IDENTIFICAÇÃO DE ASSUNTOS EM COMENTÁRIOS POSTADOS EM REDES SOCIAIS

Gabriel Medeiros Jospin

Prof. Dr. Luciano Antonio Digiampietri

Escola de Artes, Ciências e Humanidades / Universidade de São Paulo

gabriel.jospin@usp.br; digiampietri@usp.br

Objetivos

Atualmente, milhões de pessoas produzem conteúdo online que pode ser acessado nas redes sociais, fóruns de discussão online e aplicativos de trocas de mensagens. A grande quantidade e variedade de informações produzidas trouxe novos desafios e oportunidades para a área de análise de dados. Uma delas é a identificação automática do assunto (ou tópico) que está sendo discutido, atividade que pode ser usada para auxiliar na análise de tendências ou identificação de surtos de doenças, aumento de criminalidade ou de incêndios florestais.

O objetivo deste projeto foi testar diferentes abordagens que ajudam na visualização, organização ou identificação dos assuntos que são tratados em comentários de redes sociais, realizando testes em mensagens postadas no Twitter. O projeto teve caráter exploratório testando e analisando os resultados de técnicas de visualização, agrupamento e extração de tópicos de conjuntos de dados textuais.

Este projeto está contextualizado em um projeto maior que visa a monitorar a publicação de comentários no Twitter relacionados à incidência de incêndios florestais.

Métodos e Procedimentos

Este projeto foi organizado em quatro etapas: Revisão bibliográfica de trabalhos publicados; Escolha do ferramental, baseada nos trabalhos correlatos; Especificação e implementação de técnicas para extração de características e classificação de assunto utilizando modelos agrupadores e baseada nas características extraídas; Realização de testes e validação dos resultados obtidos. As tecnologias utilizadas neste projeto foram a linguagem de programação *Python*, além das bibliotecas Pandas, Numpy, Unicodedata, Re, Sklearn, e Natural Language ToolKit. Utilizou-se do Google Codelab para desenvolver o código de forma interativa entre as partes, sem a necessidade de um ambiente específico de desenvolvimento.

Primeiramente, foi realizada a extração de tweets de contas oficiais na rede social Twitter e a busca por *hashtags* específicas, de forma a associar os dados com nossa pesquisa de interesse, neste caso a exploração de ferramentas para categorização de tweets com testes aplicados à identificação de incêndios florestais. Com estes textos em mão foi construído um pequeno classificador para verificar se o tweet tinha ou não relação com queimadas. Além disso, foi construída uma tabela de frequência para uma análise qualitativa das mensagens filtradas. Partindo para o uso de ferramentas não supervisionadas foi construído um classificador K-means a fim de dividir em conjuntos temáticos os posts, e validados por meio da classificação de ter ou não relação com queimadas. Por fim, utilizando modelos de Topic Modeling, buscamos determinar o assunto de cada conjunto por meio dos tópicos mais relevantes.

Resultados

Durante o desenvolvimento deste projeto, foram coletados 6.450 posts das contas de usuário selecionados e 409 comentários utilizando as *hashtags*. Para o conjunto total de postagens, apenas 407 (5,93%) eram relacionadas a queimadas ou incêndios florestais. Porém, se observados apenas os comentários com as

hashtags selecionadas, observamos que 342 (83,86%) estavam relacionadas a queimadas. As primeiras análises realizadas utilizaram métodos quantitativos, isto é, buscou-se a temática geral dos posts extraídos, inicialmente com base na frequência relativa das palavras:

Tabela 1 - Frequência relativa de algumas palavras

	@defesacivil	@mpdemt	Hashtags
queimadas	0,00	0,07	9,04
mptm	0,00	8,71	0,00
civil	7,19	0,13	0,09
fumaça	0,52	0,04	8,27
amazônia	0,00	0,15	7,99

(Autor, 2023)

Na primeira etapa foi realizada uma clusterização utilizando o algoritmo k-means. Primeiro foi feita uma filtragem para a remoção de outliers. Depois, estimou-se a quantidade de clusters ideal por meio da distância entre clusters, obtendo o valor ótimo em 6 clusters diferentes. Em seguida utilizamos uma abordagem de *Topic Modeling* (ou modelagem de tópicos), que se diferencia da busca de palavras mais frequente, pois primeiro agrupa o conjunto de postagens por semelhanças e depois extrai quais os principais grupos de palavras formados para ser considerado o tema geral do conjunto. Para o problema em questão, a escolha do topic modeling serve para caracterizar os agrupamentos construídos na clusterização. A partir deste, se obteve o seguinte resultado:

Tabela 3 - Topic Modeling dos Clusters

Grupo	#1	#2	#3
0	chuvas	deste	ano
1	amazonas	queimada	feira
2	queimadas	amazônia	agosto
3	queimadas	amazônia	setembro
4	brasil	centro	oeste
5	queimadas	povo	souza

(Autor, 2023)

Conclusões

Neste trabalho foi estudado, especificado, implementado e testado uma abordagem para detecção de temáticas de posts em português relacionados com a temática de queimadas. Além da construção de um pequeno banco de dados próprio e inédito, rotulado manualmente. Com base nestas características, conseguimos concluir a quantidade de posts relacionados com a temática. Em seguida, foram extraídas, inicialmente, características ligadas aos comentários (n-gramas de palavras). Usando modelos de agrupamento, criamos seis grupos e identificamos os principais temas destes grupos, descobrindo assim subtópicos de cada agrupamento. Um algoritmo de classificação usando uma rede MLP foi desenvolvido, o qual foi capaz de acertar a identificação de 92,3% das postagens como relacionadas ou não a incêndios florestais (considerando o conjunto de postagens obtido a partir de hashtags de interesse).

Consideramos, assim, que os resultados iniciais foram bastante satisfatórios. Pretende-se, como trabalhos futuros, extrair novas características dos comentários, bem como construir uma base de dados maior, o que pode ser bastante útil para o classificador que pode estar sendo limitado pelo pequeno número de instâncias de treinamento.

Referências Bibliográficas

[1] ALDAYEL, Abeer; MAGDY, Walid. **Stance Detection on Social Media: State of the Art and Trends.** Information Processing & Management, Volume 58, Issue 4, 2021.

[2] Liu, L., Tang, L., Dong, W. et al. **An overview of topic modeling and its current applications in bioinformatics**. SpringerPlus 5, 1608, 2016.

[3] Marques-Toledo CdA, Degener CM, Vinhal L, Coelho G, Meira W, et al. **Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level.** PLoS Negl Trop Dis 11(7): e0005729, 2017.

[4] S. Salvador and P. Chan. **Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms**. 16th IEEE ICTAI, pp. 576-584, 2004.

[5] XAVIER, F. et al. **Análise de redes sociais como estratégia de apoio à vigilância em saúde durante a Covid-19.** Estudos Avançados. 34 (99), 2020.