Modelagem de tópicos aplicada à recomendação na Amazon

Antonio Carlos Farias Ferreira
Departamento Acadêmico de Informática (DAINF)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
antfer@alunos.utfpr.edu.br

***Abstract*—Este trabalho investiga as relações entre as descrições textuais de produtos vendidos online e os padrões de consumo dos usuários. Para isso, foi aplicada modelagem de tópicos sobre as descrições utilizando o BERTopic, permitindo identificar temas latentes associados a aspectos técnicos, gêneros de conteúdo e nichos específicos. A partir desses tópicos, foram construídos perfis de usuários com um modelo de regressão logística. Também foram explorados três métodos de fatoração matricial (ALS, SVD e NMF) para predizer o consumo de tópicos. Os resultados mostram que os tópicos extraídos são eficazes para representar preferências e gerar recomendações relevantes. Por fim, realizou-se um agrupamento dos usuários com base nos perfis encontrados, mas não foram observadas diferenças significativas nas avaliações das categorias mais e menos consumidas. Esses achados indicam que a modelagem de tópicos é útil para capturar padrões de consumo, embora não tenha refletido de forma consistente nas avaliações atribuídas pelos consumidores.**

***Index Terms*—Recomendação baseada em conteúdo, perfil de usuário, modelagem de tópicos, processamento de linguagem natural, sistemas de recomendação, E-commerce**

I. Introdução

A descrição do produto é um elemento comum em páginas de venda online. Ela dá suporte aos consumidores na decisão de compra, apresentando as principais características da mercadoria anunciada. Por isso, espera-se que as descrições dos produtos consumidos por um usuário sejam capazes de fornecer informações que revelem interesses e preferências.

Esse trabalho tem por objetivo explorar as relações semânticas entre as descrições de produtos vendidos online e o consumo. A modelagem de tópicos é uma técnica de Processamento de Linguagem Natural que se utiliza de aprendizado de máquina não supervisionado para extrair palavras-chave de um conjunto de dados de texto para descobrir temas latentes. Nesse sentido, houve um enfoque em como os tópicos extraídos a partir dessa técnica podem ser utilizados para fazer a recomendação de outros produtos, se orientando à responder as seguintes questões:

1) Que tópicos emergem da descrição dos produtos?
2) Esses tópicos são capazes de descrever relações com a compra de outros produtos?
3) Esses tópicos podem se relacionar com as categorias do produto e refletir aspectos de sua avaliação pelos consumidores?

Dessa forma, o trabalho pode ser dividido em três partes. Primeiramente, é apresentada a modelagem de tópicos realizada. Em seguida, é apresentado o desenvolvimento de um modelo de recomendação de produtos baseado em perfis de consumo construídos a partir dos tópicos extraídos. Também foi avaliada a eficácia de três modelos de fatoração matricial em predizer o consumo de outros tópicos pelos usuários a partir dos tópicos já consumidos. Por fim, é feito um agrupamento dos usuários com base nos seus perfis para comparar a avaliação de cada categoria consumida pelos diferentes grupos.

Foi possível predizer o consumo de outros produtos e tópicos pelos usuários de forma eficaz, mas não foram identificadas relações entre os tópicos consumidos pelos usuários e a forma como eles avaliam cada categoria.

II. Trabalhos relacionados

A abordagem apresentada em [2] utiliza de uma modelagem de tópicos baseada em Word2Vec em posts de uma mídia social para extrair tópicos que alimentam um modelo de perfilamento baseado em regressão logística, semelhante ao utilizado neste trabalho, demonstrando como tópicos podem ser agregados para enriquecer os perfis de usuário.

Abordagens de filtragem colaborativa com fatoração matricial foram utilizadas por [4] e [5] para fazer recomendações. Nos dois casos o método consiste em reconstruir valores de interação em uma matriz usuário-item.

Identificando a densidade das interações como um obstáculo comum em sistemas de recomendação, [4] propõe uma variação do SVD orientada a densidade. Enquanto [5] explora a capacidade de predição da fatoração matricial em um sistema do nicho de cursos universitários.

III. Processamento de dados

A. Conjunto de dados

Foi utilizado um grande conjunto de dados de avaliações e produtos da Amazon na categoria de filmes e televisão fornecido por [1]. Para que o processamento fosse possível com o hardware disponível, optou-se por utilizar um recorte sobre um subconjunto, em que cada usuário tinha ao menos 5 avaliações realizadas e cada produto tinha 5 avaliações recebidas. Inicialmente esse recorte apresentava 203766 produto e 3410019 avaliações.

Após a limpeza de dados, eliminando produtos sem descrições, e uma seleção dos usuários com um número de avaliações entre 50 e 400, restaram, 2278 usuário, com uma média de aproximadamente 98 avaliações, e 24725 produtos. Essa seleção de um conjunto denso de interações foi realizada na intenção de simplificar o problema e evitar obstáculos relacionados a baixa densidade de interações. A figura 1 apresenta a distribuição do número de avaliações dos usuário no conjunto de dados antes da seleção.

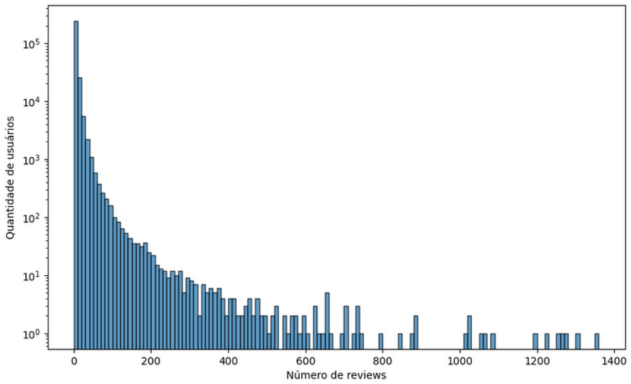

Fig. 1. Distribuição de número de avaliações por usuário

B. Modelagem de tópicos

Primeiramente, foi realizada uma modelagem de tópicos sobre as descrições dos produtos. Inicialmente, isso foi feito utilizando LDA e, mais tarde, clustering de embeddings Word2Vec. Mas o modelo realmente utilizado foi o BERTopic, que apresentou a melhor coerência semântica entre os tópicos encontrados.

Além disso, o BERTopic apresenta outra vantagem para a etapa seguinte do trabalho, em que os perfis de usuários são treinados usando regressão logística: ele realiza o clustering de embeddings dos documentos e fornece as probabilidades do documento pertencer a cada tópico encontrado.

Cabe pontuar que o rigor na seleção do conjunto de dados e a experimentação dos parâmetros da ferramenta de BERTopic oferecida por [3] proporcionaram uma grande melhoria nos resultados da modelagem. E, a partir de uma abordagem empírica, definiu-se o número de palavras por tópico como 20 e o tamanho mínimo dos tópicos como 50. Além disso, a documentação da ferramenta disponibilizada por [3] sugere que a eliminação de stopwords do corpus pode prejudicar o desempenho do BERTopic. Por isso, optou-se por utilizar a ClassTFIDFTransformer para diminuir o impacto de palavras frequentes.

Essas tratativas contribuirão para a redução do conjunto de documentos que não foram classificados como pertencentes a nenhum tópico e para a eliminação de tópicos genéricos, como “filme”, “series”, “tv” e etc.

O modelo encontrou 54 tópicos, sendo que 12553 documentos foram classificados como outliers e foram descartados junto com as suas avaliações. Por isso, alguns usuário ficaram com um número pequeno de avaliações e novamente optou-se por recortar o conjunto de dados. Foram selecionados os usuários com um número de avaliações entre 20 e 180, resultando na análise de 2128 usuários, 224476 avaliações e 12178 produtos nas estapas seguintes do trabalho.

A figura 2 apresenta o distribuição das avaliações dos usuário após o corte de produtos outliers.

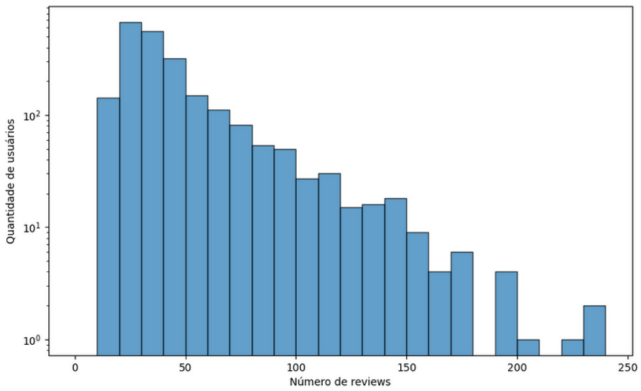

Fig. 2. Densidade de consumo por categoria em cada cluster

C. Produção de perfis e embeddings dos usuários

Nesta etapa, optou-se por um modelo baseado em regressão logística, análogo ao proposto por [2]. Para isso, foi construído um conjunto de treinamento para cada usuário com 70% dos seus produtos consumidos e um número 5 vezes maior de produtos aleatórios não consumidos. Isso foi feito para ficticiamente representar exemplos de produtos que o usuário não consumiu, uma vez que o dado de negativas de produtos não estava presente no conjunto de dados original. Como esses produtos podem nunca terem sido apresentados aos usuários, considerou-se que deveria haver uma variedade maior de exemplos de produtos não consumidos em relação aos realmente consumidos.

Como os produtos foram descritos a partir da modelagem de tópicos usando vetores de probabilidades de associação com cada tópico, o treinamento de um modelo de regressão logística para a classificação de produtos como consumidos ou não consumidos por um usuário resulta no aprendizado de um conjunto de pesos atribuídos pelo usuário para cada tópico. Ou seja, é um embedding do perfil do usuário no mesmo espaço vetorial de tópicos.

Para fins de comparação, também foi implementado um modelo ingênuo de produção de embedding dos usuários, em que é feita a média dos vetores de probabilidade de pertencimento aos tópicos dos produtos consumidos por cada usuário.

D. Predição de consumo de tópicos

Foram utilizados e comparados três algoritmos de fatoração matricial para reconstruir uma matriz de contagem de consumo usuários-tópicos: Alternating Least Squares (ALS), Singular Value Decomposition (SVD) e Non-negative Matrix Factorization (NMF). O conjunto de teste foi contruído ocultados

aleatoriamente 70% dos valores para cada usuário. Dessa forma, é reproduzido um cenário em que o usuário consumiu produtos de alguns tópicos e pretende-se predizer - ou recomendar - o consumo dos tópicos com os quais o usuário não interagiu.

E. Agrupamento de usuários

Por fim, foi realizado um clustering dos perfis dos usuários utilizando K-means, com um parâmetro arbitrário de 20 clusters. Em seguida, foram comparadas as notas dadas aos produtos consumidos pelos grupos encontrados.

IV. Resultados

A. Avaliação da modelagem de tópicos

Foram encontrados 54 tópicos com uma coerência semântica satisfatória. É possível perceber desde temáticas amplas, como: terror, música, ficção científica, história, romance, comédia, animações, esporte e super-heróis; aspectos técnicos, como formato da mídia (DVD, VHS, fita, disco e etc), duração, condição do produto, se é um box de coleção, linguagem, região, formato de áudio e etc; até nichos mais específicos, como japão, scooby-doo, Doctor Who, Star Trek e etc.

Dessa forma, no que se refere a pergunta 1, aspectos técnicos e gêneros emergiram como tópicos.

B. Avaliação da construção de perfis dos usuários

Para a avaliação do modelo de perfis dos usuários, foi construído um conjunto que incluía produtos consumidos ocultados no treino e os produtos não consumidos presentes no treino. Para cada produto do conjunto de avaliação de cada usuário foi atribuído um score igual a similaridade cosseno entre os embbedings do produto e o usuário, assim construindo um ranking sobre o qual foi aplicada uma métrica comum para a avaliação de algoritmos de recomendação chamada Normalize Discounted Cumulative Gain (NDCG), atribuindo uma relevância binária para cada posição - 1 para os produtos realmente consumidos e 0 para os produtos não consumidor. O NDCG é útil, porque ele atribui maiores valores para produtos realmente consumidos à medida que eles se posicionam mais próximos do topo do ranking, de forma que um ranking perfeito tem valor 1.

Discounted Cumulative Gain (DCG) é definido por (1) e o NDCG é definido pela (2) [2].

$$DCG_p = \sum_{i=1}^{p} \frac{liked_i}{\log_2(i+1)} \tag{1}$$

$$nDCG = \frac{DCG}{IDCG} \tag{2}$$

Os resultados médios dos modelos estão apresentados na tabela 1.

Os resultados do modelo baseado em regressão logística são bons e consideravelmente superiores aos do modelo baseado em média. Isso significa que o modelo foi capaz de captar os interesses dos usuários e fazer recomendações relevantes, demonstrando que é possível predizer com eficácia satisfatória o consumo de outros produtos a partir da análise dos tópicos que o usuário consome.

TABLE I
Resultados de nDCG@k para diferentes modelos.

Modelo	nDCG@1	nDCG@5	nDCG@10
Baseado em regressão logística	0.8472	0.8167	0.7877
Baseado em média	0.5663	0.5703	0.5660

C. Avaliação da predição de consumo de tópicos

Para avaliar e comparar os 3 modelos, utilizou-se da métrica Normalized Mean Absolute Error (NMAE), definida por (3).

$$NMAE = \frac{\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\bar{y}} \tag{3}$$

A Tabela 2 apresenta os resultados dessa métrica para cada modelo. Todos os modelos tiveram um erro muito baixo, o que significa que eles conseguiram reconstruir de forma eficaz os valores ocultados.

TABLE II
Resultados de NMAE para diferentes modelos de fatoração matricial.

Modelo	NMAE
SVD	0.0085
NMF	0.0087
ALS	0.0081

D. Avaliação do agrupamento dos usuários

Após o agrupamento dos usuários foi feita a contagem de consumo de cada categoria de produto dentro dos clusters. A figura 3 mostra a distribuição da proporção de consumo de cada categoria dentro dos clusters. Com base neste gráfico, considerou-se como categorias mais consumidas aquelas que representassem uma parcela maior que 20% das categorias consumidas pelo clusters. Como categorias menos consumidas, considerou-se aquelas que representam uma parcela inferior a 10%.

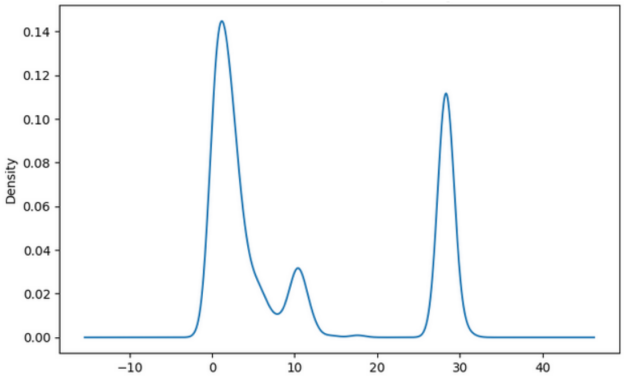

Fig. 3. Distribuição de número de avaliações por usuário

A média de avaliação das categorias foi de 4.057 para as mais consumidas e 4.061 para as menos consumidas. Logo, com uma diferença tão pequena, não é possível que os produtos de categorias mais consumidas por um grupos são avaliados diferente dos produtos de categorias menos consumidas. Logo, o modelo foi incapaz de identificar uma relação entre as categorias e os tópicos que refletisse em uma diferenciação na avaliação dos produtos.

V. Conclusão

A modelagem de tópicos foi capaz de encontrar temáticas relacionadas tanto a aspectos técnicos dos produtos, como de gênero do conteúdo. Além disso, foram encontrados tópicos ainda mais amplos, como exercícios físicos, música, história, cultura japonesa, desenhos animados e etc. Além de tópicos mais específicos, como Scooby-Doo, Doctor Who e Star Trek.

Os tópicos consumidos pelos usuários foram capazes de produzir um perfil que capturou a relação com a compra de outros produtos, confirmando que os tópicos extraídos podem ser utilizados para fazer recomendações de outros produtos. Além disso, os modelos de predição de consumo de tópicos apresentou um erro muito pequeno, demonstrando que os tópicos extraídos podem ser usados para predizer o consumo de outros tópicos. Dessa forma, os tópicos extraídos foram capazes de descrever a relação com a compra de outros produtos.

O agrupamento dos usuários com base no perfil não foi capaz de apresentar uma variação significativa entre as avaliações de categorias mais e menos consumidas pelos grupos, de forma que não foi encontrado uma relação entre os tópicos e as categorias que refletisse na avaliação pelos consumidores.

VI. Limitações e trabalhos futuros

Primeiramente, nota-se que os dados utilizados foram bastante restritos. Um conjunto de 2128 usuários com interações bastante densas não traduz a maioria dos cenários da vida real. Por isso, um trabalho futuro com um conjunto de dados mais amplo e com uma densidade menor será útil para uma melhor compreensão das questões levantadas neste trabalho.

References

[1] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proc. EMNLP*, Hong Kong, 2019, pp. 188–197.

[2] A. Alekseev and S. Nikolenko, "Word embeddings for user profiling in online social networks," *Computación y Sistemas*, vol. 21, no. 2, pp. 203–226, Jun. 2017.

[3] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022.

[4] X. Guan, C.-T. Li, and Y. Guan, "Matrix factorization with rating completion: An enhanced SVD model for collaborative filtering recommender systems," *IEEE Access*, vol. 5, pp. 27668–27678, 2017.

[5] D. Shah, P. Shah, and A. Banerjee, "Similarity based regularization for online matrix-factorization problem: An application to course recommender systems," in *Proc. IEEE TENCON*, Penang, Malaysia, 2017, pp. 1874–1879.