FERRAMENTAS PARA A MINERAÇÃO DE POSICIONAMENTOS EM FÓRUNS DE DISCUSSÃO ONLINE

Gabriel Medeiros Jospin

Laura Steinert de Freitas

Prof. Dr. Luciano Antonio Digiampietri

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

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

Objetivos

É inegável a influência das redes sociais na sociedade moderna. Segundo o artigo de Fontenelle et al. (2020), cerca de 64,7% dos brasileiros tomam suas decisões políticas baseados em informações obtidas pela internet, e a grande maioria obteve a mesma a partir do uso de redes sociais.

Este aumento da participação política nas redes sociais, porém, tem causado um evento de extremização da política. Como as redes sociais aproximam pessoas com interesses e gostos parecidos, têm surgido bolhas nas redes sociais que extremizam essas opiniões (Bacarella, et al., 2018).

Com isso, é cada vez mais importante estudar a saúde das redes sociais (identificação da formação de bolhas ou câmaras de eco, detecção de notícias falsas e da atuação de *bots*). Uma das bases para conseguir estudar a formação de bolhas em redes sociais é a detecção de posicionamento de usuários a partir de comentários em redes sociais.

O presente trabalho tem como objetivo apresentar uma ferramenta de detecção de posicionamento de comentários em português a partir de notícias publicadas em redes sociais do tipo fórum, como o Reddit. Essa ferramenta utiliza como base as palavras utilizadas em comentários e a relação das mesmas com as notícias compartilhadas.

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 predição de posicionamento utilizando classificadores 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*, *Sklearn* e *nltk.* Além disso, 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, foram rotulados os comentários de duas notícias postadas em uma comunidade do Reddit. A primeira com a temática das eleições de 2022, a segunda em um cenário mais amplo falando sobre motoristas por aplicativo. Em seguida, foram removidas dos comentários as palavras sem valor semântico atrelado (*stopwords*). Das mensagens e notícias filtradas foi calculada a distância cosseno dos textos a fim de se obter a similaridade entre os textos. Além da similaridade por palavras, também foi realizada a representação dos comentários por n-gramas e cômputo da frequência relativa TF-IDF dos n-gramas. Estes dados foram apresentados aos classificadores a fim de se obter um resultado mais preciso na predição e os resultados obtidos foram apresentados na próxima seção.

Resultados

Ao todo foram realizados 120 testes, correspondendo ao produto de cinco classificadores, quatro estratégias para extrair características do texto (unigrama, bigrama, trigrama e concatenação das três), duas formas de valorar as características (contagem ou TF-IDF) e três estratégias de redução de dimensionalidade (nenhuma, seleção de atributos com K-Best e PCA). Para cada caso de testes foram medidas, usando validação cruzada com três subconjuntos, a acurácia, a medida F1 e a área sob a curva ROC. Devido a quantidade de resultados obtidos, serão apresentados apenas os melhores resultados considerando cada conjunto de técnicas e métodos.

A tabela 1 apresenta os melhores resultados obtidos para cada uma das métricas utilizadas. Destaca-se que a melhor acurácia e o desempenho na medida F1 foram obtidas pela Rede Neural MLP, enquanto a AUC teve um resultado melhor com o classificador SVM. Observa-se que os melhores resultados de acurácia e AUC foram obtidas por bigramas, enquanto a medida F1 tem melhor desempenho com trigramas. Em termos de ponderação de n-gramas, as soluções com MLP utilizaram a contagem, já a solução com SVM utilizou TF-IDF. Por outro lado, ao se considerar a redução de dimensionalidade, as melhores soluções em termos de acurácia e AUC não utilizaram nenhuma estratégia de redução. Já o maior valor da medida F1 foi atingido utilizando Análise de componentes principais (PCA).

Tabela 1: Resultado dos melhores classificadores

Classificador	**Acurácia**	**F1**	**AUC**
MLP, bigrama, contagem., nenhuma	**0,752**	0,525	0,633
MLP ,trigrama, contagem PCA	0,612	**0,571**	0,45
SVM,bigrama,TF-IDF, nenhuma	0,739	0,425	**0,677**

Conclusões

Neste trabalho, foi estudado, especificado, implementado e testado um algoritmo para detecção de posicionamento de comentários em português realizados em relação a notícias jornalísticas.

Além da construção de um pequeno banco de dados próprio e inédito, rotulado manualmente, foram extraídas, inicialmente, características ligadas aos comentários (n-gramas de palavras), bem como foi calculada a distância cosseno entre o comentário e a notícia.

Com base nestas características, um classificador do tipo MLP foi capaz de identificar com 75,2% de acurácia o posicionamento dos comentários.

Embora o algoritmo ainda não apresente uma acurácia muito alta, os resultados iniciais foram considerados bastante satisfatórios. Pretende-se, como trabalhos futuros, extrair novas características dos comentários (utilizando, por exemplo, *word embeddings* ou transformadores), 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] Carolina Alves Fontenelle, Conceição Souza. **Redes sociais: A internet assume papel preponderante nas eleições presidenciais de 2018**. Revista eletrônica dos discentes da Escola de Sociologia e Política da FESPSP [Internet]. 2020 [cited 2022 May 2];1(13):29-42. from:http://revistaalabastro.fespsp.org.br/index.php/alabastro/article/view/274

[2] BACCARELLA, C. V., WAGNER, T. F., KIETZMANN, J. H., & MCCARTHY, I. P. (2018). **Social media? It's serious! Understanding the dark side of social media**. European Management Journal, 36(4), 431–438. doi:10.1016/j.emj.2018.07.002

[3] **Maioria dos bolsonaristas arrependidos votaria em Lula no 2º turno, diz Datafolha**. Carta Capital. 2021 acesso em 2 de maio de 2022.