", há o mesmo radical, "mar", e há o mesmo sentido léxico, mas em "a nota dela é alta" e "ela nota aquilo" há o mesmo radical "nota" mas não há o mesmo sentido semântico.

Essa abordagem permite a construção de uma *bag of words* (BOW), a qual consiste num vetor de tamanho n , onde n é a quantidade de termos sem repetição do documento *stemizado*. Essa BOW indexa cada termo no qual as entradas desses termos são o número de vezes que cada termo apareceu no documento antes de ser *stemizado*.

A etapa de pré-processamento pode ser resumida no seguinte Algoritmo 2.

Algoritmo 2 – Algoritmo de pré-processamento

```

1 for tuite in lista de tuítes:
2     aplica expressões regulares ao tuíte
3     tokeniza o tuíte
4     remove stop words do tuíte
5     for palavra in tuíte tokenizado, sem stop word e sem lixo:
6         bag of words [] <- stemiza palavra

```

3.3.2 A pesagem dos termos

A pesagem dos termos de um documento pré-processado numa BOW é o processo de achar alguma métrica que possa dizer a frequência com que cada termo aparece no documento. As métricas mais populares são as variações da *term frequency inverse document frequency*, tf_idf , onde:

$$tf_idf = tf * idf \quad (1)$$

tf = número de vezes que o termo aparece no documento

⁴ <http://www.nltk.org>

⁵ Retirado de http://www.nltk.org/howto/portuguese_en.html

$$tf\text{normalizado} = \frac{\text{número de vezes que o termo aparece no documento}}{\text{número de palavras no documento}}$$

$$idf = \frac{1}{\log \frac{n}{df}}, n \text{ é o número de documentos}$$

$$df = \text{número de documentos que contêm o termo}$$

Entretanto, neste trabalho, a métrica escolhida para o cálculo da frequência foi a do tf , frequência pura do termo, pois essa abordagem permite uma verificação mais clara sobre a frequência de palavras numa *bag of words*, pois é a simples contagem de aparições que uma palavra teve no conjunto de documentos. De modo que medidas, como a quantidade de documentos que há num conjunto e a quantidade de palavras de um único documento, não influenciem na contagem de aparições da palavra.

3.4 A tarefa de pós-processamento

Ao término da tarefa de pré-processamento, ainda há a necessidade de reconhecer quais padrões existem nos dados fornecidos. Para tal, é necessário a aplicação da tarefa de pós-processamento, que tem por finalidade analisar os dados tratados e identificar os padrões presentes neles, de modo a permitir a extração de informação.

3.4.1 O agrupamento dos documentos pré-processados

O agrupamento consiste em reunir sob um mesmo conjunto os dados que possuem algum grau de semelhança entre si. Esse grau de semelhança é dito similaridade. Para dados dentro de um conjunto essa similaridade é aumentada, em detrimento da similaridade entre esses dados e aqueles pertencentes a outros conjuntos. A finalidade de um algoritmo de agrupamento é realizar divisões no conjunto de dados fornecidos e criar subgrupos. Não é finalidade de um agrupador classificar ou aplicar rótulos a tais subgrupos. Um subgrupo é chamado pela literatura de *cluster* (LAROSE, 2005). Essa etapa será referida adiante como "etapa da clusterização", uma vez que *cluster* é equivalente a "grupo", ou "conjunto", ou "subgrupo".

A clusterização é uma etapa importante e frequentemente aplicada no processo de Mineração de Dados. A partir dos *clusters* formados é possível, então, fornecê-los como dados de entrada para processos seguintes, de modo a realizar buscas por padrões dentro de cada grupo, segundo Larose (2005).

Um dos principais objetivos da Mineração de Dados é buscar por padrões nos dados fornecidos. Sendo essa uma subárea do Aprendizado de Máquina⁶, a busca por padrões pode ser executada de duas formas: (i) supervisionado e (ii) não-supervisionado.

⁶ Aprendizado de Máquina (*Machine Learning*) consiste no estudo de algoritmos que aprendem através de erro e experiência, como realizar determinada tarefa ao invés de buscar soluções através de regras predeterminadas.

O (i) aprendizado supervisionado consiste em fazer com que um algoritmo busque por padrões semelhantes àqueles aos quais foram usados em seu treinamento de reconhecedor de padrões. Portanto, é preciso fornecer (treinar) um algoritmo para que esse seja capaz de buscar e identificar certos padrões. Desse modo, um algoritmo treinado para buscar um padrão X, só e somente, buscará pelo padrão X em todos os dados que forem aplicados a ele.

Análogo à forma (i), o (ii) aprendizado não-supervisionado visa reconhecer padrões em dados fornecidos. A diferença surge no treinamento do algoritmo, o qual buscará pelos padrões, pois um algoritmo que consiste em aprendizado não-supervisionado é treinado a reconhecer por si só os padrões nos dados de entrada. De modo que nenhum rótulo é atribuído aos dados fornecidos para esse algoritmo, uma vez que esse identificará intrinsecamente os devidos rótulos para os dados (LAROSE, 2005).

A clusterização, dentre outros métodos, por exemplo, pode ser realizada através de (i) clusterização por *k-means*, (ii) clusterização hierárquica, e por (iii) clusterização por *k-means* esférico. Cada método aplica algoritmos diferentes e métricas para a similaridade distintas.

Segundo Larose (2005), a clusterização é um método não-supervisionado, pois ao agrupar os dados em *clusters* por meio de uma similaridade, já é equivalente a identificar um certo padrão nos dados e isso é realizado pelo algoritmo treinamento prévio. A clusterização, seja ela *k-means* ou hierárquica, precisa responder às seguintes questões: (a) Como medir a similaridade? (b) Como padronizar ou normalizar valores? e (c) Quantos *clusters* os dados de entrada formarão? A questão da similaridade pode ser respondida pela métrica da distância. Uma vez que cada dado pode ser considerado um vetor num plano cartesiano. Portanto, a distância pode ser calculada de diversas formas, como a distância Euclidiana, a *cityblock*, ou a *Minkowski*.

Ainda segundo o autor, a métrica de distância mais usual é a Euclidiana, que é dada por:

$$d_{euclidiana} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}, \quad (2)$$

onde i é uma dimensão do dado, x e y são pares de dados, e i vai até m , que é a dimensão (número de atributos) dos dados.

A normalização dos dados ocorre para que nenhum *cluster* ou variável tenha valores que predominem sobre os outros na análise (LAROSE, 2005). A normalização pode ocorrer pela Normalização Máximo-Mínimo, em (3) por exemplo:

$$X' = \frac{X - \min(X)}{\text{Tamanho}(X)} \quad (3)$$

O algoritmo *k*-means é expresso no Algoritmo 3, segundo Larose (2005).

Algoritmo 3 – Algoritmo do *k*-means.

```

1 k = numero de clusters desejado pelo usuario
2 Escolher aleatoriamente k elementos para serem os centros dos k clusters
3 while convergencia = 0:
4   for elemento in dataset:
5     Encontre o centro do cluster mais proximo
6     Ck = elemento
7   for cluster in Ck:
8     Encontre o centroide do cluster
9     Atualize o centro do cluster pelo centroide
10    if Ck nao mudaram de centroide:
11      convergencia = 1

```

No algoritmo acima, se os *clusters* não mais mudam de centróide, então a convergência foi atingida e os centróides corretos foram encontrados. O cálculo do centróide é o cálculo de centro de massa dos elementos. Suponha uma série N de dados em R^2 , onde $[(x_1, y_1), \dots, (x_N, y_N)]$, os centróides são calculados como na Eq. (4):

$$c_{xy} = \left(\sum_{i=1}^N \frac{x_i}{N}, \sum_{i=1}^N \frac{y_i}{N} \right) \quad (4)$$

A clusterização hierárquica gera um dendrograma⁷ que indicará os *clusters* nos quais os dados foram colocados, de modo que é possível verificar o ponto onde um *cluster* vira sub-*cluster* de outro. Essa estrutura (o dendrograma) pode ser construída através de métodos divisivos ou aglomerativos (LAROSE, 2005).

O método aglomerativo inicia o processamento colocando cada elemento num *cluster* de si mesmo, ou seja, cada elemento já é considerado um *cluster* em si próprio. Já pelo método divisivo, todos os elementos são inicializados num primeiro grande *cluster*, que é considerado aquele que conterá a maior dissimilaridade entre os elementos e, então, os elementos são reagrupados recursivamente através das diferenças entre as dissimilaridades (LAROSE, 2005).

Novamente segundo Larose (2005), o método hierárquico mais comumente aplicado é o aglomerativo. Nesse tipo de clusterização é preciso calcular, para além das distâncias entre os elementos e os centróides dos *clusters*, as distâncias entre os *clusters*. Existem alguns critérios (*Single Linkage*, *Complete Linkage* e *Average Linkage*) que podem ser levados em conta quando se pretende determinar a distância entre dois *clusters* arbitrários.

O *Single Linkage* se baseia na menor distância entre todos os elementos no *cluster* C_0 e no *cluster* C_1 . Ou seja, considera a similaridade entre os elementos de C_0 e C_1 , o que

⁷ Um dendrograma é uma representação diagramática ramificada que mostra a relação grupal entre elementos (<https://www.dicio.com.br/dendrograma/>)

pode levar ao agrupamento de elementos não similares. O *Complete Linkage* se baseia na maior distância entre os elementos de C_0 e C_1 , ou seja, ao contrário do *Single Linkage*, esse considera a dissimilaridade entre os elementos de dois *clusters*. Por fim, há o *Average Linkage* que se baseia na distância média entre os elementos em C_0 de C_1 . Essa abordagem evita os valores máximos e mínimos, de modo a diminuir a interferência de valores extremos em cada *cluster*.

Nesse trabalho aplicou-se a clusterização hierárquica com o *average linkage*. Já para clusterização particional, aplicou-se uma variante do k -means, que é o k -means esférico. A razão para se aplicar ambas formas de clusterização foi primariamente didática, mas a visualização dos *clusters* gerados, por si só, também contribuiu na obtenção de informação.

3.4.2 A clusterização esférica

Clusterização esférica é uma variação do k -means, com a diferença de que a métrica da distância é calculada pela chamada distância do cosseno (5), onde \vec{u} e \vec{v} são pares de vetores. Dado um conjunto de dados, no caso desse trabalho, documentos (tuítes), há a geração de uma *bag of words* na etapa de pré-processamento. Essa estrutura permite considerar os documentos como vetores num espaço vetorial (HORNIK et al., 2012).

$$d_{\text{cosseno}}(\vec{u}, \vec{v}) = 1 - \cos(\vec{u}, \vec{v}) \quad (5)$$

Seja X uma matriz $M \times N$, onde M é a quantidade de documentos e N é a quantidade de termos. Cada documento é um vetor num espaço R^N , então para cada par de vetores (\vec{u}, \vec{v}) , a medida do ângulo entre esses vetores pode representar a similaridade entre os documentos.

A eq. (6) determina o ângulo entre os vetores \vec{u} e \vec{v} . Quanto menor o ângulo θ resultante, maior a semelhança entre os vetores e mais próximo de zero estará a dissimilaridade, calculada em (5).

$$\cos(\theta) = \cos(\vec{u}, \vec{v}) = \frac{\langle \vec{u} \cdot \vec{v} \rangle}{\|\vec{u}\| \cdot \|\vec{v}\|} \quad (6)$$

A aplicação da eq. (5) para todo par de vetores em X gera um Z , que é a matriz de dissimilaridades; de dimensão $M \times M$, onde:

$$Z = \begin{bmatrix} d_{\vec{u}_1 \vec{v}_1} & d_{\vec{u}_1 \vec{v}_2} & d_{\vec{u}_1 \vec{v}_3} & \dots & d_{\vec{u}_1 \vec{v}_M} \\ d_{\vec{u}_2 \vec{v}_1} & d_{\vec{u}_2 \vec{v}_2} & d_{\vec{u}_2 \vec{v}_3} & \dots & d_{\vec{u}_2 \vec{v}_M} \\ \dots & \dots & \dots & \dots & \dots \\ d_{\vec{u}_M \vec{v}_1} & d_{\vec{u}_M \vec{v}_2} & d_{\vec{u}_M \vec{v}_3} & \dots & d_{\vec{u}_M \vec{v}_M} \end{bmatrix}$$

, sendo $d_{\vec{u}, \vec{v}}$ a dissimilaridade entre os vetores \vec{u} e \vec{v} . O cálculo da dissimilaridade sempre respeitará a eq. (7):

$$d_{\vec{u}_i, \vec{v}_j} = \begin{cases} 0 & i = j \\ d_{\text{cosseno}}(\vec{u}_i, \vec{v}_j) & i \neq j \end{cases} \quad (7)$$

Um algoritmo para o cálculo das dissimilaridades entre os vetores é apresentado no Algoritmo 4.

Algoritmo 4 – Algoritmo para o cálculo da dissimilaridade.

```

1 for documento_I in bag_of_words:
2     for documento_J in bag_of_words:
3         d_cosseno(documento_I, documento_J)

```

Por fim, é realizada a redução de dimensionalidade em Z , de modo a convertê-la de uma representação de vetores em R^N , tal que N é o número de termos na *bag of words* X , para uma representação de vetores em R^2 , de modo que seja possível visualizar os dados num plano cartesiano $O(x, y)$. Isso foi implementado por meio da função MDS (*Multi-dimensional Scaling*) do pacote Manifold, da biblioteca SKLearn. Essa função trabalha de modo a agrupar os elementos multi-dimensionais através de suas similaridades ou dissimilaridades, a fim de representar essas diferenças ou similaridades num valor de distância dentro do espaço geométrico.

3.4.3 O problema da complexidade espacial e temporal no estudo de caso

A aplicação de um objeto de estudo para esse trabalho permite visualizar os conceitos teóricos anteriormente mencionados. O estudo de caso permite materializar a teoria e tornar mais tangível todos os conceitos até então abstratos. Entretanto, com a mudança de um paradigma teórico para um prático, há o surgimento de problemas até então ignorados pelas questões teóricas.

Esse trabalho utiliza como estudo de caso 11 seguradoras. São considerados também todos os seguidores dessas seguradoras com contas públicas e com algum tuíte publicado na plataforma do Twitter. São considerados até 200 tuítes por seguidor.

De acordo com os algoritmos apresentados anteriormente podemos verificar a complexidade de cada um. O Algoritmo 1 possui uma complexidade de $O(N^2)$, ao se considerar listas de seguradoras e de seguidores dessas seguradoras de tamanhos anormalmente grandes, mas como esse trabalho se propõe a estudar somente 11 contas de seguradoras, a complexidade desse algoritmo cai para $O(N)$, onde N é o número de seguidores dessas 11 contas de seguradoras e que pode ter um valor muitíssimo alto.

O Algoritmo 2 por sua vez, tem complexidade de $O(N^2)$, pois o número de tuítes e de *tokens* nesses tuítes podem ter valores igualmente altos. Assim como o Algoritmo 4), mas nesse caso N representa a quantidade de tuítes somente, mas como esse calcula

a dissimilaridade entre documentos, por isso $O(N^2)$. Como visto, fica evidente que esse trabalho está sob uma complexidade $O(N^2)$, o que não representa algo bom. Ainda mais quando se trata de *Big Data*.

Apesar disso, o algoritmo mais completo, que melhor representa o que foi implementado nesse trabalho, é o 5.

Algoritmo 5 – Algoritmo para geração da matriz X.

```

1 for seguradora in Seguradoras:
2     for seguidor in seguidores da seguradora:
3         for tuite in seguidor:
4             for termo in tuite:
5                 Realize o preprocessamento e gere X

```

Isso leva à seguinte equação:

$$\Omega = \zeta * \varphi * \tau \quad (8)$$

, onde ζ é a quantidade de contas de seguradoras inseridas na base, φ é o total de seguidores das contas de seguradoras inseridas na base e τ é o total de tuítes inseridos na base. Ω representa o problema desse trabalho, uma vez que é através dele que a complexidade pode ser descrita.

Através da etapa de pré-processamento, com a criação da matriz de documento-termo X , que é $M \times N$, onde M é o número de documentos na *bag of words* e N é o número de termos de todos os documentos fornecidos nessa *bag of words*, há a necessidade de se adicionar mais uma variável à Ω , que será λ , onde essa representa a quantidade de termos numa dada *bag of words*. O problema Ω então se torna a equação (9)

$$\Omega = \zeta * \varphi * \tau * \lambda \quad (9)$$

Contrariamente às outras variáveis em Ω , não é possível saber com exatidão o valor de λ , embora já se tenha que $\zeta_{max} = 11$, $\varphi_{max} = 101.084$ e $\tau_{max} = 10.544.880$. Então temos que Ω será um problema em função de λ . A equação (10) expressa o cálculo de Ω .

$$\Omega(\lambda) = \zeta_{max} * \varphi_{max} * \tau_{max} * \lambda = 11 * 101.084 * 10.544.880 * \lambda \approx 10^{13} * \lambda \quad (10)$$

Iterar entre elementos na ordem de 10^{13} é uma tarefa custosa. O pior caso espacial acontece quando há a criação da matriz X , pois as linhas serão τ e as colunas λ , o que gera uma matriz anormalmente grande e de dimensionalidades dependentes da quantidade de termos nos documentos.

Ainda no pior caso, e considerando o Algoritmo 5, para λ é difícil estimar um valor, mas considerando uma média de 10 palavras por tuíte, $\lambda_{medio} = 10$, e 200 tuítes por seguidor ($\tau = 200$) e uma média de aproximadamente 10.100 seguidores por seguradoras da base de dados ($\varphi_{medio} = \frac{\ell}{\zeta} \approx 9.189$), teríamos a geração de uma matriz X , MxN , onde $M * N = (\tau * \varphi_{medio}) * (\tau * \varphi_{medio} * \lambda_{medio}) = \tau^2 * \varphi_{medio}^2 * \lambda_{medio} = (200)^2 * (9.189)^2 * 10$. Assim, X será uma matriz com aproximadamente $3 * 10^{13}$ elementos e Z será uma matriz de $\tau^2 * \varphi_{medio}^2 = (200)^2 * (9.189)^2 \approx 3 * 10^{12}$ elementos.

Como X é a matriz que armazena a relação de frequência entre termo e documento, os elementos das intersecções entre termo e documento é um número absoluto. Portanto, na situação exposta acima, X teria números inteiros na ordem de 10^{13} . Sendo um número inteiro equivalente a quatro bytes, X seria uma estrutura de dados de $4 * 10^{13}$ bytes.

A questão da dimensionalidade é comum na área de Mineração de Dados, que é uma sub-área do *Big Data*. Realizar processamentos grandiosos em máquinas usuais é demasiadamente custoso, tanto temporal quanto espacialmente. Para tanto, a solução proposta foi de encontro a buscar a redução de dimensionalidade, onde se considerou $\varphi = [1, 50]$. Ou seja, limitou-se as análises aos 50 primeiros seguidores de cada seguradora.

Com a proposta acima, temos $X = (\tau^2 * \varphi_{max}^2) * \lambda_{medio} = (200)^2 * (50)^2 * 10 = 10^9$, ou seja, X possui elementos na ordem de 10^9 . Nessa perspectiva, Z , teria $\tau^2 * \varphi_{max}^2 = (200)^2 * (50)^2 \approx 10^8$.

No pior caso, há uma redução próxima à metade nas ordens de grandeza. Logo, a proposta de se limitar o número de seguidores processados se mostra eficiente, num cenário onde os processamentos ocorreram em máquinas aquém de serem super computadores.

4 Resultados

Esse capítulo tem por objetivo expor os resultados obtidos através das etapas e processos expostos em capítulos anteriores.

4.1 Análise para geração dos resultados

Usou-se como objeto de estudos as contas de perfil no Twitter das seguintes seguradoras listadas: Porto Seguro; Bradesco Seguros; Itaú Corretora; SulAmérica Seguros; Caixa Seguradora; Seguros Unimed; Allianz Brasil; Liberty Seguros; HDI Seguros Brasil; Mapfre Brasil e Tokio Marine Seguros.

Cada seguradora listada, após a busca e inserção na base de dados dos seus respectivos seguidores, possui uma importância única no banco de dados, o que acaba tornando uma mais importante que outras. A métrica dessa importância se dá pela fração na base representada pela respectiva seguradora, que é exibida na Tabela 2. A Tabela 3 exibe a porcentagem daqueles seguidores que seguem mais de uma seguradora ao mesmo tempo.

Tabela 2 – Porcentagem de distribuição dos seguidores cadastrados

Perfil analisado	Número de seguidores	Porcentagem da base de dados
CaixaSeguradora	5.507	5,60%
tokiomarine_cor	41.250	41,8%
allianz_brasil	6.955	7,06%
portoseguro	22.727	23,07%
MAPFRE_BR	2.932	2,98%
HDISegurosBr	9.769	9,92%
itaucorretora	16.311	16,57%
bradescoseguros	10.649	10,81%
libertyseguros	5.654	5,74%
Sulamerica	10.260	10,41%
segurosunimed	1.724	1,75%

Da Tabela 1, Capítulo 2, temos que as maiores seguradoras da base são *tokiomarine_cor*, *portoseguro* e *itaucorretora*. Todas juntas representam cerca de 81,3% dos seguidores inseridos. É interessante visualizar como se compartam os seguidores dessas três contas. Para isso, a Figura 6 ilustra a configuração desses três conjuntos.

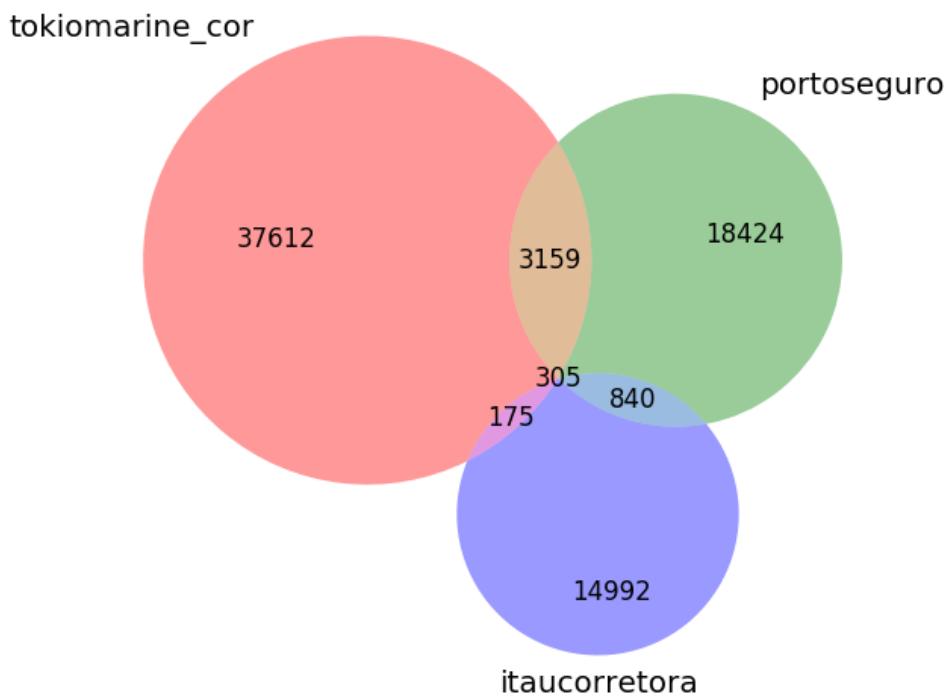


Figura 6 – Diagrama de Venn para as três maiores contas da base.

Disso é possível entender que embora essas três sejam aquelas com maior número total de seguidores, essas não são aquelas que mais compartilham seguidores na base de dados, vide Tabela 3. Por exemplo, as contas *bradescoseguros* e *portoseguro* compartilham cerca de aproximadamente 5,10% dos seguidores inseridos (que é a maior porcentagem de seguidores compartilhados). Esse valor é maior do que a porcentagem dos seguidores que simultaneamente seguem *portoseguro* e *itaucorretora*, aproximadamente 1,16%, ou *portoseguro* e *tokioarine_cor*, aproximadamente 3,5%. Isso pode significar uma maior proximidade entre o conteúdo digital de *portoseguro* e *bradescoseguros*, ou mesmo atividades num mesmo ramo no segmento de mercado segurador.

O processo de análise dos agrupamentos dos tuítes dos seguidores das contas objeto de estudo consistiu na geração de uma *tag cloud* e de alguns gráficos e imagens. Inicialmente, foi realizada a redução de dimensão da matriz X (gerada pela etapa de pré-processamento), que ainda poderia conter termos que são irrelevantes para a análise: por exemplo, palavras com poucas aparições, que não são relevantes dentro de um conjunto com mais palavras com mais aparições. Para tanto, foram calculadas as distribuições dos quartis que ajudaram a encontrar um ponto de corte que pudesse reduzir os dados corretamente, de acordo com as singularidades de cada conjunto de documentos de uma seguradora.

Tabela 3 – Porcentagem dos seguidores compartilhados entre seguradoras

	CaixaSeguradora	tokiomarine_cor	allianz_brasil	portoseguro	MAPFRE_BR	HDISegurosBr	itaucorretora	bradescoseguros	libertyseguros	Sulanamerica	segurosunimed
CaixaSeguradora	100%	0,7270%	0,4326%	1,3576%	0,3209%	0,8377%	0,3635%	1,1748%	0,7311%	1,2936%	0,1970%
tokiomarine_cor	0,7270%	100%	1,1149%	3,5174%	0,8722%	4,9694%	2,7904%	3,1265%	3,3478%	2,2803%	0,2508%
allianz_brasil	0,4326%	1,1149%	100%	1,4947%	0,6874%	1,2774%	1,1484%	1,0895%	1,2662%	5,9219%	0,4722%
portoseguro	1,3576%	3,5174%	1,4947%	100%	1,0672%	4,7430%	1,1626%	5,1035%	3,2727%	0,9575%	0,2041%
MAPFRE_BR	0,3209%	0,8722%	0,6874%	1,0672%	100%	1,0002%	0,1615%	0,8966%	0,8042%	0,3584%	0,3584%
HDISegurosBr	0,8377%	4,9694%	1,2774%	4,7430%	1,0002%	100%	0,4793%	3,6332%	3,7337%	0,3534%	0,6194%
itaucorretora	0,3635%	0,4874%	0,2457%	1,1626%	0,1615%	0,4793%	100%	1,0357%	0,0822%	2,8046%	4,7044%
bradescoseguros	1,1748%	2,7904%	1,1484%	5,1035%	0,8966%	3,6332%	1,0357%	100%	0,4001%	3,2544%	0,2843%
libertyseguros	0,7311%	3,1265%	1,0895%	3,2727%	0,8042%	3,7337%	0,3534%	2,8046%	100%	3,2544%	0,4417%
Sulanamerica	1,2936%	3,3478%	1,2662%	5,9219%	0,9575%	0,6194%	4,7044%	0,4001%	0,2843%	0,4417%	100%
segurosunimed	0,1970%	0,2803%	0,2508%	0,4722%	0,2041%	0,3584%	0,0822%	0,4001%			

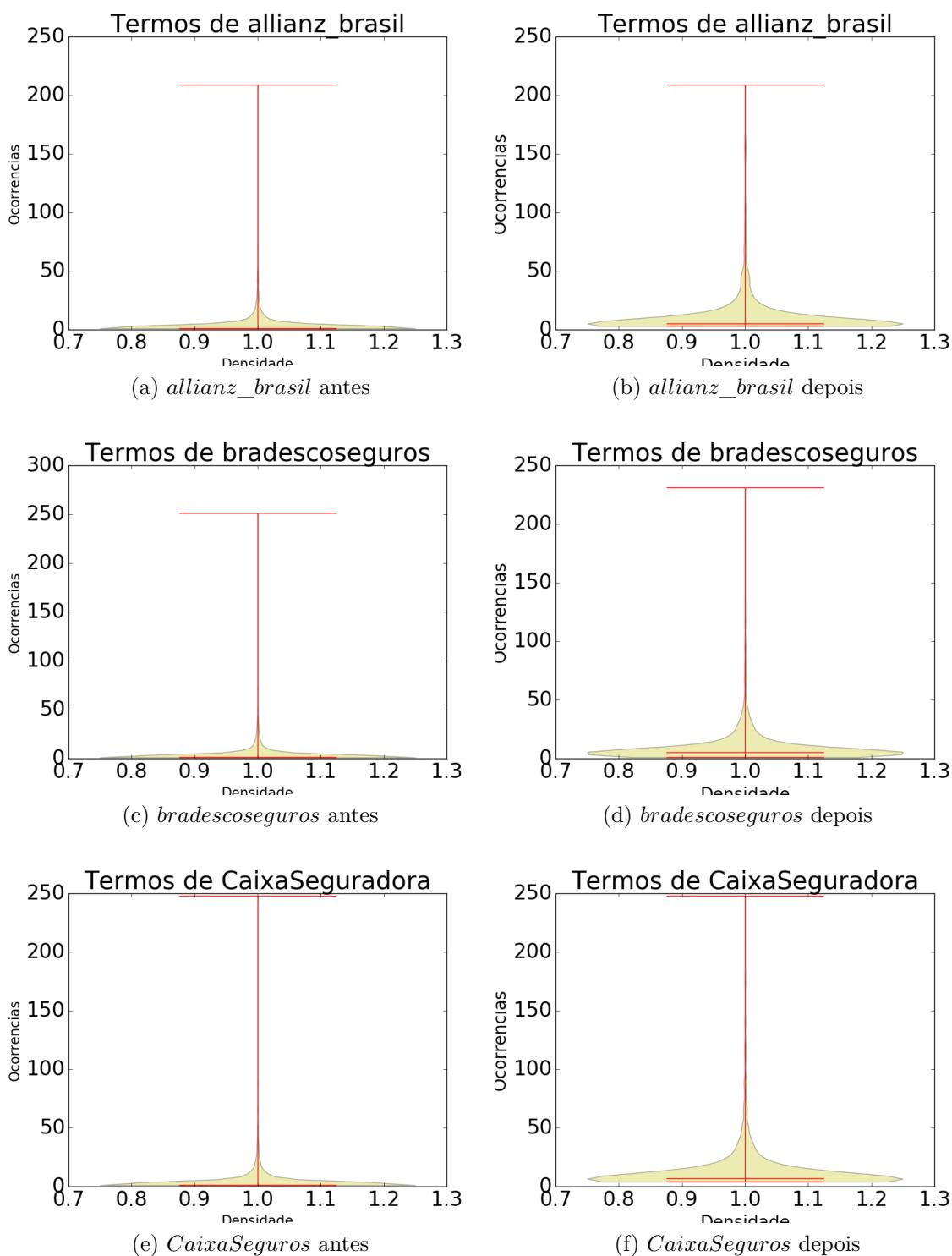


Figura 7 – Figura comparativa de dimensionalidade para cada conta.

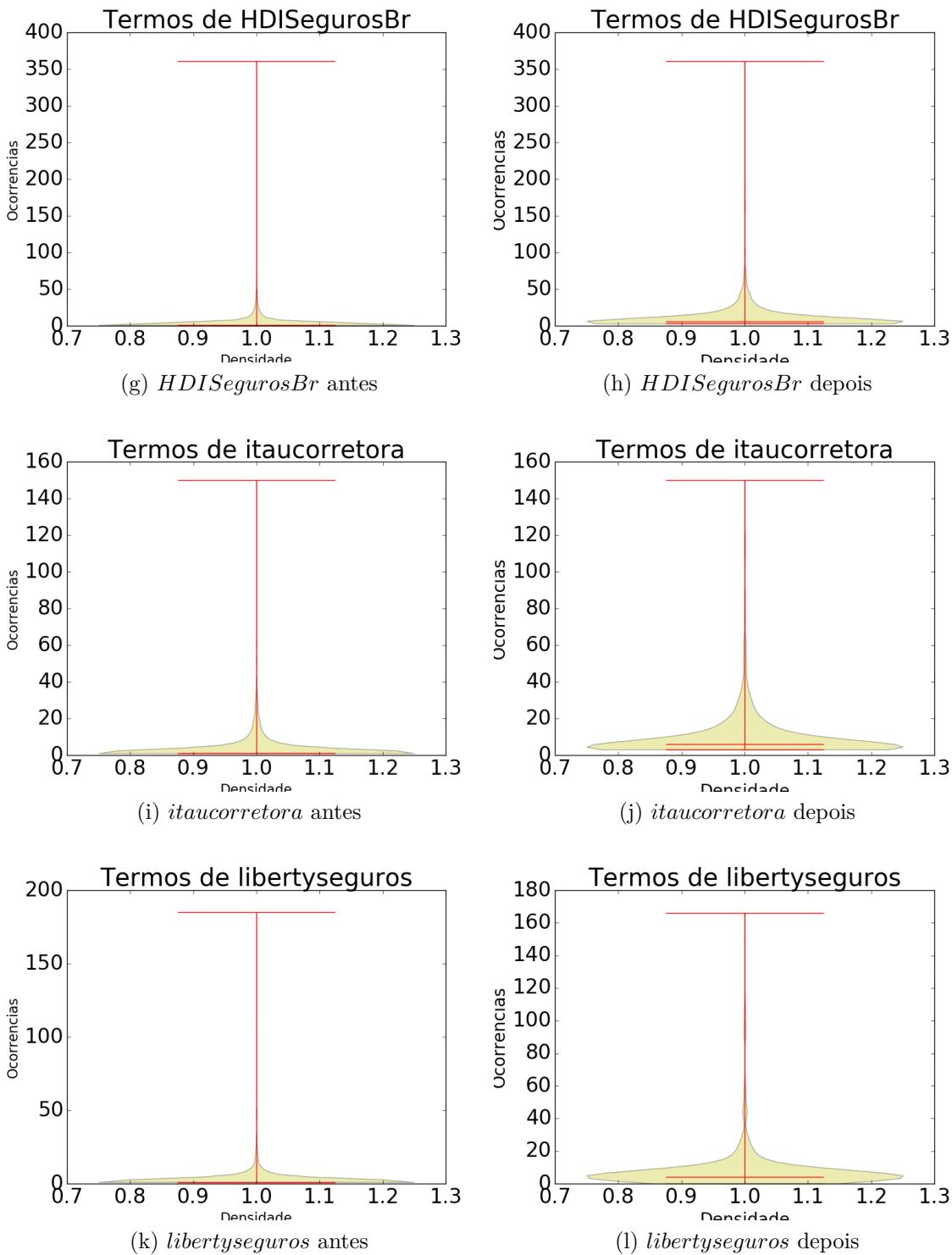


Figura 7 – Figura comparativa de dimensionalidade para cada conta - *continuação*.

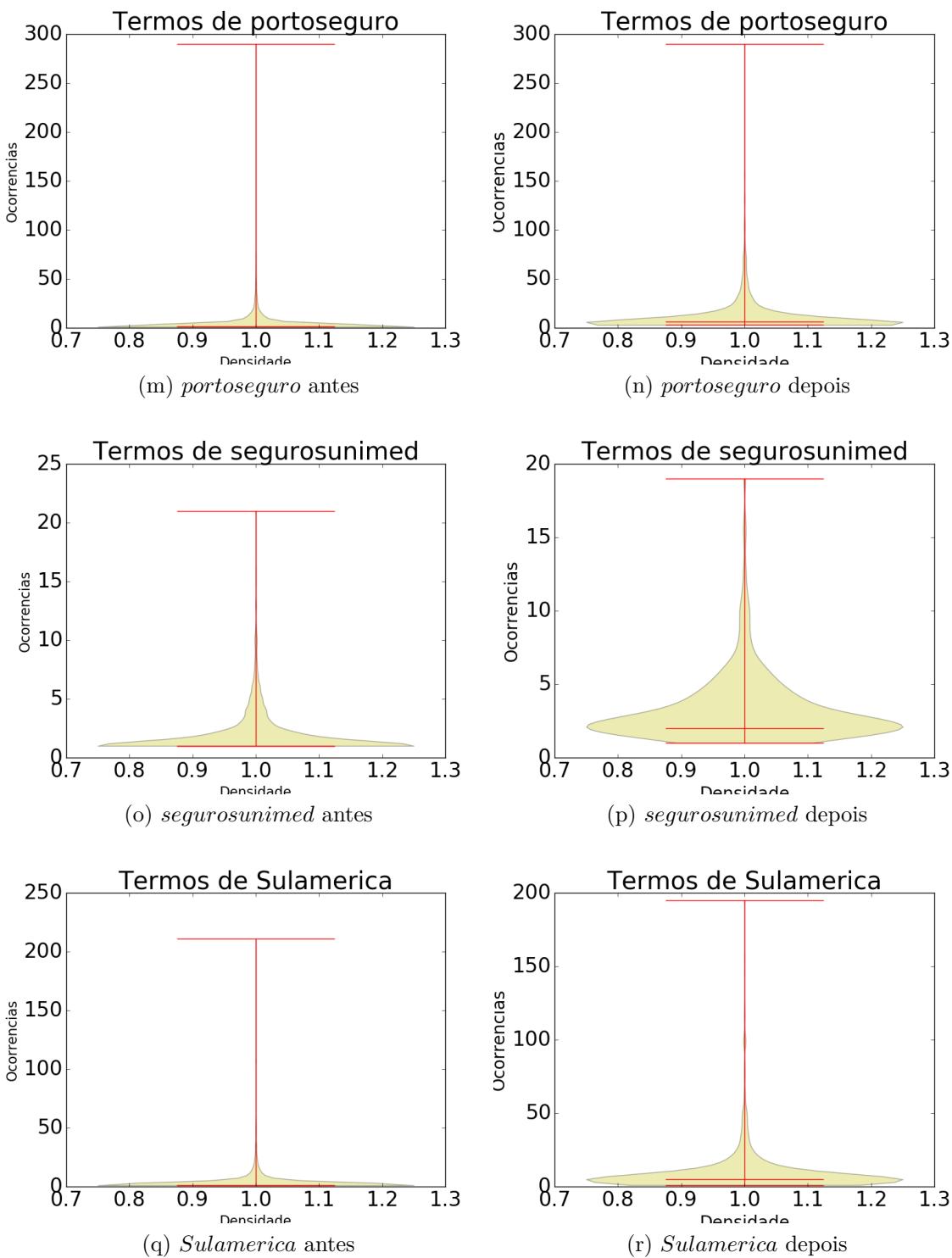


Figura 7 – Figura comparativa de dimensionalidade para cada conta - *continuação*.

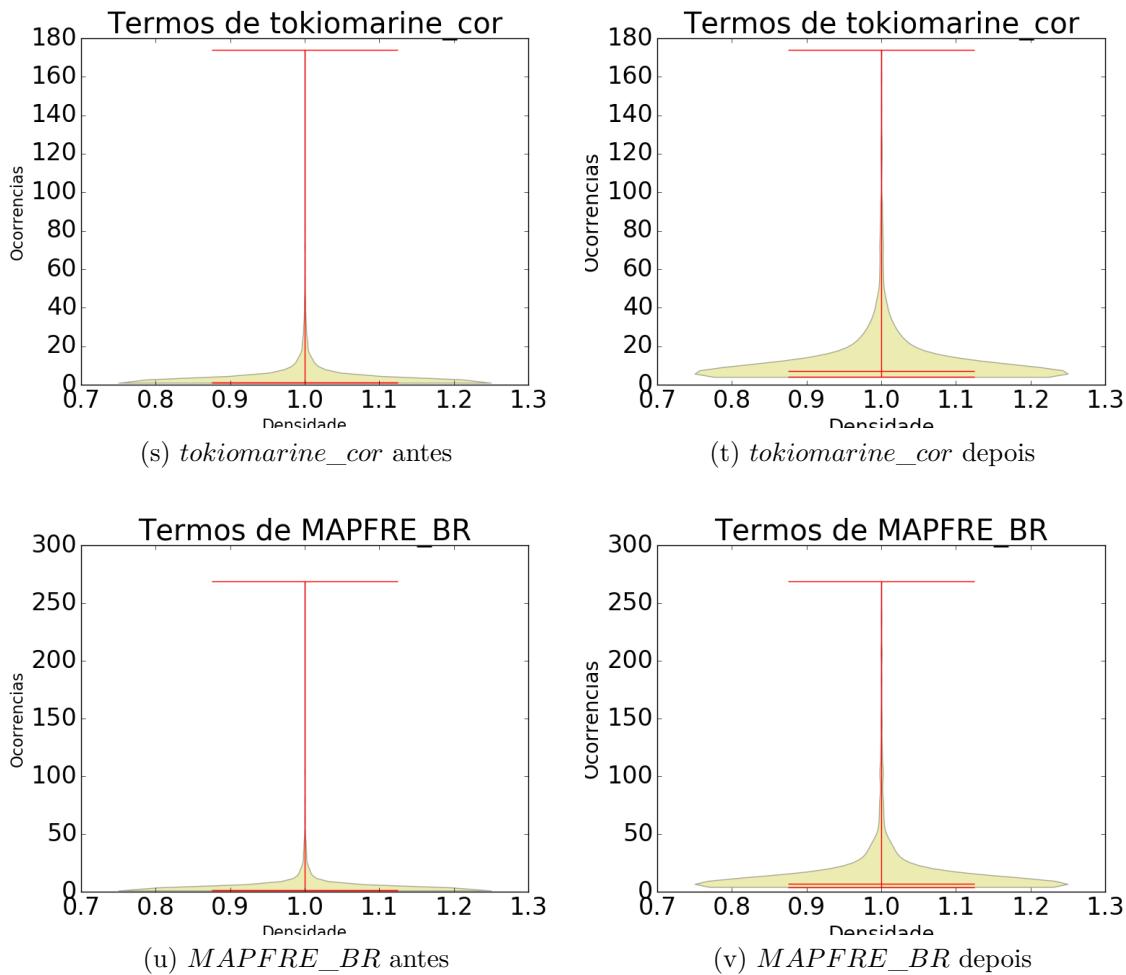


Figura 7 – Figura comparativa de dimensionalidade para cada conta - *continuação*.

Os gráficos em *violinplots* na Figura 7 ilustram como se dá a distribuição de documentos antes e depois da aplicação da redução de dimensionalidade. A primeira coluna representa o conjunto original, com todos os tuítes; a segunda coluna representa o conjunto recortado do primeiro, mas só com documentos que são relevantes para análise, de acordo com o discutido anteriormente.

Os gráficos dispostos na primeira coluna da Figura 7 auxiliam na visualização das distribuições dos termos usados pelos seguidores de uma conta de seguradora; o uso desse tipo de gráfico é mais elucidativo do que o de um *boxplot*, pois a densidade de elementos numa determinada faixa é explícita. Nos *violinplots*, a curva, no eixo das abscissas, representa a quantidade de termos e, no eixo das ordenadas, a quantidade de aparições desse termo na *bag of words*. Os novos conjuntos de termos, ou seja, os recortes dos conjuntos originais possuem a distribuição representada pelos *boxplot* da Figura 8.

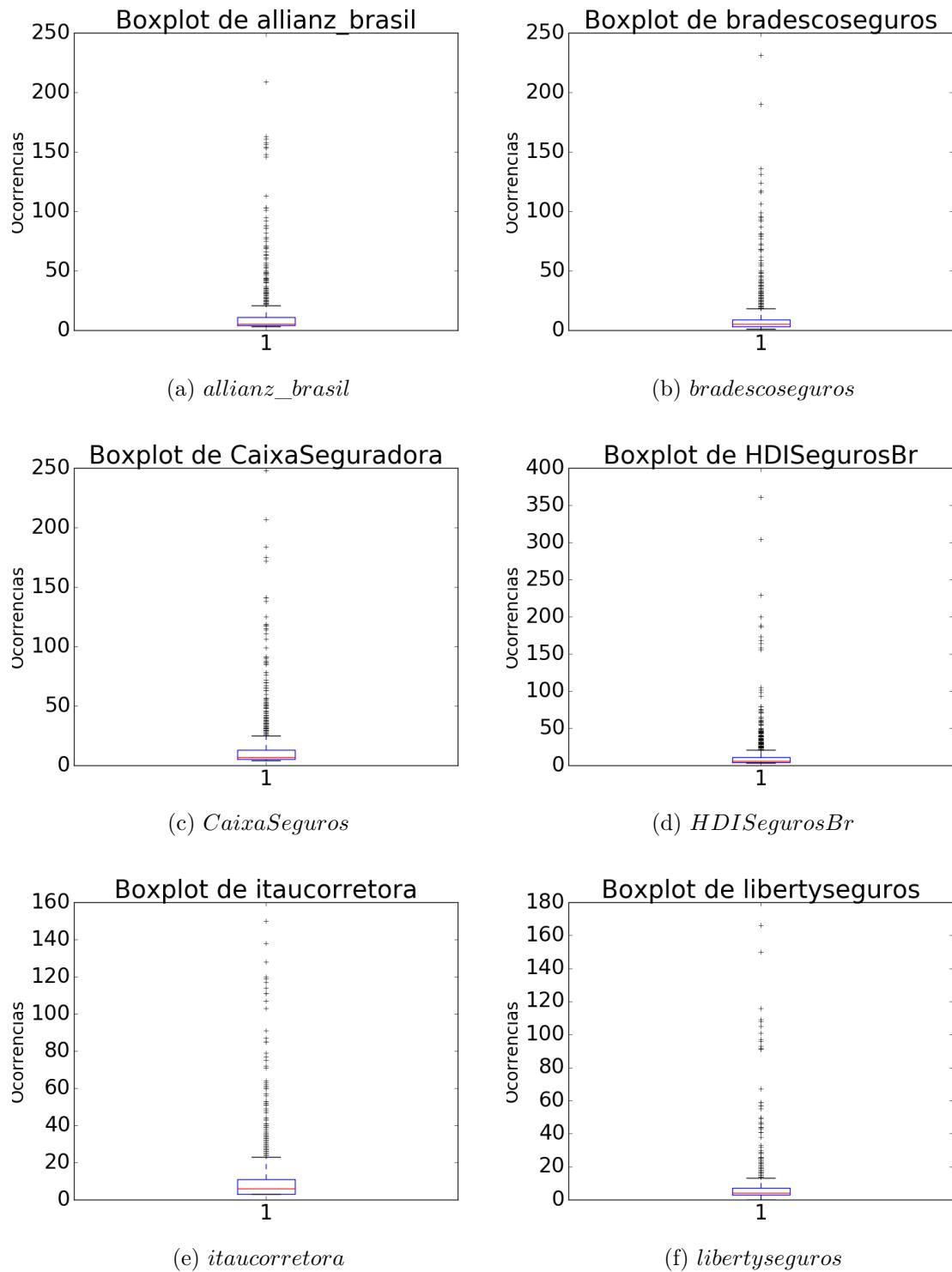


Figura 8 – Figura com os *boxplots* para cada conta após a redução de dimensão.

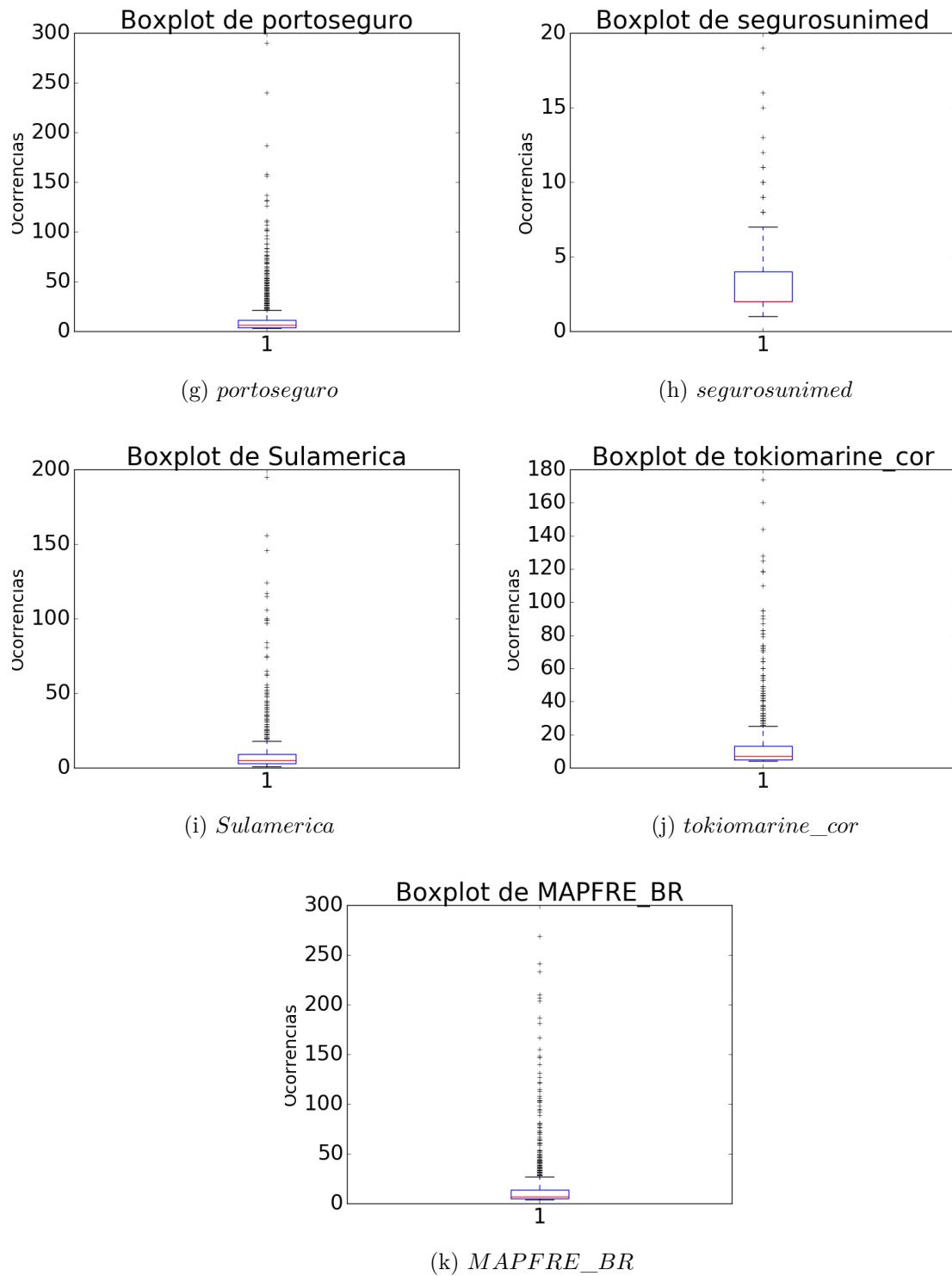


Figura 8 – Figura com os *boxplots* para cada conta após a redução de dimensão - *continuação*

Os *outliers* representados na Figura 8 não são um problema para a análise, pois uma vez que as *stop words* foram removidas, assim como os lixos (HTML, URLs, menções a usuários, etc...), não há (ou não deveria haver) palavras irrelevantes restantes. Para tanto, são os *outliers* que acabam sendo efetivamente importantes para essa análise. Nos

boxplots, o comprimento da caixa indica o mesmo que as curvas no *violinplots*, a densidade de termos naquela faixa.

Essa abordagem realiza o corte dos termos abaixo do terceiro quartil do *boxplot* do conjunto original, pois é a faixa que contém a maior quantidade de documentos e, portanto, a faixa que melhor permite reduzir a dimensionalidade da matriz original X . A Figura 9 ilustra a frequência dos termos recortados e que são relevantes para o pós-processamento.

É possível visualizar nos *boxplots* da Figura 8 que, após a redução de dimensionalidade, a maioria dos termos ainda possui poucas aparições. Ou seja, mesmo recortando da amostra original 25% dos termos mais frequentes, ainda resta nesse quartil a ocorrência de termos que são pouco usados, mas que ainda assim são mais utilizados do que 75% da amostra original. Ilustrando que, essas contas de seguradoras analisadas possuem grande heterogeneidade léxica nos tópicos utilizados pelos seus seguidores. Ou seja, seus seguidores tratam de diversos assuntos diferentes.

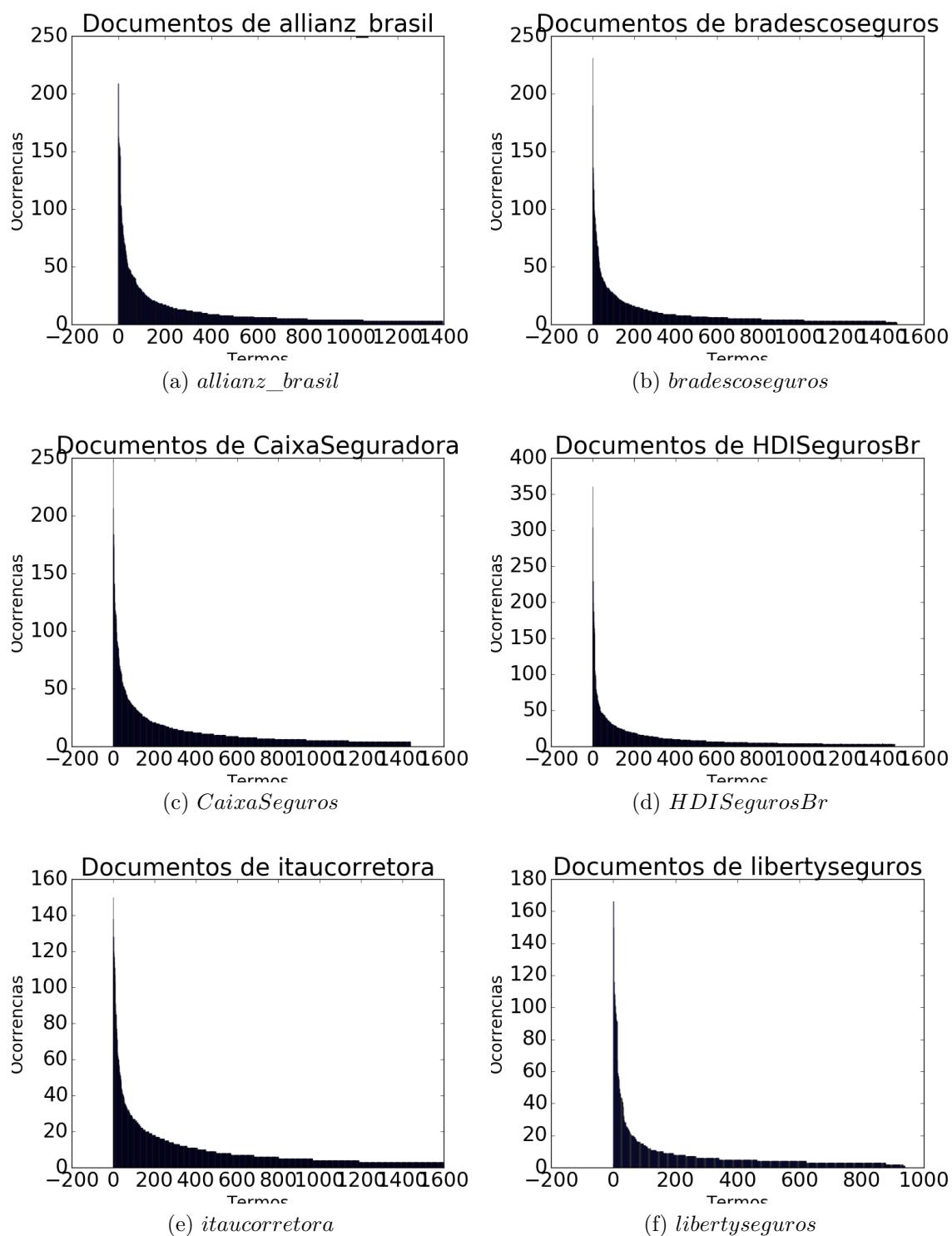


Figura 9 – Figura com a distribuição dos termos de conta após a redução de dimensão.

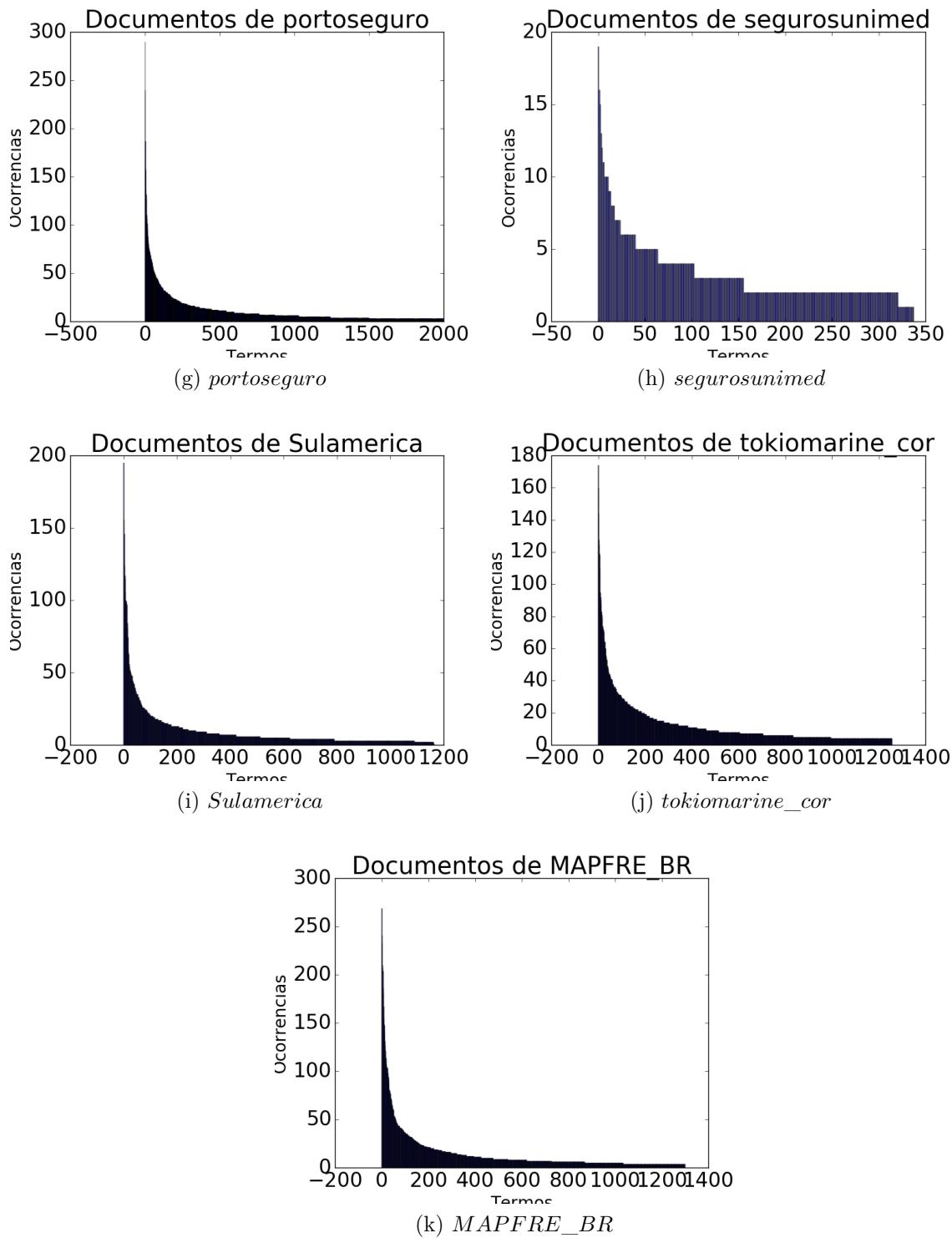


Figura 9 – Figura com a distribuição dos termos de conta após a redução de dimensão - *continuação*.

Feito isso, o próximo passo realizou a construção da matriz X' , que é a matriz X reduzida, onde ambas são $M \times N$, sendo M é o número de documentos (tuítes) dos seguidores de uma seguradora e N é o número de termos únicos desses documentos. A etapa seguinte consiste na geração de uma matriz Z , com as similaridades entre os documentos

da matriz X' . Por fim, são realizadas as clusterizações hierárquica e esférica, de modo que seja possível comparar os resultados dos agrupamentos realizados por cada tarefa de clusterização.

4.2 Análise dos resultados obtidos

A aplicação da etapa de pós-processamento permitiu gerar os *clusters* e a *tag cloud*, que foi extraída das análises *intra-cluster* a partir da análise *extra-cluster*. Essas análises serão discutidas abaixo.

4.2.1 A análise extra-cluster

A análise *extra-cluster*, neste trabalho, consistiu numa listagem dos 25 maiores *clusters*. A métrica de análise aqui é a quantidade de documentos agrupados dentro de um *cluster*. Essa etapa é executada com a simples contagem dos documentos e com a ordenação dos grupos numa lista. Com isso, é possível focalizar as análises apenas em grupos mais relevantes, pois se num grupo há documentos que são similares, então um grupo com mais documentos é mais relevante do que aquele com menos documentos, pois significa que possivelmente há mais tuítes que tratam do mesmo assunto. Antes de tudo, foi preciso gerar os *clusters*. Na tarefa de clusterização utilizou-se as abordagens esférica e a hierárquica.

A clusterização esférica permitiu gerar os gráficos mostrados na Figura 10. A tarefa de clusterização esférica gerou grupos diferentes da clusterização hierárquica, como pode ser visto na Figura 12.

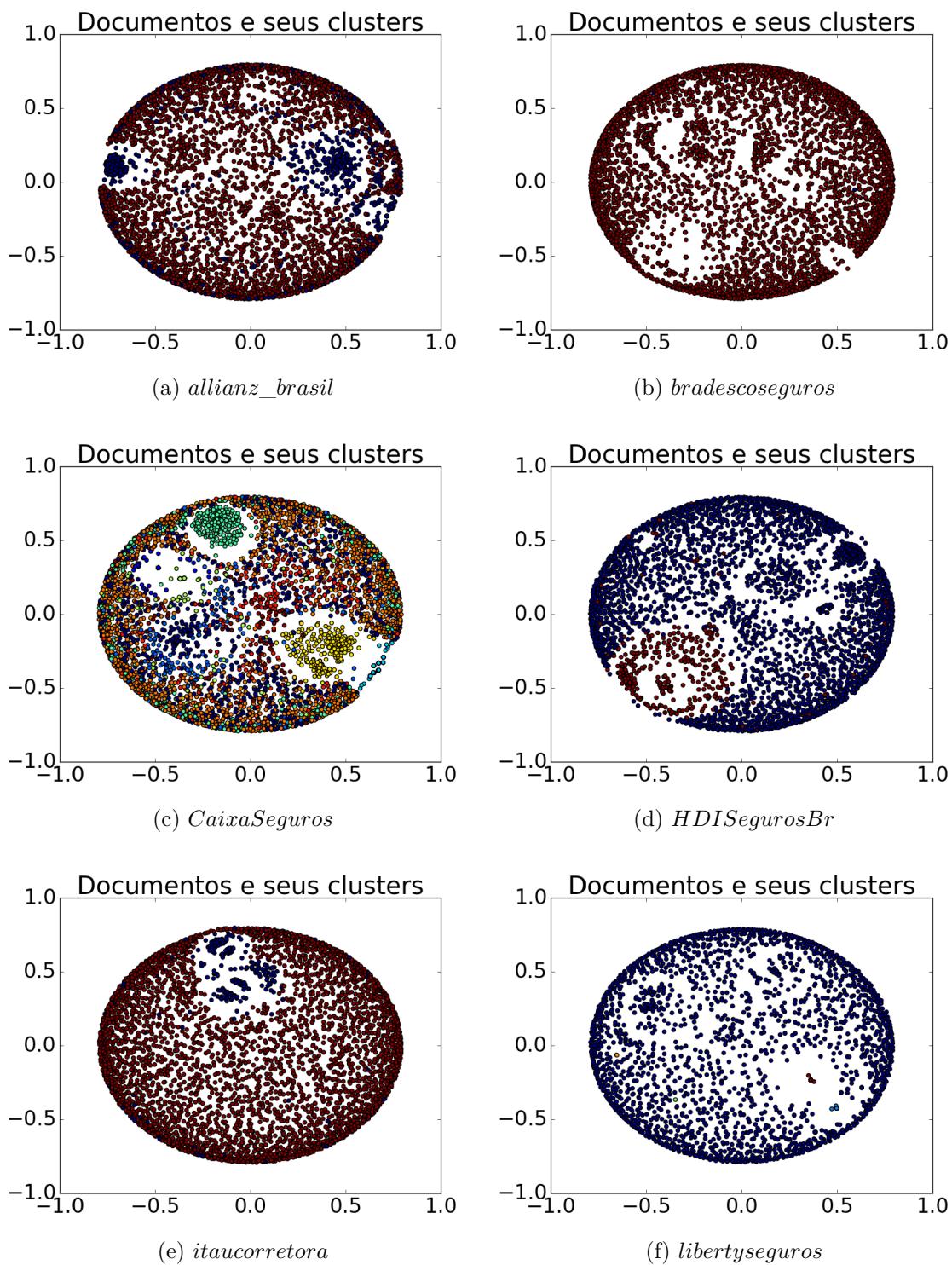
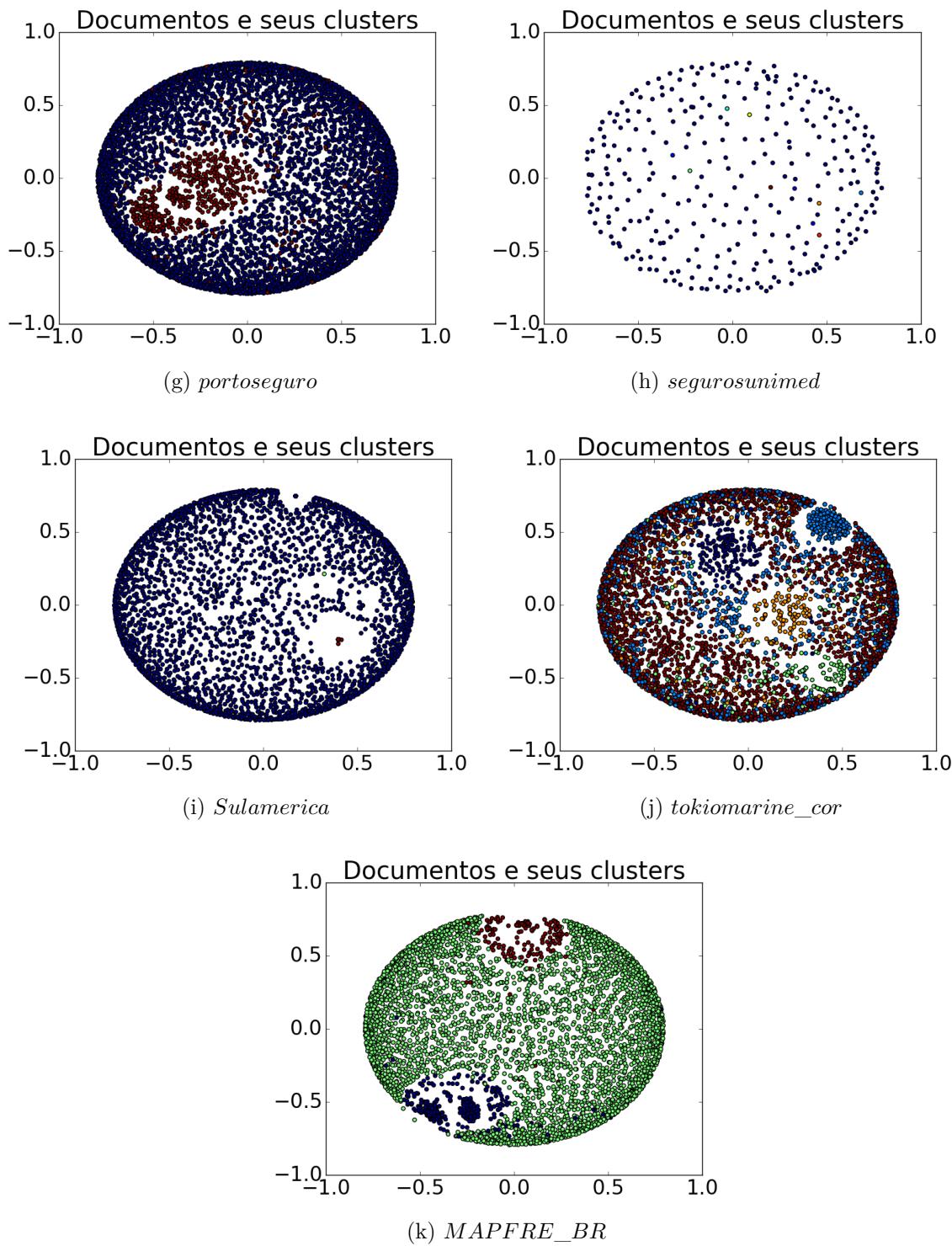


Figura 10 – Figura de clusterização particional para cada conta.

Figura 10 – Figura de clusterização particional para cada conta - *continuação*.

O cálculo do k foi possível por meio da função `silhouette_score()` do pacote *Metrics*, da biblioteca *SKLearn*, que calcula o coeficiente de silhueta. A função `silhouette_score()` permite descobrir o quanto bem os dados estão agrupados dentro de um *cluster*, por meio da distância interna dos dados no *cluster* (a) e da distância média dos *clusters* mais próximos (b), dado uma amostra de dados. O coeficiente de silhueta (s) é dado por $s = \frac{b-a}{\max(a,b)}$.

O que a função *silhouette_score()* faz, é realizar $n - 1$ agrupamentos de n dados. Iniciando em 2 (dois) vai até $n - 1$. No fim, para cada configuração de agrupamento gerado é armazenado seu respectivo s médio. Quanto maior o s médio de um agrupamento, mais denso ele é, ou seja, mais bem agrupado estão os dados.

Entretanto, esse cálculo é custoso temporalmente, por isso, nesse trabalho, foi delimitado o intervalo $[2, 10]$ para a busca do melhor grupo. O menor número de grupos possíveis é dois, por motivos triviais. O limite máximo de 10 foi escolhido arbitrariamente, por ser um valor que permite a busca por um intervalo relativamente grande de tentativas e que não consume tanto tempo de execução.

Essa abordagem verifica o quão "denso" são os *clusters* a partir de uma tentativa de agrupamento. Isso permite desconsiderar a maior variação entre cada configuração de agrupamento, pois o que importa, nesse caso, é a configuração que possui o maior valor absoluto de "densidade".

Na Figura 11, os valores no eixo das abscissas representam as configurações de agrupamento tentadas e no eixo das ordenadas qual foi seu respectivo s . O que importa é saber qual foi o maior s e não qual foi a maior variação de s , pois esse valor já indica o quão correto está aquela configuração x de acordo com a disposição dos seus *clusters*.

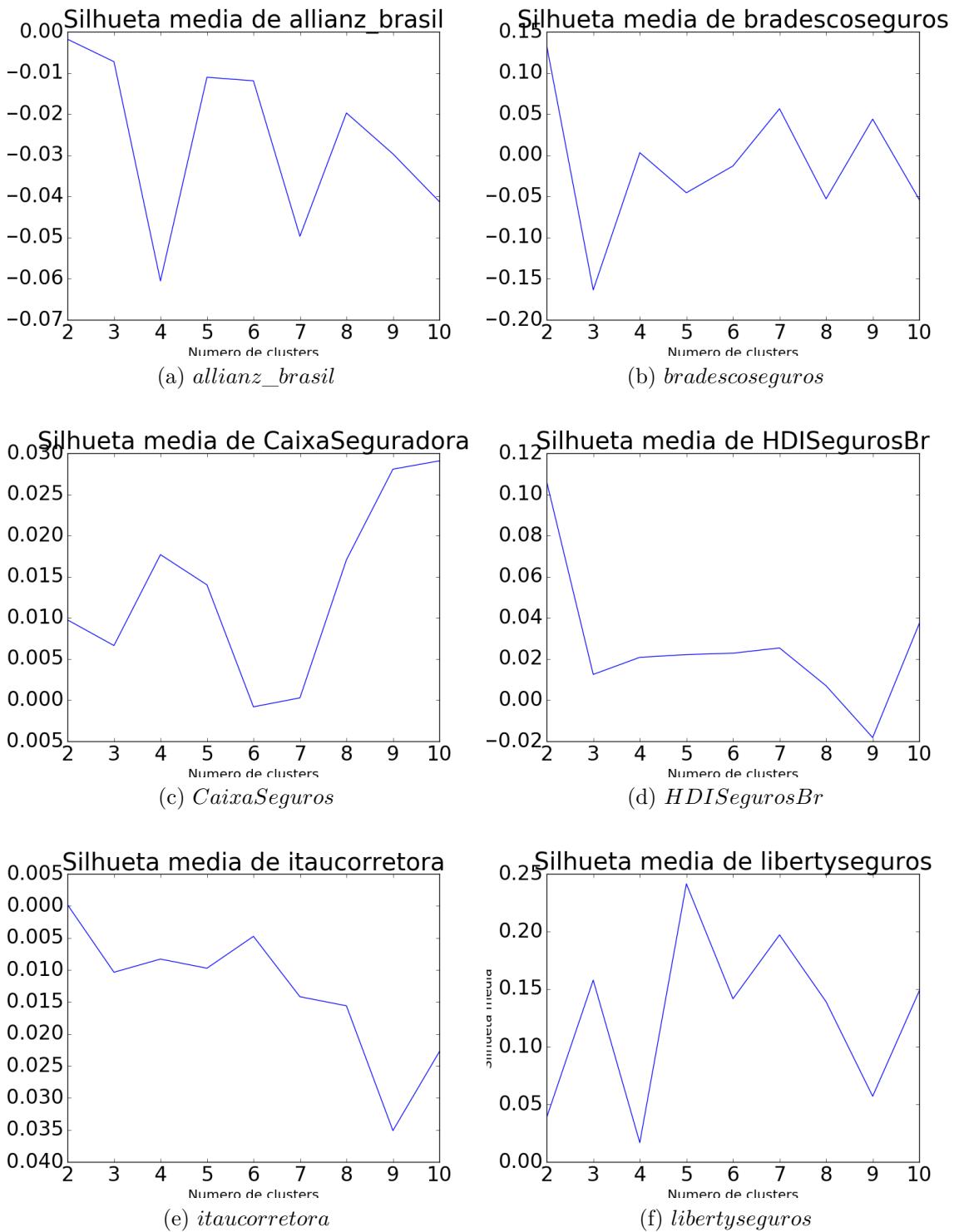


Figura 11 – Figura de clusterização particional para cada conta.

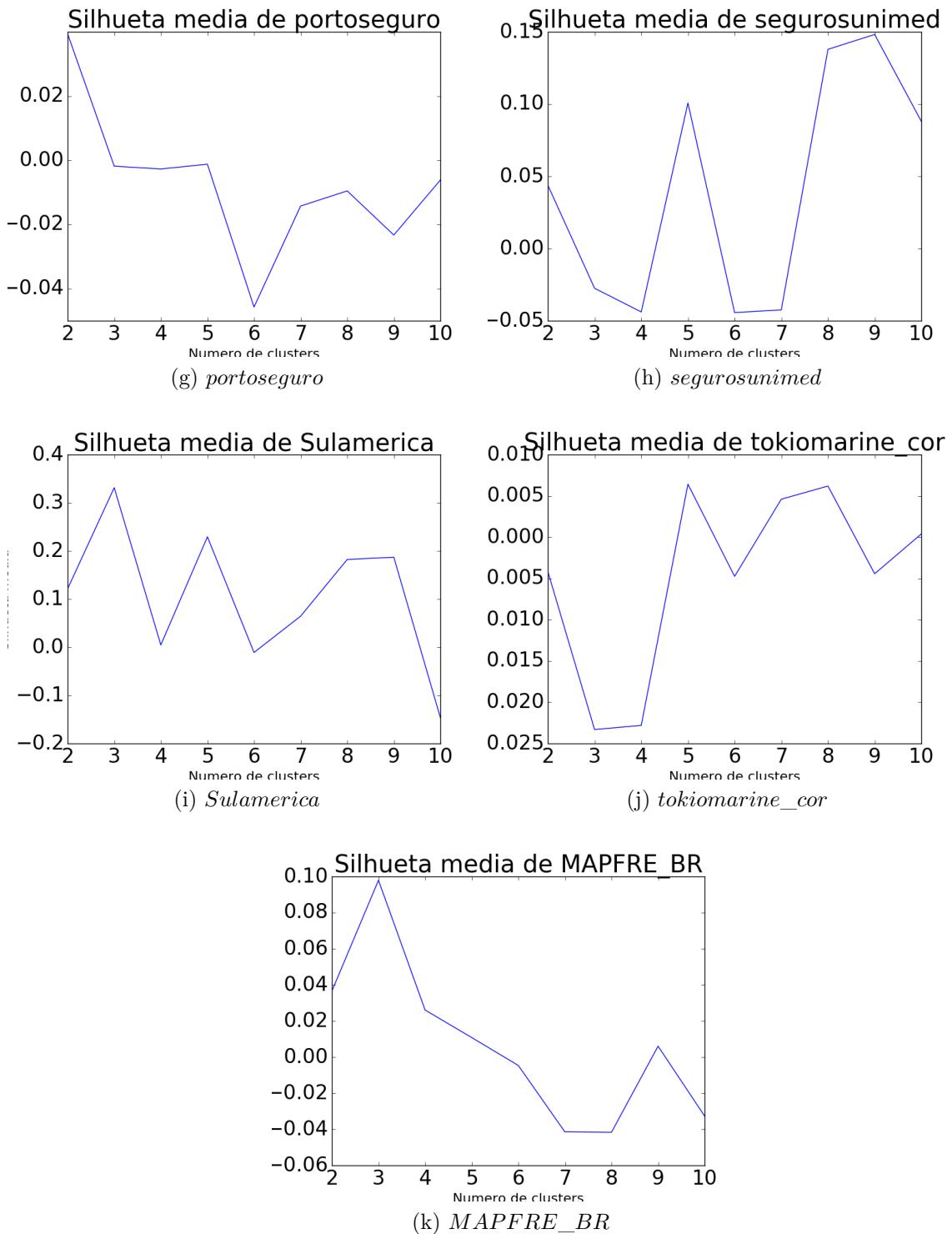


Figura 11 – Figura de clusterização particional para cada conta - *continuação*.

O intervalo $[2, 10]$ permitiu achar o k dentro desse intervalo que melhor agruparia os documentos dos seguidores das seguradoras. Em alguns casos, como por exemplo nas Figuras 11(c), 11(f), 11(h), 11(i), 11(j) e 11(k), o k ideal foi maior que dois, portanto, foi o agrupamento mais conciso que a clusterização por k_means esférico foi capaz de calcular em 10 tentativas.

Tabela 4 – Comparaçāo da quantidade de *clusters* gerados por diferentes mētods

Conta analisada	Mētodo particional	Mētodo hierárquico
<i>CaixaSeguradora</i>	10	814
<i>tokiomarine_cor</i>	5	753
<i>allianz_brasil</i>	2	714
<i>portoseguro</i>	2	1191
<i>MAPFRE_BR</i>	3	794
<i>HDI SegurosBr</i>	2	736
<i>itaucorretora</i>	2	871
<i>bradescoseguros</i>	2	704
<i>libertyseguros</i>	5	417
<i>Sulamerica</i>	3	587
<i>segurosunimed</i>	9	123

Quando hā um agrupamento mais conciso, onde os grupos aparecem bastante definidos, hā o indicativo de que aquele grupo possui documentos com bastante similaridade entre si, como é o caso das Figuras 10(c), 10(g) ou 10(k), por exemplo. O que nāo ocorre quando o agrupamento é visualmente mais esparsos, pois isso significa que aquele agrupamento gerado nāo é o que melhor define o grupo, ou seja, nāo é aquele que melhor afasta documentos com alta dissimilaridade entre si, como pode ser visto nas Figuras 10(b), 10(f), 10(h) ou 10(i), por exemplo.

Se a tarefa de clusterização agrupa documentos de acordo com sua similaridade com outros documentos, entāo é de se esperar que um grupo de documentos tenha termos iguais, ou seja, naquele grupo hā um termo que é o mais frequente. Disso é possivel entender que um grupo tem tópicos, uma vez que os termos que ali estāo sāo tópicos.

Consequentemente, hā um tópico principal num grupo, portanto, os grupos indicam quais os tópicos mais utilizados pelos seguidores de uma conta analisada. Assim, os gráficos na Figura 10 indicam os grupos em cores distintas, de modo que cada ponto colorido é um documento; um ponto azul indica um grupo que trata de um tópico *A* e um ponto vermelho indica um grupo que trata de um tópico *B*, por exemplo. A Figura 10 (b) indica que os seguidores da conta *bradescoseguros* falam somente de dois assuntos (sendo que um desses assuntos é muito mais utilizado do que o outro), enquanto que os seguidores da conta *CaixaSeguros* (Figura 10(c)) tratam de 10 assuntos.

Como a clusterização particional esférica é custosa temporalmente foi necessário limitar a área de busca pelo *k* ideal. Entretanto, a tarefa de clusterização hierárquica se mostrou mais eficiente no cálculo pela melhor configuração de *clusters* e por isso foi a que permitiu descobrir o maior númeroo possível de grupos que melhor agrupavam os documentos. Ou seja, foi somente pela clusterização hierárquica que foi possível descobrir realmente quais sāo os grupos de assuntos dos seguidores de cada conta.

Os gráficos na Figura 12 representam a clusterização hierárquica para cada conta objeto de estudo. É possível notar que a quantidade de grupos descobertos por clusterização hierárquica, no caso da conta *bradescoseguros*, é radicalmente superior àquele descoberto pela clusterização particional esférica por *k*-means.

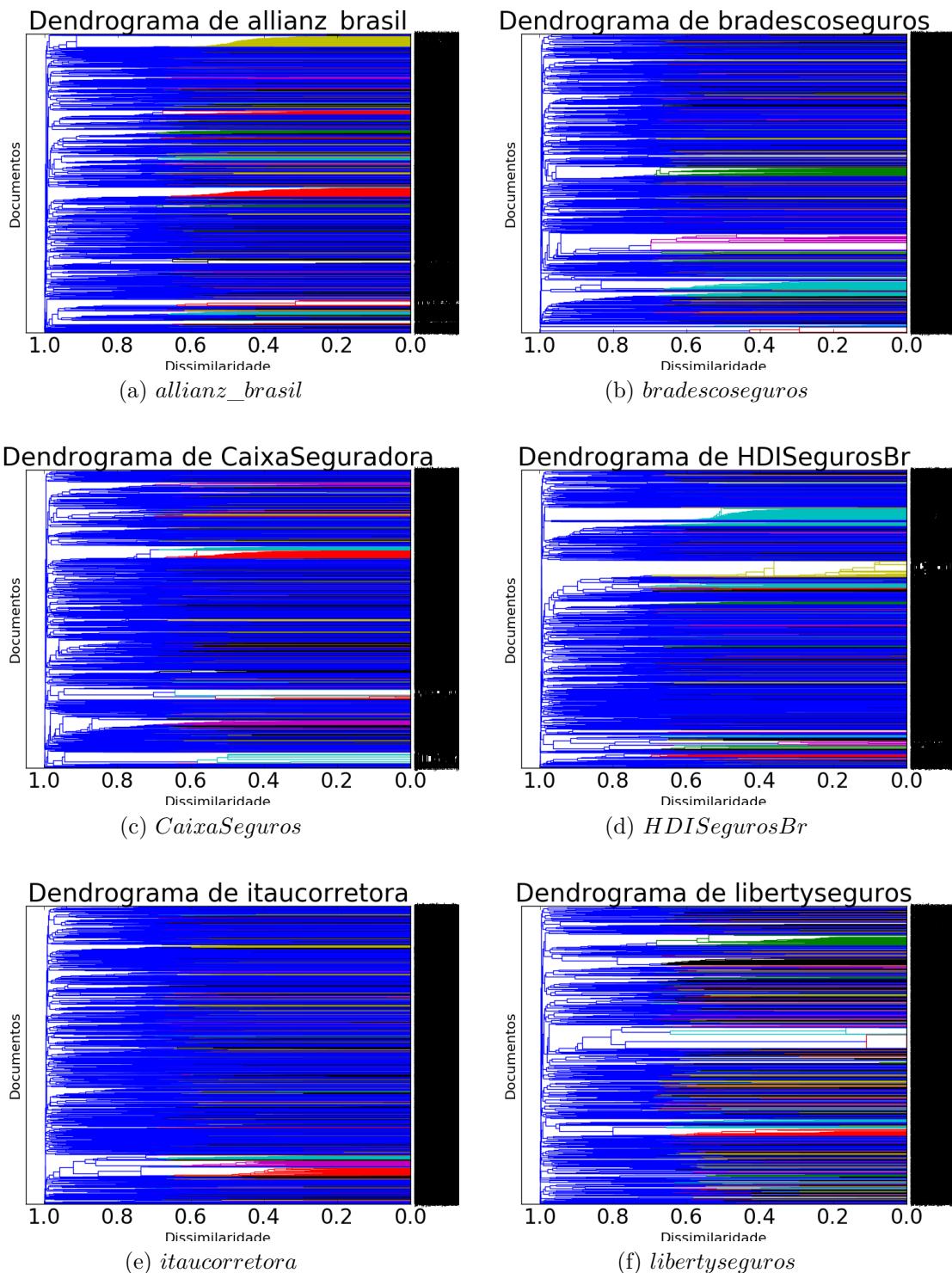


Figura 12 – Figura de clusterização hierárquica para cada conta.

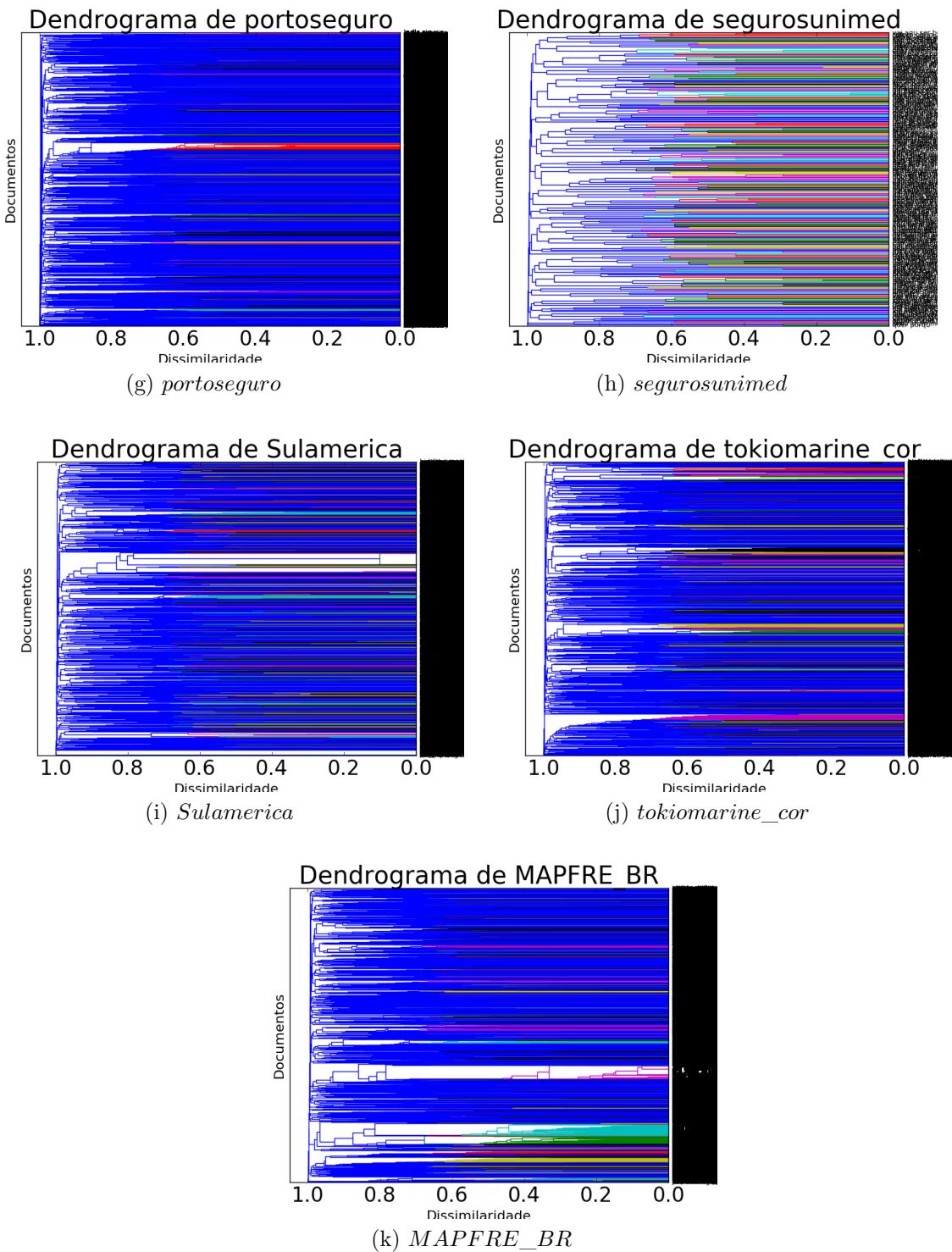


Figura 12 – Figura de clusterização hierárquica para cada conta - *continuação*.

A visualização é imprecisa devido à quantidade de grupos gerados. Por isso, para facilitar a visualização, foi recortado desses grandes dendrogramas os 25 maiores *clusters* (grupos), que, como exposto anteriormente, são aqueles que contêm as maiores quantidades de documentos agrupados. Com essa abordagem foi possível visualizá-los de um modo mais elucidativo.

Essa abordagem permite focar apenas nos grupos mais relevantes, no sentido quantitativo, que por sua vez permite obter o conhecimento sobre quem são os seguidores que *postaram* aqueles documentos, ou seja, essa abordagem permite buscar nos grupos com mais documentos, qual é o tópico daquele *cluster* e quem são as pessoas que reproduzem esse tópico. Os gráficos mostrados na Figura 13 são a focalização nos maiores *clusters* de cada conta de seguradora.

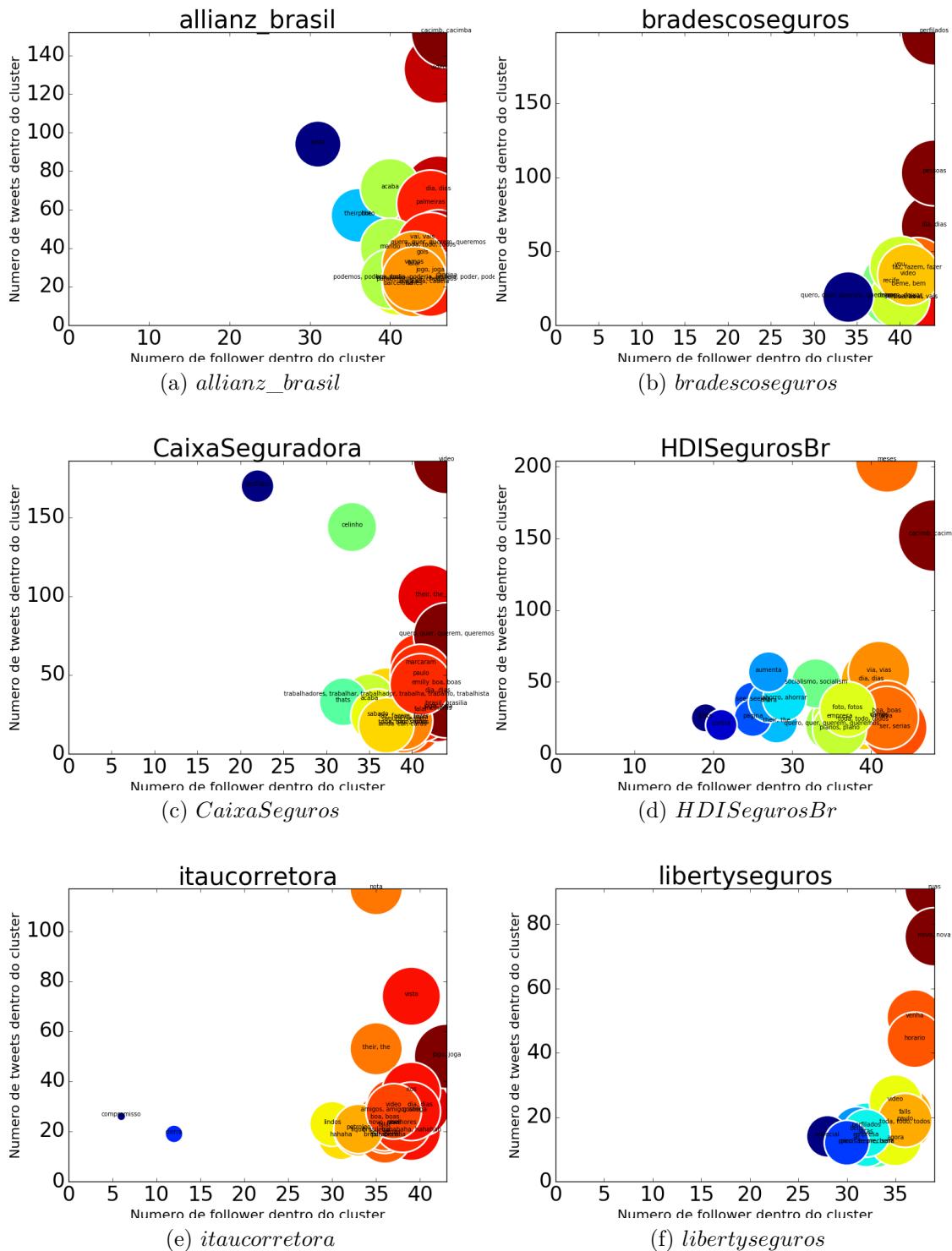
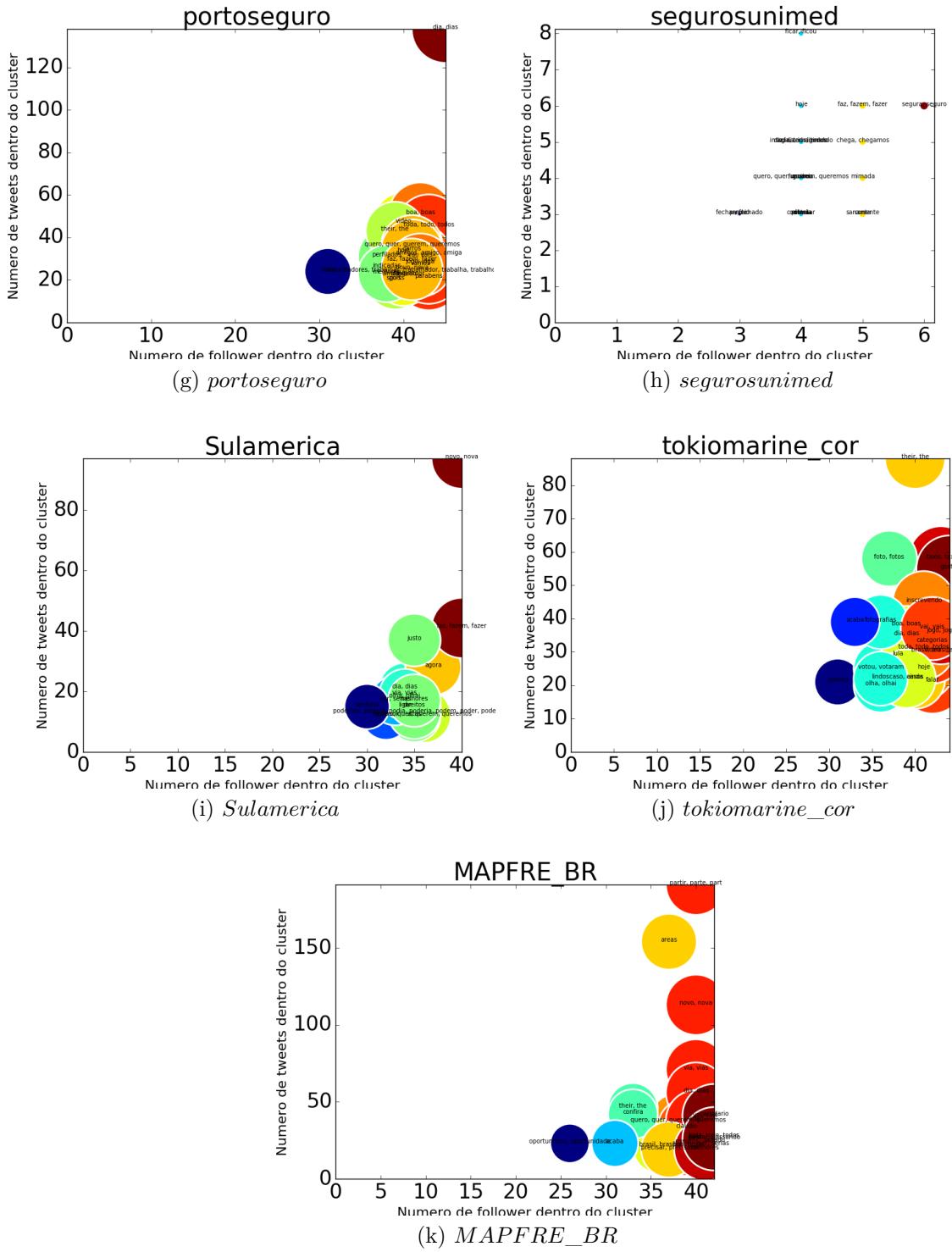


Figura 13 – Figura de gráfico de bolha para cada conta.

Figura 13 – Figura de gráfico de bolha para cada conta - *continuação*.

Os *clusters* representados nos gráficos da Figura 13 são as bolhas (ou círculos); a cor de cada *cluster* (círculo) fica mais avermelhada quanto mais seguidores um *cluster* tem, e mais azulado quanto menos tem; o diâmetro do círculo fica maior à medida que a quantidade de termos dentro desse *cluster* aumenta, e menor conforme essa quantidade diminui. O rótulo de cada círculo na Figura 13 é o tópico mais frequente dentro de cada

cluster.

4.2.2 A análise intra-cluster

A análise *intra-cluster*, neste trabalho, consistiu em verificar quais são os termos mais relevantes dentro de um *cluster*, onde buscou-se o tópico principal de um dado grupo. A métrica neste caso, é contar quantas aparições determinada palavra teve nos documentos que ali estão agrupados. As palavras com maior aparição são mais relevante do que as com menor aparição e assim foi possível gerar as *tag clouds* para cada conta (Figura 14).

A ideia de uma *tag cloud* é explicitar as palavras que são relevantes dentro de um conjunto dado. Essa relevância é mostrada no tamanho da fonte usada para desenhar a palavra. Palavras mais frequentes (relevantes) são representadas em tamanhos maiores, enquanto que as menos relevantes possuem tamanhos menores. Analisando a Figura 14 é possível notar semelhanças e diferenças entre cada conta analisada.



Figura 14 – Figura de nuvem de palavras para cada conta.



Figura 14 – Figura de nuvem de palavras para cada conta - *continuação*.

O tópico "vídeo" pode ser visto nas Figuras 14(a), 14(b), 14(c), 14(e), 14(f) e 14(g). Essa informação permite entender que os seguidores das contas representadas nessas figuras citadas são mais inclinados a discutirem sobre "vídeo" nas redes sociais e possivelmente são consumidores de mídias. Tal informação permite discriminar os seguidores das contas *HDI SegurosBr*, *segurosunimed*, *Sulamerica*, *tokiomarine_cor*

e *MAPFRE_BR* como possíveis não consumidores de materiais de mídias.

Entretanto, o tópico "segura, seguro" que tem mais relação com a área de atuação de mercado dessas contas, apenas aparece nos seguidores da conta *segurosunimed*. Isso pode significar que esses seguidores estão mais interessados em falar sobre seguros nas redes sociais do que os seguidores das demais seguradoras.

Outra análise possível é verificar como se comportam os seguidores das três maiores seguradoras da base de dados, em número de seguidores. As contas *portoseguro*, *tokiomarine_cor* e *itaucorretora* possuem tópicos únicos e diferentes, compartilham alguns tópicos mas a dissonância de assuntos é maior do que a semelhança.

A diferença nas *tag clouds* de cada conta pode ser um produto da diferença do ramo de atuação no qual cada seguradora possui maior difusão, além de poder ser um produto de coordenadas geográficas também, assim como um produto do *marketing* digital de cada seguradora.

Uma possível aplicação dos conhecimentos obtidos com a geração das *tag clouds* é permitir que uma empresa entenda de um modo mais eficaz sobre o quê falam seus seguidores na plataforma social do Twitter, lembrando-se que esses seguidores podem ser potenciais clientes. Assim como permite analisar o perfil e as necessidades de segurados de suas concorrentes e ajustar suas estratégias de mercado.

Dadas as palavras em destaque, é possível perceber que para cada seguradora analisada há grupos diferentes que, em maioria, falam de coisas diferentes uns dos outros, mesmo sendo seguidores de um mesmo segmento do mercado. Ações de *marketing*, por parte das empresas, podem ser melhor direcionadas tendo-se como base quais são as palavras (ou tópicos) mais frequentemente utilizadas pelos seus seguidores. E isso permite que a empresa fique mais próxima de seus clientes.

4.2.3 Geovisualização dos dados

Uma forma de visualização dos dados interessante é a geovisualização, que consiste em utilizar mapas para visualização de dados que se referem a posições geográficas. Essa abordagem permite o conhecimento geográfico dos dados a serem analisados, neste caso: onde se localizam os seguidores das contas de estudo das seguradoras, de modo a visualizar sua dispersão territorial.

A geração dos mapas foi feita com base nas posições de latitude e longitude, a partir das localizações disponibilizadas pelos próprios seguidores; essa informação diz onde, no mundo, esse usuário está. Entretanto, essa localização não possui um grau de confiabilidade alto, uma vez que essa informação não é coletada automaticamente pela plataforma do Twitter, mas é fornecida pelo próprio usuário, de modo que esse dado fica completamente dependente da boa-vontade do usuário de informar, ou não, o local correto

de onde está.

A implementação foi possível por meio das bibliotecas GMplot e GeoPy. A primeira, uma biblioteca que permite desenhar os pares de latitude e longitude nos mapas desenvolvidos pelo Google, de modo a torná-los interativos. A segunda, uma biblioteca que permite a fácil conversão de endereços, cidades e países em pares de latitude e longitude.

A abordagem foi, primeiro descobrir os pares de coordenadas geográficas para cada localização disponível pelos usuários seguidores das contas das seguradoras e depois, por fim, simplesmente desenhá-los no mapa. Cada conta de seguradora ganhou uma cor e na Figura 15 é possível visualizar a localização de cada seguidor. A Figura 16 permite uma visualização de mapa de calor, onde exibe a provável região da maior incidência de seguidores para cada conta de seguradora.

A partir dessas informações, uma empresa é capaz de redirecionar seus projetos de acordo com as distribuições de seus seguidores. É possível identificar, por exemplo, em qual região se concentra a maior parte de seus potenciais clientes, permitindo o desenvolvimento de estratégias de *marketing* focalizadas nessas regiões.

Dos mapas expostos nas Figuras 15 é possível visualizar, que num quadro geral, os seguidores das contas analisadas estão distribuídos pelas regiões centro-sul do Brasil. Algo caracterizado pelo desenvolvimento econômico e populacional dessas regiões. A região Norte do país carece de representatividade, uma das causas pode ser a questão populacional, uma vez que é a região menos habitada do país.

A questão econômica também pode ser um fator relevante na análise da distribuição geográfica desses seguidores. Ter bens assegurados só acontece mediante a posse de bens que podem ser assegurados. Grupos socio-economicamente mais vulneráveis são aqueles que tendem menos a se preocupar com a questão de seguros, enquanto que grupos socio-economicamente não vulneráveis são aqueles que podem tender a se preocupar com essa questão e adquirir alguma apólice de seguros.

Essa tese fica mais evidente ao se analisar os mapas das Figuras 16, cujos pontos quentes estão majoritariamente sobre as regiões Sudeste e Sul do país, que são as mais ricas e desenvolvidas da União. Ainda é possível visualizar que a conta *tokiomarine_cor* é a que possui maior distribuição territorial de seguidores. Essa é a conta com o maior número de seguidores na base de dados desse trabalho. Já a conta *portoseguro* é a que possui menor variabilidade de lugares no mapa, embora essa seja a segunda conta, em números de seguidores inseridos na base de dados, uma explicação pode ser que talvez nem todos seus seguidores viabilizaram dados referentes à geolocalização e por isso seu mapa carece de mais dados que representem efetivamente a coleção real de posições geográficas da conta estudada.

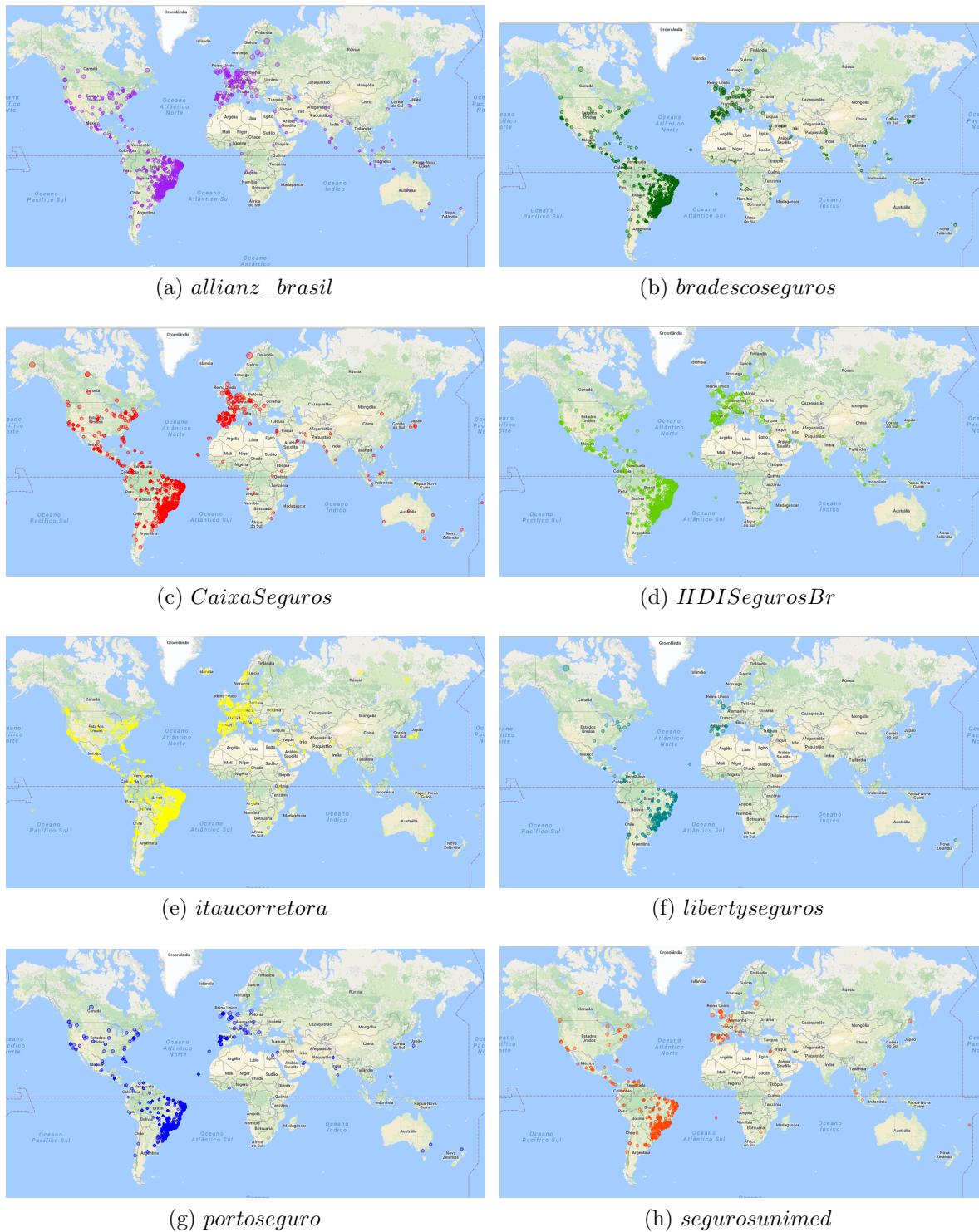


Figura 15 – Mapa de localização dos seguidores das contas de estudo.

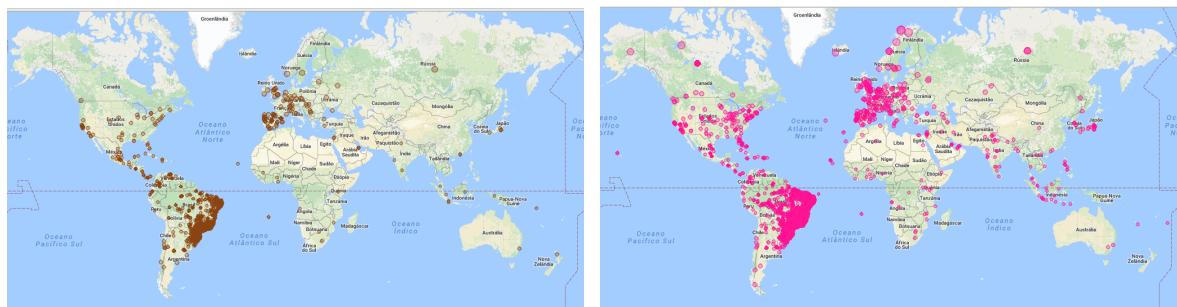
(i) *Sulamerica*(j) *tokiomarine_cor*(k) *MAPFRE_BR*

Figura 15 – Mapa de localização dos seguidores das contas de estudo - *continuação*.

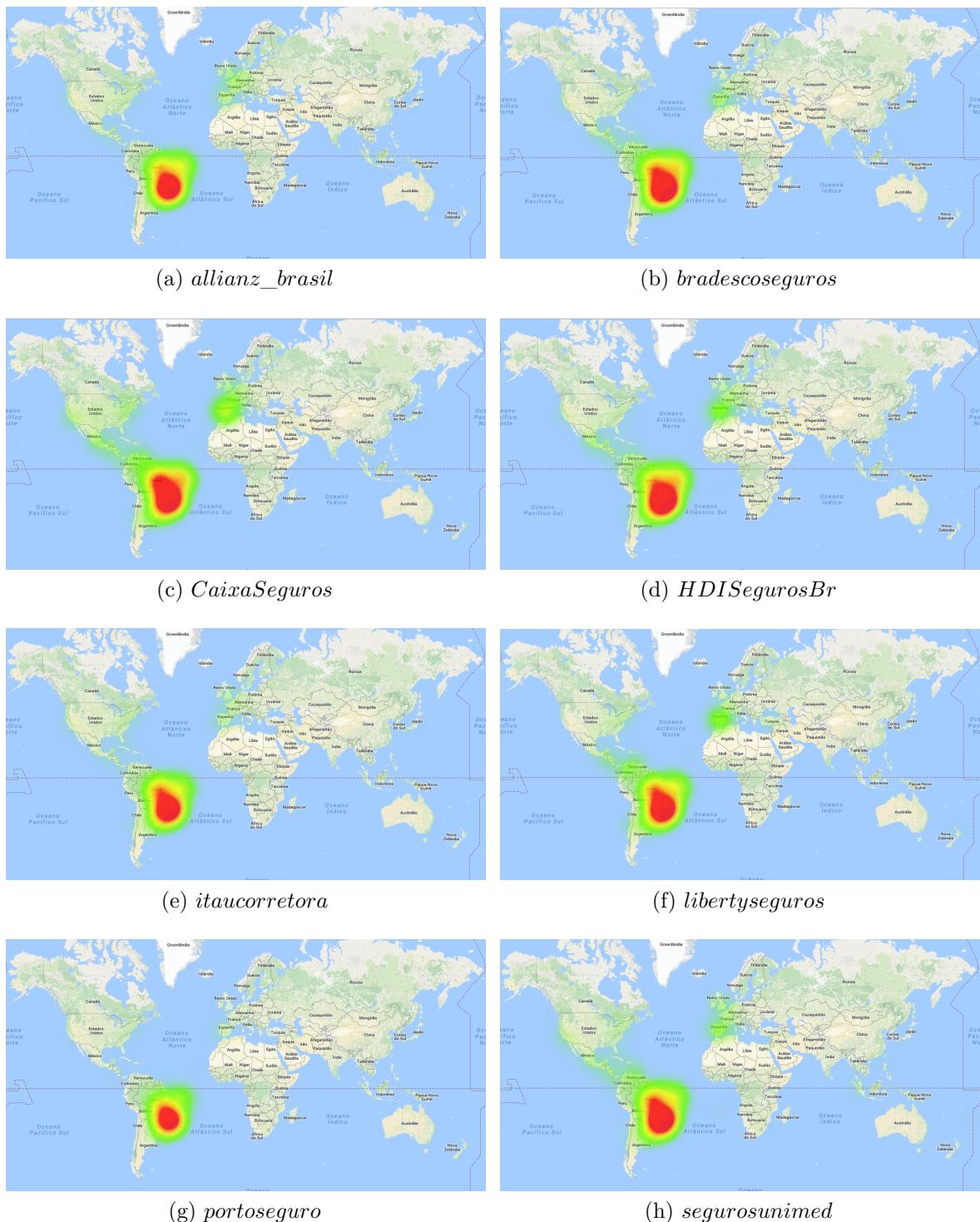


Figura 16 – Mapa de calor da localização dos seguidores das contas de estudo.

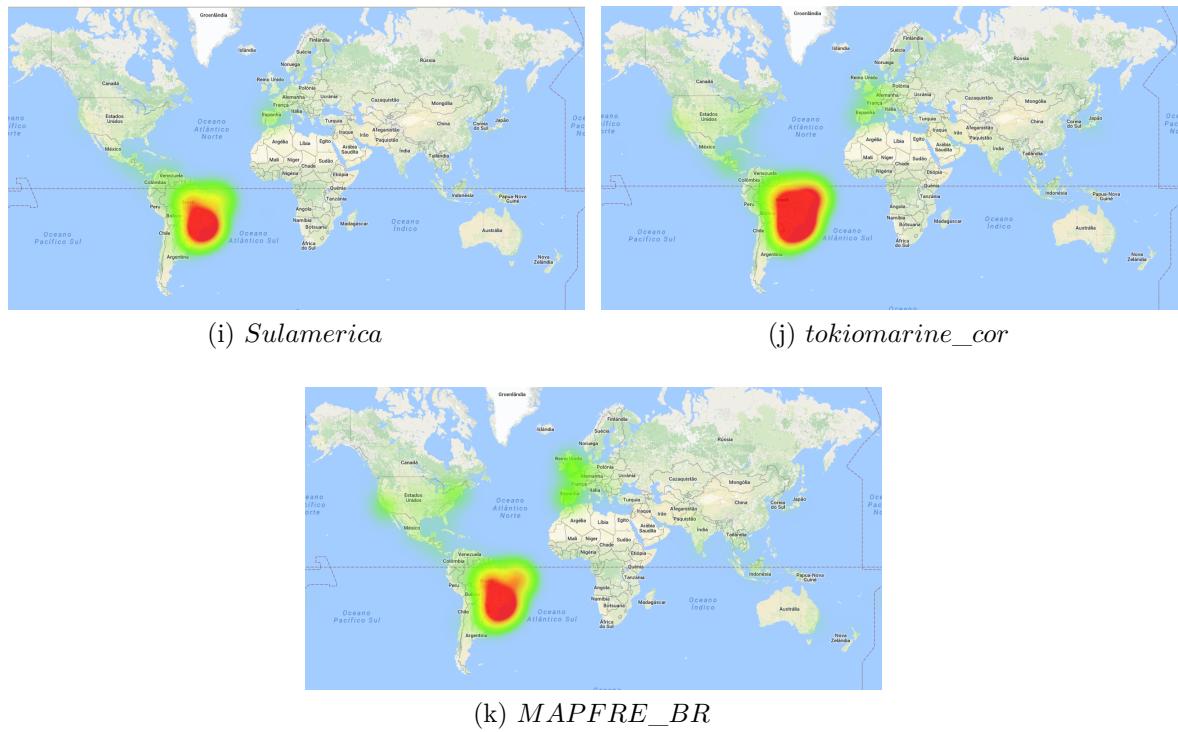


Figura 16 – Mapa de calor da localização dos seguidores das contas de estudo - *continuação*.

5 Conclusões

O objetivo proposto por esse trabalho era o de tornar possível a extração de conhecimento a partir de dados dispersos na plataforma virtual do Twitter. Para tanto, foi utilizado uma série de etapas e tarefas da mineração de dados como ferramenta de apoio a extração desse conhecimento.

O objeto de estudo foi o mercado segurador e, assim, o objetivo passou a ser: agrupar os tuítes dos seguidores de uma determinada seguradora a fim de visualizar quais são os tópicos mais reproduzidos por esses seguidores, permitindo que uma empresa (ou qualquer interessado) possa descobrir sobre o quê seus seguidores no Twitter falam sobre.

Essa proposta é relevante, pois o acesso à Internet é crescente e a cada vez mais existem dispositivos (computadores) que se conectam à rede mundial. Essa conexão inevitavelmente faz com que dados *offline* se tornem *online*, aumentando a quantidade de dados que permitem descobrir conhecimento. Essa revolução digital está apenas no começo e a *Internet of Things* (IoT) já é quase uma realidade, onde as técnicas para a análise de dados serão bastante necessárias.

As técnicas de aprendizado de máquina (e além) são mais do que importantes, são imprecindíveis para o futuro que está por vir. Casas digitais, carros autônomos, reconhecimento automatizado de tumores em imagens médicas são exemplos de como a grande área de Inteligência Artificial ajudará a humanidade num futuro não tão distante.

As simples etapas e tarefas abordadas nesse trabalho mostram o poder que tarefas de aprendizado de máquina podem possuir e como podem ajudar na obtenção de conhecimento extraído do mundo real, dinâmico, como forma de melhorar a vida das pessoas.

Com isso, pode-se verificar a validade do método e da abordagem tomadas por esse trabalho como formas de se obter real vantagem competitiva, seja em segmentos de mercado ou não, pois a obtenção de conhecimento já é, por si só, adquirir vantagem competitiva. Portanto, a proposta desse trabalho verifica-se relevante, uma vez que os processos, tarefas e algoritmos podem ser generalizados para quaisquer casos de estudo. E todos eles permitem obter conhecimento substancial sobre o conjunto dos possíveis clientes de uma empresa, desde que ambos estejam na plataforma social virtual aqui abordada.

Referências

- CHARTERED INSURANCE INSTITUTE. *Who invented insurance and why?* 2017. (acessado 12-junho-2017). Disponível em: <<http://www.cii.co.uk/membership/new-starters/who-invented-insurance-and-why/>>.
- FGV. *Número de smartphones em uso no Brasil chega a 168 milhões.* 2016. (acessado 20-abril-2017). Disponível em: <<http://www1.folha.uol.com.br/mercado/2016/04/1761310-numero-de-smartphones-em-uso-no-brasil-chega-a-168-milhoes-diz-estudo.shtml>>.
- FLORIDI, L. *Philosophy and computing: an introduction.* 1^a edição. ed. Londres, Reino Unido: Routledge, 1999.
- FRANCA, T. C.; OLIVEIRA, J. Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013. *III Brazilian Workshop on Social Networks and Mining*, 2014.
- GODFREY, D. et al. A case study in text mining: interpreting twitter data from world cup tweets. *arXiv preprint 1408.5427*, Estados Unidos, 2014.
- HADDI, E.; LIU, X.; SHI, Y. The role of text pre-processing in sentiment analysis. *Information Technology and Quantitative Management*, v. 17, p. 26–32, 2013.
- HORNIK, K. et al. Spherical k-means clustering. *Journal of Statistical Software*, Estados Unidos, v. 50, 2012.
- INVESTOPEDIA. *Commodity.* 2016. (acessado 09-maio-2017). Disponível em: <<http://www.investopedia.com/terms/c/commodity.asp>>.
- INVESTOPEDIA. *Value-Added.* 2016. (acessado 09-maio-2017). Disponível em: <<http://www.investopedia.com/terms/v/valueadded.asp>>.
- LAROSE, D. T. *Discovering knowledge in data: an introduction to Data Mining.* 1^a edição. ed. Estados Unidos: John Wiley and Sons, Inc., 2005.
- LÉVY, P. *Cibercultura.* 3^a edição. ed. São Paulo: Éditions Odile Jacob, 1997.
- MATHINA, K. T.; SHANTHI, I. E.; NANDHINI, K. Applying clustering techniques for efficient text mining in twitter data. *International Journal of Data Mining Techniques and Applications*, v. 4, p. 25–28, 2015.
- MATTOS, J. M. *A sociedade do conhecimento, da teoria de Sistemas à Telemática.* 1^a edição. ed. Brasília: Editora Universidade de Brasília, 1982.
- MELO, E. F. L.; FRANKLIN, S. L.; VIANNA, P. R. M. F. Monitoramento de redes sociais na regulação do mercado de seguros e previdência. 2015.
- PETRÓ, L. A. Relacionamento nas redes sociais virtuais: análise da inserção do mercado de seguros no twitter. 2010.

- PEWINTERNET. *Demographics of Social Media Users in 2016*. 2016. (acessado 20-abril-2017). Disponível em: <<http://www.pewinternet.org/2016/11/11/social-media-update-2016/>>.
- RIBEIRO, R. O. A.; TAVARES, T. G. B.; COHEN, D. O. Análise de usuários que conversam sobre cerveja no twitter. *Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia*, São Paulo, 2014.
- RUSSEL, M. A. *Mining the social web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub and More*. 2^a edição. ed. Estados Unidos: O'Reilly Media, 2014.
- SANTOS, W. P. Análise dos tweets sobre a black friday através da mineração de textos e análise de sentimentos. 2016.
- SILVA, P. F. et al. *O Desenvolvimento do E-Commerce no Brasil*. 2016. (acessado 09-maio-2017). Disponível em: <<http://www.administradores.com.br/artigos/academico/o-desenvolvimento-do-e-commerce-no-brasil/101304/>>.
- TOMAÉL, M. I.; ALCARÁ, A. R.; CHIARA, I. G. D. Das redes sociais à inovação. *Ciência da Informação*, Brasília, v. 34, p. 93–104, 2005.
- TSAI, C. W. et al. Data mining for internet of things: A survey. *IEEE Communications Society*, v. 16, p. 77–97, 2014.
- TUDO SOBRE SEGUROS. *Fatos e indicadores do mercado*. 2017. (acessado 12-junho-2017). Disponível em: <<http://www.tudosobreseguros.org.br/portal/pagina.php?l=267>>.
- VALOR ECONÔMICO. *As 50 maiores seguradoras - ramos gerais*. 2016. (acessado 12-junho-2017). Disponível em: <<http://www.valor.com.br/valor1000/2014/ranking50maioresseguradoras>>.
- WIGAND, R. T. Electronic commerce: definition, theory and context. *The Information Society*, Estados Unidos, v. 13, p. 1–16, 1997.