Inferência Automática de Nível Calórico de Receitas Culinárias Através de Técnicas de Aprendizagem de Máquina

Larissa F. S. Britto[1], **Luciano D. S. Pacífico**[2], **Teresa B. Ludermir**[1]

[1]Centro de Informática – CIn
Universidade Federal de Pernambuco – UFPE – Recife, PE – Brasil

[2]Departamento de Computação (DC)
Universidade Federal Rural de Pernambuco (UFRPE) – Recife, PE – Brasil

lfsb@cin.ufpe.br, luciano.pacifico@ufrpe.br, tbl@cin.ufpe.br

Abstract. *In this work, a tool for the automatic inference of calorie content in cooking recipes is proposed, through a Text Classification approach. This tool will be part of a Recommendation System under development, to assist health professionals and users in general to elaborate healthy diets.*

Resumo. *Neste trabalho, uma ferramenta para inferência automática do nível calórico de receitas culinárias é proposta, através de uma abordagem de Classificação de Textos. A ferramenta fará parte de um Sistema de Recomendação em desenvolvimento, para o auxílio a profissionais de saúde e usuários em geral na elaboração de dietas saudáveis.*

1. Introdução

A alimentação desempenha um papel essencial na vida do ser humano. A alimentação adequada e saudável pode trazer inúmeros benefícios à saúde e bem estar de um indivíduo, enquanto uma dieta inapropriada e descontrolada pode contribuir para o surgimento de distúrbios e doenças como obesidade, diabetes tipo 2, doenças cardiovasculares e câncer [Afshin et al. 2019]. O cenário atual gerado pela pandemia da Covid-19 acentua a condição de insegurança alimentar e nutricional por fatores como o excesso do consumo de alimentos considerados prejudiciais à saúde [Martinelli et al. 2020].

Diversas áreas do conhecimento humano buscam dar sua contribuição para a melhoria dos hábitos alimentares da população. Na Computação, Sistemas de Recomendação para sugestão de alimentos, receitas e menus completos têm sido propostos, a maioria deles, com o objetivo de recomendar dietas que possam contribuir positivamente para os hábitos alimentares do usuário [Yera Toledo et al. 2019, Jiang et al. 2019].

Como parte integrante de um Sistema de Recomendação de Receitas Culinárias em desenvolvimento, iniciado nos trabalhos [Oliveira et al. 2019, Britto et al. 2020a, Britto et al. 2020b], que além de gostos pessoais, habilidades culinárias, restrições alimentares e ingredientes disponíveis, leve em consideração também o fator nutricional do prato, neste trabalho será proposta uma ferramenta de inferência automática do nível calórico de receitas através da Classificação de Textos (CT). Para escolher o melhor classificador para compor esse módulo de inferência, alguns dos principais modelos adotados na literatura de CT serão avaliados e comparados. Espera-se que o desenvolvimento do sistema possa auxiliar profissionais de saúde e usuários em geral na criação de dietas

mais saudáveis, balanceadas nutricionalmente e adequadas às suas necessidades diárias. O trabalho está dividido como segue. Na próxima seção (Seção 2), é descrita em detalhes a metodologia adotada para o desenvolvimento da ferramenta proposta. Na Seção 3, a configuração experimental e os resultados obtidos serão discutidos. Por fim, na Seção 4, as conclusões do trabalho e linhas de trabalhos futuros são apresentados.

2. Metodologia

Nesta seção, as etapas para desenvolvimento da ferramenta proposta são descritas em detalhes, como ilustrado na Figura 1. A primeira etapa é a **Obtenção dos Dados**. A base de dados adotada foi extraída do *website* Food.com [Majumder et al. 2019]. Os dados utilizados neste trabalho são as *informações nutricionais* e os *modos de preparo* das receitas. Os modos de preparo contém textos instrucionais sobre o preparo das receitas, nos quais podem ser encontrados não só os ingredientes, mas também os preparos feitos nos mesmos, que podem alterar seus níveis calóricos. Através das informações nutricionais é feita a **Anotação das Classes**. As receitas são categorizadas em duas classes referentes ao seu nível calórico [U.S. Food and Drug Administration 2020]: *baixa caloria* (29891 receitas) e *alta caloria* (89791 receitas). Para evitar que o desbalanceamento das receitas impacte na performance dos classificadores, é feito o **Balanceamento dos Dados**, resultando em um total de 59782 na base de dados final (29891 receitas por classe).

Os modos de preparo das receitas são encontrados em linguagem natural. Para transformar esses textos em conjuntos mais significativos de dados é necessária uma etapa de **Pré-processamento**. Essa etapa incluiu a conversão dos textos para *lower case*, onde todas as letras são convertidas para a forma minúscula, fazendo com que termos idênticos representados de formas diferentes, como, por exemplo, “Bake”e “bake”, sejam considerados um único termo. Além disso é feita a *remoção de pontuações e caracteres especiais*.

Através da **Extração de Características** é possível transformar textos brutos em dados numéricos suportados pelos classificadores. Um dos métodos de extração de característica mais populares na literatura de CT [Kowsari et al. 2019] foi adotado neste trabalho: o *Term Frequency-Inverse Document Frequency* (TF-IDF) [Britto et al. 2020b]. O TF-IDF mensura quão importante um termo é, tentando diminuir assim a influência de termos que ocorrem com uma grande frequência, mas que possuem pouca relevância. Por fim, as características extraídas dos modos de preparo das receitas passam pelo processo de **Classificação**, a fim de identificar automaticamente a que classe tais receitas pertencem. Alguns dos classificadores mais populares na tarefa de CT foram adotados: Naive Bayes (NB), Regressão Logística (*Logistic Regression* - LR) e Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM) [Kowsari et al. 2019].

3. Avaliação Experimental

Nesta seção, os resultados experimentais obtidos serão apresentados. Com o objetivo de analisar a capacidade da inferência de nível de calorias adotando uma abordagem de Classificação de Textos em linguagem natural, três classificadores popularmente adotados na literatura de CT tiveram seus desempenhos analisados ao serem testados na base de dados proposta. As estatísticas finais da base de dados podem ser vistas na Tabela 1. Para a avaliação experimental, um esquema de validação cruzada com 10 *folds* foi adotado, e repetido por dez vezes, tentando assim, evitar resultados obtidos por sorte. Para a avaliação

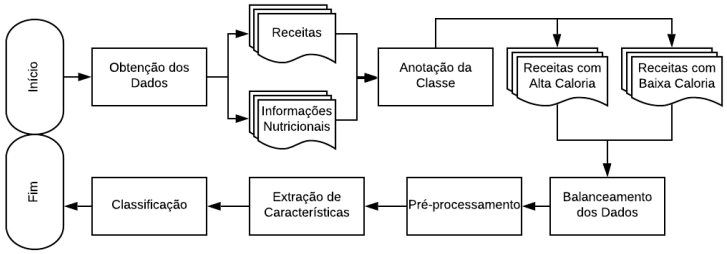

Figura 1. Etapas da abordagem proposta.

Tabela 1. Estatísticas da base de dados proposta.

	Todas as Receitas	Alta Caloria	Baixa Caloria
Número total de Receitas	59782	29891	29891
Tamanho do vocabulário	22990	17116	15773
Média de palavras por receitas	94.17	114.24	74.10
Média de sentenças por receitas	14.62	17.76	11.49
Média de palavras por sentença	6.60	6.56	6.64

dos modelos, quatro métricas provenientes da literatura de Aprendizagem de Máquina foram adotadas, são elas: Acurácia, Precisão, Revocação e F-*Measure*. Os resultados experimentais são apresentados na Tabela 2.

Considerando uma análise empírica nos resultados obtidos, podemos observar que todos os classificadores foram capazes de obter resultados satisfatórios. Dos classificadores testados, o RL alcançou os melhores resultados na maioria das métricas adotadas, alcançando uma acurácia de teste média de 81.44%, sendo seguido pelo SVM, que com resultados muito próximos, chegou a alcançar 81.25% de acurácia. Com um desempenho significativamente inferior na maioria das métricas, o NB obteve uma acurácia de 75.38%, porém teve o menor tempo médio de execução (3.35 segundos).

4. Conclusões

Neste trabalho, foi proposta uma abordagem de Classificação de Textos para a construção de uma ferramenta de inferência automática do nível calórico de receitas. Uma avaliação experimental demonstrou a eficiência da abordagem proposta, onde todos os classificadores testados foram capazes de obter resultados satisfatórios. Através de uma análise empírica, foi possível destacar o desempenho dos classificadores Regressão Logística e Máquinas de Vetores de Suporte, que obtiveram uma acurácia de teste média maior que 81%, sendo boas opções para compor o módulo de inferência proposto. Como trabalhos futuros, pretendemos analisar como diferentes informações nutricionais podem ser inferidas a partir dos modos de preparo. Além disso, pretedemos integrar a ferramenta de inferência de informações nutricionais ao Sistema de Recomendação de Receitas Culinárias em desenvolvimento, e avaliar com profissionais de saúde e usuários a qualidade das recomendações feitas. Por fim, o Sistema de Recomendação será disponibilizado ao

Tabela 2. Resultados Experimentais.

	Acurácia	Precisão	Revocação	F-*Measure*	Tempo
NB	0.7538	0.7295	**0.8068**	0.7662	**3.3502**
RL	**0.8144**	**0.8209**	0.8043	**0.8125**	4.1511
SVM	0.8125	0.8180	0.8040	0.8109	3.8470

público, como ferramenta de auxílio à elaboração de dietas saudáveis, nutritivas e adequadas às restrições dos usuários.

Agradecimentos

Os autores gostariam de agradecer à FACEPE, ao CNPq e à CAPES.

Referências

Afshin, A., Sur, P., Fay, K., Cornaby, L., Ferrara, G., Salama, J., Mullany, E., Abate, K., Cristiana, A., Abebe, Z., Afarideh, M., Aggarwal, A., Agrawal, S., Akinyemiju, T., Alahdab, F., Bacha, U., Bachman, V., Badali, H., and Badawi, A. (2019). Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 393:1958–1972.

Britto, L. F. S., Pacífico, L. D. S., and Ludermir, T. B. (2020a). Geração automática de receitas culinárias para pessoas com restrições alimentares. In *Anais do XXXIX Concurso de Trabalhos de Iniciação Científica*, pages 61–70, Porto Alegre, RS, Brasil. SBC.

Britto, L. F. S., Pacífico, L. D. S., and Ludermir, T. B. (2020b). Inferência automática do nível de dificuldade em receitas culinárias usando técnicas de processamento de linguagem natural. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 104–115, Porto Alegre, RS, Brasil. SBC.

Jiang, H., Wang, W., Liu, M., Nie, L., Duan, L.-Y., and Xu, C. (2019). Market2dish: A health-aware food recommendation system. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2188–2190, New York, NY, USA. Association for Computing Machinery.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4).

Majumder, B. P., Li, S., Ni, J., and McAuley, J. (2019). Generating personalized recipes from historical user preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5976–5982, Hong Kong, China. Association for Computational Linguistics.

Martinelli, S. S., Cavalli, S. B., Fabri, R. K., Veiros, M. B., Reis, A. a. B. C., and Amparo-santos, L. (2020). Strategies for the promotion of healthy, adequate and sustainable food in brazil in times of covid-19. *Revista de Nutrição*, 33.

Oliveira, E. G., Britto, L. F. S., Pacífico, L. D. S., and Ludermir, T. B. (2019). Recomendação e geração de receitas baseada na substituição de ingredientes. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 238–249, Porto Alegre, RS, Brasil. SBC.

U.S. Food and Drug Administration (2020). How to understand and use the nutrition facts label. https://www.fda.gov/food/new-nutrition-facts-label/how-understand-and-use-nutrition-facts-label.

Yera Toledo, R., Alzahrani, A. A., and Martínez, L. (2019). A food recommender system considering nutritional information and user preferences. *IEEE Access*, 7:96695–96711.