A Simple Q&A Search Engine in the Context of Coronavirus Pandemic

Leila Fabiola Ferreira
Team Evolution
Postgraduate Program in Applied Computing
Federal Technological University of Paraná – UTFPR
Curitiba – PR, Brazil
https://orcid.org/0000-0001-8506-203X

Mateus Cichelero da Silva
Team Evolution
Postgraduate Program in Applied Computing
Federal Technological University of Paraná – UTFPR
Curitiba – PR, Brazil
https://orcid.org/0000-0002-1311-1709

***Abstract*—This paper presents a simple search engine implementation that retrieves data from pandemic-related questions based on Tf-idf scores and Multinomial Naive Bayes algorithm classification.The application returns information from several input questions about the pandemic, showing the most relevant answer calculated by the system and also suggests a related scientific article for reading. The datasets used in this approach were obtained from websites of Brazilian Ministry of Health, Fiocruz Foundation, Laura PA Digital and CORD-19.**

***Index Terms*—Information Retrieval, Search Engine, Naive Bayes, Tf-idf, Coronavirus Pandemic**

I. Introduction

Currently, the whole world is experiencing one of the most critical situations involving infectious disease and this new scenario has generated thousands of doubts and concerns related to several topics involving this serious pandemic, the new coronavirus and its consequent disease called COVID-19.

These outbreaks of viral diseases are nothing new, and some recent cases such as the severe acute respiratory syndrome (SARS) in 2002, the H1N1 swine flu in 2009, the Ebola virus in 2014 and the Zika virus in 2015 can be cited [1]. However, each new outbreak brings new doubts automatically, especially when a high risk of contagion and high mortality rates exists due to the scenario.

In this context, there are many engineering approaches that can be applied as tools to answer society's needs and doubts. An alternative to mitigate the lack of information and facilitate the search to obtain answers on a given subject are search engines that use the concepts of information retrieval. The definition of information retrieval is very broad, but considering only the academic field, it can be defined as the technique for obtaining information in a collection of documents of unstructured data, mainly texts, that satisfy a particular search [2].

Based on these concepts, this paper explores the use of natural language processing techniques (NLP) to pre-process textual content and its many applications, exploratory data analysis (EDA) being one of them, as an important step for those who are implementing the system to understand the dataset properly [2]. To achieve this, NLP libraries and APIs in Python programming language were used for data manipulation, such as NLTK and Spacy. Also, this resources and libraries were used to apply Multinomial Naive Bayes classification algorithm to return the most suitable answer for the query and also calculate the Tf-idf scores and cosine similarity between the same input query and available articles abstracts to suggest a reading. Finally, the application interface was developed using Streamlit, an open-source app framework, through which it is possible, in a quick and simplified way, to create applications in Python.

The other subjects in this article are organized into the following topics: Section II presents research works developed in the same area. Then the data, pre-processing methods and models that were used during the development of this work are described in section III. Moving on, the results are discussed in section IV, and possible points for improvement for future work are suggested in the last section.

II. Related Work

This section aims to describe similar research works developed in the same area. As mentioned in previous sections, this work is presented in the context of a practical application and of a case study analysis. Thus, the relevance and practical impact found in this type of solution in related publications found are also highlighted.

The pandemic period presented interesting challenges and problems for researchers in information retrieval, NLP and machine learning areas in general. An example of publication with a high relation degree to the theme of this work can be found in [3], which also makes use of CORD-19 dataset for construction and validation of an information retrieval system based on two main modules: a hybrid semantics/keywords recovery subsystem that takes an input query and returns a list of the thousand most relevant documents and a reordering subsystem that improves the relevance score system of the selected documents. The work makes use of deep learning models (Siamese-BERT) as well as those based on keywords (BM25, Tf-idf). In addition, interesting techniques such as text augmentation for insufficient data and description and use of different systems validation techniques are also applied and serve as inspiration for the team's developments.

Another article of interest [4] addresses broader aspects of the motivations, use and importance of such systems in the context of pandemic. Specialized chatbots can represent unique tools for applications aimed at dissemination of information (area in which the work presented here is found), symptom monitoring, mental health support and behavioral changes (recommendation of social distance and use of masks, for instance) . On the other hand, several challenges are also raised, such as the massive proliferation of fake news and how small errors can escalate intensely when using systems of this type.

Finally, in [5] the authors analyze the performance of Q&A systems, also using the CORD-19 base, through the use of pre-trained language models (GPT-2) together with different representation strategies to filter and retain relevant sentences for responses: Term Frequency - Inverse Document Frequency (Tf-idf), Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT), and Universal Sentence Encoder (USE). In addition, the authors describe a list of main related works, as well as suggestions of methodology for qualitative analysis of systems, which are relevant to this work.

III. Data, Data Pre-processing and Models

This section aims to briefly describe the data, pre-processing steps and models used during the development of this work.

A. Data

Three main textual databases were analyzed to elaborate this work:

- **Laura PA Digital** [6]: Database of real questions (in Portuguese) about the pandemic sent to a chatbot during the period from March 2020 to March 2021 corresponding to more than 1.26 million recorded messages. However, in the context of this work, only doubt-related topics messages were applied.
- **COVID-19 FAQ from official Brazilian sources**: Frequently asked questions and answers database assembled from a web scraping of official pages of Brazilian Ministry of Health [7] and Portal Fiocruz [8]. It was applied as a source of answers in the context of the Q&A application.
- **CORD-19** [9]: Base of scientific research on the virus and the pandemic. Here, a set of more than 160 thousand unique articles was sampled covering different topics and represented by their titles, abstracts and URLs. The base was used in the application as a reading recommendation, where each research question will also return a reading suggestion showing the title and URL to access the suggested article.

B. Data Pre-processing

Textual corpus in different languages (Brazilian Portuguese and English) were used. As described in the previous sections, the data flow of this application starts with an input question in Portuguese and must return both the answers resulting from the application of classifier/search engine in Portuguese and the suggested article from the CORD-19 database in English. To make this possible, the initial question is also translated into English and processed following the techniques applied to different Corpus.

After conducting the web scraping from official government sources, the answers were manually labeled in 43 different categories and more than 500 questions from Laura PA Digital database were related to the same categories created, thus characterizing the dataset for the supervised training of Naive Bayes classifier.

From this, a series of common NLP pre-processing techniques were applied, as seen in [10]: tokenization of terms, removal of special characters, accents, punctuation, stopwords and then lemmatization. In the latter case, pre-trained language models were applied specifically for Portuguese [11] (based on the UD Portuguese Bosque treebank), and English [12], according to the language of the processed database.

As a final processing step, it was necessary to transform the originally textual content into a numerical representation format, which can then be processed and understood by the applied models. This is done here by transforming the matrix of token counts into a Tf-idf matrix representation (term-frequency times inverse document-frequency) [13].

The purpose of this representation, which differentiates it from the simple use of absolute frequencies of a term appearance in a document, is to decrease the impact of tokens that often appear in a Corpus and carry little information, thus emphasizing rare terms with greater informative value.

Equation (1) represents its general formula, where the term frequency (number of times a term appears in a given document) is multiplied by the idf component, represented in (2), with n being the total number of documents in a set of documents and df(t) the number of documents in the document set containing the term t. In the context of the application, the resulting Tf-idf vectors are also normalized by the Euclidean norm, described in (3).

$$tf - idf(t,d) = tf(t,d) * idf(t) \tag{1}$$

$$idf(t) = log\left(\frac{1+n}{1+df(t)}\right) + 1 \tag{2}$$

$$v_{norm} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + ... + v_n^2}} \tag{3}$$

C. Models

With the pre-processing step completed, the Tf-idf representation matrices were used to fit the classification model and as a basis for queries in the search engine.

The Multinomial Naive Bayes algorithm was used to create a multiclass classifier that took the Tf-idf vectors as input features and returned the class of the question asked.

The equation (4) presents the base classification rule for the algorithms of this family (y representing the class of the

analyzed instance, $\hat{y}$ the estimated class and x_i the representative features). For the multinomial case, (5) presents the set of vectors θ, where n is the number of features (in this case the vocabulary size), θ_{yi} is the probability P $(x_i|y)$ that feature i appears in a sample of class y, N_{yi} is the number of times feature i appears in class y in the training set, N_y is the total count of features in class y and α corresponds to the smoothing priors term ($\alpha = 1$: Laplace smoothing, $\alpha < 1$: Lidstone smoothing)

$$\hat{y} = arg \ \max_{y} \ P(y) \prod_{i=1}^{n} P(x_i|y) \tag{4}$$

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{5}$$

From the returned class, the system presents the pre-registered answer for the question class.

Finally, this information retrieval system for a Q&A application and suggested articles makes use of the cosine similarity for ranking and return the most suitable document.

The cosine similarity calculates the L2-normalized dot product of vectors as shown in (6) (considering x and y row vectors). It is called cosine similarity, because Euclidean (L2) normalization projects the vectors onto the unit sphere, and their dot product is then the cosine of the angle between the points denoted by the vectors [15]. Thus, its result can be used to estimate which documents are closest to the query in the analyzed dimensional space.

In both cases the input question is taken as a search query. In the case of Q&A, tokenization of sentences is applied in the Corpus of official responses and these are considered the search documents; in CORD-19, each abstract of the sampled articles is considered as a search document.

$$k(x, y) = \frac{xy^T}{\|x\| \, \|y\|} \tag{6}$$

IV. Results

This section presents the main results for the developments carried out.

Fig. 1, Fig. 2 and Fig. 3 present the results of the performance metrics of the Naive Bayes Classifier in the validation dataset (considering hold-out of 75%/25%). Despite the limited size of the dataset, this classifier obtained robust metrics, especially considering its simplicity and the use of standard hyperparameters.

In addition, an important result of this work is the application itself, available at [16]. As can be seen in Fig. 4, the web app is capable of processing a question in Portuguese as an input, returning the indicated answers through the application of classifier and the search engine and also returning a suggested scientific article from the CORD-19 database.

Therefore, through the application of basic NLP techniques, information retrieval and machine learning, its possible to develop a simple but functional system as a proof of concept.

accuracy 0.8482142857142857

	precision	recall	f1-score
definicao_coronavirus	0.60	1.00	0.75
acoes_doente	1.00	1.00	1.00
covid_coronavirus	1.00	1.00	1.00
infeccao	0.60	1.00	0.75
grupo_risco	0.60	1.00	0.75
virus_superficie	1.00	1.00	1.00
uso_mascara	1.00	1.00	1.00
seguranca_vacina	1.00	1.00	1.00
imunidade_vacina	1.00	1.00	1.00
vacina_mutacoes_virus	0.00	0.00	0.00
vacina_efeito_adverso	1.00	1.00	1.00
vacina_periodicidade	1.00	1.00	1.00
vacina_tempo_imunidade	0.00	0.00	0.00
vacinado_transmissao	1.00	1.00	1.00
quem_vacina_primeiro	1.00	1.00	1.00
vacina_quem_deve_tomar	0.00	0.00	0.00
vacina_adultos_criancas	0.00	0.00	0.00
vacina_infectados	1.00	1.00	1.00
vacina_obrigatoria	1.00	1.00	1.00
vacina_duas_doses	0.75	1.00	0.86
vacina_eficiencia	1.00	1.00	1.00
escolha_vacina	1.00	0.50	0.67
outras_vacinas_juntas	1.00	1.00	1.00

Fig. 1. Accuracy and classification performance metrics by classes - 1.

doses_diferentes_vacinas	1.00	1.00	1.00
vacinado_exterior	0.67	1.00	0.80
sem_carteira_sus	1.00	1.00	1.00
controle_vacinacao	1.00	1.00	1.00
capacidade_vacinacao_sus	1.00	1.00	1.00
comorbidade_atestado	1.00	0.67	0.80
comprar_vacina	1.00	1.00	1.00
uso_emergencial	1.00	1.00	1.00
vacina_gravida	1.00	1.00	1.00
tempo_imunidade	1.00	0.67	0.80
variantes_virus	0.50	0.67	0.57
identificacao_variantes	0.38	1.00	0.55
reinfeccao	1.00	0.67	0.80
origem_coronavirus	1.00	1.00	1.00
animais_coronavirus	0.60	1.00	0.75
sintomas	1.00	0.75	0.86
antibioticos	1.00	1.00	1.00
hidroxicloroquina	1.00	1.00	1.00
aumentar_imunidade	0.00	0.00	0.00
testes_coronavirus	0.00	0.00	0.00

Fig. 2. Classification performance metrics by classes - 2.

micro avg	0.85	0.85	0.85
macro avg	0.78	0.81	0.78
weighted avg	0.82	0.85	0.82

Fig. 3. Classification general performance metrics.

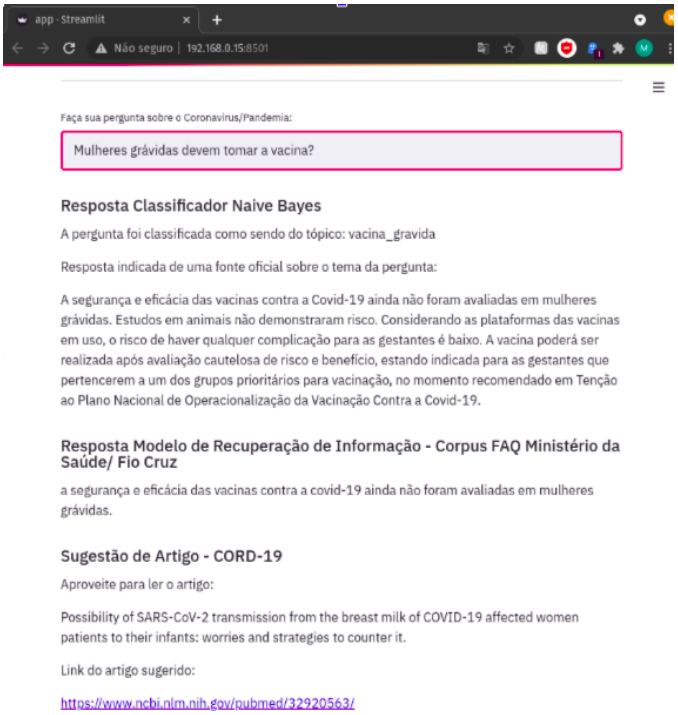

Fig. 4. Application interface.

V. Future Work

Despite the good performance metrics presented by the classifier, it should be considered that there are still few annotated examples of each class for training in the dataset. The ideal scenario would be to continue expanding the labeled base, perhaps applying some more automated method, such as bootstrapping from what has already been created manually.

Also, it is possible to explore alternatives to optimize the models and calculations applied in order to decrease the processing time and memory usage. In addition, there is the possibility of testing other classification models and document ranking methods to compare the performance of each one, as the application of more common and easy-to-apply examples was a limitation for the designing of this system.

An important point to be discussed detailed in future work is the explainability of predictions performed by the classifier. That can serve as a basis for better understanding of the system's operation and to analyze of what can be further improved.

References

[1] G. Khan, M. Sheek-Hussein, A. R. Al Suwaidi, K. Idris, F. M. Abu-Zidan, Novel coronavirus pandemic: A global health threat, Turk J Emerg Med. 2020 Apr-Jun; 20(2): 55–62. Published online 2020 May 27.

[2] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[3] Esteva, A., Kale, A., Paulus, R. et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. npj Digit. Med. 4, 68 (2021). https://doi.org/10.1038/s41746-021-00437-0

[4] Miner, A.S., Laranjo, L. & Kocaballi, A.B. Chatbots in the fight against the COVID-19 pandemic. npj Digit. Med. 3, 65 (2020). https://doi.org/10.1038/s41746-020-0280-0

[5] D. Oniani and Y. Wang, "A Qualitative Evaluation of Language Models on Automatic Question-Answering for COVID-19," presented at the BCB '20: 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Sep. 2020, doi: 10.1145/3388440.3412413.

[6] "Home," Laura. [Online]. Available: https://laura-br.com/. [Accessed: 18-Apr-2021].

[7] "Perguntas e Respostas," Ministério da Saúde. [Online]. Available: https://www.gov.br/saude/pt-br/coronavirus/perguntas-e-respostas. [Accessed: 01-May-2021].

[8] "Coronavírus: Perguntas e respostas," Fiocruz. [Online]. Available: https://portal.fiocruz.br/coronavirus/perguntas-e-respostas. [Accessed: 02-May-2021].

[9] Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: The Covid-19 Open Research Dataset. Preprint. ArXiv. 2020;arXiv:2004.10706v2. Published 2020 Apr 22.

[10] Bird, S., Klein, E. and Loper, E., 2009. Natural Language Processing with Python. Sebastopol: O'Reilly Media, Inc.

[11] "Portuguese · spaCy Models Documentation," Portuguese. [Online]. Available: https://spacy.io/models/pt#pt_core_news_md. [Accessed: 03-May-2021].

[12] "nltk.stem package," nltk.stem package - NLTK 3.6.2 documentation. [Online]. Available: http://www.nltk.org/api/nltk.stem.html#module-nltk.stem.wordnet. [Accessed: 04-Apr-2021].

[13] "TfidfTransformer," scikit. [Online]. Available: https://bit.ly/3v3exhQ [Accessed: 01-May-2021].

[14] "1.9. Naive Bayes," scikit. [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html. [Accessed: 06-May-2021].

[15] "Cosine similarity," scikit. [Online]. Available: https://scikit-learn.org/stable/modules/metrics.html#cosine-similarity. [Accessed: 10-May-2021].

[16] "Máquina de Busca no Contexto da Pandemia COVID-19: um estudo de caso para aplicações Q&A", Equipe Evolution. [Online]. Available: https://covid19-qa-ptbr.herokuapp.com/ [Accessed: 10-May-2021].