Especulação Mística. Uma abordagem de Clusterização e Busca Semântica na aproximação de preço em cartas de Magic: The Gathering

Rodrigo Marques Duarte[1], **André de Lima Salgado**[2], **Paula Figueira Cardoso**[3]

[1]Departamento de Automática – Universidade Federal de Lavras (UFLA)
Lavras, MG – Brasil

[2]Departamento de Ciência da Computação – Universidade Federal de Lavras (UFLA)
Lavras, MG – Brasil

[3]Faculdade de Computação – Universidade Federal do Para (UFPA)

rodrigo.duarte2@estudante.ufla.br, andre.salgado@ufla.br, pcardoso@ufpa.br

Abstract. *Magic: The Gathering (MTG) is a collectible card game that combines visual and textual elements. The release of a new collection has a direct impact on the game, leading to uncontrolled speculation on the prices of the new cards. This article presents an innovative approach to price speculation in MTG cards using clustering algorithms and semantic search. The technique allows for predicting initial prices with minimal information upon the card release and shows effectiveness in forecasting price ranges. It can also be applied to the stock market to predict the impact of news and used to prevent fraud and inflated prices by comparing existing samples.*

Resumo. *Magic: The Gathering (MTG) é um jogo de cartas colecionáveis que combina elementos visuais e textuais. O lançamento de uma nova coleção gera impactos diretos no jogo, e com isso especulações nos preços das novas cartas são tomadas sem menor controle. Este artigo apresenta uma abordagem inovadora para a especulação de preços em cartas de MTG utilizando algoritmos de clusterização e busca semântica. A técnica permite prever preços iniciais com informações mínimas no lançamento das cartas e mostra eficácia na previsão de faixas de preços, podendo ser aplicada a mercado de ações a fim de prever impacto de notícias e usada para prevenir golpes e preços exacerbados ao comparar amostras já existentes.*

1. Introdução

"*Magic: The Gathering (MTG)*" é um dos jogos de cartas colecionáveis mais antigos em existência [Guinness World Records 2024], com uma história que abrange três décadas e uma base de jogadores que ultrapassa 10 milhões [Draftsim 2024]. Diversos estudos já exploraram o MTG e seu mercado especulativo, como os realizados por [Pawlicki et al. 2014, Fink et al. 2015, Weber 2021]. Contudo, há um aspecto menos abordado: o MTG é, em sua essência, um jogo onde a linguagem desempenha um papel central, seja ela visual ou literária. Entender como essas informações são comunicadas e processadas pode oferecer insights valiosos para o Processamento de Linguagem Natural

(*NLP*). Com o lançamento de novas coleções, cartas têm seu preço especulado sem controle e agente controlador. Para solucionar este problema, um algoritmo para comparação da carta em lançamento com cartas já existentes foi desenvolvido. Possibilitando alertar sobre uma possivel escolha alternativa já existente de menor custo assim como previnir precificações abusivas. Neste trabalho, abordamos a problemática da classificação de preço utilizando um algoritmo de clusterização e busca semântica com o auxílio de ferramentas de NLP e modelos de linguagem. Na Seção 1, descreve-se a metodologia e recursos utilizados. Na Seção 3, apresentam-se os resultados, e por fim, na Seção 4, tecemos as considerações finais.

2. Metodologia

Para validar a proposta apresentada, foram retiradas três faixas de preço de 26.079 cartas por meio de um algoritmo de extração de conteúdo do site [Ligamagic 2024]. Essas faixas de preço foram utilizadas para construir a base de dados, que foi complementada com dados fornecidos por [Scryfall 2024]. Após a coleta, os dados passaram por um processo de limpeza para remover cartas com, custo de mana, poder e resistência dependentes de mecânicas do jogo. Na Figura 1, apresenta-se a modelagem do problema.

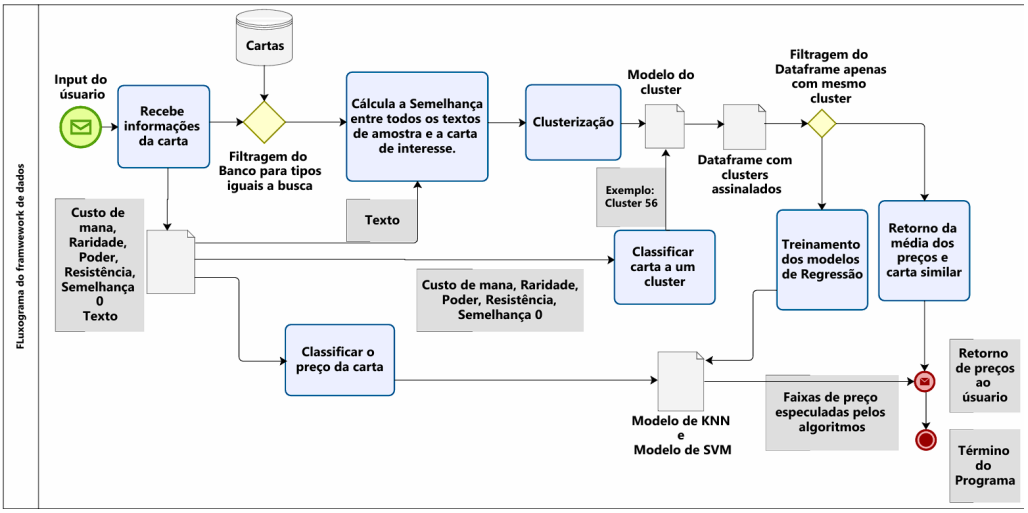

Figura 1. Fluxograma do Programa

A modelagem resultou em um algoritmo de recuperação de informação aumentada, que filtra o dataset e clusteriza cartas com atributos e textos semelhantes, permitindo a especulação de preços desde o momento de lançamento da carta, mesmo com informações mínimas [Fink et al. 2015]. A seguir, descrevemos o funcionamento do algoritmo.

2.1. Entrada de Busca, Seleção e Cálculo de Semelhanças

Inicialmente, o usuário fornece informações conhecidas sobre a carta no momento de seu lançamento, como custo de mana, raridade, poder, resistência, texto, pontos de lealdade, entre outros. Essas informações são usadas para criar um objeto que será passado pelos modelos subsequentes. O dataset é filtrado para conter apenas cartas do mesmo tipo da carta de interesse. Em seguida, calcula-se a similaridade por meio da métrica de Jaccard.

2.1.1. Métrica de Jaccard

A métrica de Jaccard é uma medida de similaridade entre dois conjuntos, calculada como o tamanho da interseção dos conjuntos dividido pelo tamanho da sua união [Manning et al. 2008]. É amplamente utilizada em problemas de comparação de textos e na análise de similaridade entre documentos. A fórmula da métrica de Jaccard é dada por:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

O valor de $J(A, B)$ varia entre 0 e 1, onde 1 indica que os conjuntos são idênticos, e 0 indica que não há elementos em comum.

2.1.2. SentenceTransformer e Mineração de Parafrase

A biblioteca *SentenceTransformer* é uma ferramenta para a criar representações vetoriais de sentenças que preservam relações semânticas. Um dos principais usos dessa biblioteca é na mineração de paráfrases, que envolve a identificação de sentenças que transmitem ideias semelhantes, mesmo quando formuladas de maneiras diferentes. Para realizar essa tarefa, o *SentenceTransformer* utiliza modelos pré-treinados como *BERT* e *RoBERTa* para converter sentenças em vetores de alta dimensão [Liu et al. 2019, Devlin et al. 2019]. Essas representações vetoriais são, então, comparadas entre si utilizando medidas de similaridade, como a similaridade de cosseno, para determinar o quão semanticamente próximas duas sentenças estão. O fluxo básico de trabalho envolve:

1. **Tokenização**: A sentença é dividida em tokens e convertida em uma sequência de embeddings.
2. **Transformação**: A sequência de tokens é passada através de camadas do modelo (e.g., BERT), resultando em uma representação vetorial densa da sentença.
3. **Mineração de Paráfrase**: As representações vetoriais das sentenças são comparadas usando similaridade de cosseno para identificar pares de sentenças que são paráfrases.

Em seguida, é calculada a similaridade por métrica de Jaccard entre a carta de interesse e as cartas no dataset filtrado, utilizando vetores gerados pelo modelo *SentenceTransformers*. Essa métrica de similaridade é crucial para a clusterização, que agrupa cartas com textos semelhantes conforme exemplo a seguir: [Scryfall 2024b, Scryfall 2024a]

1. *"Whenever one or more other Rabbits, Bats, Birds, and/or Mice you control enter, scry 1. Other Rabbits, Bats, Birds, and Mice you control get +1/+1."*;
2. *"Other Squirrels you control get +1/+1."*.

A análise dos textos revelou uma similaridade de 0,005679 entre as sentenças, além de um padrão claro nas paráfrases, como o aumento de poder e resistência de certas criaturas, onde a semântica é semelhante. Incorporar esse algoritmo de busca semântica ao *K-means* pode aprimorar o desempenho do algoritmo, pois permitirá uma clusterização mais eficiente, agrupando melhor as cartas com textos mais semelhantes.

2.1.3. Clusterização e Classificação de Pertencimento da Carta de Entrada

A clusterização visa agrupar cartas com textos semelhantes, utilizando a métrica de Jaccard como parâmetro. Para isso, foi usada a biblioteca do *sklearn* (*slkearn.clustering* e *sklearn.pipeline*) para construir a *pipeline* de clusterização. Os dados foram normalizados usando o algoritmo *MinMax* para otimizar o cálculo de distâncias no *K-means*. Foram definidos 170 clusters para prevenir a formação de agrupamentos extensos e reduzir os efeitos de *overfitting*. O coeficiente de silhueta, usado para validar a disposição dos centróides, foi de 0,6514. Com o modelo treinado, a carta de interesse é atribuída ao cluster mais adequado, filtrando o dataset para usar apenas amostras desse cluster na especulação do preço.

3. Resultados

Com base no cluster final, podemos retornar a média dos preços na amostra e determinar faixas de preço para a carta de interesse. Além disso, identificamos a carta com texto mais semelhante na amostra para verificar se já existe uma ocorrência similar do texto. As amostras são passadas por dois modelos de regressão: *Random Forest Tree (RFT)* e *KNearest Neighbor (KNN)*. Com uma divisão de 20% para teste e validação, os erros médios quadráticos e absolutos, assim como as faixas de preço previstas, são apresentados nas Tabelas 1 e 2.

RFT %	KNN %
3.672582	15.733944
78.898198	79.506967
53.006127	80.778658

Tabela 1. Erro Médio Quadrático do RFT e KNN

RFT	KNN
12.162827	10.165323
60.973797	53.930584
56.188981	55.616094

Tabela 2. Erro Médio Absoluto do SVM e KNN

Preços RFT	Preços KNN
5.41250	0.256667
13.40635	2.203333
26.90155	8.093333

Tabela 3. Faixas de preço mínimo, médio e maior preço obtida para uma carta pelos 2 algoritmos

4. Considerações Finais

Este estudo apresentou uma metodologia para a especulação de preços em cartas do MTG, combinando técnicas de NLP e algoritmos de clusterização. Os resultados demonstram que o método proposto é eficaz na previsão de faixas de preço e tem potencial para ser aplicado para prever impacto de notícias em ações e em comparação de produtos onde análises de sequências textuais são relevantes. Este trabalho foi estudo piloto para aplicação da metodologia em processos de classificação de preços de cafés populares a

partir da análise do design e conteúdo de suas embalagens. Encontramos limitações no uso da solução devido à natureza dos dados das cartas e suas altas proximidades que ocasionam centroides muito próximas. Entendemos que o uso de um modelo de vetorização com contextos treinados diretamente ao jogo pode melhorar a separação semântica das amostras.

5. Agradecimentos

Agradecemos pelo apoio e financiamento Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP processo nº 202106968−3, além de CAPES, CNPq, UFLA, UFPA e FAPEMIG.

Referências

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Draftsim (2024). Mtg arena player count - how many people play? https://draftsim.com/mtg-arena-player-count/. Accessed: 2024-08-12.

Fink, D., Pastel, B., and Sapra, N. (2015). Predicting the strength of magic: The gathering cards from card mechanics. *Cs 229: Machine Learning Final Project, December 2015*.

Guinness World Records (2024). First modern trading card game. https://www.guinnessworldrecords.com/world-records/first-modern-trading-card-game. Accessed: 2024-08-12.

Ligamagic (2024). Ligamagic. Acesso em: 2 ago. 2024.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Pawlicki, M., Polin, J., and Zhang, J. (2014). Prediction of price increase for magic: The gathering cards. In *Proceedings of the Conference*, Stanford, CA. Stanford University.

Scryfall (2024). Scryfall. Acesso em: 2 ago. 2024.

Scryfall (2024a). Squirrel sovereign. Acesso em: 2 ago. 2024.

Scryfall (2024b). Valley questcaller. Acesso em: 2 ago. 2024.

Weber, D. (2021). Exploring markets: Magic the gathering - a trading card game. Working Paper 3/2021, IU Internationale Hochschule.