BRYT: Automated Keywords Extraction for OD

Umair Ahmed
University of Camerino
umair.ahmed@unicam.it

Marco Piangerelli
University of Camerino
marco.pinagerelli@unicam.it

Andrea Polini
University of Camerino
andrea.polini@unicam.it

I. Introduction

a) Motivations: In the information age, there has been substantial growth in the volume and circulation of data. As a result, numerous organizations and stakeholders have come to appreciate the significance of open data in fostering transparency and generating added value. To realize these benefits, they aim to make their data freely available.

According to the Open Knowledge Foundation, data is considered open if it can be "freely accessed, used, modified, and shared by anyone for any purpose". Most of the data portals provide mechanisms to host open data sets and make them available, but they severely lack in terms of findability. Users with less than adequate technical skills struggle with findability, and eventually, it affects the accessibility and usability of open data sets that are already available.

Findability can be enhanced through intuitive search engines and intelligent recommendation systems for laymen users. Quality metadata plays a significant role in the effectiveness of search engines, making it essential for improving findability [2].

b) Innovations: We examined 15 notable data portals in this study and discovered they lacked intelligent and intuitive search mechanisms. The analysis of these data portals is presented in Table 1. None of the portals featured a recommender system, and most of them relied on basic keyword searches, with some offering autocomplete suggestion mechanisms related to dataset titles. The search engines typically matched metadata elements such as titles, categories, or keywords. The majority of searches were not intuitive, mainly due to the lack of representative metadata for the given datasets. Most search mechanisms concentrated on metadata, and a large number of datasets lacked sufficient metadata to be easily findable. Although not all metadata elements contribute to findability, titles, keywords, and categories significantly enhance it. In this research, we emphasize keywords, given their strong association with search engines and their ability to concisely represent the document [3]. This study explores methods to automate the keyword extraction process, aiming to recommend keywords to publishers for new datasets and populate missing keywords in existing datasets.

c) Contributions: We selected the European Data Portal (EDP) as our case study for this study. The EDP is regarded as one of Europe's most extensive data portals, aggregating data from 73 other data portals in the european area. According to the findability index published by EDP, keywords and categories of datasets contribute to 60% of the findability score [1]. The details of the index are illustrated in Fig 1. Focusing on the findability score, we analyzed 37,306 datasets in the portal. Among them, 11,958 (32%) had three or fewer keywords, while the remaining 25,348 (68%) datasets had more than three. However, a significant number of these datasets exhibited non-representative traits and flaws. The flaws observed include redundant keywords, column names used as keywords, other exact metadata used as keywords, title words, and numerical data as keywords. These findings indicate that many publishers may not have paid sufficient attention to the keywords, either entering them hastily or skipping them entirely. It rendered the findability of those datasets to be difficult.

As manual keyword provision proved insufficient, we proposed generating and recommending automated representative keywords to make data more findable. The most comprehensive metadata attribute for a dataset is its description, so we used this to extract keywords for the datasets. For our experiment, we collected 69,423 metadata records from 13 different themes within the EDP. After gathering the data, we cleaned it by removing duplicates, empty entries, garbage characters, non-representative data, and items that could not be translated.

In our study, each algorithm we used, including BERT, RAKE, YAKE, TEXTRANK, and ChatGPT, generated between 2 and 10 keywords from the dataset metadata. Our proposed hybrid methodology, BRYT, combined the results of the first four algorithms and employed cosine similarity to rescore and select the top 10 keywords from the generated list. To evaluate the effectiveness of our approach, we used Gestalt pattern matching and Jaccard similarity to score the matching with the original keywords. Our results showed that 69.1% of keywords matched majorly (more than 50% or 5 keywords) while 24.7% matched minorly (less than or equal to 50% or 5). Moreover, BRYT outperformed the other algorithms in major matches (27.1%), while YAKE had better results in minor matches (35.5%). Our study established that the proposed hybrid methodology was superior to other algorithms, particularly in major matches.

II. Results

Apart from employing NLP techniques, we also proposed a hybrid methodology, BRYT, which combined the output of four algorithms (BERT, YAKE, RAKE, TEXTRANK), recalculated their relevance using cosine similarity, and selected the

TABLE I
PERFORMANCE OF ALGORITHMS DURING MAJOR MATCHES IN EACH CATEGORY

Category/Theme	**BERT**	**YAKE**	**RAKE**	**TEXT RANK**	**CHATGPT**	**Hybrid**
Agriculture, fisheries, forestry and food	7	10	0	4	7	**15**
Economy and finance	72	42	11	5	44	**82**
Education, culture and sport	17	15	1	2	14	**32**
Energy	6	6	2	2	14	**17**
Environment	52	53	10	12	81	**93**
Government and public sector	36	34	10	7	38	**53**
Health	52	40	4	13	31	**64**
International issues	3	5	1	3	2	**9**
Justice, legal system and public safety	8	3	2	2	8	**14**
Population and society	64	33	5	7	32	**88**
Regions and cities	43	73	6	2	**110**	38
Science and technology	94	104	16	18	**157**	96
Transport	15	11	5	3	12	**20**

top 10 keywords from the rescored keywords. We employed gestalt pattern matching and jaccard similarity to score the matching with the original keywords. Our evaluation of the results from the lens of each algorithm, considering each category and varying lengths of description, reflected that our hybrid methodology outperformed other algorithms in having more representative keywords.

a) Data Analysis: Initially, our analysis focused on the structure of the cleaned data, which comprised 1393 datasets across 13 themes/categories, reduced from an original count of 69423 datasets. Each dataset had one or more than one theme/category associated with it. Apart from the categories, most datasets had description lengths between 500 and 3000, while a considerable number of datasets also had greater than 3000. Following the data analysis, we moved on to evaluating our methodology.

b) General vs. Algorithmic Matches: At the beginning of our evaluation, we first analyzed our general results, which reflected that 69.1% of instances were majorly matched using one of the algorithms. In comparison, 24.7% were minorly matched and 6.2% had no matches.

While comparing the major matches of each algorithm, we found that our proposed hybrid methodology BRYT performed better than other algorithms, with 27.1% of times having more efficient or equally efficient matches, while ChatGPT was a close second with 23.8%. This percentage also entails that there were instances where other algorithms matched equally, and in that case, it counted both algorithms as an efficient/winner algorithm in that particular instance.

c) Algorithmic Matches Across Categories: Moving further in our evaluation, we analyzed our results more deeply to gain more intuitive and meaningful insights. As previously discussed, there were 13 themes/categories that data in European Data Portal belonged to, we analyzed the number of instances majorly matched, minorly matched or did not match at all in each theme/category. During our analysis, we found that in most of the categories, there were heavily major matches except “Regions and Cities”. As we analyzed further, we found that the majority of descriptions in that category were more structured than textual. Most of them defined data in terms of column and value about some informative attributes, followed by minimal descriptive text. This implied that the studied algorithms worked better with more textual descriptions.

d) Algorithmic Perspectives: Matches Across Categories and Description Lengths: After our findings about major and minor matches, we scatter-plotted them separately according to description length, categories, and matches with each algorithm. First, we plotted the instances of major matches with respect to the description lengths in each theme/category considering all algorithms. As reflected in Fig 7, most major matches went up to the description length of 4000. Except for the category of “International Issues”, which had minimal data points as major matches, most were populated in a similar pattern and the majority of them had a dense population of matches up to a description length of 2500.

During our analysis, we found some insightful findings and correlations. They entailed that BRYT performed better in most of the major matching cases where the description was in a proper textual format while ChatGPT performed better in instances where there were minor matches even if the description was in a structured format. We also found that in the bigger picture description length had no effect on the efficiency of algorithms in terms of major matches. However, in minor matches, it was more probable for them to have a description under 2000 characters than longer. Additionally, we found that minor matches were minimal in other categories except for ”Regions and Cities” and ”Science and Technology”, where there were a lot, and ChatGPT flourished in them, while BRYT mostly dominated major matches. Our analysis entailed more confidence in our proposed methodology and its implication in European Data Portal.

REFERENCES

[1] EU. European data portal. https://data.europa.eu/en, 2023. Accessed: 2023-11-28.

[2] Nazanin Firoozeh, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. Keyword extraction: Issues and methods. *Natural Language Engineering*, 26(3):259–291, 2020.

[3] Zakariae Alami Merrouni, Bouchra Frikh, and Brahim Ouhbi. Automatic keyphrase extraction: An overview of the state of the art. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 306–313, 2016.