Uma Abordagem para Detecção de Tópicos Relevantes em Redes Sociais Online

Bruno Á. Souza[1], Thais G. Almeida[1], Alice A. Menezes[1], Carlos M. S. Figueiredo[2], Fabíola G. Nakamura[1], Eduardo F. Nakamura[1]

[1]Universidade Federal do Amazonas (UFAM) – Manaus, AM – Brasil

[2]Universidade do Estado do Amazonas (UEA) – Manaus, AM – Brasil

{bruno.abia,tga,alice.menezes,fabiola,nakamura}@icomp.ufam.edu.br

cfigueiredo@uea.edu.br

Abstract. *The popularization of social networks has contributed to increase the amount of information shared daily, making these networks a source of information about diversified events. However, this information is difficult to understand, since there is a contextual diversity about these events and the high cost of processing to eliminate noises, make the process of recognition of relevant information challenging. In this context, this work propose an approach to characterize relevant information to events, through the extraction of topics in shared data on Twitter, where we used as a study scenario the phases of Lava Jato operation in 2016. In this way, we evaluated three machine learning methods (K-means, LDA and NMF) and compared pre-processing techniques for cleaning texts in order to observe if there is an improvement in algorithms performance. In addition, we use the Silhouette technique to find the best value of clusters, eliminating the need to rank relevant topics. In our results we demonstrated that our approach is able to monitor social networks to characterize events when we use NMF combined with named entity recognition.*

Resumo. *A popularização das redes sociais tem contribuído para o aumento da quantidade de informações compartilhadas diariamente, tornando estas redes uma fonte de informação sobre eventos diversos. Porém essas informações são de difícil compreensão, uma vez que há uma diversidade contextual sobre esses eventos e o custo de processar uma elevada massa de dados, tornam desafiador o processo de reconhecimento de informações relevantes. Neste contexto, este trabalho propõe uma abordagem para caracterização de informações relevantes de eventos, através da extração de tópicos em dados compartilhados no Twitter, onde utilizamos como cenário de estudo as fases da operação Lava Jato em 2016. Deste modo, avaliamos o desempenho de três métodos de aprendizagem de máquina (K-means, LDA e NMF) e comparamos técnicas de pré-processamento para limpeza dos textos com o intuito de observar se há melhora no desempenho dos algoritmos. Além disso, utilizamos a técnica Silhouette para estimar um valor de cluster sobre uma determinada amostra de dados, eliminando a necessidade de classificar quais tópicos são relevantes. Em nossos resultados demonstramos que nossa abordagem é capaz de monitorar redes sociais com o intuito de caracterizar eventos quando utilizamos o NMF combinado com reconhecimento de entidade nomeada.*

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). Os usuários das RSOs costumam compartilhar informações relacionadas ao contexto no qual estão inseridos, como por exemplo, informações de trânsito, negócios, relatos pessoais e acontecimentos diários [Ramos et al., 2016]. Isso possibilita que as RSOs funcionem como sensores para identificação de dados do mundo real, devido à quantidade de informações que são produzidas diariamente e diversidade dos dados, como: localização, horário, fotos e vídeos [Sakaki et al., 2010].

Um exemplo de Rede Social Online que disponibiliza informações compartilhadas de usuários é o Twitter[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]. O Twitter permite que os usuários façam atualizações (*tweets*) com até 140 caracteres, onde o intuito é compartilhar informações curtas e de rápida leitura. Com isso, é possível monitorar essa rede social para descobrir eventos de grande escala, onde esses acontecimentos dentro do ambiente de RSO são considerados como assuntos importantes e com grande repercussão, que servem para caracterizar fenômenos sociais (seja na esfera econômica, cultural ou política), como por exemplo, ações sobre o *impeachment* no Brasil [Souza et al., 2016] e os terremotos no Japão [Sakaki et al., 2010].

Outra aplicação em descoberta de eventos nas redes sociais, é a possibilidade de se monitorar um determinado acontecimento como, por exemplo, as fases da Operação Lava Jato da Polícia Federal, e descobrir quais são os tópicos mais relevantes de um determinado dia ou período, a fim de quantificar as pessoas envolvidas e os impactos gerados. Neste contexto, este trabalho tem como objetivo comparar técnicas não-supervisionadas de aprendizagem de máquina na tarefa de detecção de tópicos em Redes Sociais Online, com o intuito de propor uma abordagem que permita extrair informações e demonstrar a viabilidade de se utilizar o Twitter para descobrir relatos relevantes de um evento.

Partindo do cenário discutido, as principais contribuições deste artigo são: (a) avaliação de técnicas de aprendizagem de máquina (*K-means*, *Non-negative Matrix Factorization* e *Latent Dirichlet Allocation*) na tarefa de extração de tópicos em textos em português; (b) análise de duas abordagens de pré-processamento para textos, buscando eliminar ruídos existentes nos dados; (c) estudo de caso considerando os *tweets* compartilhados sobre as fases da Operação Lava Jato; e (d) a utilização da técnica de *Silhouette* para estimar a quantidade de agrupamentos em uma amostra de dados, a fim de obter um valor de *cluster* que melhor represente esse conjunto, pois como a maioria dos trabalhos estima altos valores para a extração, este processo acaba se tornando lento e custoso.

O restante deste trabalho está dividido da seguinte forma: na Seção 2 são apresentados os trabalhos relacionados à extração de tópicos em redes sociais; na Seção 3 é descrita a metodologia utilizada na condução deste trabalho; na Seção 4 são apresentados os experimentos e os resultados obtidos. Por fim, a Seção 5 apresenta a conclusão e trabalhos futuros.

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

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

2. Trabalhos Relacionados

Os estudos sobre o problema de detecção de eventos tiveram início em 1998, sendo tratado como um problema de descoberta de tópicos, onde o objetivo é o agrupamento de informações de maneira contextualizada [Allan et al., 1998]. Portanto, dado um conjunto A de textos, existe uma quantidade B de contextos (temas) 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) o processamento da quantidade de informações geradas; e (ii) os ruídos existentes, visto que nem todas as informações compartilhadas sobre um determinado tema são relevantes. Buscando solucionar estes desafios, diversos trabalhos que propõem abordagens de seleção de tópicos relevantes por palavras-chave têm surgido [Zhao et al., 2011; Diao et al., 2012; Cataldi et al., 2010; Suh et al., 2016; Katragadda et al., 2016].

Esses métodos utilizam palavras-chave para criar um agrupamento de termos e realizam a detecção de tópicos usando modelos probabilísticos, como por exemplo, o *Latent Dirichlet Allocation* (LDA), que é utilizado no trabalho de Bolelli et al. [2009] como forma de identificar tópicos relevantes. Diao et al. [2012] aplicam o LDA para identificar tópicos de acontecimentos passados em dados do Twitter. Os autores assumem que os tópicos são constituídos de palavras-chave de múltiplos *tweets*, onde o principal objetivo de sua abordagem 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. Com isso, os autores demonstram uma abordagem diferente dos trabalhos que se limitam a apenas reconhecer a informação.

No trabalho de Xu et al. [2003], os autores apresentam uma adaptação do *Non-negative Matrix Factorization* (NMF) para demonstrar a eficiência do método em extrair tópicos. O estudo mostrou que em comparação aos métodos LSI e SC, a adaptação realizada no NMF para extrair tópicos apresenta um rendimento superior de 1.16% em relação aos demais. No trabalho de Suh et al. [2016], os autores capturaram tópicos locais de sua região, a fim de detectar informações relevantes. Os autores compararam o NMF com suas variantes e o LDA. Desta forma, concluíram que o modelo utilizando o NMF 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.

A partir dos desafios citados anteriormente e das técnicas utilizadas para detecção de informações relevantes, nossa abordagem foca em analisar uma amostra de dados e poder definir qual a quantidade de *clusters* que melhor a representam, a fim de reduzir a quantidade de interações que os métodos executam para extrair tópicos. Outro fator que pode ser destacado, é a aplicação de uma arquitetura de pré-processamento e extração de tópicos que seja capaz de manipular a larga quantidade de dados capturados sobre um determinado evento, a fim de caracterizar os relatos mais significativos. Além disso, aplicamos as técnicas de extração de tópicos baseadas nos trabalhos anteriores com o intuito de validar a abordagem feita neste trabalho.

3. Método Proposto

O método proposto na elaboração deste trabalho consiste das seguintes etapas (ilustradas na Figura 1): (i) coleta de dados provenientes do Twitter para construção da base de dados; (ii) pré-processamento dos dados, a fim de retirar possíveis ruídos dos *tweets*; e (iii) seleção da melhor quantidade e extração de tópicos utilizando a técnica de *Silhouette* combinada com os métodos de aprendizagem de máquina. Nas subseções que seguem, detalhamos cada uma destas etapas.

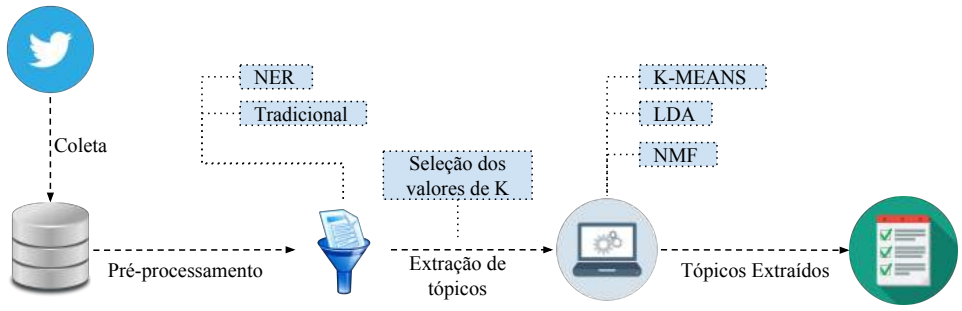


Figura 1. Arquitetura do método proposto.

3.1. Base de Dados

A base de dados utilizada neste trabalho é composta por *tweets* escritos em português, pertencentes ao período de 01/01/2016 a 20/11/2016. Os *tweets* da nossa base de dados são relativos a Operação Lava Jato, deflagrada em 2014 pela Polícia Federal do Brasil, cujo objetivo é investigar a prática de crimes financeiros e desvio de recursos públicos. Neste artigo, consideramos somente as fases deflagradas no ano de 2016 que correspondem às fases da 22^a à 36^a.

A fim de coletar *tweets* sobre a Operação Lava Jato, utilizamos a Search API da rede social Twitter[3], que permite a coleta de dados históricos. Para filtrar *tweets* relativos a determinados assuntos, a Search API utiliza uma *query* de busca, que é equivalente a uma expressão boleana envolvendo termos. No trabalho proposto, a *query* de busca especificada foi "lava jato", e resultou em uma base com um total de 652.210 *tweets*. A Figura 2 mostra a distribuição de *tweets* por mês.

3.2. 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, detalhados na próxima subseção. Para tanto, utilizamos o *framework* Apache Spark[4] que facilita a manipulação de dados em larga escala (*Big Data*). O *framework* Apache Spark baseia-se em duas das principais abstrações de programação paralela [M. Zaharia and Stoica, 2010]: *Resilient Distributed Datasets* (uma coleção de objetos particionados entre um conjunto de máquinas que podem ser reconstituídos caso uma partição seja perdida) e operações paralelas (e.g., *reduce*, *collect*, *foreach*).

[3]https://www.twitter.com
[4]http://spark.apache.org/

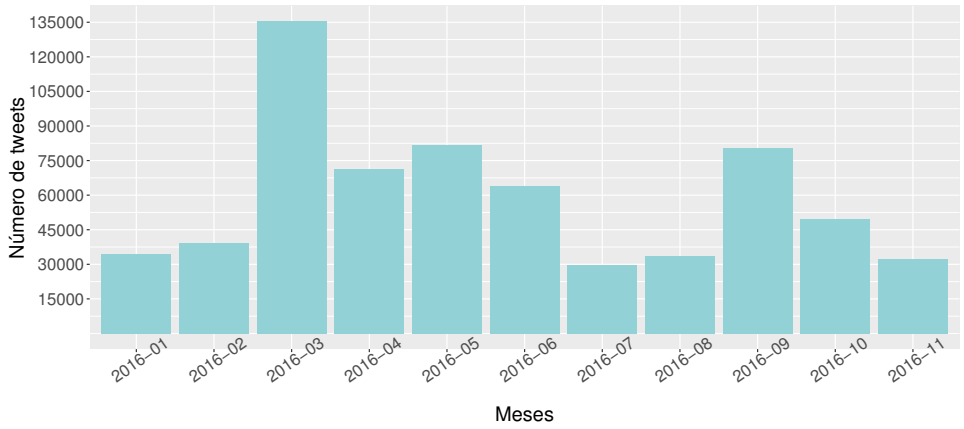


Figura 2. Distribuição de *tweets* por mês.

Neste trabalho, realizamos dois pré-processamentos distintos: tradicional e com reconhecimento de entidade. O pré-processamento tradicional (figura 3), 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) e expansão de contrações [Stiilpen Junior and Merschmann, 2016]. Após a normalização, os *tokens* que forem *stopwords* são descartados e, por fim, os afixos dos *tokens* restantes são eliminados (*stemming*).

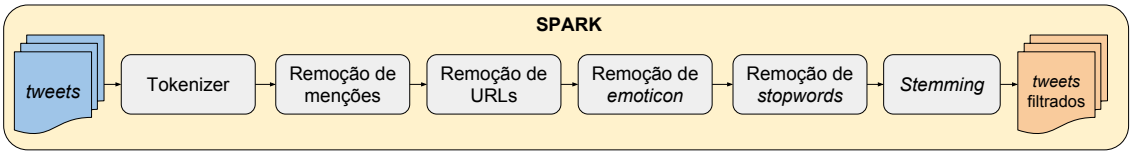


Figura 3. Pré-processamento tradicional.

O pré-processamento com reconhecimento de entidade (Figura 4), 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. O objetivo destes dois pré-processamentos é verificar se há melhora na eficiência dos métodos ao extrair tópicos, pois as matrizes de termos em cada pré-processamento possuem diferentes características.

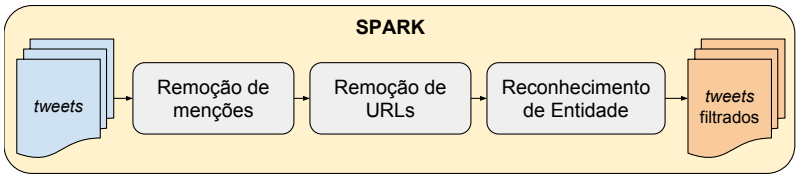


Figura 4. Pré-processamento com reconhecimento de entidade.

Após a aplicação do pré-processamento sobre a base de dados, o vetor de características de cada um dos metódos 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 dos métodos de aprendizagem de máquina e reduz a quantidade de termos com baixa frequência. Esse valor foi selecionado, após observarmos que durante os experimentos esses termos não tinham representatividade dentro da matriz de termos.

3.3. 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*[5], que baseada numa amostra de dados, permite que observemos através de uma representação gráfica qual seria o melhor valor de k para esse conjunto. Essa técnica permite a validação da consistência dos dados dentro de um agrupamento, onde em sua definição é assumido que os dados são agrupados por uma técnica que tenha como parâmetro de entrada um número K de *clusters*. Matematicamente, o *Silhouette* pode ser representado pela Equação 1:

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

onde $a(i)$ é uma representação de quão bem um dado i é atribuído a um *cluster*, e $b(i)$ a menor média de não similaridade de i em relação a qualquer outro *cluster* que o dado não seja membro. Com isso, podemos obter valores de [-1, 1]. Esses coeficientes quando estão perto de +1 indicam que a amostra está longe dos *clusters* vizinhos. Um valor 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 (conforme ilustrado na Figura 5).

Esses valores obtidos através da técnica de *Silhouette* são utilizados como parâmetros de entrada para os algoritmos de extração de tópicos, ou seja, cada número inteiro obtido corresponde a quantidade de tópicos da amostra em análise. Vale ressaltar, que essa técnica foi aplicada a cada algoritmo, levando em consideração que os mesmos possuem estratégias diferentes para alcançar o mesmo resultado.

O ***Latent Dirichlet Allocation (LDA)*** consiste em um modelo probabilístico, no qual cada documento é modelado como uma combinação de tópicos e onde cada tópico corresponde a uma distribuição multinomial sobre as palavras [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. Em nossa abordagem seguimos a estratégia adotada por Zhao et al. [2011], onde assumimos que um tópico T também está atrelado a um usuário

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

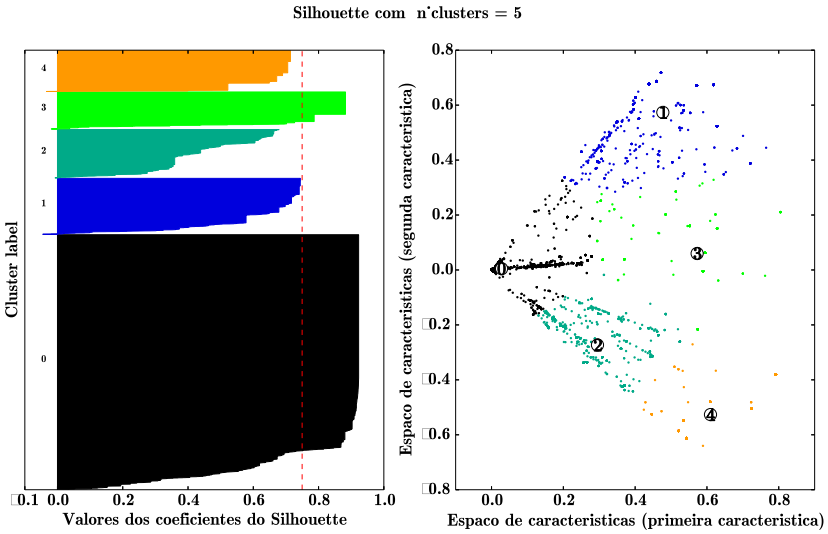


Figura 5. Exemplo do gráfico de *Silhouette* para seleçao de valores de k com k 5.

e não unicamente a uma distribuição de palavras, pois conforme observado, o modelo tradicional do LDA apresenta baixo rendimento em textos curtos [Zhao et al., 2011].

O ***K-means***. Seja X = x_1,x_2,...,x_d o conjunto de pontos d-dimensionais a serem agrupados em um conjunto de K *clusters*, C = c_k, $k = 1$,...,K. O *K-means* encontra uma partição de tal forma que o erro quadrático entre a média empírica de um *cluster* e os pontos no *cluster* sejam minimizados.

O ***Non-negative Matrix Factorization*** **(NMF)**. O NMF consiste em um algoritmo de fatorização de matrizes que encontra a fatorização 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 $m \times d$, onde m é o número de termos distintos e d é equivalente ao número de documentos em S. O NMF encontra a menor aproximação de V em termos de alguma métrica específica (e.g., norma) fatorando V no produto (WH) de duas matrizes de menor dimensões $W_{m\times k}$ e $H_{k\times n}$, 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].

4. Experimentos e Resultados

Durante esta fase conduzimos os nossos experimentos em dois cenários, onde o primeiro cenário consiste em demonstrar a utilização das técnicas de extração de tópicos, tendo como entrada uma matriz de termos que foi processada com uma arquitetura tradicional (Figura 3). O segundo cenário descreve a utilização de pré-processamento através do reconhecimento de entidades (Figura 4), onde a matriz de termos gerada, consiste unicamente de termos reconhecidos como entidades em cada *tweet*.

Para validarmos nossa arquitetura de extração de tópicos, ilustrada na Figura 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) divulgadas nos meios de comunição tradicionais ($P = \frac{topicos_{relacionados}}{Noticias}$). O objetivo desta abordagem é comprovar que monitorando o Twitter, conseguimos descobrir informações relevantes sobre um determinado acontecimento.

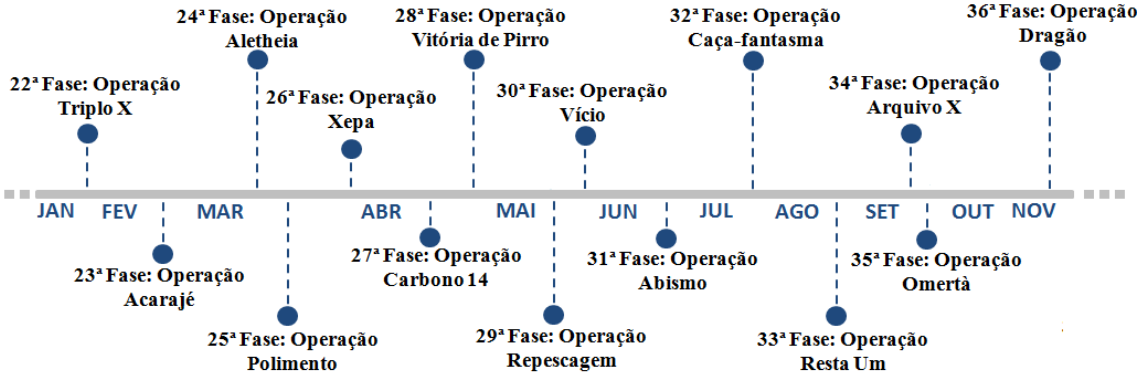


Figura 6. Linha do tempo das fases da Operação Lava Jato em 2016.

Durante os experimentos aplicamos os algoritmos de extração de tópicos aos conjuntos de dados de cada mês, pois conforme observado na Figura 6, a Operação Lava Jato da Polícia Federal, de janeiro a novembro teve fases relatadas nas redes sociais mensalmente, atingindo um total de 15 operações. Vale ressaltar que levamos em consideração todos os relatos compartilhados sobre esses eventos, desde prisões de pessoas envolvidas à ações judicias executadas pelos órgãos responsáveis.

	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 1. Valores de K observados na técnica de *Silhouette*.

Em ambos os cenários, executamos a técnica de *Silhouete* para obter a melhor quantidade de tópicos para cada amostra de dados (cada mês), conforme observado na Tabela 1, ou seja, a cada amostra de dados executamos a variação de *clusters* de 2 até 15 para realizar a verificação.

4.1. Cenário I

Neste cenário, para uma melhor compreensão de como as avaliações foram realizadas, selecionamos o mês de setembro que teve maior divergência entre os tópicos extraídos em relação as informações compartilhadas em outros canais, onde ocorreram a 34ª e a 35ª 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 [6], O Globo [7], Estadão[8] e site da Polícia Federal[9]) que mais detalhavam os acontecimentos da Operação Lava jato foram:

1. O ex-ministro Guido Mantega é preso temporariamente, mas o juiz Sérgio Moro manda soltá-lo horas depois.
2. O empresário Eike Batista declarou que, em 1º de novembro de 2012, recebeu pedido de "um então ministro e então presidente do Conselho de Administração da Petrobrás" para que fizesse um pagamento de R$ 5 milhões destinado ao PT [10].
3. Supremo Tribunal Federal manda ação contra Eduardo Cunha para Sérgio Moro.
4. O ex-ministro da Fazenda Antônio Palocci é preso na 35ª operação, por atuar de maneira direta para proporcionar vantagens à empreiteira Odebrecht.
5. Gleisi Hoffmann e Paulo Bernardo virão réus da Operação Lava Jato no Supremo Tribunal Federal.

Neste contexto, ao aplicarmos o pré-processamento apresentado na Figura 3 com os algoritmos de aprendizagem de máquina, observamos se os tópicos extraídos tinham alguma menção as notícias compartilhadas nos canais de informação. Durante esta avaliação, observamos que os tópicos extraídos pelo NMF, conforme apresentado na Tabela 2, foram os que mais se aproximaram das notícias compartilhadas por outras fontes de informação (conforme ilustrado na Figura 7(a)) .

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 2. Tópicos extraídos pelos algoritmos.

Observamos durante os experimentos com esta abordagem que a base de dados possui outras informações que não estavam no contexto direto da Lava Jato. Essas informações, que possuem maiores frequências dentro de uma *bag-of-words* acabam em sua maioria se tornando um tópico coletado pelos algoritmos. Com isso, este cenário

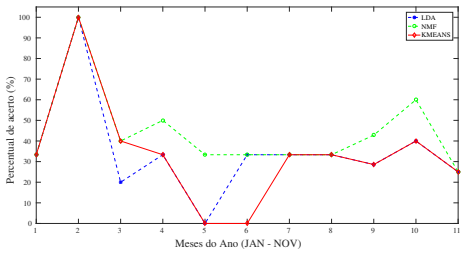
[6] http://paranaportal.uol.com.br/editoria/operacao-lava-jato/

[7] http://infograficos.oglobo.globo.com/brasil/todas-as-fases-da-operacao-lava-jato.html

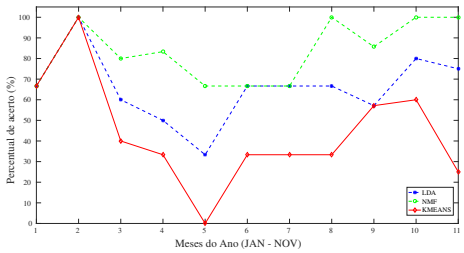
[8] http://infograficos.estadao.com.br/public/politica/operacao-lava-jato/fases/

[9] http://www.pf.gov.br/imprensa/lava-jato/fases-da-operacao-lava-jato

[10] http://politica.estadao.com.br/blogs/fausto-macedo/eike-batista-relatou-a-procuradoria-pedido-de-r-5-milhoes-ao-pt/



(a) Pré-processamento tradicional.



(b) Pré-processamento com reconhecimento de entidade.

Figura 7. Resultado da validação de tópicos comparado com canais de notícias.

apesar de ter um pré-processamento robusto, não apresentou bom desempenho, conforme observado na Figura 7(a), tornando inviável a utilização deste padrão para o monitoramento das notícias relevantes de um acontecimento, devido a captura desses ruídos pelos algoritmos aplicados neste trabalho.

4.2. Cenário II

Neste cenário, para fins de avaliação, será apresentado o mesmo mês analisado no cenário 1. Aplicamos como pré-processamento o reconhecimento de entidade nomeada, onde para documento (*tweet*) pode existir uma ou mais entidades, conforme exemplo: "*Tribunais Superiores derrubam 18 erros de Sérgio Moro.*", tendo como resultado o reconhecimento da entidade "Sérgio 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. Como as notícias ou relatos estão atreladas a uma determinada pessoa, grupos, empresas ou locais, a utilização dessa estratégia torna-se viável.

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 3. Tópicos extraídos pelos algoritmos com reconhecimento de entidade.

Durante esse pré-processamento, eliminamos entidades com frequências menores que 20, com o intuito de igualar os parâmetros de acordo com o primeiro cenário. Como resultado, observamos que o NMF apresentou um melhor desempenho (Figura 7(b)) na extração de tópicos, 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 3).

Com isso, sumarizamos nossos resultados 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 Operação Lava Jato da Polícia Federal do Brasil;

2. O pré-processamento com reconhecimento de entidade nomeada aplicado a nossa arquitetura, apresentou rendimento superior a 70% em todos os meses com o algoritmo NMF, comprovando a viabilidade de utilizar este modelo para monitorar o Twitter, com o objetivo de identificar as informações relevantes de um determinado acontecimento, pois observar um evento monitorando diretamente as pessoas envolvidas se mostrou mais aplicável;
3. 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 algoritmo *PolyGlot-Ner*;

5. Conclusão e Trabalhos Futuros

Neste trabalho, foram analisadas técnicas não-supervisionadas de aprendizagem de máquina, a fim de descobrir tópicos relevantes em dados compartilhados em redes sociais. Apesar de avanços em técnicas de detecção de eventos em *streaming*, observamos que uma aprendizagem *offline*, quando combinada com ferramentas que aceleram o processo de manipulação dos dados pode alcançar resultados no mesmo nível de outras técnicas. Com isso, comparamos dois modelos para extração de tópicos, combinando com os algoritmos de LDA, NMF e *K-means*, na tarefa de monitorarmos um determinado evento para descobrir quais foram as informações relevantes daquele mês. Além disso, reduzimos a quantidade de informações extraídas utilizando a técnica de *Silhouette*, a fim de verificar a melhor quantidade de *cluster* para uma determinada amostra de dados.

Esses modelos, foram avaliados em uma base de dados de *tweets* que relatavam informações sobre a Operação Lava Jato da Polícia Federal no ano de 2016. Os resultados mostraram que a nossa arquitetura, utilizando reconhecimento de entidade nomeada como pré-processamento, combinado com o algoritmo de NMF, são os que apresentam melhor resultado neste cenário. Como trabalhos futuros, pretendemos replicar nossos experimentos, adequando nossa arquitetura ao cenário online, além utilizar outras técnicas de *cluster*, como por exemplo, o uso de redes neurais.

6. Referências

Aisopos, F., Papadakis, G., Tserpes, K., and Varvarigou, T. (2012). 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.

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

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

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

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

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). 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.

Diao, Q., Jiang, J., Zhu, F., and Lim, E.-P. (2012). 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.

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

M. Zaharia, M. Chowdhury, M. J. F. S. S. and Stoica, I. (2010). Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*.

Ramos, P., Reis, J., and Benevenuto, F. (2016). Uma analise da polaridade expressa nas manchetes de notıcias brasileiras.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). 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.

Shahnaz, F., Berry, M. W., Pauca, V. P., and Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. In *Proceedings of the Information Processing & Management*, volume 42, pages 373–386. Elsevier.

Souza, B. A., Almeida, T. G., Menezes, A. A., Nakamura, F. G., Figueiredo, C. M., and Nakamura, E. F. (2016). 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. ACM.

Stiilpen Junior, M. and Merschmann, L. H. C. (2016). 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.

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

Xu, W., Liu, X., and Gong, Y. (2003). 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.

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