DESENVOLVIMENTO DE FERRAMENTAS PARA A MINERAÇÃO DE OPINIÕES/POSICIONAMENTOS EM FÓRUNS DE DISCUSSÃO ONLINE

Laura Steinert de Freitas

Gabriel Medeiros Jospin

Prof. Dr. Luciano Antonio Digiampietri

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

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

Objetivos

Em 2022, em todo o mundo, 95,2% do tempo que um cidadão utilizou a internet no mês foi gasto dentro de redes sociais [1]. As redes sociais e fóruns de discussão online são locais nos quais as pessoas podem expressar seus pontos de vista de forma livre em nossa sociedade [2], com cada vez mais aderência e visibilidade para marketing e veiculação de informações governamentais. Durante a pandemia de COVID-19, é possível notar um maior número de usuários de redes sociais, uma vez que o acesso restrito ao exterior de suas casas se tornou cenário comum. As plataformas permitem ao usuário expor seu ponto de vista, abrindo espaço para respostas de diferentes pessoas que podem ter opiniões distintas sobre o mesmo conteúdo, seja ele relacionado com o que está sendo debatido ou não. O presente projeto tem como objetivo o desenvolvimento de uma ferramenta para mineração de opiniões e posicionamentos em fóruns de discussão online, tal qual o “Reddit”, citado como o 11° site mais utilizado no mundo [1], utilizado para o projeto. Esta ferramenta utiliza como base dados de comentários acerca de notícias chave, aferindo quais são as principais palavras citadas e a relação entre o comentário e a notícia em questão.

Métodos e Procedimentos

O desenvolvimento deste projeto é composto por: Revisão bibliográfica sobre trabalhos correlatos; Seleção, com base na literatura correlata, de um conjunto técnicas que foram implementadas e que envolveram a extração de diferentes características (*features*); Especificação e implementação de ferramentas para o cálculo (ou extração) automático das características selecionadas e adaptação, quando necessário, das técnicas ou ferramentas encontradas na literatura; Realização de testes e validação das ferramentas implementadas. A validação das ferramentas implementadas considerou métricas tradicionais da avaliação de sistemas de classificação. Das tecnologias utilizadas para o desenvolvimento do código, utilizamos as seguintes bibliotecas do Python: *Pandas*, *Numpy*, *Unicodedata*, *Re*, *Sklearn*, e *Natural Language ToolKit*. Essas bibliotecas foram utilizadas concomitantemente com o Colab da Google, para que a equipe conseguisse trabalhar em conjunto e online no projeto, sem a necessidade de outros ambientes de desenvolvimento específicos. A comparação entre o texto da notícia e de cada um dos comentários foi medida com base na distância cosseno. Para possibilitar a avaliação da solução proposta, realizamos inicialmente a rotulação manual dos comentários tanto para verificar se estavam ou não relacionados com a notícia,, no caso dos relacionados, averiguar se havia concordância ou discordância com a notícia, mas também para verificar se estavam ou não relacionados ao ex-presidente Lula. Com os dados rotulados, os modelos desenvolvidos para classificação foram testados com base nas características (*features*) extraídas. Conforme apresentado, a

distância cosseno foi uma das características calculadas, aferindo a capacidade de opiniões se manterem relevantes para o estudo de acordo com as manchetes. Em adição, foram extraídos os n-gramas das palavras dos comentários. Este conjunto de características foi passado para o classificador para que este realizasse a classificação, por meio da construção de uma árvore de decisão, onde 70% dos dados foram utilizados para treinamento e 30% para testes, para conseguir prever a opinião do autor do comentário em relação à notícia apresentada, no caso específico, verificar automaticamente se a opinião era relacionada ao presidente Lula ou não.

Resultados

Foi utilizada a notícia [4], e os comentários providos do Reddit. Em nossos resultados, almejamos averiguar a quantidade de comentários que eram relacionados ao presidente Lula. A Árvore de Decisão foi treinada com base em 70% do conjunto de dados e foi modelada com acurácia de 81,9%. A matriz de confusão resultante dos testes é apresentada na figura 1.

		Predição	
		Relacionado	Não relacionado
Esperado	Relacionado	86	12
	Não relacionado	12	23

Figura 1: Matriz de Confusão da Predição.

Observa-se, pela matriz de confusão que apesar do conjunto de dados ser desbalanceado (mais de 72% dos comentários não são relacionados), o modelo de predição produzido foi capaz de acertar a predição de 81,9%, atingindo uma revocação da classe minoritária de 65,7% e de 87,7% da classe majoritária.

Conclusões

O projeto seguiu a metodologia prevista, com a rotulação manual de dados e extração de características chave dos comentários, além do cálculo da distância cosseno entre o comentário e a manchete avaliada.
Os resultados acerca do classificador Árvore de Decisão apresentaram acurácia satisfatória, de 81,9%. Ainda há espaço para melhorias, mas, para os fins deste projeto, podemos considerar os resultados satisfatórios, com a construção de um modelo de classificação capaz de identificar se um comentário em rede social se relaciona ou não a uma entidade nomeada (no caso, ao ex-presidente Lula). O conjunto de dados produzido possui 440 comentários, apesar de não ser um conjunto muito grande, foi possível ao classificador produzir um modelo de classificação com bons valores de acurácia e revocação (em especial da classe minoritária). Contudo, ainda há espaço para desenvolvimento e aprimoramento da solução. Possíveis passos futuros para a melhoria da solução podem envolver a extração de outras características do conjunto de dados, com o uso de transformadores/*word embeddings*, por exemplo, além da utilização de outros classificadores observados em na pesquisa bibliográfica.

Referências Bibliográficas

[1] - We Are Social and Hootsuite. **Digital 2022 Global Overview Report**. Jan, 2022.

[2] - XAVIER, F. et al. **Análise de redes sociais como estratégia de apoio à vigilância em saúde durante a Covid-19**. Estud. av. 34 (99). Maio - Agosto 2020. São Paulo, SP, Brasil.

[3] - ALDAYEL, Abeer; MAGDY, Walid. **Stance Detection on Social Media: State of the Art and Trends.** In: Information Processing & Management, Volume 58, Issue 4, July 2021, 102597, ISSN 0306-4573, Cornell University, UK.

[4] - "**Lula diz que aborto "é direito da mulher" e questão de "saúde pública""** ;Thayná Schuquel para a Revista Metrópoles. Setembro de 2021. Disponível em: https://www.metropoles.com/brasil/politica-brasil/lula-diz-que-aborto-e-direito-da-mulher-e-questao-de-saude-publica