UMA ABORDAGEM PARA SELEÇÃO DE TÓPICOS RELEVANTES EM REDES SOCIAIS ONLINE

BRUNO ÁBIA SOUZA

UMA ABORDAGEM PARA SELEÇÃO DE TÓPICOS RELEVANTES EM REDES SOCIAIS ONLINE

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Eduardo Freire Nakamura
Coorientadora: Fabíola Guerra Nakamura

Manaus
Março de 2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Souza, Bruno Ábia
S729u Uma Abordagem para Seleção de Tópicos Relevantes em Redes Sociais Online / Bruno Ábia Souza. 2017
81 f.: il. color; 31 cm.

Orientador: Eduardo Freire Nakamura
Coorientadora: Fabiola Guerra Nakamura
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Redes Sociais Online. 2. Detecção de Informação. 3. Machine learning. 4. Twitter. I. Nakamura, Eduardo Freire II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Uma abordagem para Seleção de Tópicos Relevantes em Redes Sociais"

BRUNO ÁBIA SOUZA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Freire Nakamura - PRESIDENTE

Prof. Carlos Maurício Seródio Figueiredo - MEMBRO INTERNO

Profa. Elloá Barreto Guedes da Costa - MEMBRO EXTERNO

Manaus, 24 de Março de 2017

Dedico esta dissertação a minha querida avó Raimunda de Oliveira Ábia, uma senhora simples e bondosa, que mesmo não estando mais entre nós, sempre foi uma pessoa que através de suas atitudes mostrou como a vida pode ser vivida de forma autêntica.

Agradecimentos

Apesar de ser muito difícil agradecer a todas as pessoas que tiveram paciência em me ajudar nesses dois anos. Primeiramente gostaria de agradecer a Deus por sempre estar comigo e nunca desistir, mesmo que eu faça escolhas erradas.

Agradecer aos meus pais que foram pessoas que apesar de todas as dificuldades como: Doenças, momentos de raiva e tristeza sempre me aconselharam a ter calma em tudo que eu fosse fazer. Ao meu irmão Roger Ábia e sua esposa Herika Ábia, que apesar de estarem em outro país sempre torceram por mim e me incentivaram a continuar em busca dos meus sonhos. A minha irmã Simone Ábia e seu marido Alison, por sempre estarem presentes e torcendo por mim.

Agradecer especialmente aos amigos Alice Adativa, Igor Carneiro, Thais Gomes, Rayol Neto, Luciana Mendes, Maria Ivanilse, Josiane Rodrigues, Ana Carolina Oran, Elizamary Souza e Juan Gabriel que estão comigo desde o início desta jornada. Novamente agradecer a Alice que foi uma pessoa que sempre me incentivou, revisou meus textos e mostrou uma nova perspectiva em relação sobre as dificuldades da vida.

Agradecer aos meus orientadores Eduardo Nakamura e Fabíola Nakamura pela extrema paciência durante as orientações. Sempre me desafiando a superar meus limites e desenvolver características de um verdadeiro pesquisador. Vale a pena ressaltar que a minha escolha em fazer mestrado foi devido eu sempre me espelhar no tipo de profissional que o professor Eduardo se tornou. Outra situação que vale a pena destacar, que mesmo eu batendo o carro na sua esposa, ainda continuou me incentivando desde o início do mestrado. Agradeço também à CAPES pelo suporte através de bolsa de mestrado.

Agradecer aos meus amigos Awdren Fontão, Laíza Serrão, Anderson Pimentel, Daniel Xavier, Ludimila Gonçalves e Kevin Gustavo pelas conversas de incentivo e apoio durante os momentos difíceis no mestrado. Ao Awdren novamente, pois foi uma pessoa que deu várias dicas de como melhorar minha escrita. A meu grande amigo William Suane que foi uma pessoa que nos momentos de dúvidas sobre minhas escolhas profissionais me aconselhou de uma forma incrível. Aos meus amigos Júlia

Reis, Henrique Ribeiro, João Bolzan, Skarllet Roberta, Antonio Lucas, Neia, Pablo Rodrigo, Susan Iansã, Luene Freire, Fernando Santos e a todo meu grupo de vida que estão sempre me ajudando durante meus problemas pessoais.

"Tudo o que um sonho precisa para ser realizado é alguém que acredite que ele possa ser realizado."
(Roberto Shinyashiki)

Resumo

O crescente uso das redes sociais fez com que a quantidade de informações que são compartilhadas diariamente (opiniões pessoais, notícias e eventos do tipo natural ou social) aumentasse, tornando estas redes uma fonte de informação sobre eventos. No Brasil, acontecimentos recentes, como a Operação Lava Jato conduzida pela polícia Federal e o processo de *impeachment* da Presidente são exemplos de eventos que geraram repercussões nos meios de comunicação. Esses fatos marcantes, podem ser utilizados para caracterizar quais são os acontecimentos relevantes de um determinado evento. Nas redes sociais, esses acontecimentos geram discussões, compartilhamentos e novas notícias, onde informações relevantes acabam tendo mais repercussões. Contudo, processar uma elevada massa de dados a fim de eliminar ruídos e reconhecer informações torna-se muito custoso. Neste contexto, este trabalho tem como objetivo demonstrar uma abordagem para caracterização de informações relevantes de eventos, através da extração de tópicos em dados compartilhados no Twitter, onde avaliamos o desempenho de três métodos de aprendizagem de máquina (*K-means*, *Latent Dirichlet Allocation* - LDA e *Non-Negative Matrix Factorization* - NMF) usados para extrair tópicos sobre as bases de dados da Operação Lava Jato e do processo de *impeachment* da presidente do Brasil em duas arquiteturas de pré-processamento diferentes (tradicional e com reconhecimento de entidade), demonstrando que é possível utilizar uma rede social como fonte de dados para descobrir os tópicos relevantes através do sensoriamento de usuários que observam um evento. Em nossos experimentos, observamos que as técnicas de pré-processamento têm influência direta sobre o resultado da extração de tópicos. Além disso, observamos que a técnica *Silhouette* nos ajudou a encontrar o melhor valor de *clusters* para uma determinada amostra de dados. Nos resultados o NMF apresentou o melhor desempenho nas duas bases de dados, tanto na tarefa de extração de tópicos quanto no tempo de execução.

Palavras-chave: Twitter, Redes Sociais, Aprendizagem de máquina, Eventos.

Abstract

The increasing use of social networks has made the amount of information that is shared daily (personal opinions, news and events of the natural or social type) increasing significantly, making social networks a source of information about events. In Brazil, recent events such as Lava Jato Operation conducted by the Federal Police and the impeachment process of the Brazil President are examples of events that have had repercussions in the traditional medias. These facts can be used to characterize the relevant informations of an event. In social networks, the events generate discussions, sharings and news posts, where relevant information have more repercussions. However, processing a large mass of data to eliminate noise and recognize the information becomes costly. In this context, this work aims to show an approach to the characterization of relevant information by extracting topics in shared data on Twitter, where we evaluate the performance of three methods of machine learning (K-means, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF)) used to extract topics on Lava Jato's databases and the impeachment process of the Brazilian president in two different preprocessing architectures (traditional and with entity recognition), demonstrating that it is possible to use the social networks as a data source to discover the relevant topics through the sensing of users that observe an event. In our experiments, we observed that the preprocessing techniques have a direct influence on the result of the topics extraction. In addition, we observed that the Silhouette technique helped us to find the best value of clusters for a given data sample. In the results the NMF presented the best performance in both databases, both in the topic extraction task and in the execution time

Keywords: Twitter, Social Networks, Machine Learning, Events.

Lista de Figuras

1.1 Quantidade (milhões) de usuários por redes sociais. 2
1.2 Quantidade de *tweets* coletados sobre o processo de *impeachment* da Presidente do Brasil durante o período de Janeiro a Maio [Souza et al., 2016]. . 3

2.1 Etapas do processo de DCBD [Fayyad et al., 1996]. 10
2.2 Exemplo do gráfico de *Silhouette* para seleçao de valores de k. 17
2.3 Arquitetura do Spark [Correia, 2016]. 19

4.1 Arquitetura da abordagem proposta. 27
4.2 Pré-processamento tradicional. 29
4.3 Pré-processamento com reconhecimento de entidade. 30
4.4 Modelo de comparação dos tópicos extraídos com canais de notícias. 32

5.1 Distribuição de *tweets* por mês sobre a Operação Lava Jato no período de 01/01/2016 à 20/11/2016 34
5.2 Exemplo de aplicação da técnica de *Silhouette* no mês de setembro. 35
5.3 Visualização da Função de *Silhouette* na base de dados da Operação Lava Jato. 36
5.3 Visualização da Função de *Silhouette* na base de dados da Operação Lava Jato (cont.). 37
5.4 Linha do tempo das fases da Operação Lava Jato em 2016. 39
5.5 Percentual de acerto utilizando pré processamento tradicional. 40
5.6 Percentual de acerto utilizando reconhecimento de entidade. 41
5.7 Comparação da quantidade de atributos do vetor de características das arquiteturas de pré-processamento. 42
5.8 Tempo de execução coletado do primeiro cenário. 44
5.9 Tempo de execução coletado do segundo cenário. 44
5.10 Distribuição dos dados durante o processo de *impeachment*. 46
5.11 Percentual de acerto no primeiro cenário. 48

5.12 Percentual de acerto no segundo cenário. 49
5.13 Tempo de execução coletado do primeiro cenário. 50
5.14 Tempo de execução coletado do segundo cenário. 50

Lista de Tabelas

2.1 Conjunto de exemplos no formato atributo valor. 13

3.1 Resumos dos trabalhos relacionados. 26

5.1 Valores de K observados na técnica de *Silhouette*. 38
5.2 Tópicos extraídos pelos algoritmos. 41
5.3 Tópicos extraídos pelos algoritmos com reconhecimento de entidade. 43
5.4 Configuração do servidor para execução dos experimentos. 43
5.5 Valores de K observado na técnica de *Silhouette*. 47

Sumário

Agradecimentos **ix**

Resumo **xiii**

Abstract **xv**

Lista de Figuras **xvii**

Lista de Tabelas **xix**

1 Introdução **1**

1.1 Problema 3
1.2 Objetivos 4
1.3 Principais Contribuições 4
1.4 Publicações 5
1.5 Organização da Dissertação 5

2 Fundamentação Teórica **7**

2.1 Redes Sociais 7
2.2 Twitter 8
2.3 Descoberta de Conhecimento e Mineração de Dados 9
2.4 Aprendizagem de máquina 11
2.4.1 Algoritmos supervisionados para classificação 13
2.4.2 Algoritmos não-supervisionados 14
2.5 Análise dos componentes principais 15
2.6 Silhouette 16
2.7 Representação dos dados (TF-IDF) 17
2.8 Apache Spark 18
2.9 Considerações Finais 19

3 Trabalhos Relacionados **21**
3.1 Estado da Arte 21
3.1.1 Métodos de documento central (*Document-pivot*) 22
3.1.2 Métodos de característica central (*feature-pivot*) 23
3.2 Sínteses dos trabalhos relacionados 25
3.3 Considerações Finais 26

4 Abordagem Proposta **27**
4.1 Visão Geral da Abordagem Proposta 27
4.2 Etapa 1: Seleção e Coleta de dados 28
4.3 Etapa 2: Pré-Processamento 29
4.4 Extração de Tópicos 30
4.5 Avaliação da abordagem 31
4.6 Considerações Finais 32

5 Experimentos e Resultados **33**
5.1 Experimento 1 - Operação Lava Jato da Polícia Federal 33
5.1.1 Base de dados 34
5.1.2 Seleção dos Valores de K 34
5.1.3 Extração dos Tópicos 38
5.1.4 Discussão dos Resultados do Experimento I 42
5.2 Experimento 2 - *Impeachment* da Presidente do Brasil 45
5.2.1 Base de Dados 45
5.2.2 Seleção dos Valores de K 46
5.2.3 Extração de Tópicos 47
5.2.4 Discussão dos Resultados do Experimento II 49

6 Considerações Finais **53**
6.1 Limitações da abordagem 54
6.2 Trabalhos Futuros 54

Referências Bibliográficas **55**

Capítulo 1

Introdução

Uma rede social é composta por conjuntos de pessoas ou grupos com algum padrão de contato ou interação entre si [Atefeh and Khreich, 2013]. Uma forma de instanciar o conceito de redes sociais para a interação entre usuários a partir de uma ferramenta online é a Rede Social Online (RSO). As RSO permitem aos usuários compartilharem: fotos, vídeos, mensagens, criar comunidades e manterem-se conectados. Para maior conectividade nas redes sociais online, há um subconjunto de pessoas que utilizam os dispositivos móveis, tais como *tablets* e *smartphones*, facilitando o compartilhamento de dados, independente da circunstância ou localização geográfica [Wives, 2013].

Os usuários das RSO costumam compartilhar informações relacionadas ao contexto no qual estão inseridos, como por exemplo, informações de trânsito, negócios, relatos pessoais e acontecimentos diários. Isso possibilita que as RSO e seus usuários funcionem como sensores para identificação de dados do mundo real, devido a quantidade de informações que serão produzidas diariamente e diversidade dos dados, como: localização, horário, fotos e vídeos [Sakaki et al., 2010].

O Twitter é um exemplo de rede social que disponibiliza informações compartilhadas de usuários[1]. Segundo um levantamento feito em 2016, o Twitter contabilizou aproximadamente 313 milhões de usuários que produziram mais de 400 milhões de *tweets* por dia[2] (Figura 1.1). O Twitter permite que os usuários façam atualizações (*tweets*) com até 140 caracteres, onde o intuito é compartilhar informações curtas e de rápida leitura. Nos últimos anos a plataforma tem sido observada e utilizada para pesquisas devido a recursos fornecidos pelo *microblog* que facilitam o acesso aos dados (API) [Sakaki et al., 2010, Ozdikis et al., 2012, Cheng et al., 2010, Kropivnitskaya et al., 2016, Souza et al., 2016], diferentemente de outras redes socais como o Facebook

[1]https://www.twitter.com

[2]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

e Whatsapp que impossibilitam tal recurso.

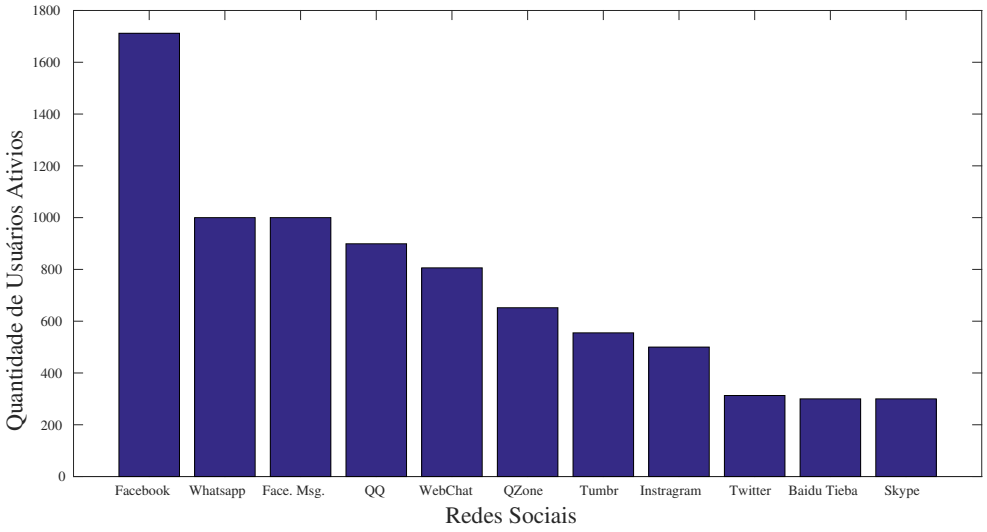


Figura 1.1: Quantidade (milhões) de usuários por redes sociais.

Com a popularidade das redes sociais e a forma como revolucionaram a comunicação entre indivíduos, diversos trabalhos como Sakaki et al. [2010], Ozdikis et al. [2012], Cheng et al. [2010], Kropivnitskaya et al. [2016], Almeida et al. [2016], tem surgido com o intuito de analisar as redes sociais, tendo como foco demonstrar o impacto na sociedade e como pode-se utilizar os dados gerados nesses ambientes para a produção do conhecimento [Ellison et al., 2007].

Nesse contexto, podemos observar que o Twitter é uma das redes sociais que viabiliza essa análise devido à disponibilidade de dados e diversidade de informações. Com isso, é possível monitorar essa rede social para a descoberta de tópicos e eventos de grande escala como por exemplo, movimentações políticas, as eleições nos Estados Unidos, os terremotos no Japão [Sakaki et al., 2010] e demais acontecimentos dentro de um espaço e tempo.

Em se tratando de eventos naturais, ou seja, quando ocorre um evento físico perigoso, tais como sismos, desabamentos, furacões, inundações, incêndios ou qualquer outro fenômeno natural que provoca danos extensos à propriedade ou fazem vítimas, é possível monitorar dados compartilhados no Twitter [Sakaki et al., 2010, Earle et al., 2012, Abel et al., 2012, Kropivnitskaya et al., 2016] para notificar os departamentos de segurança, fornecendo informações sobre quais eventos estão ocorrendo.

Outra aplicação em descoberta de eventos nas redes sociais, é a possibilidade de monitorarmos um determinado acontecimento de grande porte, como por exemplo, o processo de *impeachment* da presidente do Brasil que gerou altos índices de atividades

nas redes sociais (conforme ilustrado na Figura 1.2), e descobrir quais são os tópicos mais relevantes de um determinado dia ou período, a fim de quantificar as informações que tiveram destaque dentro das redes sociais.

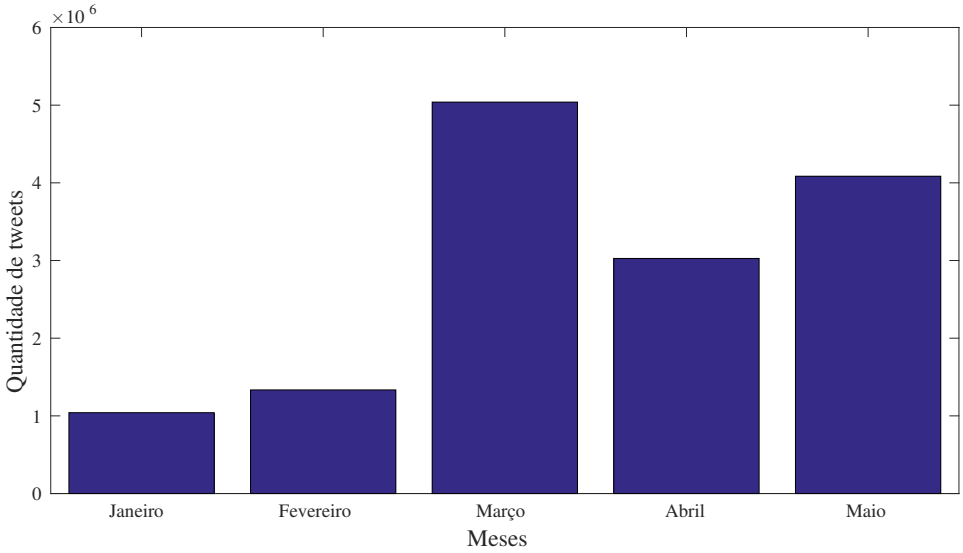


Figura 1.2: Quantidade de *tweets* coletados sobre o processo de *impeachment* da Presidente do Brasil durante o período de Janeiro a Maio [Souza et al., 2016].

1.1 Problema

As redes sociais podem ser utilizadas como uma grande fonte de dados para descoberta de eventos do mundo real, pois como constantemente os usuários estão compartilhando dados sobre diversos aspectos das suas vidas e de acontecimentos que têm influência direta ou indireta no seu cotidiano, estes contribuem para o aumento diário do volume de dados que são gerados nessas redes sociais. Analisando essa larga massa de dados, podemos quantificar um determinado evento e saber quais novos acontecimentos foram gerados por meio do mesmo.

Analisando os trabalhos da literatura, podemos destacar métodos como os que foram criados por Sakaki et al. [2010, 2013], Katragadda et al. [2016], Sankaranarayanan et al. [2009], Katragadda et al. [2016], Petrović et al. [2010], Sayyadi et al. [2009] que desenvolveram técnicas consideradas estado da arte na detecção de eventos, onde os autores tem como objetivo monitorar dados compartilhados nas redes sociais para detectar temas que estejam sendo citados por vários usuários. Essa caracterização é feita através de filtros estatísticos que combinam diversos contextos, a fim de criar grupos (*clustering*) que estejam relatando sobre um determinado tema.

Observamos que as técnicas descritas pelos autores apenas notam a existência de acontecimentos, ou seja, sem fornecer detalhes ou características do evento, pois limitam-se a apenas reconhecer um acontecimento. Outro fator observado, é durante o processo de análise da quantidade de agrupamentos que devem ser criadas para representar o evento. Além disso, observamos que os métodos descritos não apresentam meios que possibilitem manipular dados em larga escala. Esse contexto encontrado nos trabalhos se encaixa no problema de descoberta de tópicos descritos por Allan et al. [1998], onde podemos selecionar técnicas não supervisionadas de Aprendizagem de Máquina para tentar solucionar o problema de como extrair conteúdos relevantes de dados que estejam ligados a eventos de grande porte, a fim de quantificar os tópicos úteis que melhor descrevam os acontecimentos diários, mensais ou por faixa de tempo.

1.2 Objetivos

O objetivo geral deste trabalho é propor e avaliar uma metodologia eficaz, que baseada em técnicas não supervisionadas de aprendizagem de máquina aplicada a dados obtidos das redes sociais online, possamos identificar e quantificar tópicos relevantes de um determinado evento.

Para atingir o que está sendo descrito neste trabalho, os seguintes objetivos específicos devem ser cumpridos:

1. Avaliar a performance dos métodos *Latent Dirichlet Allocation* (LDA), *Non-negative Matrix Factorization* (NMF) e *K-means*, a fim de verificar qual melhor extrai tópicos relevantes das bases coletadas;
2. Verificar a viabilidade do uso de ferramentas de *Big Data* para extração de tópicos relevantes em um cenário *offline*.
3. Verificar/Avaliar se metodologia proposta apresenta resultados superiores as abordagens existentes no estado da arte, através de comparações;
4. Demonstrar a viabilidade de se utilizar o Twitter como fonte de dados para se caracterizar eventos.

1.3 Principais Contribuições

A combinação de técnicas de aprendizagem de máquina, ferramentas de *Big Data*, Redes Sociais Online e eventos de grande porte que ocorreram no Brasil, nos possibilitou gerar as seguintes contribuições:

1. A construção de uma metologia que utiliza ferramentas de *Big Data* (Apache Spark) para extrair tópicos relevantes de uma grande massa de dados, comprovando a viabilidade de se utilizar extração de tópicos em ambientes *offline*;

2. A comparação de técnicas de aprendizagem de máquina na execução de extração de tópicos em diferentes bases de dados;

3. A comparação de arquiteturas de pré-processamento de dados, a fim de demonstrar que ao utilizarmos reconhecimento de entidade nomeada, ao invés de processamento padrão, podemos obter melhores resultados;

4. A utilização da técnica de *Silhouette* para selecionar o valor da quantidade de *clusters* para uma amostra de dados;

1.4 Publicações

O trabalho realizado nesta dissertação permitiu a publicação do artigo: For or Against?: Polarity Analysis in Tweets About Impeachment Process of Brazil President no WEBMedia 2016, onde utilizamos métodos de Análise de Sentimentos como forma de caracterizar um evento [Souza et al., 2016].

Como publicações secundárias tivemos o artigo: Sentiment Analysis of Portuguese Comments from Foursquare, onde aplicamos uma metodologia similar ao artigo já citado, porém em outro domínio de aplicação.

1.5 Organização da Dissertação

O Capítulo 2, apresenta os conceitos necessários para melhor compreensão e entendimento dos aspectos gerais que compõem a abordagem proposta. O Capítulo 3 apresenta revisão da literatura relacionada à área de abrangência deste trabalho. Além disso, apresenta pontos importantes sobre a identificação de tópicos relevantes em redes sociais nos diversos cenários aos quais foram aplicados. O Capítulo 4 apresenta a metodologia, técnicas e ferramentas utilizadas para o desenvolvimento da pesquisa. O Capítulo 5 apresenta os resultados alcançados nas bases de dados, através da aplicação da abordagem proposta. Por fim, o Capítulo 6 apresenta nossas conclusões, limitações do método e trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta conceitos fundamentais para o entendimento da abordagem proposta para o monitoramento e caracterização dos eventos. Na Seção 2.1 serão apresentados conceitos sobre redes sociais. Na Seção 2.2 são apresentadas características gerais do Twitter. Na Seção 2.3 são apresentados os conhecimentos gerais de Descoberta de Conhecimento em base de dados e mineração de dados. A Seção 2.4 apresenta uma breve descrição dos algoritmos de aprendizagem de maquina utilizados. A Seção 2.5 apresenta a técnica de redução de dimensionalidade aplicada a nossa abordagem. A Seção 2.6 apresenta a técnica de *Silhouette*, utilizada para analisar a qualidade dos agrupamentos criados. A Seção 2.7 serão apresentados conceitos para representação dos dados que são fornecidos aos algoritmos de aprendizagem de máquina. Na Seção 2.8 é apresentada a ferramenta *Spark*, utilizada para o processamento de dados em larga escala

2.1 Redes Sociais

Uma rede social são conjuntos de pessoas ou grupos com algum padrão de contato ou iteração entre si [Wives, 2013]. Os padrões de amizade entre esses indivíduos, relações e os casamentos entre famílias são exemplos de redes sociais que foram estudadas durante o passado [Newman, 2003]. Os conceitos de Redes Socais Online começaram a evoluir a partir dos trabalhos de Wellman et al. [1996], onde segundo o autor uma rede social tem um mesmo suporte as pessoas como uma redes de computadores, pois interligam seus usuários assim como as máquinas, assim criando um ambiente de interação.

Partindo do estudo de [Wellman et al., 1996], podemos considerar os usuários que utilizam redes sociais online como fornecedores de dados, pois a partir do momento que realizam seu cadastro, inicia-se o processo de interação, onde os mesmos

podem compartilhar seus pontos de interesses, selecionar amigos e eleger sobre quais temas gostariam de ser notificados. A partir dessas interações, os usuários começam a compartilhar informações como fotos, vídeos, mensagens e suas localizações. Tornando-os também sensores ativos, pois através dos seus dados é possível descobrir opiniões, emoções e outras informações do cotidiano.

Existe uma rede social em especial, o Twitter, que permite essas interações citadas, através de um *microblog* [Java et al., 2007]. Esse serviço derivado dos *blogs*, abreviatura de *web log*, que são exatamente diários de texto, imagens, vídeos e outras mídias. Criando um ambiente que possibilite o compartilhamento de informações curtas e de rápida leitura.

2.2 Twitter

O Twitter é um *microblog*, com mais de 320 milhões de usuários produzindo mais de 400 milhões de *tweets* por dia[1]. O Twitter permite que seus utilizadores façam atualizações, que são chamados de *tweets*, com no máximo 140 caracteres, através da pergunta: *What are you doing?* A limitação dessa quantidade de caracteres, segundo Atefeh and Khreich [2013] é uma característica do *microblog*, que permite que a informação seja fácil de consumir e rápida para divulgar. Os *tweets* são publicados em tempo real, ou seja, a cada atualização que o usuário faça, seja de um *smartphone*, *Web* ou aplicações terceiras, suas informações são compartilhadas para seus seguidores. A portabilidade apresentada pelo Twitter é uma das razões que torna essa rede social bastante popular.

O Twitter é também uma rede social que permite que seus usuários criem seus perfis para que possam interagir com qualquer usuário sem a necessidade de pedir autorização ou uma conexão recíproca por padrão, mas o usuário dispõe de uma funcionalidade de poder modificar as configurações de privacidade do seu perfil, selecionando quem pode visualizar os conteúdos compartilhados e quem pode segui-lo. Esse modelo de rede social cria um relacionamento assimétrico [Brzozowski and Romero, 2011], onde o *microblog* não estipula um número máximo de pessoas que um determinado usuário deve seguir, sendo mais um ponto forte na rede social.

As configurações padrões dos perfis do Twitter, fazem com que as *timelines* (histórico de mensagens do usuário) dos usuários sejam públicas, ou seja, um usuário possa saber o que outro está sentindo ou fazendo de forma aberta, já do ponto de vista cientifico: possibilita a análise dados compartilhados, ou seja, o Twitter se torna uma base de informações, onde podemos observar dados publicados sobre acontecimentos

[1] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

do mundo real, sejam desastres ou eventos sociais, pois está Rede Social Online disponibiliza recursos (API) que permitem a coleta e manipulação de dados compartilhados pelos perfis públicos.

Observando o que foi dito, podemos considerar os usuários do Twitter como sensores sociais, onde seus *tweets* podem ser utilizados para descobrir várias informações úteis, assim como estudar as interações entre usuários como foi feito por Java et al. [2007]. Neste trabalho, como já foi dito, utilizaremos o Twitter como fonte de dados para aplicarmos a nossa metodologia de extração tópicos relevantes sobre um determinado evento de grande porte.

2.3 Descoberta de Conhecimento e Mineração de Dados

A Descoberta de Conhecimento em Base de Dados (DCBD, em inglês *KDD - Knowledge Discovery DataBase*) é uma resposta ao processo de extração de informação em banco de dados, sejam comerciais ou científicos [Miller, 2004]. Com a evolução da tecnologia, aumentou a possibilidade de locais para compartilhar informações, e com isso, novas fontes de dados podem ser usadas para extração de novos conhecimentos. Observando esse contexto, a diversidade de informações fornecidas por esses ambientes, apresentam heterogeneidade entre as fontes de dados, onde técnicas tradicionais de consulta não conseguem recuperar essas informações [Wives, 2013].

O processo de descoberta tem incentivado várias empresas a investirem em tecnologias que sejam capazes de realizar buscas de informações, antes escondidas, e transformá-las em conhecimento útil, onde possam ajudar em tomadas de decisões ou respostas fundamentais para pesquisas ou na busca de conhecimento.

O termo mineração de dados, que popularmente é usado como "descoberta de conhecimento", é apenas um componente, do processo que é a descoberta de conhecimento em base dados.

A mineração de dados tem o objetivo de transformar dados em informação [Wives, 2013]. A DCBD se encaixa como um processo de mais alto nível, onde a mineração é responsável por transformar esses dados em conhecimento. A Figura 2.1 apresenta todas fases da descoberta de conhecimento em base de dados.

A descoberta de conhecimento apresenta a ideia de que uma informação dentro uma base de dados está oculta em formato de padrões. Esses padrões são propriedades e relacionamentos não aleatórios que são válidos, novos, úteis e de análise compreensível [Wives, 2013]. Onde, *Válido* é o padrão genérico o suficiente para ser aplicado a novos

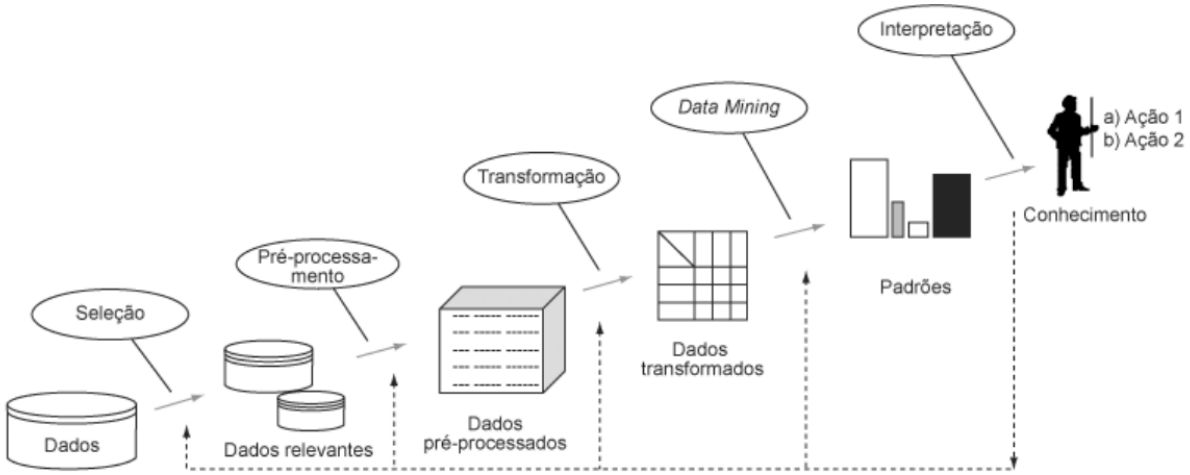


Figura 2.1: Etapas do processo de DCBD [Fayyad et al., 1996].

dados. *Novo*, seria um padrão ainda não identificado ou inesperado. *Útil* é um padrão que leva a tomada de decisão. *Compreensível* é um padrão simples e que possa ser percebido por qualquer indivíduo [Fayyad et al., 1996].

A funções desenvolvidas pelo DCBD ultrapassam as análises estatísticas tradicionais para conseguir acomodar os dados que não são favoráveis a esse tipo de análise. As análises estatísticas geralmente utilizam uma quantidade de dados numéricas sem ruídos. Em contraste, os dados coletados nas empresas são geralmente não numéricos, ruidosos e na sua maioria incompletos. Esses dados também são coletados, para que após a continuação do processo do DCBD, seja possível descobrir novos conhecimentos. A DCBD consiste de princípios e técnicas de estatística, aprendizagem de máquina (*machine learning*), reconhecimento de padrões, busca numérica e visualização científica para receber os novos dados encontrados por meio de técnicas específicas [Miller, 2004].

A descoberta de conhecimento é fortemente indutiva se comparada à análise estatística tradicional. Os modelos estatísticos apresentados atualmente são de confirmação, forçando que o responsável informe qual modelo deseja utilizar, com base em alguma informação na literatura ou alguma tese que já tenha testado essa informação. Por outro lado, os padrões considerados interessantes que são descobertos na mineração de dados são por definição, difíceis de serem conclusivos na sua identificação a priori [Miller, 2004, Wives, 2013].

A DCBD se apresenta como uma ferramenta para explorar domínios que são difíceis de perceber somente com as capacidades de interpretação humana. Observando o que foi dito pela literatura, podemos observar que o DCBD se encaixa melhor na fase

inicial do processo dedutivo, quando o pesquisador formula ou ajusta teorias, baseando-se em fatos ordenados e observações do mundo real. Quando se deseja buscar uma infinidade de informações, a estatística tradicional não pode competir com a ampla, mas difusa, descoberta de conhecimento em base dados. [Miller, 2004, Wives, 2013].

Os algoritmos de mineração de dados amplamente usados são: classificação, regras de associação, identificação de grupos (*clustering*) e a identificação de padrões sequenciais. Esses algoritmos possibilitam solucionar problemas que estão relacionados a tipos de dados específicos e com relações bem definidos.

Um dado que tem se tornado muito popular pelo fácil acesso e que demandam novos algoritmos para sua extração de conhecimento são os oriundos de redes sociais. O grande diferencial desses dados é a forma como existem relações entre as informações, pois como provêm de usuários que utilizam essas redes, a riqueza dos dados se torna ainda maior.

Alguns avanços na área de mineração de dados em redes sociais que podem ser citados são:

- Análise de sentimentos;
- Sistemas de recomendação;
- Formação de comunidades;
- Detecção de tópicos;
- Computação afetiva.

Em suas características essas áreas possuem aplicações de técnicas de aprendizagem de máquina na execução de atividades, e a cada dia novos trabalhos têm explorado essas técnicas de aprendizagem, a fim de avaliar o rendimento desses algoritmos dentro de novos domínios.

2.4 Aprendizagem de máquina

Segundo Alpaydin [2010], aprendizagem de máquina é a programação de computadores para otimização dos critérios de desempenho usados em exemplos de dados ou experiências passadas. Apresentado um modelo definido com alguns parâmetros, o processo de aprendizagem é a execução de um programa de computador que otimiza essas entradas do modelo, usando os dados de treinamento ou experiências passadas, onde esse modelo fornecido pode ser *preditivo* para fazer previsões de valores futuros, ou *descritivos* que podem ganhar conhecimento sobre um dado, ou ambos.

Aprendizagem de máquina utiliza em seus processamentos teoria de construção de modelos matemáticos, pois essas teorias têm como objetivo realizar o princípio de inferência denominado de indução, no qual conseguimos obter conclusões genéricas a partir de um conjunto particular de exemplos fornecidos ao nosso modelo [Monard and Baranauskas, 2003]. O aprendizado indutivo pode ser classificado em dois tipos principais: aprendizagem supervisionada e não-supervisionada. Nesta pesquisa focamos em apresentar e avaliar técnicas que trabalhem com aprendizagem não supervisionada.

Com o propósito de facilitar o entendimento de alguns termos utilizados em aprendizado de máquina, e assim melhorar o entendimento das próximas seções deste trabalho, apresentamos uma lista de conceitos, segundo Mitchell et al. [1997]:

1. *Exemplo (padrão, instância)* - Uma informação única no mundo do qual um modelo poderá aprender, ou sobre qual modelo pode ser usado (exemplo: predição). Muitos trabalhos de aprendizagem de máquina descrevem exemplos através de vetores de características;

2. *Característica (atributo, variável)* - uma quantidade de informações que descrevem um exemplo. Um atributo tem seu domínio definido pelo seu tipo, assim descrevendo quais valores que ele pode assumir.

3. *Vetor de características* - uma lista de características que descreve um exemplo, assumindo que a quantidade de características definida pelo exemplo é proporcional a quantidade de dimensões, ou seja, n-dimensional (conforme ilustrado na Tabela 2.1).

4. *Classe* - no aprendizado supervisionado em problemas de classificação, todo exemplo possui pelo menos um atributo especial denominado rótulo ou classe, que descreve o fenômeno de interesse.

5. *Conjunto de exemplos (dados)* - é composto por um número de exemplos (padrões) com seus respectivos valores de atributos. No caso do aprendizado supervisionado, a cada exemplo também é associada uma classe como apresentado na Tabela 5.4. Onde esse conjunto é divido em dois subconjuntos: **conjunto de treinamento**, utilizado para o aprendizado do conceito e o **conjunto de teste** utilizado para a medição do grau de efetividade do conceito aprendido.

6. *Acurácia (taxa de erro)* - é a taxa de predições corretas (ou incorretas) realizada pelo modelo para um determinado conjunto de dados. A acurácia é geralmente estimada utilizando um conjunto independente de testes, que ainda não foi usado

Tabela 2.1: Conjunto de exemplos no formato atributo valor.

	X_1	X_2	. . .	X_m	Y
T_1	x_{11}	x_{12}	. . .	x_{1m}	y_1
T_2	x_{21}	x_{22}	. . .	x_{2m}	y_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
T_n	x_{n1}	x_{n2}	. . .	x_{nm}	y_n

em nenhum momento durante a fase de aprendizado. Existem outros meios para validar a acurácia por meio mais complexos, como por exemplo, *cross-validation* e *bootstrap*.

7. *Holdout* - o estimador *holdout* divide os exemplos em uma porcentagem fixa de exemplos p para treinamento $(1 - p)$ para teste, considerando normalmente $p > 1/2$.

8. *Cross-Validation* - este estimador é um meio termo entre os estimadores **holdout** e **leave-one-out**, onde os exemplos são divididos aleatoriamente em r partições mutuamente exclusivas (*folds*) de tamanho aproximadamente igual a $\frac{n}{r}$ exemplos. Os exemplos nos $(r - 1)$ *folds* são usados para treinamento e a hipótese indutiva é testada nos *folds* remanescentes. Esse processo é repetido r vezes, cada vez considerando um *fold* diferente para teste.

2.4.1 Algoritmos supervisionados para classificação

A aprendizagem supervisionada tem como entrada um conjunto de dados (x_i,y_j), onde x_i representa um valor de um conjunto fornecido e y_j representa o rótulo atribuído a x_i. Tendo posse dessas informações, deve-se produzir um classificador que formalmente é um exemplo par (x, $f(x_i)$), também chamado modelo, preditor ou hipóteses, que possa ser capaz de predizer um determinado rótulo ($f(x_i)$) para novos dados [Lorena and de Carvalho, 2007].

O processo de tentar prever uma amostra que um classificador faz é chamado de teste. O classificador selecionado também pode ser descrito como uma função f, que tem como entrada um x e realiza uma previsão y. Os rótulos também, chamados de classes, representam o interesse sobre o qual se deseja realizar uma determinada previsão.

Cada instância é descrita por um vetor de características, ou atributos, e um determinado rótulo associado a cada entrada x_i, pois o objetivo da aprendizagem in-

dutiva é construir um classificador que possa determinar de forma correta a classe de novos exemplos que ainda não foram rotulados, ou seja, realizar uma predição.

2.4.2 Algoritmos não-supervisionados

Na aprendizagem não supervisionada não necessitamos de um conhecimento sobre o ambiente que estamos analisando, ou seja, não é necessário que existam exemplos rotulados. Quando estamos trabalhando com esse tipo de aprendizagem, os algoritmos aprendem a representar (agrupar) as entradas que são fornecidas de acordo com uma medida de qualidade.

Neste trabalho propomos a utilização dos seguintes métodos de aprendizagem de máquina não-supervisionada para extrair tópicos da base de dados: *Latent Dirichlet Allocation*, *K-means* e *Non-negative Matrix Factorization*. As implentações consideradas neste trabalho são as disponíveis na API *Scikit learn*[2]. Abaixo os métodos citados são brevemente descritos.

***Latent Dirichlet Allocation* (LDA)**. Este método consiste em um modelo probabilístico, no qual cada documento é modelado como uma combinação de tópicos, onde cada tópico corresponde a uma distribuição multinomial sobre as palavras [Blei et al., 2003b, Bolelli et al., 2009].

A distribuição documento-tópico e tópico-palavra aprendidas pelo LDA, por meio de inferência Bayesiana, descrevem os melhores tópicos por documento e as palavras mais descritivas para cada tópico. Uma definição formal para um tópico, é uma distribuição sobre um vocabulário fixo. Por exemplo, o tópico da genética tem palavras sobre genética com alta probabilidade e o tópico de biologia evolutiva tem palavras sobre biologia evolutiva com alta probabilidade. Matematicamente, podemos representar o LDA de acordo com a Equação 2.1, onde a matriz $\beta = [\beta_{ij}]$ é a responsável pela parametrização da distribuição e β_{ij} significa a probabilidade de observarmos uma palavra chave j em um tópico i. Com isso, é possível observar que o LDA assume que sempre que uma palavra w for observada, a mesma estará associada a um tópico z de tamanho N.

$$P(\omega|\alpha,\beta) = \sum_{z} P(\omega|z,\beta)P(\omega|z,\alpha) \tag{2.1}$$

O ***K-means*** tem como principal objetivo dividir os M pontos em N dimensões dentro de K agrupamentos. Neste contexto, seja X = x_1,x_2,...,x_n o conjunto de n pontos d-dimensionais a serem agrupados em um conjunto de K *clusters*, C = c_k, k = 1,...,K.

[2]http://scikit-learn.org/stable/

O *K-means* encontra um partição de tal forma que o erro quadrático entre a média empírica de um *cluster* e os pontos no *cluster* sejam minimizados. Seja u_k a média de um cluster c_k. O objetivo do K-means é minimizar a soma do erro quadrático de todos os K *clusters* [Jain, 2010]. Matematicamente podemos representá-lo de acordo com a Equação 2.2:

$$J(C) = \sum_{k=1}^{k} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{2.2}$$

Embora o K-means tenha sido proposto em 1955, este algoritmo ainda é utilizado devido a sua eficiência, simplicidade e fácil implementação [Jain, 2010].

***Non-negative Matrix Factorization* (NMF)**. O NMF consiste em um algoritmo de fatoração de matrizes que encontra a fatoração positiva de uma matriz positiva recebida como entrada [Xu et al., 2003]. Seja S o conjunto de documentos representados por uma matriz V de dimensão mxn, onde m é o número de termos distintos e n é equivalente ao número de documentos em S. O NMF encontra a menor aproximação de V em termos de alguma métrica específica fatorando V no produto (WH) de duas matrizes de menor dimensões W (mxk) e H (kxn), onde k é o número de tópicos. Cada coluna de W é um vetor base que contém uma codificação semântica ou um conceito de V e cada coluna de H contém uma codificação da combinação linear dos vetores bases que aproxima a correspondente coluna de V [Shahnaz et al., 2006].

2.5 Análise dos componentes principais

A análise dos componentes principais (do inglês, *Principal Components Analysis* - PCA) é um método não supervisionado que utiliza processamento ortogonal para converter componentes que possivelmente estão relacionados em conjunto de valores linearmente separáveis denominados de componentes principais. Em suas aplicações, esses métodos estão interessados em encontrar um mapeamento a partir das entradas no espaço d-dimensional para um novo espaço ($k < d$)-dimensional, com perda mínima de informação [Alpaydin, 2010].

Matematicamente, o PCA é definido por uma matriz X^T, onde cada linha n dessa matriz representa um conjunto de características, e cada uma das m colunas é um documento particular da base de dados. A decomposição em valores da matiz X é representada por $W\Sigma V^T$, onde W é a matriz de autovetores da matriz de covariância XX^T, Σ é uma matriz diagonal quadrada retangular e V é a matriz de autovetores de X^tX. Assim, o PCA que preserva a dimensionalidade pode ser representado pela

Equação 2.3:

$$Y^T = X^T W = V\Sigma^T W^T W = V\Sigma^T \tag{2.3}$$

O PCA é comumente usado para redução de dimensionalidade em dados, pois apresenta melhor transformação de ortogonalidade, porém essas técnicas acabam se tornando custosas [Jolliffe, 2002].

2.6 Silhouette

Silhouette é um método de interpretação e validação da consistência dentro de agrupamentos de dados [Rousseeuw, 1987]. Sua análise pode ser usada para estudar a distância de separação entre os *clusters* resultantes de uma determinada base de dados. Em sua definição é assumido que os dados são agrupados por uma técnica que tenha como parâmetro de entrada um conjunto de K *clusters*, onde para cada dado i, seja $a(i)$ a média de não similaridade de i em relação aos outros dados do mesmo *cluster*. Com isso, podemos dizer que $a(i)$ é uma representação de quão bem i é atribuído a um *cluster* c, ou seja quanto menor o valor da representação melhor será a associação [Rousseeuw, 1987]. Definimos então a diferença de similaridade média do ponto i para um *cluster* c, como sendo a média da distância de i para todos os outros pontos de c.

Seja $b(i)$ a menor média de não similaridade de i em relação a qualquer outro *cluster* que o dado não seja membro. Com isso podemos definir um *Silhouette* através da Equação 2.4:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.4}$$

Outra forma de representar a Equação 2.4 é através da Equação 2.5

$$s(i) = \begin{cases} 1 - a(i)/b(i), & Se \quad a(i) < b(i) \\ 0, & Se \quad a(i) = b(i) \\ b(i)/a(i) - 1, & Se \quad a(i) > b(i) \end{cases} \tag{2.5}$$

A técnica conforme observamos Equação 2.5 apresenta um intervalo de valores de [-1, 1]. Esses coeficientes quando estão perto de +1 indicam que a amostra está longe dos *clusters* vizinhos. Um valor de 0 indica que a amostra está perto do limite

de decisão entre dois *cluster* vizinhos e valores negativos indicam que essas amostras podem ter sido atribuídas ao *cluster* errado[3].

O gráfico de *Silhouette* conforme exemplificado na Figura 2.2 exibe uma medida de quão próximo cada ponto em um *cluster* está em relação a seus vizinhos e, portanto, fornece uma maneira de avaliar visualmente os parâmetros como o número de *clusters*.

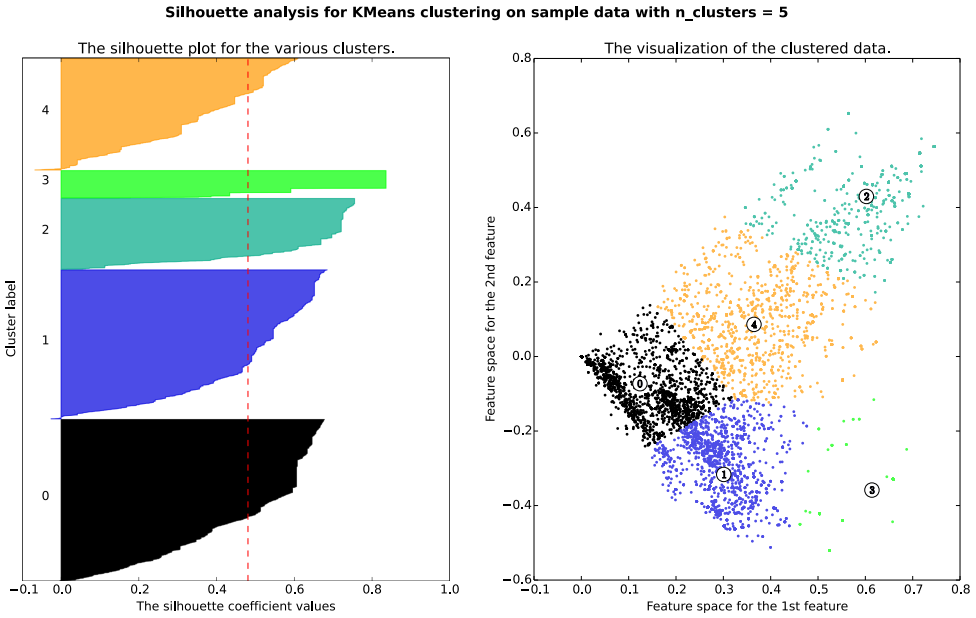


Figura 2.2: Exemplo do gráfico de *Silhouette* para seleçao de valores de k.

2.7 Representação dos dados (TF-IDF)

O modelo TF-IDF (*Term Frequency – Inverse Document Frequency*) é uma medida estatística que gera uma representação de documentos na forma vetorial [Salton and Buckley, 1988]. A frequência de um termo TF em um documento i é definida como sendo o número de vezes que o termo aparece no texto:

$$\mathrm{Tf}_{i,j} = \frac{f_{i,j}}{\sum_{k=1}^{K} f_{k,j}}, \tag{2.6}$$

onde $f_{i,j}$ é a quantidade de ocorrências no termo t_i no texto e o somatório (K) é a quantidade de termos distintos (número de termos no documento). Já a frequência

[3]http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

do documento invertida (IDF) mede a importância geral que o termo possui no texto. O IDF é definido como o logaritmo do quociente entre o número total de textos D e o número de textos que contém o termo t_i:

$$\text{IDF}_i = \log \frac{D}{|D_{t_i}|}, \tag{2.7}$$

onde D_{t_i} é o conjunto de textos que contém o termo t_i, e D é o tamanho desse conjunto de textos. O peso TF-IDF é calculado através do produto entre a Equação 2.6 e a Equação 2.7:

$$\text{TF}_{i,j} \times \text{IDF}_i \tag{2.8}$$

O peso TF-IDF é diretamente proporcional à frequência do termo no texto e inversamente proporcional à frequência do termo no conjunto de textos [Witten et al., 1999].

2.8 Apache Spark

Um dos grandes desafios do nosso trabalho é como processar uma quantidade de dados que tínhamos coletados sobre eventos que tinham acontecido no Brasil. Inicialmente observamos a biblioteca do SciKit-learn[4], que auxilia na aplicação de algoritmos de aprendizagem de máquina, mas não dava suporte a massa de dados que possuíamos. Com isso, começamos o estudo de ferramentas que facilitassem a criação de aplicações para se trabalhar com *Big Data*. Durante essa fase encontramos o Apache Spark, que é ferramenta para o desenvolvimento de aplicações que utilizam grandes quantidades de dados. O objetivo desta ferramenta é aumentar a velocidade de processamento dessas aplicações, que anteriormente não podia ser alcançado.

Com o Spark, introduziu-se o conceito de *Resilient Distributed Datasets* (RDD) [Zaharia et al., 2012] que dão suporte a aplicações que reusam os resultados gerados por outras iterações. Os RDD, segundo Correia [2016], são um conjunto de partições que podem ser armazenadas em memória volátil ou em memória persistente e que só podem ser criadas a partir de transformações, onde cada RDD é capaz de saber como foi criado a partir do seu histórico de mudanças. Este histórico serve para assegurar a tolerância a falhas, pois caso aconteçam erros, é possível através do seu histórico recriar essa RDD.

Esses RDD devem ser usados em aplicações que utilizam a mesma operação para todo o conjunto de dados. Com isso, visto que boa parte dos algoritmos de aprendi-

[4]http://scikit-learn.org/

zagem aplicam várias operações iterativamente, podemos observar que os algoritmos apresentam um desempenho melhor quando todos os dados estão em memória [Correia, 2016].

O desempenho do Spark é uma das múltiplas vantagens que ele apresenta. A configuração do ambiente é executada apenas baixando uma versão compilada do Spark e criando uma ligação com máquina Mestre (*Driver Program*). Quando o ambiente estiver interligado, basta somente submeter uma aplicação e a mesma será distribuída pelo *cluster*, conforme ilustrado na figura 2.3.

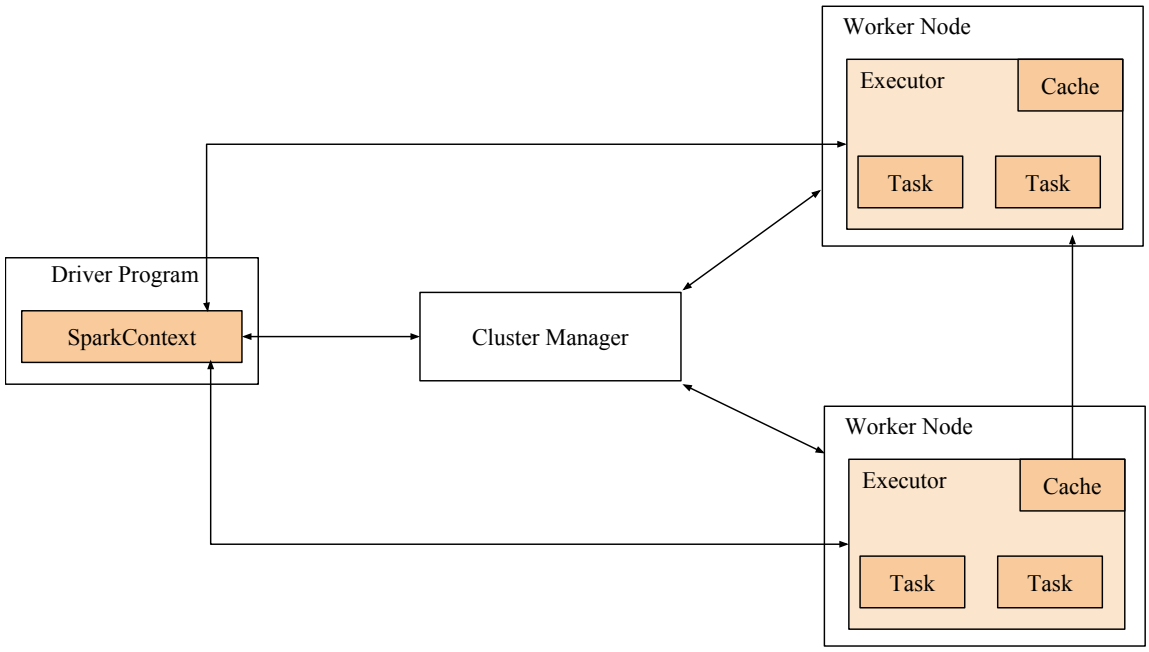


Figura 2.3: Arquitetura do Spark [Correia, 2016].

O *Driver Program* gerencia cada aplicação, designando tarefas aos *Worker Nodes*, que são responsáveis pelo processamento da aplicação. O *Cluster Manager* é responsável pela comunicação entre *Driver Program* e *Worker Nodes*.

Outro fator que contribuiu na escolha do Spark foi devido a ferramenta disponibilizar uma API para se trabalhar com aprendizagem de máquina (MLib)[5].

2.9 Considerações Finais

Neste capítulo foi apresentado um resumo dos conceitos teóricos necessários para entender o desenvolvimento de nosso trabalho. As diversas escolhas realizadas para definir

[5]http://spark.apache.org/docs/latest/mllib-guide.html

os métodos baseiam-se nestes conceitos. O ferramental teórico estudado possibilita entender os trabalhos relacionados com nossa pesquisa (Capítulo 3).

Nos próximos capítulos são apresentados os trabalhos relacionados, nosso método, os resultados da experimentação, os estudos de casos e as conclusões.

Capítulo 3

Trabalhos Relacionados

Neste capítulo é apresentado um resumo das soluções recentes (de 2003 a 2016) para o problema de identificação e extração de tópicos de eventos em redes sociais. O conjunto de oito trabalhos revisados abordam técnicas para identificar informações relevantes no Twitter por meio de agrupamentos de dados com técnicas não supervisionadas de aprendizagem de máquina. Entre alguns dos desafios enfrentados pelas abordagens estão: (1) a eliminação de ruídos existentes nas bases de dados e; (2) a maximização da taxa de acerto no reconhecimento desses termos.

3.1 Estado da Arte

O problema de como detectar eventos começou a ser estudado originalmente em 1998, sendo tratado como um problema de descoberta de tópicos [Allan et al., 1998]. Os autores propuseram um estudo piloto de comparação das abordagens *UMass*, *Dragon* e *CMU* para descobrir novos acontecimentos em um *streamming* de dados. Os experimentos foram executados em uma base de dados com 16.000 notícias do canal CNN[1]. Os resultados, apesar de apresentarem um percentual de acerto de 80% na abordagem *UMass*, apresentavam limitações no reconhecimento de novos eventos quando era necessário processar uma grande escala de dados.

Com base nessas ideias começaram a surgir novos trabalhos que relatam métodos e modelos para se extrair tópicos de redes sociais com intuito de gerar conteúdo útil [Sakaki et al., 2010, 2013, Petrović et al., 2010, Diao et al., 2012]. O conceito de tópico ou evento segundo Sakaki et al. [2010], é um acontecimento que decorre dentro de um determinado espaço e tempo, onde podem ser extraídas informações para previsão e caracterização de acontecimentos do mundo real.

[1] `http://edition.cnn.com/`

A detecção de tópicos tem como objetivo o agrupamento de informações de maneira contextualizada. Portanto, dado um conjunto A de textos, existe uma quantidade B de contextos (tema) associados a estes textos. Neste caso, B é menor ou igual a A, visto que mais de um texto pode discorrer sobre o mesmo tema. Em redes sociais, a detecção de tópicos enfrenta desafios adicionais, tais como: (i) a quantidade de informações geradas; (ii) e os ruídos existentes, visto que nem todas as informações compartilhadas sobre um determinado tema são relevantes.

Os métodos de detecção de tópicos podem ser classificados em duas categorias: métodos de documento central e métodos de característica central, abordados nas próximas seções.

3.1.1 Métodos de documento central (*Document-pivot*)

O métodos de documento central (em inglês, *Document-Pivot methods*) detectam tópicos baseado na distância semântica entre dois documentos. Neste contexto, o trabalho desenvolvido por Phuvipadawat and Murata [2010] utiliza o método de documento central, combinado com reconhecimento de entidade nomeada na representação dos documentos, para fornecer um sistema de reconhecimento de novas notícias no Twitter, onde cada *tweet* recuperado é representado por uma *bag-of-words* em conjunto com o *term frequency - inverse document frequency* (TF-IDF). Os documentos são mesclados de forma incremental considerando a similaridade textual entre os *tweets* de entrada e os *clusters* existentes. Tal abordagem também foi utilizada por Becker et al. [2011] para classificar se uma informação que está sendo citada nas redes sociais é ou não um tópico relevante. Os autores concluíram que sua abordagem tem um potencial de processamento para detectar tópicos relevantes, e propõem como trabalhos futuros, utilizar as emoções dos usuários ao escrever seus textos, pois assim podem entender a percepção dos usuários em relação aos tópicos extraídos.

Na tarefa de *First Story Detection* (FSD) discutida por Petrović et al. [2010], considerado um dos trabalhos mais completos na área de *Document-pivot*, tinha como objetivo detectar o primeiro documento que estivesse relatando um tópico relevante em uma base de dados representativa. Para alcançar este resultado os autores utilizaram o algoritmo *Locality Sensitive Hashing* (LSH) para detectar os documentos que possuíam baixa similaridade com os *clusters* existentes. Os autores demonstraram que é possível conseguir tempo de processamento constante em uma base em larga escala com 160 milhões de *tweets*, além de resultados representativos.

Um outro tipo de estratégia engloba as abordagens baseadas em grafos. O trabalho desenvolvido por Sayyadi et al. [2009] analisa as regiões dos grafos de palavras-chave

a fim de detectar novos eventos. Os documentos, ao qual os autores se referem como palavras-chave, têm o seu TF-IDF calculado, eliminando documentos que possuem baixa frequência. A partir destas frequências, um novo nó para um grafo é criado para cada palavra-chave do texto de entrada, e um limite é criado entre os nós com palavras-chaves em comum. Cada limite do grafo precisa ter uma quantidade mínima de palavras-chave, caso contrário é desprezado, além de que ele também deve possuir uma probabilidade acima de um limite determinado. Esse valor é a probabilidade de uma palavra-chave de um nó existir no outro nó do limite

Apesar dos métodos apresentados possuírem uma abordagem fundamentada e bem consolidada, recentemente têm surgido publicações como Diao et al. [2012], Cataldi et al. [2010], Suh et al. [2016], Katragadda et al. [2016] que demonstram uma abordagem de seleção de tópicos relevantes por palavras-chave. Com isso, a abordagem descrita neste trabalho se fundamenta nesta estratégia, pois levamos em consideração os resultados apresentados pelos trabalhos já citados, que superam a abordagem por *Document-pivot*.

3.1.2 Métodos de característica central (*feature-pivot*)

Esses métodos utilizam palavras-chave para criar um agrupamento de termos e realizar a detecção de tópicos [Diao et al., 2012, Cataldi et al., 2010]. Originalmente, eram utilizados os métodos de *Document-pivot* para a detecção de novos tópicos [Brants et al., 2003]. Recentemente, abordagens usando modelos probabilísticos para detectar agrupamento de termos (*feature-pivot*) como, por exemplo, o *Latent Dirichlet Allocation* (LDA) que é considerado estado da arte [Blei et al., 2003a], estão sendo utilizadas como forma de extrair tópicos relevantes [Bolelli et al., 2009].

Diao et al. [2012] aplicam o LDA para identificar tópicos em dados do Twitter. Os autores assumem que cada *tweet* representa um único tópico, enquanto um tópico é feito de palavras-chaves de múltiplos *tweets* . O modelo proposto consiste em simultaneamente capturar duas observações: (1) *tweets* publicados que tenham o mesmo tópico, e; (2) *tweets* publicados por um mesmo usuário referentes ao mesmo tópico. Os autores assumem que possuem um *stream* D de *tweets*, denominados como $d_1, d_2, .., d_D$. Cada *tweet* d_i é gerado por um usuário u_i, onde u_i é um índice entre 1 e U, e o U é o total de números de usuários.

O sistema proposto foi aplicado e analisado em um base de dados com 3.967.927 *tweets* no período de 01/09/2011 a 30/11/2011, apresentando como resultado uma média de acerto de 83.33%, onde os autores tratavam eventos esportivos, políticos e econômicos que haviam ocorrido neste intervalo de tempo e queriam comprovar que

seria possível representar todo tipo de evento com apenas 3 tópicos. Porém em seus resultados, apesar de representativos, podemos observar que o modelo não trata ruídos, o que ocasionou a detecção de tópicos irrelevantes.

Xu et al. [2003] confrontaram o *Non-negative Matrix Factorization* (NMF) com as técnicas de clusterização *Latent Semantic Indexing* (LSI) e *Spectral Clustering* (SC). O estudo mostrou que para os métodos LSI e SC, a possibilidade de capturar direções que nem se quer pertencem a região onde os documentos estão semanticamente agrupados é sempre real. Uma consequência disto, é que os métodos atribuíam pesos sem valor e impossibilitavam uma correta associação entre *clusters* que possuem relação entre si. Para o NMF essa possibilidade é falsa, visto que não há necessidade de haver ortogonalidade entre os vetores das matrizes dos documentos associados, gerando assim direções sempre presentes na "região semântica". Com isso, os autores identificam o NMF, como método que realiza extração de forma mais confiável, apresentado uma média de 75% de acerto nos tópicos extraídos quando comparado com os demais algoritmos.

Zhao et al. [2011] descrevem uma comparação entre os meios de comunição tradicionais e as redes sociais, com o intuito de demonstrar que é possível utilizar as redes sociais como fonte segura de informações. Neste contexto, os autores propõem um modelo de extração de tópicos em 3 níveis (Descobrir o tópico, categorizá-lo e identificar o tipo de tópico). Durante o processo de descoberta de tópicos os autores demonstraram uma adaptação do LDA para processamento de *tweets*, onde em seus resultados é possível observar que sua adaptação apresenta um resultado de 0,17% melhor que o LDA tradicional.

Jin et al. [2011] descreveram a implementação de uma abordagem que supera as limitações de extração de tópicos em textos curtos do LDA. Além disso, os autores demonstram que sua adaptação apresenta bom rendimento em textos curtos e longos, quando comparado com os outros métodos. Em seus experimentos, os autores aplicaram os métodos em 3 bases de dados distintas, utilizando métricas como Entropia e *Normalized Mutual Information* (NMI) para avaliar a qualidade do agrupamento dos dados. Ao analisar a abordagem utilizada para encontrar a quantidade de tópicos de uma determinada amostra de dados os autores concluíram que sua estratégia é ineficiente em relação aos demais métodos, porém sua abordagem apresenta uma entropia de pelo menos 0,128% menor que os demais métodos.

Suh et al. [2016] capturaram tópicos locais de sua região, a fim de detectar informações relevantes. Os autores compararam o NMF com suas variantes e o LDA. Eles concluíram que o modelo proposto superava os demais, tanto na captura de tópicos relevantes, como também na velocidade de processamento, tornando o método viável para captura de tópicos em tempo real e em larga-escala, apresentando um resultado

que aponta uma média de 40% de acerto maior que os variantes do LDA. Os autores concluíram que a adaptação feita em sua abordagem consegue detectar tópicos em bases de dados que algoritmos tradicionais não conseguem.

Chakraborti and Dey [2016] avaliaram como o algoritmo *K-means* se comporta com uma abordagem multi-nível. A cada nível, calcula-se a quantidade de grupos de documentos, de modo que no fim desta etapa infere-se novamente um valor de agrupamento que também conta com os *cluster* não convergentes. A abordagem se mostrou válida para o uso em aplicações de negócios e outros cenários que possuam uma base de dados em larga escala. Com estas abordagens de gerenciamento de grandes base de dados, podemos avaliar as principais informações extraídas de uma base textual e obter os assuntos, tópicos e entidades relacionadas. Essa abordagem descrita pelos autores, demonstra como reduzir o erro na previsão do valor ideal de *clusters* para uma determinada base de dados, e com isso melhorar a qualidade dos agrupamentos criados.

O trabalho de Katragadda et al. [2016] utiliza um modelo de análise de tópicos (TDT) para detectar eventos no Twitter. Os autores mostram os desafios de se aplicar esta estratégia num método de detecção em tempo real. Ao ocorrer um evento, observou-se que algumas palavras são constantemente publicadas, gerando uma anomalia. Tendo identificado as palavras com comportamento não usual, o modelo extrai os eventos dos grafos que contêm essas palavras, já pré-processadas, com uma combinação de métodos de extração de seções do grafo e de clusterização, tornando assim o evento detectável. Os autores concluíram que seu modelo é capaz de detectar o que é um novo evento e o que é a continuação de um acontecimento. Vale ressaltar, que os autores destacam o baixo custo computacional do seu modelo, além de conseguir detectar tópicos em longos períodos de tempo.

3.2 Sínteses dos trabalhos relacionados

A Tabela 3.1 resume os aspectos fundamentais dos trabalhos relacionados neste capítulo, como forma de representação dos dados, algoritmos utilizados e média de acerto na tarefa de detecção de tópicos. Os trabalhos indicados com * foram utilizados como *baseline*, a fim de comparar os resultados. O trabalho apresentado com "–" nos resultados não deixou explícito o valor alcançado.

Autor	**Algoritmo**	**Métrica**	**Resultado**
Xu et al. [2003]*	*NMF*	Acurácia	89%
Sayyadi et al. [2009]	*cosine similarity*	Acurácia	–
Petrović et al. [2010]	*LSH*	DET	0.70
Zhao et al. [2011]*	*Twitter-LDA*	Avaliação Manual	67.5%
Jin et al. [2011]	*γ-DLDA*	Entropia	0.553
Diao et al. [2012]	TimeUserLDA	Precisão de K	1.00
Chakraborti and Dey [2016]	*K-means*	APS	0.10
Katragadda et al. [2016]	*EDO*	*F-Score*	61%

Tabela 3.1: Resumos dos trabalhos relacionados.

3.3 Considerações Finais

Os trabalhos atuais citados ao longo deste capítulo apresentam diversas abordagens na tarefa de detecção de tópicos em redes sociais. No entanto, não foi definida nenhuma abordagem que tenha sido aplicada à tarefa de caracterizar eventos, ou seja, quantificar todos os tópicos relevantes sobre um determinado acontecimento dentro de um espaço e tempo e que seja capaz de processar dados em larga escala. A dificuldade para definir essa aplicação ocorreu em virtude do custo para ser executado em memória principal, ou seja, a preferência por processamento em tempo real é uma alternativa dos autores para este problema. Além disso, modelos como Jin et al. [2011], Zhao et al. [2011] apresentam formas ineficientes para seleção da quantidade de agrupamentos para uma determinada amostra de dados, ou seja, em sua maioria os valores selecionados são acima da quantidade de tópicos que se deseja descobrir.

Em nossa proposta selecionamos dois eventos de larga escala para serem analisados em uma abordagem derivada dos trabalhos apresentados por Xu et al. [2003], Zhao et al. [2011], Diao et al. [2012], onde são utilizados os algoritmos propostos nesses trabalhos para execução da extração tópicos. Outros fatos que podem ser destacados como diferencial em nossa abordagem é a utilização da técnica de *Silhouette* para identificar o melhor valor de tópicos em um conjunto de dados, e a capacidade de nossa arquitetura conseguir processar uma quantidade de dados, justificando a utilização da ferramenta *Spark*, pois os dados utilizados em nossos experimentos encontram-se em escala de *Big-data*. Vale ressaltar que os trabalhos apresentados possuem métricas diferentes para avaliar suas abordagens. Em nossa avaliação de tópicos extraídos, utilizamos uma abordagem similar a proposta por Zhao et al. [2011], onde comparamos os tópicos extraídos com canais de notícias e retiramos uma acurácia desta comparação, visto que o autor desenvolveu um trabalho similar ao proposto neste trabalho.

Capítulo 4

Abordagem Proposta

Nesta capítulo iremos apresentar a descrição da solução que está sendo utilizada ao longo deste trabalho, a fim de quantificar os tópicos mais relevantes dos seguintes eventos: as 15 fases da Operação Lava Jato em 2016 e do decorrer do processo de *impeachment* da presidente da República Federativa do Brasil. Além de demonstrar como foi feita a validação da nossa abordagem.

4.1 Visão Geral da Abordagem Proposta

Nossa abordagem consiste em extrair tópicos relevantes que estejam relacionados a um determinado evento de larga escala, utilizando dados das redes sociais. Através desse trabalho demonstramos que é possível monitorar e analisar a grande massa de dados fornecidos pelas redes sociais, para descobrir informações sobre a repercussão de um determinado evento e que novos fatos surgem, sejam de impacto mundial, nacional ou local.

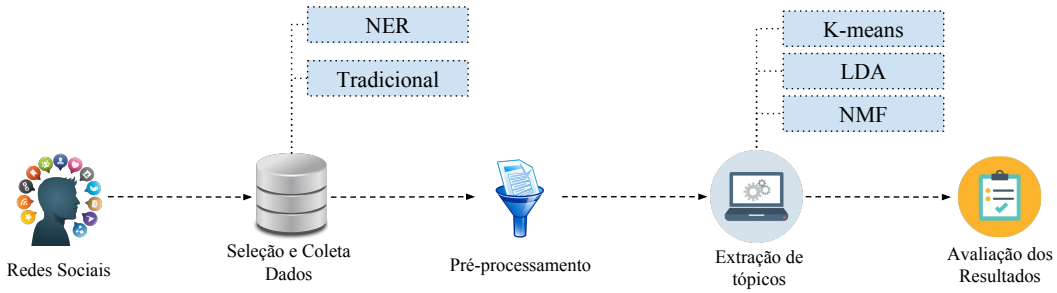


Figura 4.1: Arquitetura da abordagem proposta.

A abordagem proposta na elaboração deste trabalho consiste das seguintes etapas

(ilustradas na Figura 4.1):

1. Seleção e Coleta de dados provenientes do Twitter para construção da base de dados;

2. Pré-processamento dos dados em larga escala, a fim de retirar possíveis ruídos dos *tweets*, além de eliminar atributos irrelevantes ao modelo;

3. Extração de tópicos pelos métodos de aprendizagem de máquina não-supervisionados;

4. Avaliação dos Resultados.

4.2 Etapa 1: Seleção e Coleta de dados

O Twitter possui, além dos serviços que são destinados ao usuário final, diversas APIs que auxiliam os desenvolvedores a integrarem serviços do *microblog* às suas aplicações. Esse ecossistema de desenvolvimento fornecido pelo Twitter o torna uma das redes sociais mais utilizadas para pesquisas e outras aplicações. Podemos destacar a API do Twitter como uma das principais alavancas do serviço, pois nos primeiros anos gerava 91% dos *tweets* [Wives, 2013].

Neste trabalho, estamos utilizando a *Search API*[1], que nos permite coletar dados dos últimos 7 dias sobre eventos que estão ocorrendo ou já ocorreram. A *Search API* permite identificar *tweets* específicos que estejam relatando um determinado acontecimento, através de palavras chaves e outros parâmetros que podem ser fornecidos. E por meio desse serviço, podemos recuperar diversas informações que são úteis. Durante a fase de coleta, identificamos as palavras-chaves referentes aos eventos, limitamos a coleta a apenas *tweets* em português e criamos um *script* que armazenava toda estrutura de dados fornecida pela API, mas como nossa abordagem se baseia em análise textual de informações compartilhadas nas redes sociais, selecionamos os seguintes itens, baseado nos trabalhos já citados:

- Nome de identificação do usuário na rede social;
- Data de criação do *tweets*;
- Texto dos *tweets*;
- Geo-localização.

[1]https://dev.twitter.com/rest/public/search.

Um dos problemas encontrados durante a fase de coleta de dados, foram *tweets* que apresentavam as mesmas palavras que identificamos como chaves de um determinado evento (as 14 fases da Operação Lava Jato), mas não estavam relatando o mesmo acontecimento, como por exemplo "*Vou levar meu carro no melhor Lava Jato de SP* ". Para isso precisamos garantir que os dados estão sendo fornecidos para o modelo de aprendizagem, estejam unicamente relatando o evento em questão. Para identificar esses ruídos, observamos que tais informações fora do contexto apresentavam a mesma sentença em todos os *tweets* (swarm.com). Com isso, foi necessário criar uma fase de pré-processamento que terá como objetivo eliminar esses dados ruidosos, além de descartar termos não representativos aos algoritmos, como: preposições, artigos, *URLs*, menções e emoticons.

4.3 Etapa 2: Pré-Processamento

Após a etapa de coleta dos dados, os *tweets* foram pré-processados para remoção de possíveis ruídos e redução do número de atributos dos vetores de características dos métodos não-supervisionados, já detalhados no Capítulo 2. Para tanto, utilizamos o *framework* Apache Spark[2] que facilita a manipulação de dados em larga escala (*Big Data*). Tal ferramenta foi necessária porque a quantidade de dados não permitia uma execução dos algoritmos em memória principal.

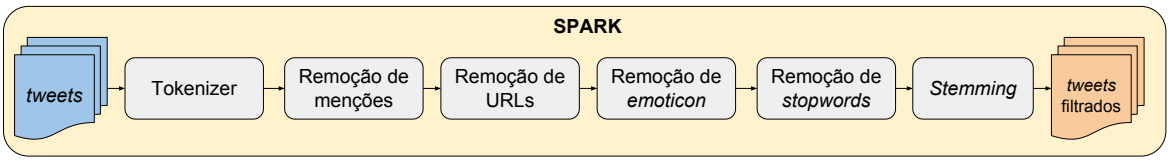


Figura 4.2: Pré-processamento tradicional.

Neste trabalho, realizamos dois pré-processamentos distintos (tradicional e com reconhecimento de entidade) com o objetivo de verificar se há melhora na eficiência dos métodos ao extrair tópicos. O pré-processamento tradicional (Figura 4.2), consiste em separar os *tweets* em *tokens* (segmentos de sentenças), para logo em seguida, remover as menções, URLs e *emoticons* de cada um deles. Posteriormente, os *tokens* são normalizados, isto é, são submetidos a transformações (e.g., tratamento de pontuação, limpeza de caracteres especiais), correções (e.g., ortográficas, morfológicas) e expansão de contrações [Stiilpen Junior and Merschmann, 2016]. Após a normalização, aqueles *tokens* que forem *stopwords* (palavras que podem ser considerados irrelevantes para

[2] `http://spark.apache.org/`

o contexto estudado) são descartados e, por fim, os afixos dos *tokens* restantes são eliminados (*stemming*).

O pré-processamento com reconhecimento de entidade (Figura 4.3), por sua vez, consiste em primeiramente remover as menções e as URLs dos *tweets*. Em seguida, as entidades (substantivos próprios) presentes nos *tweets* são extraídas por meio do algoritmo *Polyglot-Ner* [Al-Rfou et al., 2015]. Tal algoritmo modela a tarefa de reconhecimento de entidades como um problema de classificação a nível de palavra.

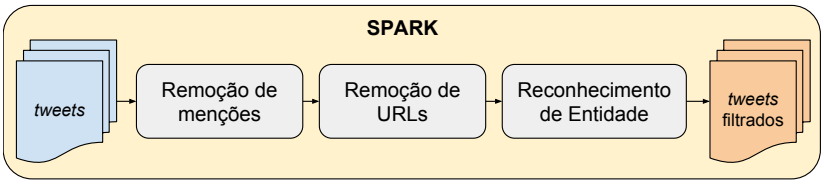


Figura 4.3: Pré-processamento com reconhecimento de entidade.

Após a aplicação do pré-processamento sobre a base de dados, o vetor de características de cada um dos métodos não-supervisionados é modelado como uma *bag-of-words*, em que cada *tweet* t é representado por um vetor p_1, p_2,..., p_n, onde p_i é a frequência ponderada da palavra i no *tweet* t normalizada pela frequência da palavra i na base de dados. Tal modelagem de vetor de característica é análoga a proposta por Aisopos et al. [2012]. Entretanto, em nossa modelagem, caso a frequência ponderada de uma palavra i no *tweet* t seja menor que vinte, tal palavra é eliminada da *bag-of-words*, isto permite a redução da dimensão do vetor de características e reduz a influência de ruídos (termos de pouca relevância). Tal valor foi utilizado, pois após vários testes para seleção desta quantidade de *clusters*, observamos que em ambas as estratégias de pré-processamento, tínhamos palavras ou entidade não representativas, que após serem eliminadas, não apresentavam alteração sobre os tópicos extraídos. Vale ressaltar, que assumimos como premissa que uma informação é relevante se possuir uma elevada frequência na base de dados.

Após eliminado os termos indesejáveis, cálculos das frequências dos documentos e construção do vetor de características, submetemos ao processo de treinamento e extração dos tópicos relevantes na base de dados.

4.4 Extração de Tópicos

Neste trabalho utilizamos os seguintes métodos de aprendizagem de máquina não-supervisionada para extrair tópicos da base de dados: *Latent Dirichlet Allocation*, *K-means* e *Non-negative Matrix Factorization*. Para cada um dos métodos, consideramos

a quantidade de *clusters* recomendados pela técnica de *Silhouette*[3], ou seja, baseado numa amostra de dados, podemos observar visualmente qual seria o melhor valor de k para esse conjunto. Vale ressaltar, que este processo de seleção de valores não está automatizado, com isso se faz necessário uma intervenção manual dentro da abordagem para visualizar qual seria o valor apropriado para uma amostra de dados. Após esta fase, este valor selecionado é utilizado como parâmetro de entrada para a nossa abordagem, pois os algoritmos selecionados exigem um número de tópicos que se deseja criar [Xu et al., 2003, Diao et al., 2012].

As implementações dos algoritmos utilizados neste trabalho, permitem que seja possível extrair a quantidade ideal de tópicos, além da quantidade de termos que se deseja visualizar em cada tópico (denominados de *NGram*). Neste contexto, representamos um tópico relevante por *trigram* (três termos), pois observamos em nossos experimentos que este valor apresentou maior ganho de informação. Esse valor foi obtido, quando buscamos a combinação de termos que nos ajudasse a representar um tópico como sendo um documento, ou seja, esta combinação com *trigram* permita realizar tal filtragem. O objetivo da seleção deste valor é representar cada tópico extraído como uma única informação, que será validada com outras fontes de informação.

A fase de extração de tópicos foi executada em períodos diferentes, pois cada base de dados apresenta características particulares, por exemplo, na base de dados que corresponde as 15 fases da Operação Lava Jato da Polícia Federal, observamos que mensalmente aconteciam relatos de novas movimentações deste evento (conforme ilustrado na Figura 5.4), e com isso dividimos nossos dados de forma proporcional aos acontecimentos, para que posteriormente realizássemos as avaliações dos tópicos extraídos com outras fontes de dados.

4.5 Avaliação da abordagem

A nossa abordagem foi avaliada levando em consideração a arquitetura de pré-processamento aplicada, tempo de execução dos algoritmos e média de acerto dos tópicos extraídos. Aplicamos nossa abordagem em duas bases de dados, com o intuito de verificar qual abordagem de pré-processamento que combinada com os algoritmos, apresenta maior média de acerto, quando comparada com outras fontes de informação.

Durante os experimentos executamos os algoritmos selecionados para extração de tópicos em ambas as arquiteturas, a fim de coletar o tempo de execução dos algoritmos e os tópicos criados, para que fosse possível calcular a média do tempo. Vale ressaltar,

[3] `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`

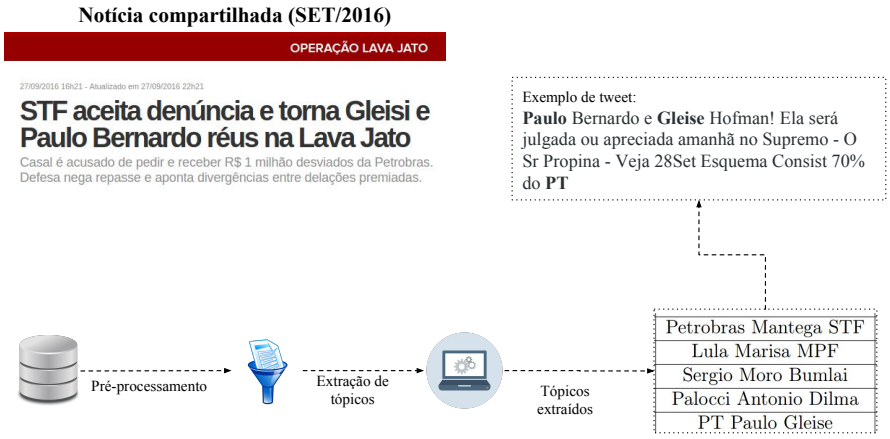


Figura 4.4: Modelo de comparação dos tópicos extraídos com canais de notícias.

que os eventos selecionados tinham quantidades de subeventos diferentes, ou seja, a cada período analisado por nossa abordagem é necessário executar todo processo de extração de tópicos.

As médias de acerto foram calculadas levando em consideração se os tópicos que foram extraídos possuem alguma similaridade com as informações que foram disponibilizadas nas mídias como jornais, sites e *blogs* de entidades jornalísticas (conforme ilustrado na Figura 4.4). Assumimos que todas as informações que foram compartilhadas por esses canais de notícias tiverem alguma citação nas redes sociais, pois a partir desta premissa, podemos realizar a comparação entre as informações compartilhadas pelos canais de notícias e as redes sociais.

4.6 Considerações Finais

Neste capítulo apresentamos a abordagem aplicada ao longo deste trabalho. Vale ressaltar, que nossa abordagem necessita de um parâmetro de entrada, referente a quantidade de tópicos que se deseja extrair de uma base de dados. Esse valor pode ser obtido através da técnica de *Silhouette*, e está sendo aplicado em ambas as arquiteturas que utilizamos em nosso trabalho. Após obtido esse valor, nossa abordagem extrai os *trigram* para cada tópico, sendo mais tarde comparado com as informações compartilhadas em canais de notícias para se obter a média de acerto.

Os resultados referentes à nossa abordagem podem ser observados no Capítulo 5, onde foram executados testes em duas bases de dados distintas, cada uma referente a um determinado evento pré-selecionado, baseado na repercussão desses acontecimentos nas redes sociais.

Capítulo 5

Experimentos e Resultados

Esta capítulo descreve como foram realizados os experimentos e quais resultados foram alcançados na extração de tópicos relevantes. Os experimentos foram organizados em 2 subseções, onde cada seção corresponde a base de dados de um evento. De cada evento estão sendo avaliados: os tópicos extraídos em ambas as arquiteturas de pré-processamento, tempo de execução dos algoritmos e o valor de *clusters* para cada amostra de dados.

5.1 Experimento 1 - Operação Lava Jato da Polícia Federal

A Operação Lava Jato é uma série de investigações da Polícia Federal do Brasil que teve início em 17 de março de 2014, cujo objetivo consiste em investigar a prática de crimes financeiros e desvio de recursos públicos, onde já foram realizados mais de cem mandatos de busca e apreensão, de prisão temporária, de prisão preventiva e de condução coercitiva. É considerada como o maior processo de investigação contra a corrupção do país. No ano de 2016, a Operação Lava Jato teve 15 fases que envolveram entidades importantes da política brasileira.

Nas redes sociais, a Operação Lava Jato teve repercussão massiva, onde existiram embates de opiniões entre usuários[1], e como consequência é gerado uma grande escala de dados para serem analisados, tornando viável a realização deste estudo.

[1]http://www.jornaldotocantins.com.br/editorias/estado/judici%C3%A1rio-1.456286/o-efeito-da-lava-jato-nas-redes-sociais-1.1171504

5.1.1 Base de dados

Para realização dos experimentos, utilizamos o Twitter como fonte dos dados para geração da base. No escopo do trabalho, selecionamos somente os *tweets* em português. A coleta foi realizada no período de 01/01/2016 à 20/11/2016, referente a todas as operações da Lava Jato deflagradas pela Polícia Federal em 2016, ou seja, da 22ª à 36ª fase. Nesse contexto, utilizamos unicamente a palavra "lava jato" para coleta dos dados, com isso, obtivemos 652.210 *tweets*. A figura 5.1 mostra a distribuição desses *tweets* por mês.

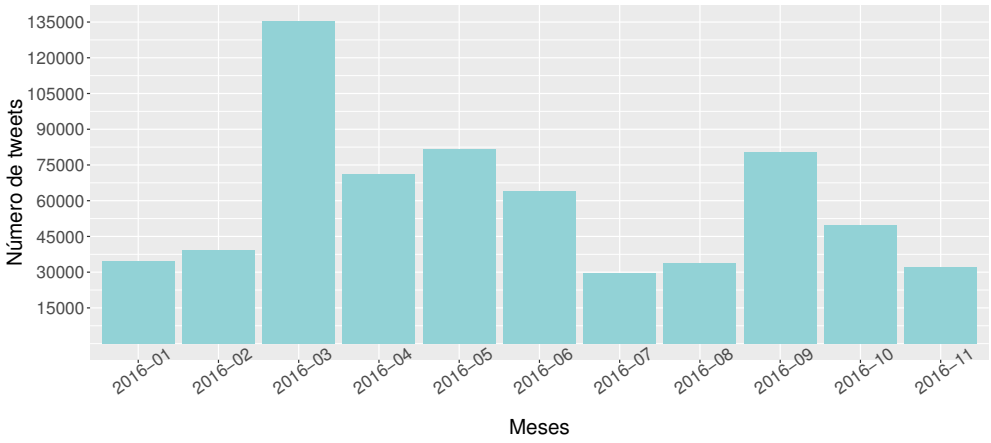


Figura 5.1: Distribuição de *tweets* por mês sobre a Operação Lava Jato no período de 01/01/2016 à 20/11/2016 .

5.1.2 Seleção dos Valores de K

Para a seleção dos valores de K e avaliação dos agrupamentos criados, aplicamos a técnica *Silhoutte*. Durante esta etapa, aplicamos esta técnica nas duas abordagens que estão sendo descritas neste trabalho em conjuntos com os algoritmos de extração de tópicos. Neste cenário, para o melhor entendimento da técnica utilizada selecionamos o mês de setembro (mês com maior divergências) da base de dados da Operação Lava Jato e aplicamos a técnica com o algoritmo de *K-means* para exemplificar o processo executado. Selecionamos esta amostra de dados, devido ser a mesma que será utilizada para descrever os cenários de extração de tópicos na Subseção 5.1.3.

O procedimento para a execução desta técnica, consiste em fornecer os vetores de características a qual se deseja criar os agrupamentos. Após a criação deste vetor, realizamos uma redução de dimensão nos dados utilizando o PCA, pois a matriz de termos criada pelo TF-IDF estava em uma escala de alta dimensão. Em seguida, criamos os agrupamentos com o *K-means* com um valor inicial igual a 2, para que

fosse calculado a média de *Silhoutte* para este valor inicial. Com a média calculada utilizamos a ferramenta MatplotLib[2] para visualização do resultado, conforme ilustrado na Figura 5.2.

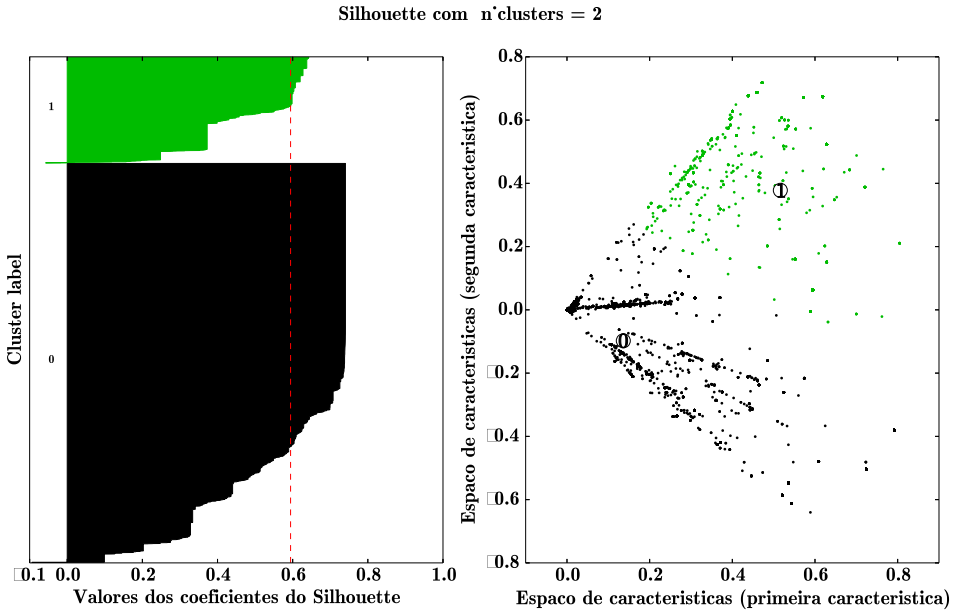


Figura 5.2: Exemplo de aplicação da técnica de *Silhouette* no mês de setembro.

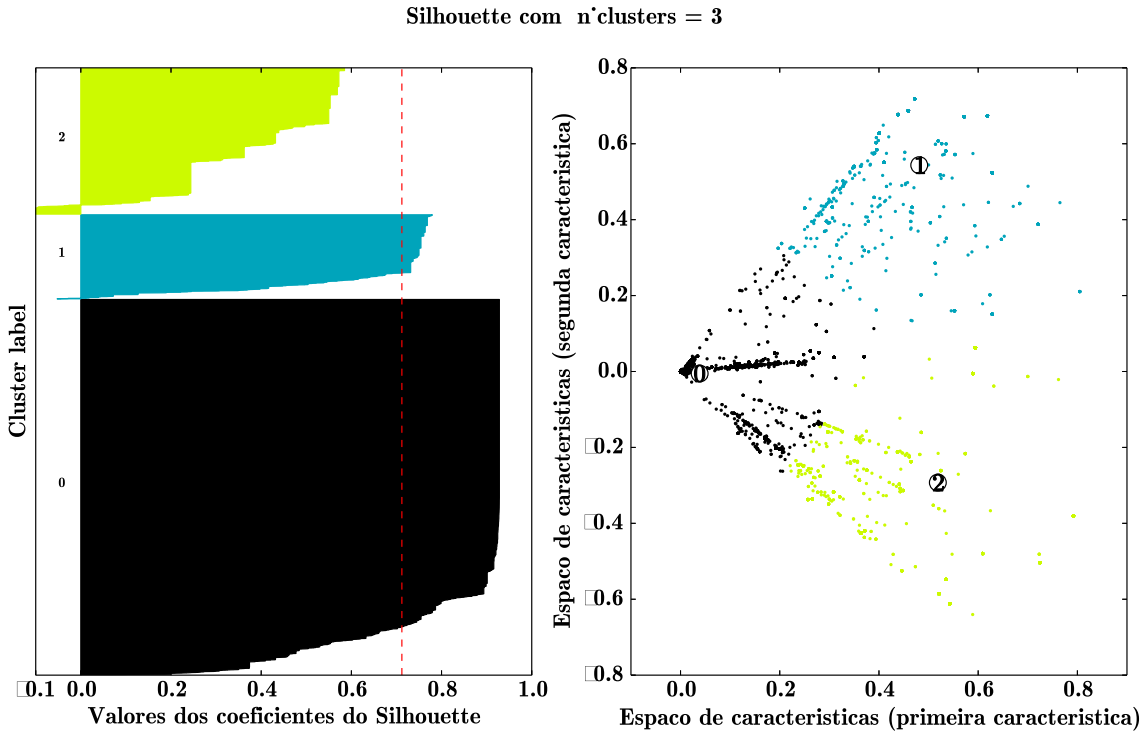
Vale ressaltar, que este procedimento é incremental, ou seja, para cada amostra de dados selecionada, testamos este conjunto de dados em uma variação de valores de k de 2 até 15, com o intuito de verificar qual apresenta melhor agrupamento (conforme ilustrado na Figura 5.3).

Como resultado, verificamos neste mês o valor de K que será fornecido como parâmetro de entrada para a extração de tópicos é igual a 5, pois em relação as premissas já explicadas no Capítulo 2, esse valor atende .

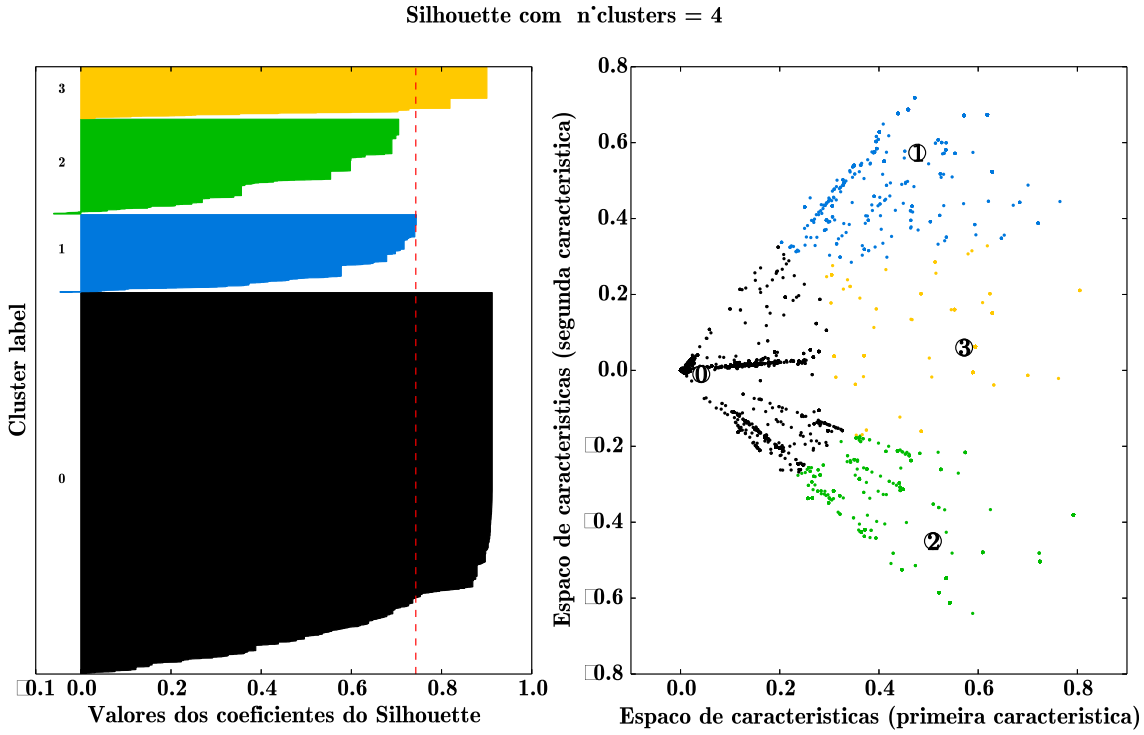
Os demais resultados alcançados nas amostras da base de dados da Operação Lava Jato da Polícia Federal, podem ser observados na Tabela 5.1, visto que esta técnica foi aplicada para verificar se haveria diferença nos valores. Um comportamento observado durante esta extração foi o fato dos algoritmos LDA e NMF em todos os meses, apresentarem valores de quantidade de tópicos iguais.

Em nossa abordagem consideramos os valores de K que foram observados na técnica de *Silhouette*, pois esses valores nos ajudam a determinar a quantidade de tópicos que serão extraídos de cada amostra de dados. Quando comparamos os tópicos extraídos com outras fontes de informações, verificamos unicamente se resultados

[2]http://www.matplotlib.org/

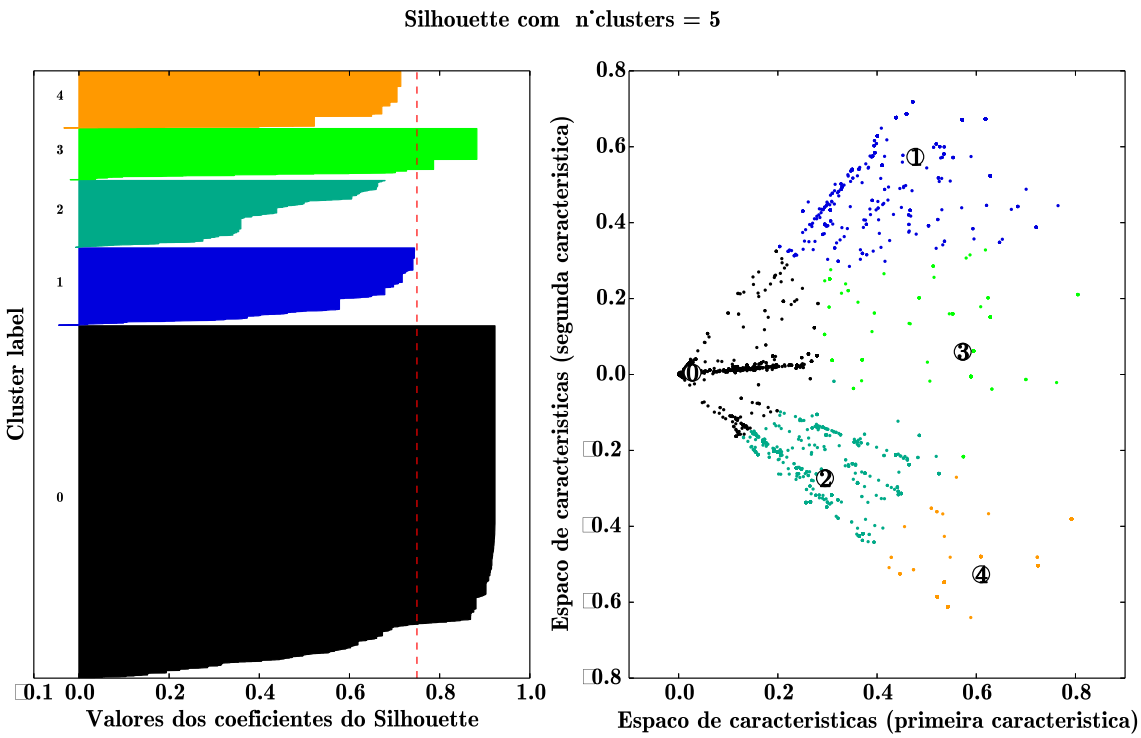


(a) Distribuição dos *clusters* com K=3

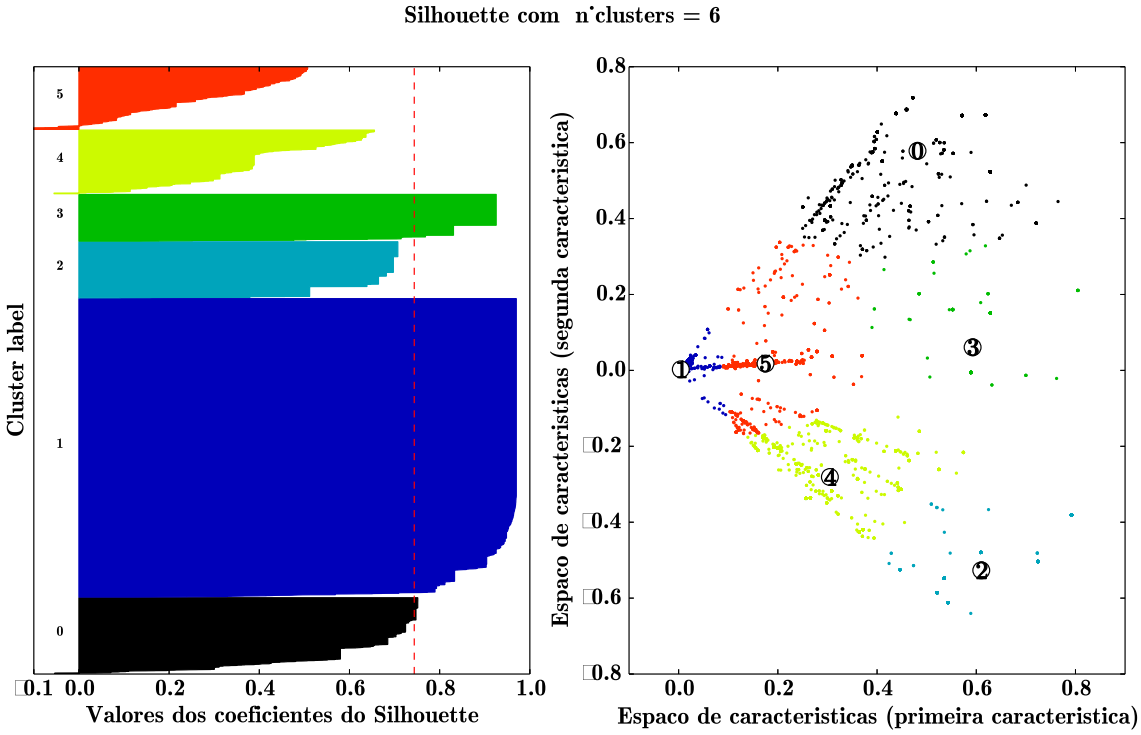


(b) Distribuição dos *clusters* com K=4

Figura 5.3: Visualização da Função de *Silhouette* na base de dados da Operação Lava Jato.



(c) Distribuição dos *clusters* com K=5



(d) Distribuição dos *clusters* com K=6

Figura 5.3: Visualização da Função de *Silhouette* na base de dados da Operação Lava Jato (cont.).

	pré-processamento tradicional			pré-processamento com reconhecimento de Entidade		
Meses	**LDA**	**NMF**	**K-means**	**LDA**	**NMF**	**K-means**
Jan	5	5	5	4	4	7
Fev	4	4	5	5	5	7
Mar	5	5	9	5	5	8
Abr	5	5	7	5	5	7
Mai	5	5	6	5	5	6
Jun	7	7	6	6	6	6
Jul	4	4	5	5	5	5
Ago	5	5	6	3	3	5
Set	5	5	5	5	5	5
Out	7	7	8	8	8	6
Nov	5	5	7	4	4	6

Tabela 5.1: Valores de K observados na técnica de *Silhouette*.

apresentados pelos algoritmos contêm algum acontecimento relacionado ao evento em estudo. Em casos que a quantidade de notícias é inferior a quantidade de tópicos extraídos pelos algoritmos, consideramos a diferença como sendo ganho de informação, pois como as técnicas criam tópicos a partir de elevadas frequências é possível observar se uma informação é relevante.

5.1.3 Extração dos Tópicos

Durante esta fase conduzimos os nossos experimentos em dois cenários, onde o primeiro cenário consiste em descrever a utilização das técnicas de extração de tópicos, tendo como entrada uma matriz de termos que foi processada com uma arquitetura tradicional (Figura 4.2). O segundo cenário descreve a utilização de pré-processamento através do reconhecimento de entidades (Figura 4.3), onde a matriz de termos gerada, consiste unicamente de termos reconhecidos como entidade em cada *tweet*.

Para validarmos nossa arquitetura de extração de tópicos, ilustrada na Figura 4.1, realizamos uma comparação dos tópicos extraídos por cada algoritmo com canais de notícias, e verificamos se as informações compartilhadas nas redes sociais sobre a Operação Lava Jato em um determinado mês, são equivalentes a outras fontes de informação. Essa comparação gera um percentual de acerto para cada algoritmo, onde dividimos o total tópicos extraídos pelos algoritmos que correspondiam as notícias compartilhadas, pelo total de informações (notícias) divulgados nos meios de comunição tradicionais ($P = \frac{topicos_{relacionados}}{Noticias}$). O objetivo desta abordagem é comprovar que moni-

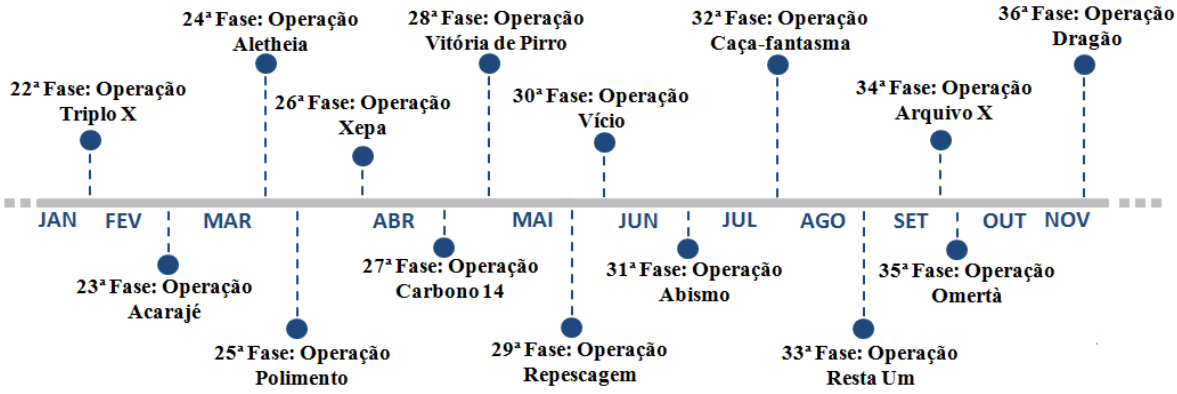


Figura 5.4: Linha do tempo das fases da Operação Lava Jato em 2016.

torando o Twitter, conseguimos descobrir informações relevantes sobre um determinado acontecimento.

Durante os experimentos aplicamos os algoritmos de extração de tópicos aos conjuntos de dados de cada mês, pois conforme observado na Figura 5.4, a operação Lava Jato da Polícia Federal, de janeiro a novembro teve fases relatadas nas redes sociais mensalmente, atingindo um total de 14 operações. Vale a pena ressaltar que levamos em consideração todos relatos compartilhados sobre esses eventos, desde prisões de pessoas envolvidas à ações judicias executadas pelos órgãos responsáveis.

Cenário I

Neste cenário, para um melhor entendimento de como as avaliações foram realizadas, selecionamos o mês que teve maior divergência entre os tópicos extraídos em relação as informações compartilhadas em outros canais, ou seja, o mês de setembro, onde ocorreu a trigésima quarta e trigésima quinta fase da Operação Lava Jato com uma amostra de 80.577 *tweets*.

Neste período, os acontecimentos de maior impacto segundo as fontes de informação Paraná Portal [3], O Globo [4], Estadão[5] e site da Polícia Federal[6] foram:

1. O ex-ministro Guido Mantega é preso temporariamente, mas o juiz Sérgio Moro manda soltá-lo horas depois;

[3] http://paranaportal.uol.com.br/editoria/operacao-lava-jato/
[4] http://infograficos.oglobo.globo.com/brasil/todas-as-fases-da-operacao-lava-jato.html
[5] http://infograficos.estadao.com.br/public/politica/operacao-lava-jato/fases/
[6] http://www.pf.gov.br/imprensa/lava-jato/fases-da-operacao-lava-jato

2. O empresário Eike Batista declarou que, em 1º de novembro de 2012, recebeu pedido do então ministro e presidente do Conselho de Administração da Petrobrás para que fizesse um pagamento de R$ 5 milhões destinado ao PT;

3. Supremo Tribunal Federal manda ação contra Eduardo Cunha para Sérgio Moro;

4. O ex-ministro da Fazenda Antônio Palocci foi preso na 35ª operação, por atuar de maneira direta para proporcionar vantagens à empreiteira Odebrecht;

5. Gleisi Hoffmann e Paulo Bernardo são réus da Operação Lava Jato no Supremo Tribunal Federal.

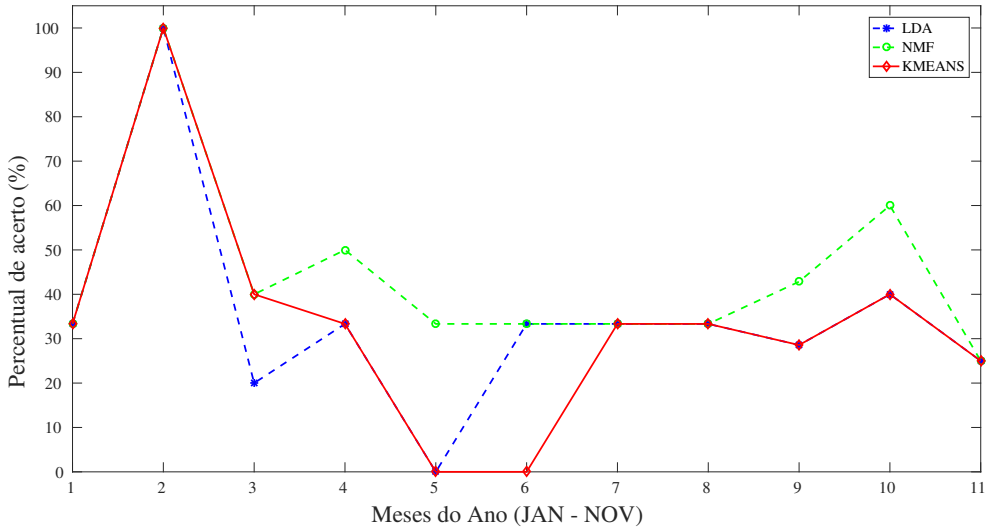


Figura 5.5: Percentual de acerto utilizando pré processamento tradicional.

No cenário de estudo observamos que os tópicos extraídos pelo NMF, conforme apresentado na Tabela 5.2, foram os que mais se aproximaram das notícias compartilhadas por outras fontes de informação. Os tópicos encontrados correspondem a mais de 40% de acerto nesta comparação. Enquanto os demais algoritmos, apresentaram um rendimento inferior a 40%.

Observamos durante os experimentos com este modelo, que apesar de o pré-processamento possuir múltiplas etapas com o objetivo de reduzir os ruídos, a base de dados possui outras informações classificadas como ruídos, pois não estavam no contexto direto da Lava Jato. Essas informações, que possuem maiores frequências dentro uma *bag-of-words* acabam em maioria se tornando um tópico coletado pelos algoritmos. Com isso, este cenário apesar de ter um pré-processamento com múltiplas etapas, não apresentou bom desempenho, conforme observado na Figura 5.5, tornando

Tópicos LDA	Tópicos NMF	Tópicos Kmeans
Marisa Lula operação	diz Lula operação	youtube video gostei
Guido prisão horas	fase Mantega operação	lula denuncia reu
video youtube gostei	justiça temer ministro	Bernardo Gleise réu
abafar operação Moro	Lula Moro réu	Lula operação diz
lula denuncia denunciado	youtube gostei video	preso Mantega nova

Tabela 5.2: Tópicos extraídos pelos algoritmos.

enviável a utilização deste padrão para o monitoramento das notícias relevantes de um acontecimento, devido a captura desses ruídos pelos algoritmos aplicados neste trabalho.

Cenário II

Neste cenário, para fins de avaliação, será apresentado o mesmo mês analisado no Cenário 1. Neste contexto, aplicamos como pré-processamento o reconhecimento de entidade nomeada, onde para cada documento (*tweet*) pode existir uma ou mais entidades, conforme exemplo abaixo:

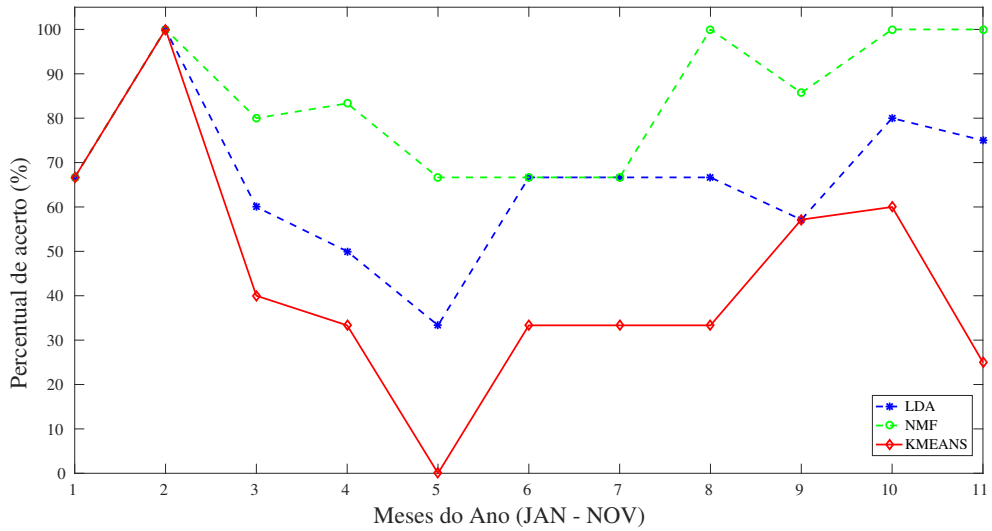


Figura 5.6: Percentual de acerto utilizando reconhecimento de entidade.

"*Tribunais Superiores derrubam 18 erros de Sergio Moro.*", tendo como resultado o reconhecimento da entidade "Sergio Moro". Com isso, nossa *bag-of-words* pode ser considerada como uma *bag-of-entities*, visto que essa regra foi aplicada a toda base de dados, com o objetivo de reconhecer quais foram as entidades mais citadas em um

determinado mês, pois como as notícias ou relatos estão atreladas a uma determinada pessoa, grupos, empresas ou locais, torna-se viável a utilização dessa estratégia.

Durante esse pré-processamento, eliminamos entidades com frequências menores que 20, com o intuito de reduzir o tamanho da matriz de termos e igualar com o primeiro cenário. Quando comparado a quantidade de atributos do vetor de características das arquiteturas de pré-processamento, é possível observar na Figura 5.7 que a arquitetura de reconhecimento de entidade de nomeada consegue reduzir o valor de atributos dentro do vetor de forma representativa. Podemos observar nos experimentos que essa quantidade tem influência direta nos resultados alcançados.

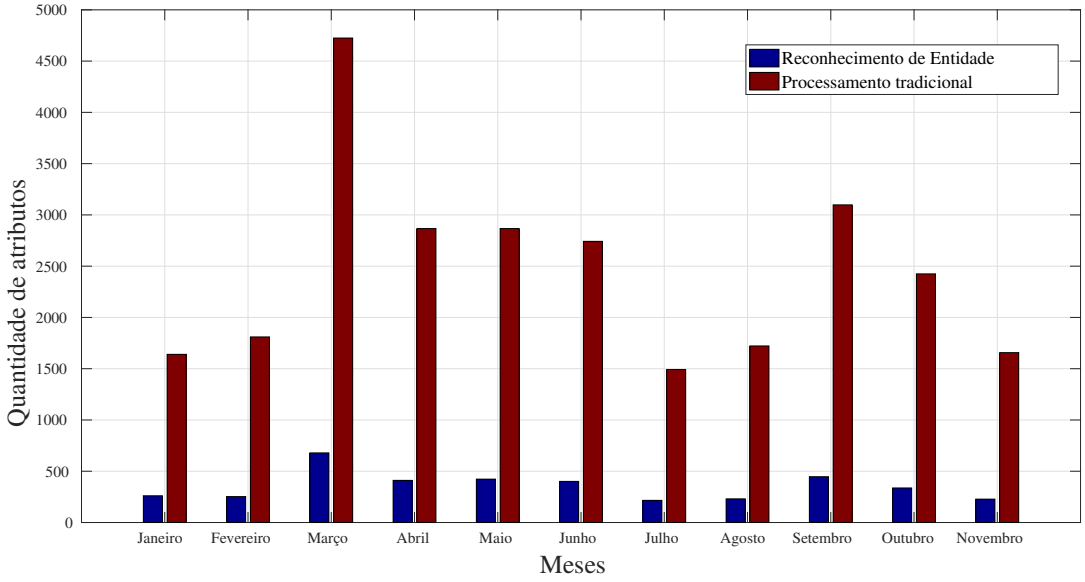


Figura 5.7: Comparação da quantidade de atributos do vetor de características das arquiteturas de pré-processamento.

Após a fase de pré-processamento, realizamos a extração dos tópicos e observamos que o NMF apresentou um desempenho acima de 80% de acerto na extração de tópicos (conforme ilustrado na Figura 5.6), quando comparado com os canais de notícias, pois além de eliminar os ruídos da base dados, também reconheceu os indivíduos chaves de cada notícia (Tabela 5.4).

5.1.4 Discussão dos Resultados do Experimento I

Os resultados observados na base de dados da Operação Lava Jato da Polícia Federal do Brasil, indicaram o NMF como sendo o algoritmo que melhor extraí tópicos relevantes quando comparados com outros canais de notícias, apresentando um rendimento médio

LDA	NMF	KMEANS
MPF STF Moro	Petrobras Mantega STF	Palocci Antonio PF
Palooci Eike Renan	Lula Marisa MPF	Lula PT Mantega
PT Cunha PSDB	Sergio Moro Bumlai	Moro Lula Sergio
Lula Marisa Temer	Palocci Antonio Dilma	Lula PT Curitiba
Sergio Dilma Curitiba	PT Paulo Gleise	Lula Marisa MPF

Tabela 5.3: Tópicos extraídos pelos algoritmos com reconhecimento de entidade.

acima de 80%. Vale ressaltar que a *bag-of-words* em conjunto com o TF-IDF que são fornecidos para os algoritmos eram iguais, mas a forma como é inferido o cálculo das probabilidades dessas funções geram a diferença nos resultados.

Quando realizamos a comparação entre as abordagens de pré-processamento, podemos observar a redução da quantidade de atributos nos vetores de características, onde já era esperado que o reconhecimento de entidade apresentasse valores inferiores, porém ao computarmos o tempo de execução do processo de extração de tópicos, observamos que o reconhecimento de entidade apresenta maior custo computacional para a aplicação do processo. É possível observar nas Figuras 5.8 e 5.9 o tempo que os processos alcançaram. Vale ressaltar que o tempo apresentado equivale a todo processo para a extração dos tópicos, que se inicia na leitura da base de dados até a extração dos tópicos. Para alcançar estes valores os experimentos desta base de dados foram executados em uma máquina com a seguinte configuração:

Configuração do Servidor	
Processador:	Intel(R) Xeon(R) CPU E5-4617 0 @ 2.90GHz
Arquitetura:	64bits
Memória:	20GB
S.O:	Ubuntu 16.04
HD:	50gb

Tabela 5.4: Configuração do servidor para execução dos experimentos.

Com isso, sumarizamos nossos resultados nesta base de dados da seguinte forma:

1. O pré-processamento tradicional aplicado a nossa abordagem, apresentou um baixo rendimento na tarefa de extração de tópicos para o monitoramento de notícias sobre a Operação Lava Jato da Polícia Federal do Brasil, pois mesmo realizando um pré-processamento com múltiplas etapas, não apresentou um bom rendimento na tarefa de extração de tópicos;

2. O pré-processamento com reconhecimento de entidade nomeada aplicado a nossa arquitetura, apresentou um bom rendimento, comprovando a viabilidade de uti-

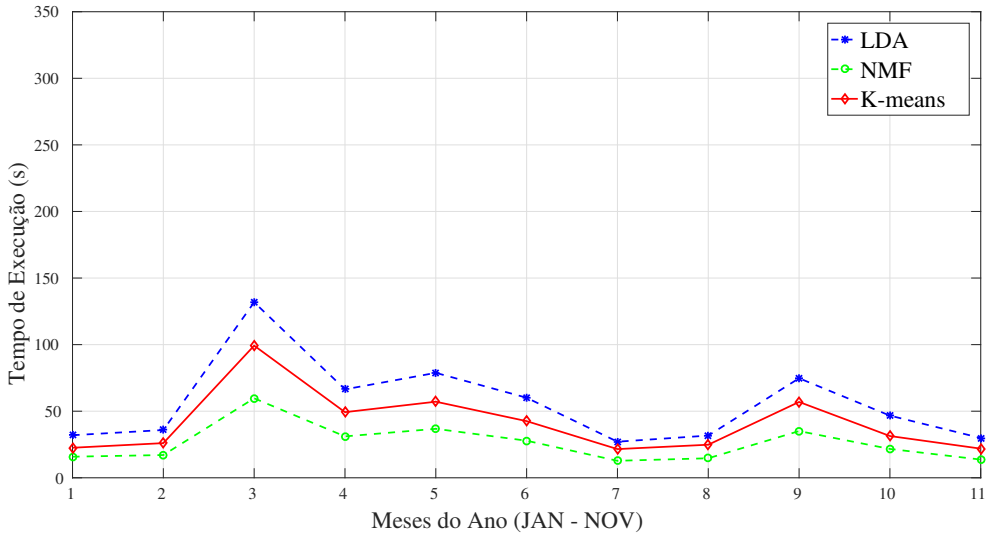


Figura 5.8: Tempo de execução coletado do primeiro cenário.

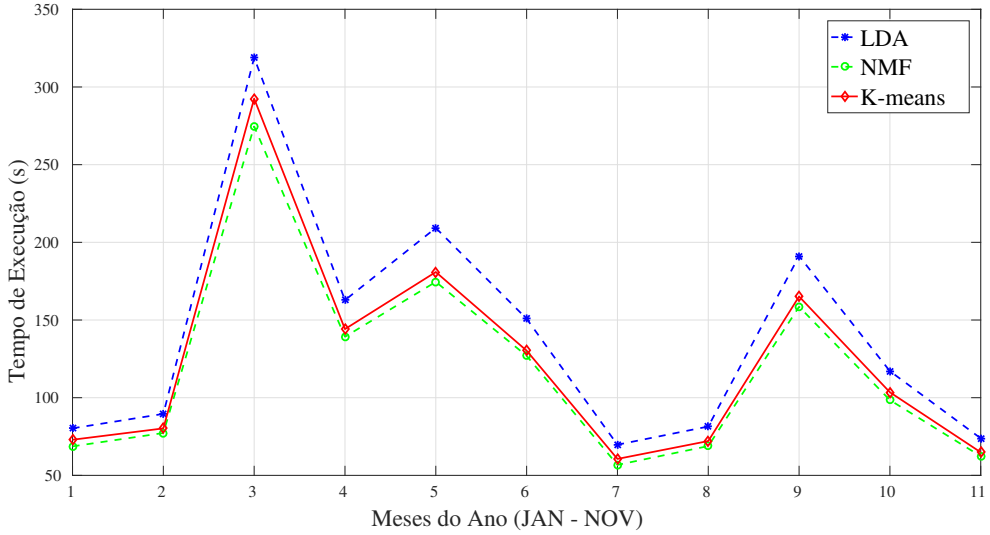


Figura 5.9: Tempo de execução coletado do segundo cenário.

lizar este modelo para monitorar o Twitter, com o objetivo de identificar as informações relevantes de um determinado acontecimento;

3. O algoritmo de NMF apresentou melhor rendimento, sendo 1,27% melhor que os outros algoritmos utilizados. O baixo rendimento no primeiro cenário é devido matriz de características formada possuir ruídos, dificultando o processo de extração do NMF. No segundo cenário, o NMF apresentou um desempenho de pelo menos 1,26% melhor que os demais, devido a matriz gerada levar em consideração apenas as entidades citadas nos texto, reduzindo a quantidade de atributos do

vetor de características;

4. Apesar dos resultados alcançados na abordagem com pré-processamento de reconhecimento de entidade serem mais representativos que a outra abordagem, o mesmo consome mais recursos computacionais para a execução da extração de tópicos, onde identificamos que o ponto de maior consumo é a classificação da entidade feita pelo *PolyGlot*;

5. O tempo de execução do LDA cresce proporcionalmente de acordo com a quantidade de dados utilizada para extração de tópicos multiplicada pela quantidade de iterações feitas, ou seja, quanto mais dados possuir o LDA proporcionalmente ficará mais custoso o processo. Com isso, podemos observar o porquê do LDA ter sido o processo mais custoso para extrair tópicos.

5.2 Experimento 2 - *Impeachment* da Presidente do Brasil

O processo de *impeachment* da Presidente do Brasil teve início em 2 de dezembro de 2015, quando o então Presidente da Câmara dos Deputados, Eduardo Cunha, aceitou a denúncia de crime de responsabilidade fiscal contra Dilma Roussef, feita pelo procurador de justiça Hélio Bicudo e os advogados Miguel Júnior e Janaina Paschoal. O processo se encerrou no dia 31 de agosto de 2016, onde o resultado foi a destituição de Dilma Rousseff do cargo.

5.2.1 Base de Dados

A base de dados utilizada neste experimento é composta por *tweets* escritos em português, pertencentes ao período de 01/01/2016 a 30/11/2016. Os *tweets* da nossa base de dados são relativos ao processo de *impeachment* da presidente do Brasil, Dilma Rousseff, onde monitoramos o processo um mês após sua abertura até dois meses após a destituição da Presidente.

Durante a fase de coleta de dados, utilizamos palavras chaves que observamos ter uma relação com o cenário que estávamos analisando, como por exemplo: *impeachment*, Dilma, Lula, Pedaladas Fiscais, Lava Jato, Golpe, PT, PSDB e PMDB. Essas palavras foram observadas por meios de canais de notícias e redes sociais, onde diariamente uma grande escala de dados eram compartilhados nesses ambientes, relatando opiniões a favor ou contra o andamento deste processo. Esses níveis de polaridade foram estudados

em trabalhos anteriores [Souza et al., 2016], onde os autores demonstram as polaridades dos *tweets* compartilhados durante os cinco primeiros meses do processo. Vale ressaltar, que a base de dados utilizada neste experimento, foi a mesma aplicada na análise de polaridade, porém retiramos os dados que estavam referentes a Operação Lava Jato e inserimos mais dados referentes ao processo de *impeachment*, visto que o trabalhou se limitou a analisar até 19/05/2016, ou seja, após a semana de votação do *impeachment* no Senado Federal.

Os novos dados inseridos para realização deste experimento foram obtidos através da continuação da coleta até o final de novembro. Durante a construção desta base, eliminamos os *tweets* referentes ao mês de janeiro de 2015, visto que não tínhamos uma quantidade representativa de dados para realizarmos a análise. A distribuição final da base de dados pode ser observada na Figura 5.10.

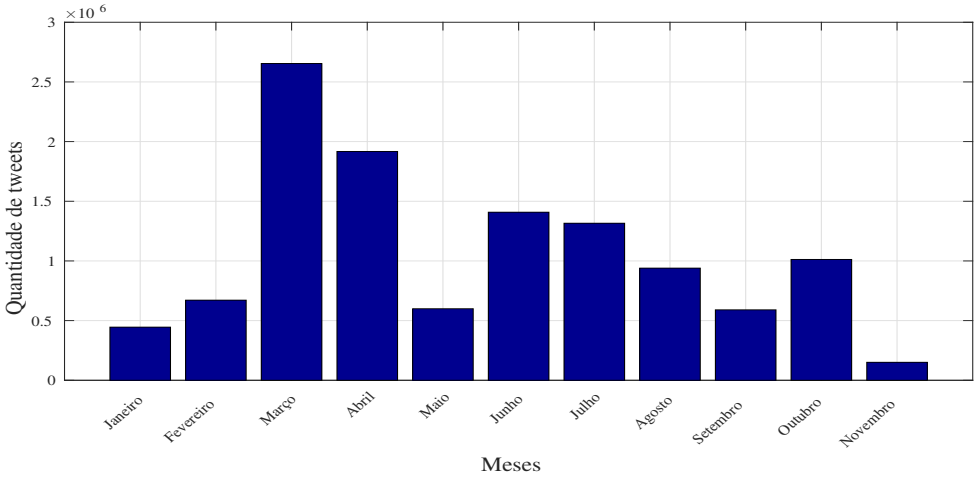


Figura 5.10: Distribuição dos dados durante o processo de *impeachment*.

5.2.2 Seleção dos Valores de K

Para seleção dos valores de K, utilizamos o procedimento realizado no último experimento correspondente à seção 5.1, onde selecionamos uma amostra de dados para ser analisada e aplicamos a técnica de *Silhouette* para observar visualmente qual seria o valor que melhor representa a quantidade de *clusters* para a amostra selecionada.

Os valores obtidos na Tabela 5.5, correspondem aos valores observados na técnica de *Silhouette*. Durante a execução desta etapa e do experimento anterior, foi necessário combinar as funções disponibilizadas pelo Spark de programação paralela com o *Silhouette*, pois os dados desta base de dados estavam em uma escala que não permitia seu processamento com memória principal. Vale ressaltar que a utilização dessas

funções não alteram os resultados visto sua função é unicamente processar dados em larga em escala.

	pré-processamento tradicional			pré-processamento com reconhecimento de Entidade		
Meses	**LDA**	**NMF**	**K-means**	**LDA**	**NMF**	**K-means**
Jan	5	5	5	6	6	6
Fev	7	4	5	5	5	7
Mar	6	6	7	6	6	8
Abr	6	6	7	6	6	7
Mai	7	7	7	6	6	8
Jun	5	5	6	6	6	6
Jul	4	4	7	6	6	6
Ago	5	5	6	5	5	5
Set	5	5	5	5	5	5
Out	4	4	5	4	4	6
Nov	3	3	5	4	4	5

Tabela 5.5: Valores de K observado na técnica de *Silhouette*.

Os valores obtidos com essa função foram utilizados como parâmetros de entrada para nossa abordagem durante este experimento. Caso os valores fossem diferentes, conforme visualizado na Tabela 5.5, que alguns casos o algoritmo de *K-means* apresentou valores acima do NMF e LDA, optamos por manter o valor obtido como parâmetro de entrada.

5.2.3 Extração de Tópicos

Para validarmos a abordagem proposta aplicada a esta base de dados, realizamos a mesma comparação dos tópicos extraídos de cada algoritmo com canais de notícias, conforme descrito na Seção 5.1.

Durante os experimentos, aplicamos os algoritmos de extração de tópicos aos conjuntos de dados correspondentes a cada mês durante o processo de *impeachment* e três meses após a finalização do julgamento, com o intuito de descobrir quais foram os tópicos relevantes durante e após o evento estudado. Com isso, a melhor forma de analisar os dados coletados, seria aplicando a abordagem proposta a cada mês, visto que os marcos relevantes do *impeachment* foram relatados mensalmente.

Os experimentos nesta base de dados foram divididos em dois cenários, onde o primeiro cenário corresponde ao pré-processamento tradicional combinados com os algoritmos selecionados e o segundo cenário representa os resultados obtidos aplicando

reconhecimento de entidade nomeada. Vale ressaltar, que neste experimento os mês de fevereiro não apresentou notícias compartilhadas em outras fontes de informação, desta forma a analise comparativa não foi feita no referido mês.

Cenário I

Neste cenário de estudo, aplicamos o pré-processamento tradicional combinado com os algoritmos LDA, NMF e *K-means*. Durante a experimentação e análise dos resultados, observamos que esta abordagem apresentou baixo rendimento (conforme ilustrado na Figura 5.11) em relação a comparação feita dos tópicos extraídos com os canais de notícias.

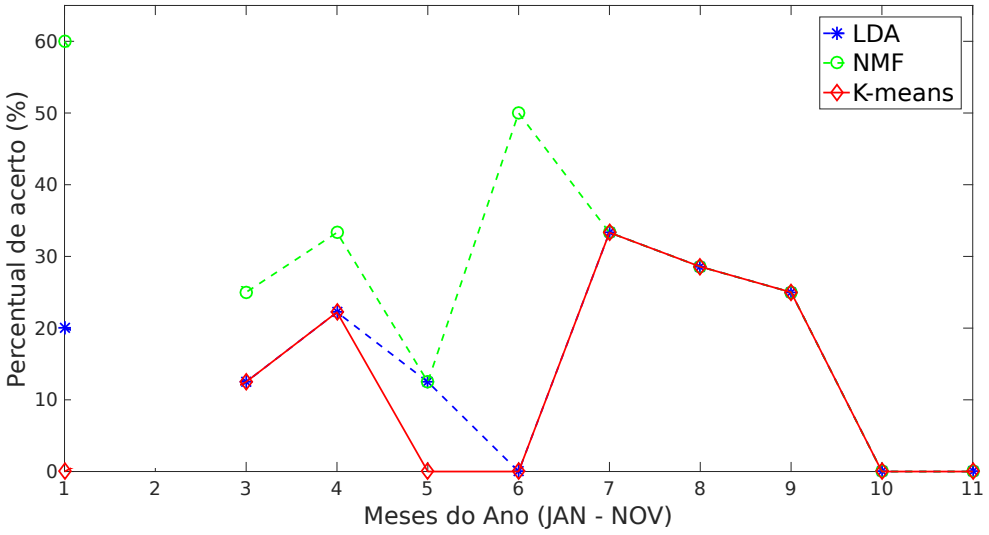


Figura 5.11: Percentual de acerto no primeiro cenário.

Um comportamento observado sobre este experimento, foi que mesmo as notícias sendo compartilhadas mensalmente os usuários continuavam comentar sobre assuntos de outros meses. Essa análise foi observada, pois durante os meses de março a junho os tópicos eram similares.

Como resultado final deste cenário, o NMF apresentou melhor rendimento em relação aos demais algoritmos, porém com uma média de acerto de 30%.

Cenário II

No segundo cenário observamos que os resultados foram mais expressivos em relação a primeira abordagem, porém o mesmo comportamento analisado no primeiro cenário,

pode ser visto nesta abordagem. Com isso, os resultados de março a junho, ficam a abaixo de 50%.

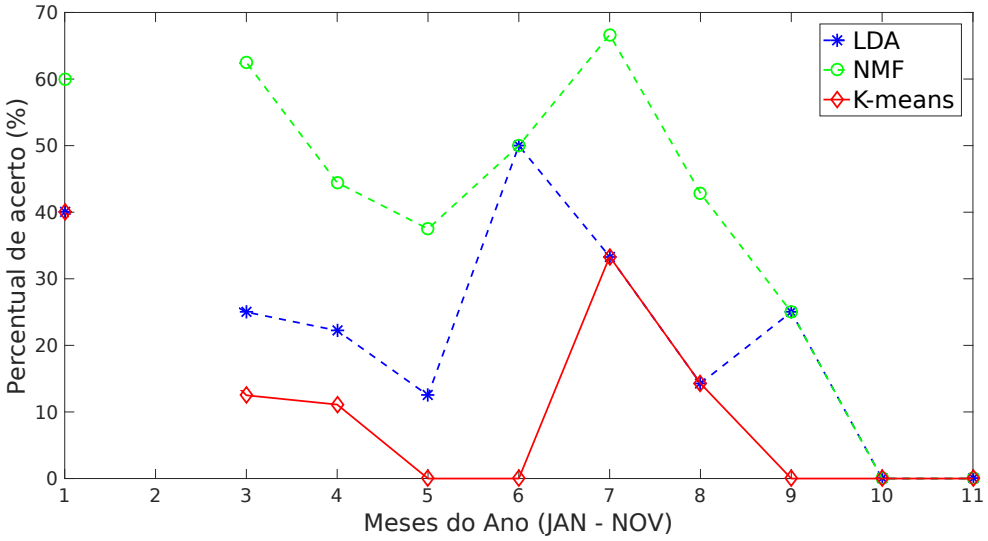


Figura 5.12: Percentual de acerto no segundo cenário.

Como resultado final o NMF, apresentou um rendimento de 45% se analisarmos a média de todos os meses. O baixo rendimento apresentado nos meses de outubro e novembro, ocorreu devido a terem sido compartilhados apenas uma notícia por mês, e conforme observarmos os tópicos extraídos, essas informações não tiveram repercussão nas redes sociais, ou seja, após o período de *impeachment*, os usuários das redes sociais não mantiveram seus níveis de atividades após uma nova fase do evento ter sido relatada.

5.2.4 Discussão dos Resultados do Experimento II

Os resultados observados na base de dados sobre o processo de *impeachment* da presidente do Brasil, indicaram o NMF como sendo o algoritmo que melhor extrai tópicos relevantes quando comparados com outros canais de notícias, apresentando um rendimento acima de 45%.

Um fator observado durante os experimentos desta base de dados foi tempo de execução do processo de extração dos tópicos, pois diferente da quantidade de dados utilizada para processar as informações referentes a Operação Lava Jato da Polícia Federal, os dados referentes ao processo de *impeachment* estão em escalas maiores, ou seja, o tempo de execução para realizar a operação aumentou, gerando uma demora na extração dos tópicos. Com isso, foi possível observar limitação no modelo *offline*, pois

mesmo utilizando ferramentas que auxiliam a manipulação de dados em alta escala o tempo de execução ainda é elevado.

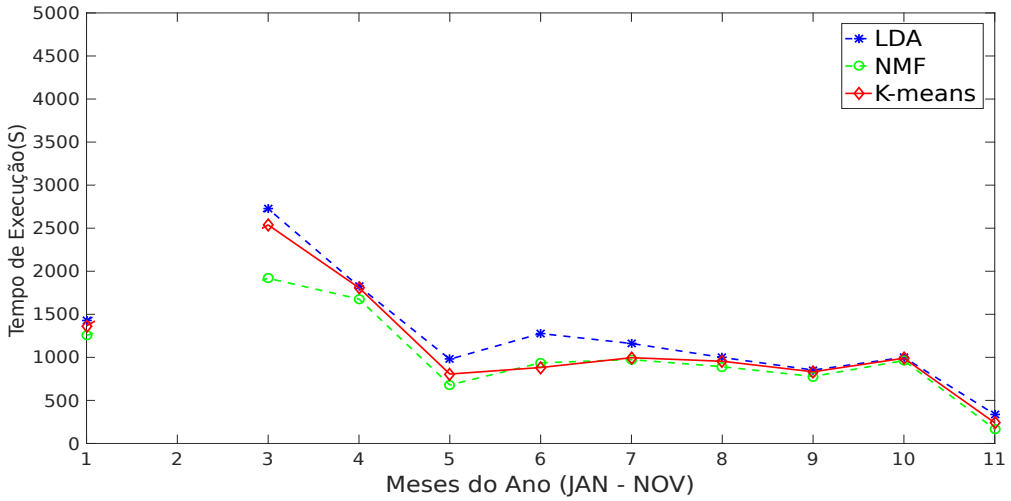


Figura 5.13: Tempo de execução coletado do primeiro cenário.

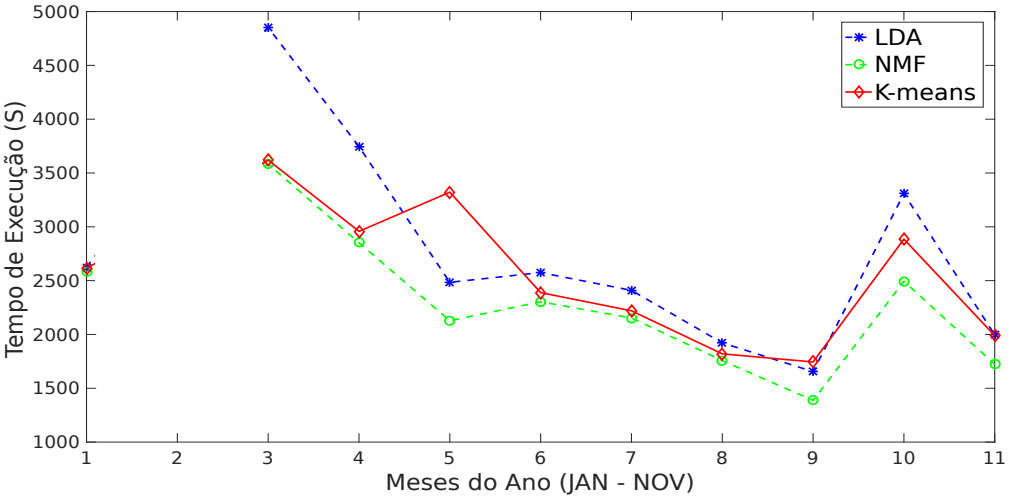


Figura 5.14: Tempo de execução coletado do segundo cenário.

Com isso, sumarizamos nossos resultados neste experimento da seguinte forma:

1. O pré-processamento tradicional aplicado a nossa arquitetura, apresentou um baixo rendimento na tarefa de extração de tópicos para o monitoramento de notícias sobre a processo de *impeachment* da presidente do Brasil;

2. O algoritmo de NMF apresentou melhor rendimento no primeiro e no segundo cenário, porém sendo inferiores a 50%, pois observamos uma mudança de com-

portamento nos usuários que relatavam opiniões sobre o evento ocorrido, gerando baixos índices quando comparado os tópicos extraídos com os canais de notícias;

3. A abordagem proposta apesar de apresentar resultados com NMF acima de 45% no monitoramento dos tópicos, exige um tempo de espera acima de uma hora quando é necessário processar dados em grande escala.

Capítulo 6

Considerações Finais

Neste trabalho abordamos o problema de seleção de tópicos para realizar o monitoramento de informações relevantes de um determinado evento. O desafio desta extração no cenário *offline* é como processar uma grande quantidade de dados em um tempo hábil. Por este motivo, conduzimos nossos experimentos propondo uma comparação de abordagens utilizando um *framework* que auxilia o processamento de dados, visando extrair tópicos relevantes das redes sociais, para que seja possível monitorar esses ambientes, a fim de quantificar os subeventos gerados através da aplicação desta abordagem.

Em nossa abordagem, aproveitamos a capacidade de processar uma grande escala de dados para obter os tópicos com maiores frequências, e compará-los com canais de notícias, a fim de descobrir se as informações que repercutem em outros canais de comunição são similares às redes sociais. Em nosso trabalho, consideramos quatro etapas principais para obtermos os resultados: coleta dos dados, pré-processamento, seleção dos valores k e extração dos tópicos. Essas etapas foram consolidadas tomando como base o levantamento bibliográfico do estado-da-arte.

Em nossos experimentos, comparamos a abordagem proposta em duas bases de dados, onde cada uma apresentava características individuais. Em todos os experimentos, a abordagem que utiliza reconhecimento de entidade nomeada como pré-processamento e combinada com o algoritmo NMF apresentou melhor desempenho, porém a extração da entidade de um texto tem um tempo de execução elevado, que torna esta abordagem lenta, quando a base de dados a qual será executada possui dados em grande escala. Esse comportamento pode ser observado no experimento dois referente ao processo de *impeachment*, onde o processamento de três milhões de *tweets* do mês de março teve um elevado tempo de execução.

6.1 Limitações da abordagem

Mesmo além da limitação encontrada na quantidade de dados que são necessários para representar um evento, observamos que o método utilizado para o reconhecimento de entidade nomeada (*Polyglot*) torna nossa abordagem lenta, onde mesmo apresentando resultados representativos, o gasto computacional do método, exige um *hardware* mínimo que seja capaz de implementar a abordagem.

Outra limitação desta abordagem é a necessidade de aplicar todo o processamento descrito neste trabalho em cada amostra de dados que se deseja descobrir os tópicos, ou seja, durante o procedimento descritos nos experimentos referentes ao processo de *impeachment*, executamos as duas abordagens nos 10 meses analisados, e com isso, gerando um alto custo para treinando dos algoritmos, pois nossa abordagem foi proposta pensando em um cenário *offline*.

6.2 Trabalhos Futuros

Como trabalhos futuros, pretendemos adaptar nossa abordagem para que seja capaz de ser executada num cenário em *streamming*, assim reduziríamos o gasto computacional para o processamento dos tópicos. Outra abordagem, seria aplicar algoritmos de detecção de comunidade e conceitos de redes complexas para inferir as informações relevantes de base de dados.

Outro fator observado, é que durante a filtragem de um *tweet* que representa o tópico extraído, não consideramos a polaridade do texto, ou seja, se um documento possuir o *trigram* que representa o tópico extraído, o mesmo será filtrado e apresentado ao leitor independente do polaridade do documento.

Referências Bibliográficas

F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web*, pages 305–308. ACM, 2012.

F. Aisopos, G. Papadakis, K. Tserpes, and T. Varvarigou. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 187–196. ACM, 2012.

R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada*. SIAM, 2015.

J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. 1998.

T. G. Almeida, B. A. Souza, A. A. Menezes, C. M. Figueiredo, and E. F. Nakamura. Sentiment analysis of portuguese comments from foursquare. In *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*, Webmedia '16, pages 355–358, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4512-5. doi: 10.1145/2976796.2988180.

E. Alpaydin. *Introduction to machine learning*. The MIT Press Cambridge, Massachusetts London, England, 2010.

F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 2013.

H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11:438–441, 2011.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003a. ISSN 1532-4435.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003b.

L. Bolelli, Ş. Ertekin, and C. L. Giles. Topic and trend detection in text collections using latent dirichlet allocation. In *European Conference on Information Retrieval*, pages 776–780. Springer, 2009.

T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 330–337. ACM, 2003.

M. J. Brzozowski and D. M. Romero. Who should i follow? recommending people in directed social networks. In *ICWSM*, 2011.

M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM, 2010.

S. Chakraborti and S. Dey. Multi-level k-means text clustering technique for topic identification for competitor intelligence. In *Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on*, pages 1–10. IEEE, 2016.

Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

A. M. Correia. *Aprendizagem automática em larga escala nas redes sociais para a descoberta de ameaças de segurança*. PhD thesis, UNIVERSIDADE DE LISBOA, 2016.

Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 536–544. Association for Computational Linguistics, 2012.

P. S. Earle, D. C. Bowden, and M. Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012.

N. B. Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in knowledge discovery and data mining. 1996.

A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31 (8):651–666, 2010.

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 775–784. ACM, 2011.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

S. Katragadda, S. Virani, R. Benton, and V. Raghavan. Detection of event onset using twitter. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 1539–1546. IEEE, 2016.

Y. Kropivnitskaya, K. F. Tiampo, J. Qin, and M. A. Bauer. Real-time earthquake intensity estimation using streaming data analysis of social and physical sensors. *Pure and Applied Geophysics*, pages 1–19, 2016.

A. C. Lorena and A. C. de Carvalho. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67, 2007.

H. J. Miller. Geographic data mining and knowledge discovery. *Handbook of Geographic Information Science*, pages 352–366, 2004.

T. M. Mitchell et al. Machine learning, 1997.

M. C. Monard and J. A. Baranauskas. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1:1, 2003.

M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2): 167–256, 2003.

O. Ozdikis, P. Senkul, and H. Oguztuzun. Semantic expansion of hashtags for enhanced event detection in twitter. In *Proceedings of the 1st International Workshop on Online Social Systems*. Citeseer, 2012.

S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.

S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123. IEEE, 2010.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931, 2013.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM, 2009.

H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *Icwsm*, 2009.

F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42 (2):373–386, 2006.

B. A. Souza, T. G. Almeida, A. A. Menezes, F. G. Nakamura, C. M. Figueiredo, and E. F. Nakamura. For or against?: Polarity analysis in tweets about impeachment process of brazil president. In *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*, Webmedia '16, pages 335–338, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4512-5. doi: 10.1145/2976796.2988216.

M. Stiilpen Junior and L. H. C. Merschmann. A methodology to handle social media posts in brazilian portuguese for text mining applications. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 239–246. ACM, 2016.

S. Suh, J. Choo, J. Lee, and C. K. Reddy. L-ensnmf: Boosted local topic discovery via ensemble of nonnegative matrix factorization. 2016.

B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite. Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual review of sociology*, pages 213–238, 1996.

I. H. Witten, A. Moffat, and T. C. Bell. *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, 1999.

L. K. Wives. *Descobrindo eventos locais utilizando análise de séries temporais nos dados do Twitter*. PhD thesis, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2013.

W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association, 2012.

W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, pages 338–349. Springer, 2011.