Detecção de Eventos no Twitter através de Grafos de Visibilidade Natural

Fernanda Tenório[1], Eduarda T. C. Chagas[2], Pedro H. Barros[2], Heitor S. Ramos[2]

[1] Laboratório de Computação Científica e Análise Numérica (LaCCAN)
Universidade Federal de Alagoas (UFAL) – Maceio, AL – Brazil

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

fernanda.tenorio@ctec.ufal.br

{eduarda-chagas, phbarros}@ufmg.br

heitor@dcc.ufmg.br

Abstract. *The Internet provides an ever-increasing volume of data and information, helping us to better understand its users and the environment around them. One of the approaches used to detect and understand the events that occur in the world has been an analysis of social networks, such as Twitter, used in this article. Thus, considering the change in the dynamics of data behavior after the presence of an event, we propose a new detection method based on the computation of complex network metrics applied to the bigram extracted from the content of tweets, identifying events through system dynamics changes. To validate our approach, we used two sets of data collected by [Aiello et al. 2013], in which we observed satisfactory results when compared with the techniques already present in the literature.*

Resumo. *A Internet vem nos fornecendo cada vez mais dados e informações, ajudando a compreender melhor os seus usuários e o ambiente que os rodeiam. Uma das abordagens usadas para detectar e compreender eventos que ocorrem ao redor do mundo vem sendo a análise de redes sociais, como o caso do Twitter, usado no presente artigo. Assim, considerando a mudança da dinâmica do comportamento dos dados após a presença de um evento, propomos um novo método de detecção baseado no cálculo de métricas de redes complexas aplicadas aos bigram extraídos do conteúdo de tweets, identificando eventos por meios de mudanças de dinâmica do sistema. Para validar nossa proposta usamos dois conjuntos de dados coletados por [Aiello et al. 2013], no qual observamos resultados satisfatórios quando comparados com as técnicas já presentes na literatura.*

1. Introdução

Nos últimos anos, a popularização da Internet nos fornece grandes quantidades de dados. Ao se tornar mais acessível ao público, tornou-se crescente o número de usuários que possuem como finalidade a comunicação e o entretenimento. Assim, foi imprescindível o crescimento de mídias sociais e o processo de globalização da informação, transformando a internet em uma valiosa fonte de dados, que ao serem tratados e analisados, podem nos

revelar importantes características do comportamento social, político e econômico dos usuários nas redes.

Sendo a 12^{o} rede social com maior número de membros ativos, o Twitter possui 100 milhões de usuários ativos diários e cerca de 6.000 *tweets* por segundo, atingindo mais de 500 milhões de *tweets* por dia[1]. O aumento no fluxo de informações na internet e o consequente o crescimento de mídias sociais, dentre elas o Twitter, se dá principalmente pela oferta de smartphones mais baratos no mercado, cobertura Wi-Fi acessível e expansão das redes de 3G e 4G.

Permitindo que os usuários se comuniquem por meio de micro-mensagens, de até 280 caracteres, o Twitter se tornou uma poderosa ferramenta para detectar e compreender eventos que ocorrem ao redor do mundo. [Dou et al. 2012] define um evento como "uma ocorrência que causa mudanças no volume de dados de texto que discutem o tópico associado em um momento específico". Logo, eventos podem ser vistos como um resumo sucinto dos fluxos de informações nas mídias sociais, revelando a evolução de fenômenos sociais ao longo de um determinado período de tempo.

A análise das relações existentes entre o uso de palavras-chave em um intervalo de tempo, eventos e as respostas dos usuários aos eventos são capazes de revelar informações sofisticadas das opiniões das massas sobre um determinado fenômeno. Embora considerada uma valiosa fonte de informações quando comparada às mídias tradicionais de notícias, devemos salientar que existem alguns desafios ao trabalhar com análise de dados gerados pelo Twitter, como por exemplo:

(i) Os algoritmos devem possuir uma abordagem escalonável para suportar a grande quantidade de dados gerado pelos *tweets*;

(ii) Os *tweets* são compartilhados em tempo real, logo possuem uma grande relação com o contexto temporal na qual se encontra inserido;

(iii) Devido a limitação de espaço em suas mensagens, os *tweets* geralmente possuem a característica de apresentar um conteúdo breve e informal, com frases não estruturadas, erros de digitação e abreviaturas, exigindo um tratamento prévio de seu conteúdo;

(iv) Nem todos os *tweets* apresentam informações úteis. Como já destacado por [Parikh and Karlapalem 2013] "metade dos *tweets* são inúteis e não transmitem nenhuma informação valiosa ". Embora tais dados não prejudiquem o tempo de processamento, eles são extremamente nocivos ao resultado final da análise.

Assim, neste trabalho apresentamos um novo método para detecção de eventos no Twitter baseado na interpretação do grafo de visibilidade, onde inferimos a presença ou não de um evento, por meio de uma técnica de aprendizagem adaptativa aplicada na análise de mudança das frequências dos bigramas[2] (duas palavras consecutivas para um dado elemento textual) presentes nos *tweets*.

Este trabalho está organizado da seguinte forma: Seção 2 apresenta os trabalhos relacionados a detecção de eventos no Twitter; Seção 3 descreve a metodologia utilizada para análise dos dados; Seção 4 apresenta os principais resultados; e Seção 5 conclui esse

[1] `https://www.omnicoreagency.com/twitter-statistics/`
[2] `https://en.wikipedia.org/wiki/Bigram`

trabalho.

2. Revisão da literatura

Há diversos estudos presentes na literatura que possuem como objetivo extrair informações e detectar e/ou sumarizar eventos nas mídias sociais de modo eficiente. Como pode ser visto a seguir, grande parte das técnicas e abordagens focam na análise dos textos compartilhados pelo Twitter usando tópicos emergentes.

[Sakaki et al. 2010] conseguiram por meio da análise de *tweets* detectar com até 96% de precisão terremotos identificados pela *Japan Meteorological Agency* (JMA) com escala de intensidade sísmica igual ou maior a 3. Neste trabalho, construíram um modelo probabilístico com base em recursos como palavras-chave, número de palavras e o contexto dos *tweets*. Cada usuário do Twitter foi modelado como um sensor e aplicado um filtro de Kalman [Evensen 2003] juntamente com um filtragem de partículas [Nummiaro et al. 2003] conseguindo assim encontrar o centro e a trajetória do local do evento.

[Mathioudakis and Koudas 2010] propuseram o *TwitterMonitor* que utilizando um algoritmo de extração de contexto baseado em [Deerwester et al. 1990] que consegue identificar a tendência das palavras-chave e agrupá-las de acordo com suas co-ocorrências, obtendo não apenas a detecção em tempo real de eventos no Twitter, como também fornecendo uma síntese precisa da descrição de cada tópico.

[Cataldi et al. 2010] aplicaram o algoritmo *PageRank* para mensurar o grau de influência de cada usuário sobre os *tweets*. Por meio da análise da energia do conteúdo, a mensagem pode então ser classificada ou não como emergente e assim definida uma janela temporal para determinar seu ciclo de vida.

[Weng and Lee 2011] propuseram EDCoW (Detecção de Eventos com Agrupamento de Sinais Baseados em *Wavelet*), onde aplicaram análise de sinais [Rosso et al. 2009] e *Wavelets* nas palavras, conseguindo descartar aquelas consideradas triviais e detectar eventos usando agrupamento de sinais das palavras em conjunto com o particionamento do gráfico baseado em modularidade.

[Li et al. 2012] propôs o *Tweetvent*, que consegue detectar segmentos de *tweets* em uma janela de tempo fixa e por meio de uma variante do algoritmo Jarvis-Patrick [Jarvis and Patrick 1973] agrupar eventos candidatos, que posteriormente são comparados com artigos da Wikipédia para constatar a veracidade e importância dos resultados obtidos.

[Parikh and Karlapalem 2013] utilizaram a análise do comportamento das frequências dos bigramas extraídos ao longo de um determinado período de tempo e usando métricas de similaridade conseguiram agrupar palavras-chave relacionadas ao mesmo evento.

[Dang et al. 2016] criaram um modelo baseado em Redes Bayesianas Dinâmicas [Murphy and Russell 2002] para detectar palavras-chave emergentes e utilizaram DBSCAN [Ester et al. 1996] se baseando na co-ocorrência para realizar o agrupamento destas palavras, identificando assim tópicos emergentes. Tal método usa como entrada informações sobre o compartilhamento de *tweets* e o relacionamento de seguidores.

[Ester et al. 1996] desenvolveram um classificador dos eventos mais prováveis de um determinado período de tempo. Para classificar e detectar eventos com bom desempenho, consideram como entrada para o algoritmo o número de eventos que está ocorrendo em um intervalo de tempo e utilizam uma métrica baseada na frequência inversa de frequência do documento (TF-IDF) do bigrama.

Sabendo que o Twitter é caracterizado pela mudança incremental da distribuição dos seus dados, uma vez que alguns usuários possuem um tempo de resposta curto a eventos, enquanto outros demoram mais para realizar o *tweet*, nossa abordagem consiste em identificar a mudança da dinâmica do sistema e assim detectar a ocorrência de eventos. Mudanças de dinâmica não devem ser analisadas usando o mesmo conjunto de técnicas aplicadas para sistemas estacionários, sendo neste caso recomendadas abordagens adaptativas [Gama et al. 2014]

Nossa proposta se difere das demais devido ao fato de considerarmos que o sistema sofre uma mudança de dinâmica, ou seja, estamos observando uma mudança de dinâmica através da estrutura do grafo de visibilidade. Logo, buscamos observar mudança de comportamento, e não somente uma detecção de anomalias (*outliers*) como as técnicas encontradas na literatura usualmente fazem.

3. Metodologia

3.1. Conjunto de Dados

Para validar nossa proposta, usamos dois conjuntos de dados coletados por [Aiello et al. 2013], onde os autores reuniram *tweets* relacionados a alguns importantes eventos mundiais que aconteceram em 2012:

1. **FA Cup**[3]: Este conjunto de dados contém *tweets* sobre a Copa de Futebol Inglês (FA Cup). Esta é a principal competição masculina de futebol no país e pertence à mais antiga associação de futebol do mundo. A Figura 1 mostra a série temporal correspondente aos *tweets* coletados sobre a FA Cup. Cada amostra contém o número de *tweets* coletados em um minuto. Em 2012, Chelsea e Liverpool jogaram a partida final, com gols de Ramirez (11') e Drogba (52') para o Chelsea e Carrol (62') para o Liverpool. Assim, o Chelsea venceu a partida por $2-1$, tendo esta duração de 90 minutos, mais 15 minutos de intervalo;
2. **The Super Tuesday Primaries**: Nos Estados Unidos, o presidente é eleito em uma eleição indireta, com o vencedor sendo determinado por eleitores do Colégio Eleitoral. *Super Tuesday* refere-se informalmente às terças-feiras no início de uma temporada presidencial dos EUA, quando o maior número de estados realiza eleições primárias. A *Super Tuesday* em 2012 aconteceu em 6 de março de 2012, com 419 delegados (18,3 % do total). Mostramos na Figura 2 a série temporal correspondente aos tweets coletados sobre a *Super Tuesday* onde cada amostra contém o número de tweets coletados em cinco minutos.

[Aiello et al. 2013] criaram esses conjuntos de dados usando as *hashtags* oficiais do evento. Eles construíram o *ground truth* verificando os principais relatórios de mídia para identificar tópicos significativos para cada conjunto de dados. Foi identificado 13 e 22 tópicos significativos para o FA Cup e *Super Tuesday*, respectivamente. Os detalhes da construção do conjunto de dados podem ser encontrados no artigo original da proposta.

[3] `https://en.wikipedia.org/wiki/FA_Cup`

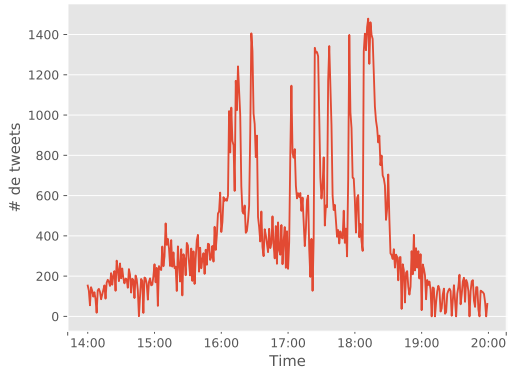

Figura 1. Série temporal do número de *tweets* para o conjunto de dados do FA Cup

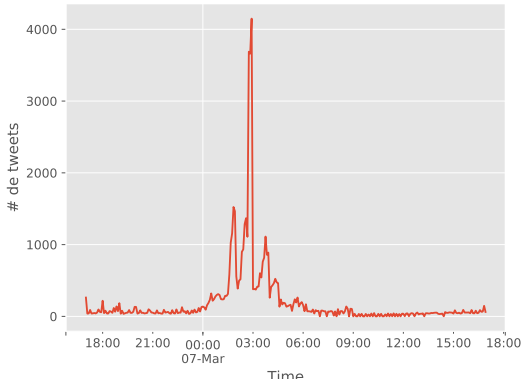

Figura 2. Série temporal do número de *tweets* para o *Super Tuesday*

3.2. Baseline

A distribuição de Poisson [Poisson 1837] é uma distribuição de probabilidade discreta que expressa a probabilidade de um número de eventos que ocorrem em um período de tempo fixo, dado que esses eventos ocorrem com uma taxa média conhecida e independentemente do tempo desde o último evento [Katti and Rao 1968]. Portanto a distribuição de Poisson é o modelo probabilístico mais comum para detectar eventos incomuns e independentes em uma unidade específica de espaço ou tempo.

Consideramos a taxa de Poisson constante dentro de uma janela de tempo (Processo de Poisson homogêneo), onde foi estimado a taxa de Poisson $\hat{\lambda}$ de um determinado tempo como a média do número de tweets coletados dentro de uma janela deslizante $P^{i:n} = \{p_i, p_{i+1}, ..., p_{i+n-1}\}$, que corresponde à frequência de todos os tweets em um horário i. Nós usamos 15 unidades de um minuto para calcular a janela deslizante. A taxa de Poisson $\hat{\lambda}$ é estimada como $\hat{\lambda} = \frac{1}{n}\sum_{j=i}^{i+n-1} p_j$.

Um evento é detectado quando observamos que a probabilidade de $k = p_{i+n}$ tweets em $P^{i:n}$ é suficientemente menor que um limite ϵ dentro de uma determinada janela deslizante. Avaliamos a probabilidade de k em $P^{i:n}$ como $\Pr(k;\hat{\lambda}) = \hat{\lambda}^k e^{-\hat{\lambda}}/k!$, com $\hat{\lambda} < k$. Nós detectamos um evento sempre que a distribuição de probabilidade (PDF) em k for menor que um determinado limite ϵ. Isso significa que k provavelmente será uma observação rara, dada a média $\hat{\lambda}$.

Devido à alta sensibilidade apresentada pela distribuição de Poisson, ou seja, é exponencial em λ, consideramos $\epsilon = 10^{-20}$ como um comportamento de anomalia. Se um evento for detectado na janela deslizante $P^{i:n}$, presumimos que o tempo $i+n$ é responsável pela anomalia, ou seja, a hora em que o evento ocorreu.

Após a detecção dos eventos, obtemos os bigramas mais frequentes, agrupando-os com base nos termos (palavras) em comum, formando as palavras-chave relacionadas ao evento detectado.

Observe que, embora a distribuição de Poisson detecte eventos sem conhecer o conteúdo dos *tweets*, nosso método difere pois analisa o assunto presente em cada *tweet*

Bigrama	Valor
game-soccer	4
sunday-game	4
fair-play	4
twitter-now	3
match-today	3
half-time	3
now-win	3
competing-teams	2
soccer-great	2
final-match	2
spirit-team	2
start-match	1

Tabela 1. Exemplo de alguns bigramas para uma dada unidade de tempo.

Bigrama	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
start-match	1	3	4	2	40	47	62	79

Tabela 2. Série temporal para o bigrama **start-match**.

por meio de bigramas.

3.3. Proposta

3.3.1. Calcular Bigrama

Utilizando a representação mostrada em [Barros et al. 2018], fizemos o mapeamento da contagem do número de bigramas por unidade de tempo. A tabela 1 mostra um exemplo para um valor de tempo t_0 no conjunto de dados do FA Cup.

3.3.2. Criação das séries temporais

À medida que o método é desenvolvido para processamento em tempo real, primeiro coletamos um conjunto de observações dentro de uma janela e depois avançamos pela janela para analisar mais dados. Portanto, primeiro determinamos o mapeamento definido na etapa anterior para a primeira janela e continuamos a determinar esses conjuntos para as janelas subsequentes.

A janela é denotada por $W^{i:n} = \{X_p\}$, onde $p \in \{t_i, t_{i+1}, \ldots, t_{i+n-1}\}$ é um intervalo de tempo dentro de W, t_i é o tempo inicial e n é o número de elementos de W, um exemplo para $f(b)$ com o bigrama b = start-match. X_p denota o conjunto de contagens para todos os bigramas no momento p. A janela, com tamanho n, desliza uma unidade de tempo, de forma que a diferença entre janelas consecutivas é exatamente uma observação, como podemos ver na tabela 3, para uma janela de tamanho $n = 4$. Observe que essa janela consideramos apenas a contagem de um bigrama específico.

Tabela 3. Abordagem da janela deslizante.

Bigrama - (start, match)								
Tempo	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
Contagem	1	3	4	2	40	47	62	79
Contagem ($W^{0:3}$)	1	3	4	2	-	-	-	-
Contagem ($W^{1:4}$)	-	3	4	2	40	-	-	-
Contagem ($W^{2:5}$)	-	-	4	2	40	47	-	-
Contagem ($W^{3:6}$)	-	-	-	2	40	47	62	-
Contagem ($W^{4:7}$)	-	-	-	-	40	47	62	79

3.3.3. Grafo de Visibilidade

O grafo de visibilidade consiste num algoritmo que mapeia séries temporais em grafos, como podemos ver na figura 3 e 4, onde vemos a transformação de uma série temporal (a esquerda) em um grafo (a direita) utilizando a abordagem do grafo de visibilidade. [Lacasa et al. 2008] estuda como os grafos são úteis na caracterização dessas séries temporais, além de mostrar que os mesmos revelam informações não triviais sobre as séries em estudo, a fim de saber se o processo que a gerou pode ser caracterizada usando a teoria dos grafos.

Ainda em [Lacasa et al. 2008], foi mostrado que os grafos de visibilidade que foram mapeados conseguem herdar estruturas das séries temporais utilizadas no mapeamento, ou seja, série periódicas são convertidas em grafos regulares, séries aleatórias em grafos aleatórios e séries fractais em grafos livre de escala.

Logo, devido ao fato de um evento modificar a dinâmica da frequência dos *tweets*, quando um evento não está ocorrendo, a série temporal e mapeada em um grafo aleatório. Na medida que o evento passa a ocorrer, o grafo se transforma num grafo *small-world*.

Com isso, dado a série temporal para um bigrama b obtido da seção 3.3.2, analisamos o grafo gerado pela transformação do grafo de visibilidade, onde se o mesmo for um grafo *small-world*, consideramos que está ocorrendo um evento.

Uma rede small-world pode ser quantificada por um coeficiente ω como visto em [Telesford et al. 2011], definido por

$$\omega = \frac{L_r}{L} - \frac{C}{C_r}$$

, onde L e L_r o comprimento médio da rede analisada e da rede aleatória, respectivamente; C e C_r é o coeficiente de agrupamento da rede analisada e da rede aleatória, respectivamente. Uma transformação simples é adotada $\omega' = 1 - |\omega|$ para obtermos valores de ω' entre 0 (rede não small-world) até 1(rede small-world).

Existem outras métricas para quantificação de coeficiente de *small-worldness* existem na literatura, por exemplo [Humphries and Gurney 2008], porém essa técnica se torna altamente influenciável pelo tamanho da rede, fazendo com isso que tenhamos preferido a utilização do coeficiente apresentado em nosso trabalho.

Sabendo que ocorreu um evento para um bigrama b, agrupamos os bigramas com termos(palavras) em comum e assim, construímos um grafo de sumarização, que representa o evento. Ou seja, nesse grafo cada vértice representa uma palavra e as aresta identifica que as palavras pertencem ao mesmo bigrama.

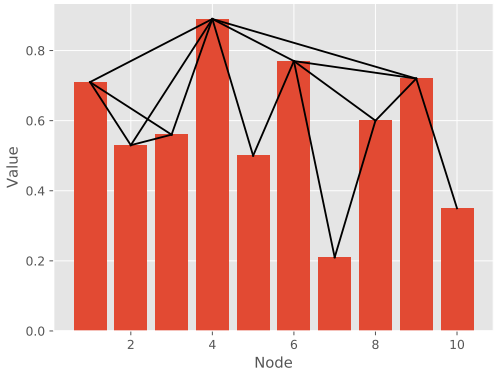

Figura 3. Série temporal

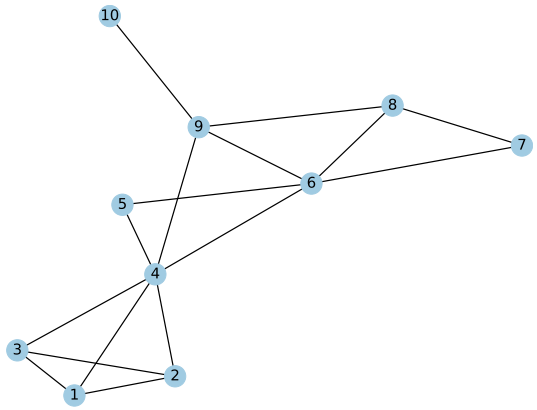

Figura 4. Grafo de visibilidade correspondente a série temporal da Figura 3

3.3.4. Particionar Grafo

Nós dividimos o grafo de sumarização encontrado na etapa anterior para identificar os eventos. Após o particionamento, consideramos que cada grafo conectado com mais de dois elementos encontrados após a partição está associado a um evento.

O Processo de Clusterização Markoviano (MCL), proposto por [Van Dongen 2000] define uma seqüência de processos estocásticos matriciais (inflação e expansão) chamados operadores para particionamento do gráfico.

A idéia principal do algoritmo MCL é simular o fluxo em um grafo normalizado, aumentando o fluxo onde a corrente é forte (muitas visitas do caminhante aleatório para um certo vértice) e abaixando o fluxo onde a corrente é fraca (poucas visitas do caminhante a um certo vértice).

O operador de expansão favorece os caminhos mais curtos, ou seja, passeios aleatórios com poucos passos, favorecendo a visita a novos grupos. Esse operador associa novas probabilidades a todos os pares de nós, diminuindo a probabilidade de caminhos longos e aumentando para caminhos curtos. Assim, o operador de expansão é responsável por permitir que o fluxo conecte diferentes regiões do gráfico.

O operador de inflação é responsável por fortalecer e enfraquecer o fluxo atual. A inflação terá então o efeito de aumentar as chances de passeios dentro do grupo e diminuir as caminhadas entre os grupos. Isso é realizado sem qualquer conhecimento prévio da estrutura de agrupamento.

Decidimos usar esse algoritmo porque ele permite que dois nós estejam em dois grupos separadamente. Em outras palavras, mesmo que dois eventos distintos tenham termos (palavras) em comum, o algoritmo pode segmentá-los de maneira satisfatória.

3.3.5. Avaliação

Para analisar nossos resultados, usamos o conjunto de dados apresentado em [Aiello et al. 2013]. Aplicamos as mesmas métricas utilizadas, para garantir que nossos resultados sejam diretamente comparáveis aos fornecidos.

- *Recall* do tópico (T-Rec): porcentagem de eventos detectados com sucesso, ou seja, a taxa positiva verdadeira para detecção de evento

$$\text{T-Rec} = \frac{\textit{ground truth} \text{ para Tópico } \cap \text{ eventos do Tópico detectados}}{\textit{ground truth} \text{ para Tópico}}.$$

- Precisão da palavra-chave (K-Prec): porcentagem de palavras-chave detectadas corretamente sobre o total de palavras-chave de um determinado evento, ou seja, a taxa real negativa para a detecção de palavras-chave

$$\text{K-Prec} = \frac{\textit{ground truth} \text{ para palavras-chave} \cap \text{ palavras-chave detectadas}}{\text{palavras-chave detectadas}}.$$

- Recall de palavras-chave (K-Rec): porcentagem de palavras-chave detectadas corretamente sobre o *ground truth* de palavras-chave de um determinado evento , ou seja, a taxa positiva verdadeira de detecção de palavras-chave

$$\text{K-Rec} = \frac{\textit{ground truth} \text{ Palavras-chave} \cap \text{ palavras-chave detectados}}{\textit{ground truth} \text{ palavras-chave}}.$$

- F_1-Score (K-Score): para melhor comparação entre a técnica, adotamos as métricas F_1-score para palavras-chave

$$\text{K-Score} = 2 \cdot \frac{\text{K-Rec} \cdot \text{K-Prec}}{\text{K-Rec} + \text{K-Prec}}.$$

Observe que essas métricas são calculadas para cada intervalo de tempo.

4. Resultados e Discussões

Logo, como o modelo depende do tamanho da janela deslizante n, nós fizemos uma investigação para determinar qual o valor dessa variável maximizaria a precisão. [Mockus 1975] propôs um novo método chamado Otimização Bayesiana que consiste em otimizar funções como uma ”caixa preta”, ou seja, quando não conhecemos seu comportamento ou a sua derivada.

O método consiste de, com alguns pontos conhecidos, determinar a forma da função através de uma regressão. Usualmente, essa predição é feita através de um Processo Gaussiano devido a algumas características (escalável para poucos pontos e não paramétrico).

Com isso, tendo como base a regressão do processo gaussiano, ou seja, a possível forma da função que queremos otimizar, é definida uma função utilitária que consiste em achar o próximo candidato para ponto otimizado. Neste trabalho é realizado a discretização para os pontos analisados (devido ao tamanho da janela ser um número inteiro). Utilizamos um ponto inicial aleatório, e posteriormente, 5 rodadas do algoritmo como podemos ver na figura 5, onde foi encontrado o valor de $n = 10$.

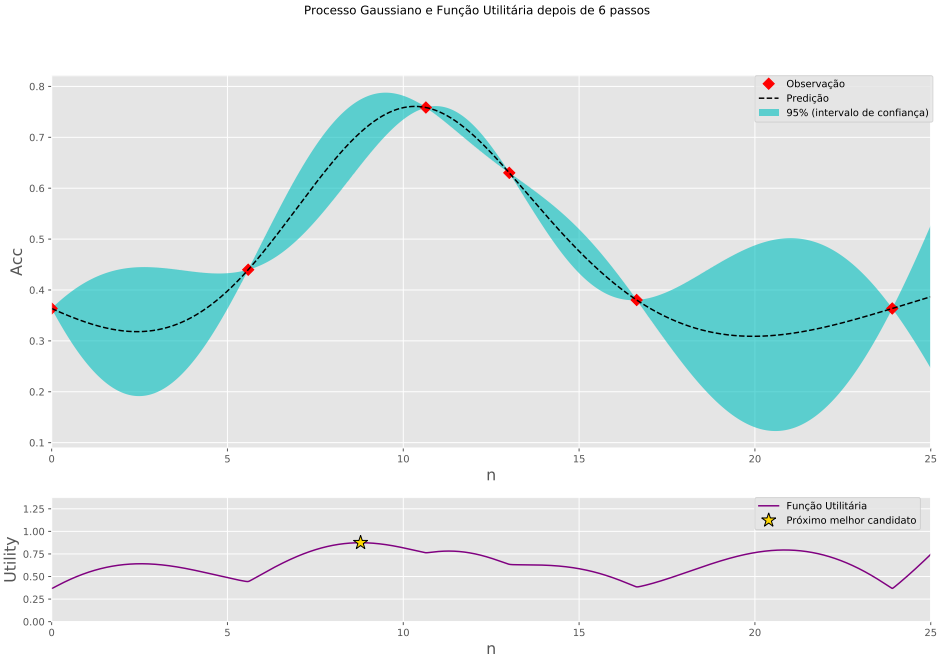

Figura 5. Optimização Bayesiana para o conjunto de dados do FA Cup

Para avaliação dos nossos resultados, nós utilizamos as métricas descritas na seção 3.3.5, como podemos ver nas tabelas 4 e 5, onde comparamos os resultados com algumas técnicas encontradas para literatura que já reportaram resultados para esses conjuntos de dados. Somente os valores de [Choi and Park 2019] e [Barros et al. 2018] foram coletados e copiados dos artigos originais. Os melhores resultados foram apresentados em negrito.

Os resultados do método de Poisson já eram esperados. Como a proposta analisa somente o número de *tweets*, sem olhar seu conteúdo, esperávamos que possuísse resultados menos expressivos e por isso, está sendo usada como *baseline*.

Para o conjunto de dados FA Cup, observamos que nossa abordagem é a segunda melhor para a métrica K-Pec e quarta melhor para K-Rec, porém na junção de ambas (K-Score) conseguimos o melhor resultado. Isso indica que nossa proposta possui um balanceamento entre K-Prec e K-Rec, indicando que a mesma consegue um resultado satisfatório para ambas. Para T-Rec nossa proposta é a terceira melhor.

Para o conjunto de dados *Super Tuesday*, nossa abordagem foi a melhor no comparativo com a métrica T-Rec (empatando com duas outras propostas). Obtivemos o quarto e terceiro melhor resultado para as métricas K-Prec e K-Rec, respectivamente. Além disso, considerando o K-Score obtivemos o terceiro melhor resultado.

5. Considerações Finais

Neste trabalho, nós utilizamos o grafo de visibilidade para detectar a ocorrência de um evento em uma rede social. Descobrimos que a ocorrência de um evento apresenta uma mudança de dinâmica que pode ser observada através de métricas em redes complexas.

Tabela 4. Comparação dos métodos usando *FA Cup dataset*

Método	T-Rec	K-Prec	K-Rec	K-Score
Poisson (baseline)	0.308	0.124	0.202	0.154
[Petrović et al. 2010] (Doc-p)	0.692	0.346	0.503	0.410
[Aiello et al. 2013] (FPM)	0.308	**0.694**	0.512	0.589
[Aiello et al. 2013] (SFPM)	0.615	0.241	0.608	0.345
[Aiello et al. 2013] (BNGran)	0.846	0.310	0.567	0.401
[Aiello et al. 2013] (GFeat-p)	0.238	0.120	0.471	0.191
[Nguyen and Jung 2017]	0.769	0.453	0.548	0.496
[Blei et al. 2003] (LDA)	0.538	0.204	**0.643**	0.310
[Weng and Lee 2011] (EDCoW)	0.384	0.312	0.357	0.333
[Choi and Park 2019] (HUPM)	**0.923**	0.320	0.600	0.417
Nossa Proposta	0.769	0.615	0.597	**0.605**
[Barros et al. 2018]	0.769	0.528	0.596	0.560

Tabela 5. Comparação dos métodos usando *SuperTuesday dataset*

Método	T-Rec	K-Prec	K-Rec	K-Score
Poisson (baseline)	0.091	0.321	0.243	0.154
[Petrović et al. 2010] (Doc-p)	0.182	0.351	0.437	0.389
[Aiello et al. 2013] (FPM)	0.136	**0.698**	0.372	0.485
[Aiello et al. 2013] (SFPM)	0.273	0.617	0.593	0.605
[Aiello et al. 2013] (BNGran)	0.364	0.522	0.613	0.564
[Aiello et al. 2013] (GFeat-p)	0.091	0.108	0.294	0.158
[Nguyen and Jung 2017]	0.409	0.612	**0.714**	**0.659**
[Blei et al. 2003] (LDA)	0.136	0.101	0.212	0.137
[Weng and Lee 2011] (EDCoW)	0.273	0.345	0.381	0.362
[Choi and Park 2019] (HUPM)	**0.455**	0.420	0.678	0.519
Nossa proposta	**0.455**	0.525	0.621	0.569
[Barros et al. 2018]	**0.455**	0.325	0.421	0.367

Com isso, nós propomos um novo método para detectar eventos no Twitter baseando nas séries temporais formadas pela frequência de palavras-chaves extraídas do conteúdo dos *tweets*. Nossa proposta não assume nenhuma informação previa a cerca do *tweet*, e consegue identificar qualquer tipo de evento, sem restrição de idioma utilizado.

Nossa proposta apresenta resultados satisfatórios quando comparadas com o estado-da-arte encontrado na literatura para os conjuntos de dados analisados.

Referências

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

Barros, P., Cardoso, I., A.F. Loureiro, A., and Ramos, H. S. (2018). Event detection in social media through phase transition of bigram entropy. In *IEEE Symposium on Computers and Communications (ISCC)*, Natal, Brazil.

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

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*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA. ACM.

Choi, H.-J. and Park, C. H. (2019). Emerging topic detection in twitter stream based on high utility pattern mining. *Expert Systems with Applications*, 115:27 – 36.

Dang, Q., Gao, F., and Zhou, Y. (2016). Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks. *Expert Syst. Appl.*, 57(C):285–295.

Deerwester, S., Duais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantics analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dou, W., Wang, X., Ribarsky, W., and Zhou, M. (2012). Event detection in social media data. In *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content*, pages 971–980.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press.

Evensen, G. (2003). The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44.

Humphries, M. D. and Gurney, K. (2008). Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PloS one*, 3(4):e0002051.

Jarvis, R. A. and Patrick, E. A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.*, 22(11):1025–1034.

Katti, S. and Rao, A. V. (1968). Handbook of the poisson distribution.

Lacasa, L., Luque, B., Ballesteros, F., Luque, J., and Nuno, J. C. (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975.

Li, C., Sun, A., and Datta, A. (2012). Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 155–164. ACM.

Mathioudakis, M. and Koudas, N. (2010). Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA. ACM.

Mockus, J. (1975). On the bayes methods for seeking the extremal point. *IFAC Proceedings Volumes*, 8(1, Part 1):428 – 431. 6th IFAC World Congress (IFAC 1975) - Part 1: Theory, Boston/Cambridge, MA, USA, August 24-30, 1975.

Murphy, K. P. and Russell, S. (2002). Dynamic bayesian networks: representation, inference and learning.

Nguyen, D. T. and Jung, J. E. (2017). Real-time event detection for online behavioral analysis of big social data. *Future Generation Computer Systems*, 66:137 – 145.

Nummiaro, K., Koller-Meier, E., and Van Gool, L. (2003). An adaptive color-based particle filter. *Image and vision computing*, 21(1):99–110.

Parikh, R. and Karlapalem, K. (2013). Et: Events from tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 613–620, New York, NY, USA. ACM.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). 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.

Poisson, S. D. (1837). Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés. *Paris, France: Bachelier*, 1:1837.

Rosso, O. A., Craig, H., and Moscato, P. (2009). Shakespeare and other english renaissance authors as characterized by information theory complexity quantifiers. *Physica A: Statistical Mechanics and its Applications*, 388(6):916 – 926.

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*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Telesford, Q. K., Joyce, K. E., Hayasaka, S., Burdette, J. H., and Laurienti, P. J. (2011). The ubiquity of small-world networks. *Brain connectivity*, 1(5):367–375.

Van Dongen, S. M. (2000). *Graph clustering by flow simulation*. PhD thesis.

Weng, J. and Lee, B.-S. (2011). Event detection in twitter. *ICWSM*, 11:401–408.