Comparative analysis of automatic keyword extraction algorithms: KeyBERT, YAKE and GPT-3

Luis Roberto Polo Bautista[1], Raquel Casique Vasquez[2]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación, Ciudad de México, México

[2] Instituto Politécnico Nacional, Escuela Nacional de Biblioteconomía y Archivonomía, Ciudad de México, México

lpolob2023@cic.ipn.mx, rcasiquev1900@alumno.ipn.mx

Abstract. In this paper, a comparative analysis of three algorithms used for automatic keyword extraction was performed: KeyBERT, YAKE and GPT-3. The main objective is to analyze the cosine similarity between the automatically extracted keywords and those assigned manually as part of the indexing performed by the authors of the articles. To carry out this comparative analysis, a Google Colaboratory environment was developed and a corpus of English article abstracts obtained from Scopus was used. The results revealed that GPT-3 obtained a higher cosine similarity, demonstrating efficiency in the automatic extraction of keywords and generating results similar to how a person would do it.

Keywords. Automatic keyword extraction, Cosine similarity, GPT-3.

1. Introduction

Among the lines of research in the field of Information Retrieval are the linguistic phenomena associated with the analysis of documentary content, both to organize information and to retrieve it [1]. With the increasing amount of data and information that currently exists in all digital media, the manual analysis of documents has become a complicated task. As a direct consequence, several solutions based on artificial intelligence and machine learning have emerged in recent years to facilitate these processes [2].

One of the most relevant solutions in this context is text mining. It consists of computationally analyzing the discourse of specific areas of knowledge to extract relevant information through various techniques, one of which is Natural Language Processing. This technique includes processes related to keyword identification, named entity recognition, document classification according to content, among others.

Natural Language Processing (NLP) is a branch of Artificial Intelligence (hereafter AI) focused on enabling computers to recognize human language. NLP-based systems are primarily designed to understand and interact with human speech and text. This tool incorporates numerous tasks, such as text translation, keyword extraction, topic modeling, among others. However, machine learning is required to automate these steps and provide more accurate results [3].

Machine learning is also a branch of AI, it allows a computer system to learn to perform a specific task from input data [4]. Machine learning has three approaches, i) supervised machine learning, it is based on providing the computer with enough training data previously labeled and structured by humans, to generate predictions and learn to execute specific tasks from the input data; ii) unsupervised machine learning, this approach does not need training data, as it is based on pattern recognition and identification of data structures automatically; iii) reinforcement

learning, it is based on learning through feedback from its performance.

In the context of Information Retrieval, one of the techniques of content analysis is indexing (also called indexing), this technique seeks to express the most significant information of a document through the assignment of descriptor terms and thus create a language of representation and mediation between a user and a document [5]. The conceptual representation of an object, a document, a discourse or an image facilitates its storage and retrieval in physical collections and digital environments available on the Internet. The representation is achieved through the retrievable entities of the documentary resource, the relationships between entities and with other documentary resources [6].

More specifically, indexing is the action of describing or identifying a document in relation to its content, by assigning the representation of concepts in the form of derived terms in natural language, preferably simple or compound nouns. During indexing, concepts are extracted from the document through a process of intellectual analysis and then transformed into indexing terms [7]. From the traditional point of view, this is a process that involves additional time and effort, taking into account that in some cases they must know or be familiar with the knowledge area of a document to identify and understand more accurately the vocabulary used [8].

Likewise, the indexing process can result in the assignment of erroneous descriptors that do not describe all or an essential part of the content of a document, making the information retrieval process difficult for a user [8]. However, through advances in AI and NLP, there are now tools that allow automating the document indexing process, enabling more efficient information management. In this way, the disadvantages of traditional indexing, such as the assignment of erroneous descriptors, among others, could be solved.

2. Related work

There have been several works that seek to evaluate and compare automatic keyword extraction systems. For example, in [9], the YAKE, TextRank, TF-IDF, SingleRank and RAKE algorithms were evaluated, taking into account metrics such as precision, recall and F1 score. In this study, the YAKE algorithm showed superior performance compared to the other models. In [10], TextRank, RAKE and PositionRank algorithms were compared using a corpus of research articles. The performance of the algorithms was evaluated in terms of their similarity to the keywords manually assigned by the article authors. In this case, the PositionRank algorithm performed better.

In the paper [11], a comparison between different algorithms based on TF-IDF and KEA was performed using a corpus of 100 English academic texts. The results showed that a modification of the TF-IDF algorithm generated superior performance, with an accuracy rate above 70%. In [12], the YAKE, TopicRank, MultipartiterRank, KPMiner, KEA and WINGNUS algorithms were compared using a specialized corpus in the field of electric double layer capacitors (EDLC). The evaluation metrics used were Jaccard similarity, Cosine and Cosine with vector. In this case, the MultipartiterRank algorithm showed better results in terms of similarity to the keywords provided by experts.

As can be seen, there are currently several tools and algorithms that allow us to extract keywords from documents, as well as metrics that help us evaluate the performance of these tools and analyze the efficiency and similarity of keywords. In this work, YAKE, KeyBERT and GPT-3 algorithms were implemented, since these models have shown to have an accuracy above other models and since language models perform NLP tasks, it was considered that it would be suitable for this process.

3. State of the art

This section describes the state of the art of the algorithms implemented for the comparative analysis.

3.1 KeyBERT

It is based on the semantic similarity of words in order to extract keywords from a document, this algorithm is constituted from the embeddings of the

BERT (Bidirectional Encoder Representations from Transformers) model [13]. BERT is a state-of-the-art language model built through the use of neural networks that were trained from large volumes of information and unstructured data (text).

The structure of the KeyBERT algorithm consists of three main parts: i) extraction of document embeddings using BERT; ii) extraction of word embeddings; iii) use of cosine similarity [13]. In the first and second steps, the libraries scikit-learn [14] and SentenceTranformers [15] are used to vectorize and extract the embeddings of the words that make up the document discourse, and to establish their semantic relationship.

Similarly, the maximum number of n-grams (tokens) that will constitute the keywords was established. The third step consisted of implementing cosine similarity by means of the scikit-learn library [14] to determine the keywords that best represent the content of a document.

3.2 YAKE

It is an automatic keyword extraction method that employs unsupervised machine learning, based on statistical textual features extracted from individual documents to select the most relevant keywords from a text. This model does not need to be trained on a specific set of documents, nor does it depend on dictionaries, external corpora, text size, language or domain knowledge. YAKE is based on local text features and statistical information, such as co-occurrence and term frequency. This model is very robust to linguistic or domain diversity and can be easily applied to large document collections and contexts. However, it can also be implemented on individual documents, this determines that the model can work independently of the existence of a corpus [1].

The YAKE algorithm consists of five main steps: i) text pre-processing and identification of candidate terms; ii) feature extraction; iii) term score calculation; iv) n-gram generation and candidate keyword score calculation; and v) data de-duplication and classification [1]. This algorithm is a powerful solution for automatic keyword extraction based on unsupervised machine learning, as it offers great flexibility and customization for different needs and contexts.

3.3 GPT-3

GPT-3 (Generative Pre-trained Transformer 3) is a natural language model developed by the company OpenAI. It is a deep learning neural network that has been trained on a large amount of linguistic data and can generate coherent, natural text in a variety of natural language tasks. GPT-3 uses a transformer language model architecture, which allows it to process variable-length text sequences and generate continuous, coherent text in response to text input.

The model is remarkable for its size, with over 175 billion parameters, making it one of the largest and most complex language models in existence. GPT-3 has been used in a variety of natural language applications, such as machine translation, sentiment analysis, text generation, question answering and natural language understanding. It has proven to be a powerful tool for content creation and automation of complex linguistic tasks.

KeyBERT and YAKE algorithms are specific tools for document keyword identification and extraction, both tools are based on different approaches, both semantic similarity and probability. On the other hand, GPT-3 is a language model that can be used in multiple NLP tasks, such as translation, text generation and keyword extraction. Thus, the GTP-3 model was used to build a keyword-to-document extraction algorithm so that it can be compared with other existing solutions and analyze the performance of each based on cosine similarity.

4. Methodology

In this article, the KeyBERT, YAKE and GPT-3 algorithms were used to extract keywords from 200 abstracts of English-language articles retrieved from Scopus. The similarity between the automatically extracted keywords and the keywords indexed by the authors was evaluated using a comparative analysis based on cosine similarity. This analysis was performed using the Google Colaboratory environment and the Python programming language version 3.10.12.

A procedure was carried out in Google Colaboratory to develop the environment and carry

out the implementation of the algorithms. The procedure consisted of the following stages: i) Environment configuration, ii) Corpus compilation, iii) Corpus pre-processing, iv) Keyword extraction and v) Cosine similarity analysis. Each of these stages was essential to achieve an accurate analysis result.

4.1 Compilation of the corpus

The Scopus database was used to obtain a corpus of abstracts of scientific articles. The search was performed using the Query: TITLE-ABS-KEY (CENG OR CHEM OR COMP OR EART OR EART OR ENER OR ENGI OR ENVI OR MATE OR MATH OR PHYS) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (SRCTYPE, "j")). Documents were selected that included metadata associated with the title, abstract, and keywords. The search was restricted to the subject category of Physical Sciences, which includes subdisciplines such as Chemical Engineering, Chemistry, Computer Science, Energy and Engineering, among others. The search was performed on April 20, 2023 and 240,765 documents were obtained, from which the top 200 were selected and exported in a CSV file.

4.2 Pre-processing of the corpus

This activity is also called text cleaning, and corresponds to the elimination of special characters from the summaries of the corpus articles, such as punctuation marks, the conversion of all letters to lowercase and the elimination of irrelevant words, such as articles and prepositions. Then, a normalization process was carried out to reduce linguistic variability in the text, such as the elimination of verb endings or the conversion of words to singular or plural.

4.3 Keyword extraction

KeyBERT, YAKE and GPT-3 algorithms were used to extract keywords from article abstracts. These algorithms use different approaches to identify and extract keywords from a text. KeyBERT uses a Transformers-based approach, which is based on a neural network model using deep learning, while YAKE uses a probabilistic approach through word frequency and GPT-3 uses a large pre-trained language model-based approach to execute various NLP tasks. Each algorithm was applied to 200 article abstracts from a corpus to obtain a set of keywords for each.

4.4 Cosine similarity analysis

After obtaining the keywords from the article abstracts using the KeyBERT, YAKE and GPT-3 algorithms, they were compared with the keywords manually indexed by the article authors. The comparison was performed using a cosine similarity measure that allows you to assess the similarity between keywords. The cosine similarity varies between -1 and 1, where 1 indicates that two vectors are identical, 0 indicates that the vectors are completely different and independent of each other, and -1 indicates that the vectors are opposite to each other. The purpose of this analysis was to evaluate the accuracy of the keyword extraction algorithms compared to manually assigned keywords.

5. Results

This section presents the results obtained from applying the methodology presented in section 4. Table 1 shows a fragment of the cosine similarity results applied to each algorithm in relation to the first 19 article summaries.

Table 1. Cosine similarity applied to each algorithm.

No	YAKE	KeyBERT	GPT-3
1	0.223124462	0.05069 8626	0.125 872968
2	0.133 520996	0	0.374 470458
3	0.1196 61972	0	0.615 15212
4	0.177 216362	0.24905 8916	0.133 594251
5	0.066 709976	0	0.137 510644
6	0.561 033873	0.19670 3616	0.549 660602

7	0.29123131	0.124507166	0.263046631
8	0.610702778	0.067161179	0.09506179
9	0.113537637	0	0.300534068
10	0.035157978	0	0.31658844
11	0.599347079	0.685103371	0.307297128
12	0.307245645	0.333603401	0.492890119
13	0.335761693	0.214604538	0.404674941
14	0.451727746	0.14253183	0.401228131
15	0.374719869	0.359060964	0.406974901
16	0.485467044	0.384293939	0.350103347
17	0.159335771	0	0.169116577
18	0.219692048	0.231691219	0.536984666
19	0.155862407	0.11399487	0.429507127

Table 1 shows the results of the cosine analysis corresponding to 19 of the 175 total article abstracts. Originally the corpus consisted of 200 articles, however, 25 were omitted since they did not have abstracts and therefore no keywords could be identified automatically. On the other hand, the average cosine similarity of the algorithms is shown in Table 2.

Table 2. Average cosine similarity of each algorithm

YAKE	KeyBERT	GPT-3
0.2319241080152155	0.20999500988117406	0.3055814 16446062

Similarly, Fig. 4, 5 and 6 visually show the cosine similarity calculations associated with the 175 keywords.

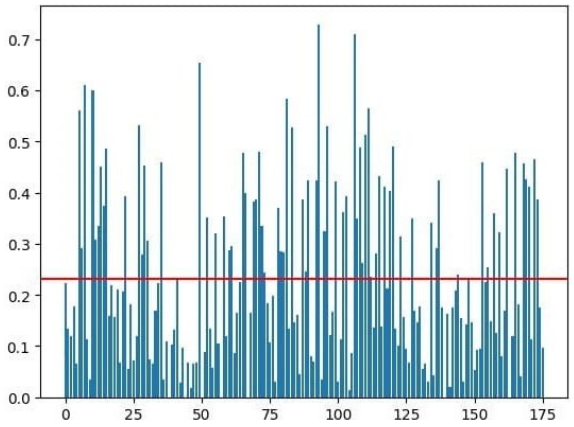

Fig. 4. Average cosine similarity YAKE

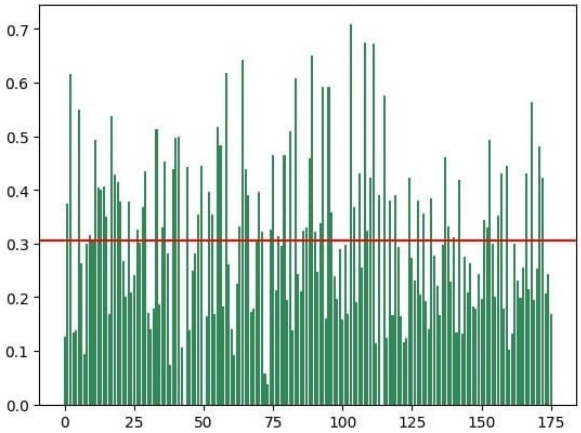

Fig. 5. Average cosine similarity GPT-3

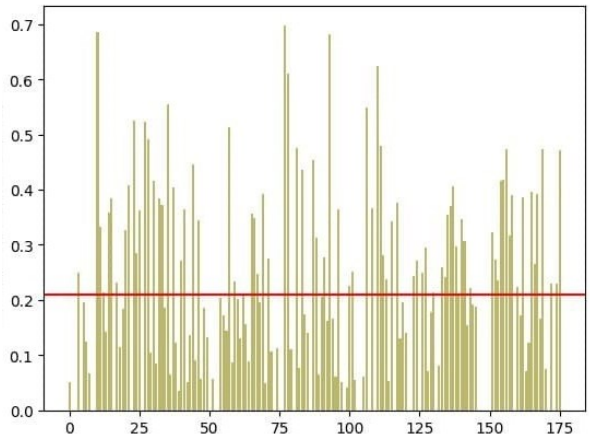

Fig. 6. Average KeyBERT cosine similarity

As mentioned above, cosine similarity varies between -1 and 1, where 1 indicates that two vectors are identical, 0 indicates that the vectors are completely different and independent of each other, and -1 indicates that the vectors are opposite to each other. Plots 4, 5 and 6 show the distribution of similarity between automatically extracted and manually indexed keywords, this distribution helps to represent the dynamics of growth and decline between cosine similarity metrics. The red line drawn on each graph represents its overall average similarity.

The GPT-3 algorithm has a higher similarity, indicating a significant similarity between both sets of keywords. Similarly, this algorithm has 82 results above its overall average, which is equivalent to 46.85% of the total. YAKE has 73 results above its overall average, equivalent to 41.71% of the total. KeyBERT has 81 results above its overall average, which equals 46.28%.

These results suggest that all algorithms have similar capabilities to identify and extract relevant keywords from text, but the GPT-3 algorithm has an advantage compared to YAKE and KeyBERT. In addition, all algorithms produce a significant number of results above the overall average. This suggests that relevant keywords can be identified with high accuracy in different contexts.

6. Conclusions and future work

The most recent studies on keyword extraction from text use models based on Transforms, which is based on a neural network model using deep learning and large pre-trained language models to be used for a large natural language processing task. Generally, these tools tend to perform better due to their ability to learn and recognize complex patterns in the training data. In this case, the most accurate algorithm based on cosine similarity was GPT-3 (0.305581416446062), followed by YAKE (0.2319241080152155) and KeyBERT (0.20999500988117406) based on their overall average similarity.

The use of these algorithms for keyword extraction can be diversified to various areas of knowledge, allowing the implementation of these models to automate the processes of information organization, in order to facilitate its management and retrieval.

In general terms, KeyBERT, YAKE and GPT-3 algorithms are an excellent option to identify and extract keywords from a document or a collection of documents, since they have similar capabilities and based on their similar average, there is no important differentiation to select a specific one, rather this would correspond to different characteristics, application area, language, extension, among others.

Future work can explore the improvement and adaptation of these algorithms to address specific challenges in different domains. For example, research can investigate how to optimize the algorithms to extract keywords from academic, scientific or literature texts, where the vocabulary and language structure may be more specialized.

Also, the combination of multiple algorithms can be considered to obtain more accurate and robust keyword extraction results. Combining different approaches and techniques can help mitigate the individual limitations of each algorithm and improve the quality of the generated keywords. In summary, the use of keyword extraction algorithms such as YAKE, KeyBERT and GPT-3 offers many opportunities in the field of information management. With the continuous development and improvement of these algorithms, it is expected that the automation of keyword extraction will become more accurate and efficient, which will contribute to better organization and retrieval of information in various areas of knowledge.

References

1. **Peña, G. A., y Peña, C. N. (2015).** Extracción de candidatos a términos de un corpus de la lengua general. Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información, 29 (67), 19–45. https://doi.org/10.1016/j.ibbai.2016.02.035
2. **Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., y Jatowt, A. (2018).** A Text Feature Based Automatic Keyword Extraction Method for Single Documents. En H. A. Pasi G., Piwowarski B., Azzopardi L. (Ed.), Advances in Information Retrieval. ECIR 2018. Lecture Notes in Computer Science (pp. 684–691). Springer International Publishing AG, part of Springer

Nature. https://doi.org/10.1007/978-3-319-76941-7_63.

3. **Yerkhassym, A., Pak, A. A., Akhmetov, I., Yelenov, A., y Gelbukh, A. (2022).** On Causality Problem in Natural Language Processing Field. Computacion y Sistemas, 26 (4), 1549-1556. https://doi.org/10.13053/CyS-26-4-4434
4. **Hurwitz, J., y Kirsch, D. (2018).** Machine Learning For Dummies. IBM Limited Edition. https://www.ibm.com/downloads/cas/GB8ZMQZ3
5. **Urbizagástegui Alvarado, R., y Restrepo Arango, C. (2011).** La ley de Zipf y el punto de transición de Goffman en la indización automática Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información, 25 (54), 71–92. https://doi.org/10.22201/iibi.0187358xp.2011.54.27482
6. **Naumis Peña, C. (2011).** La indización en la red semántica: una solución interdisciplinaria. Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información, 25 (54), 7–14. http://rev-ib.unam.mx/ib/index.php/ib/article/view/27478/25465
7. **Asociación Española de Normalización y Certificación. (1991).** Métodos para el análisis de documentos, determinación de su contenido y selección de los términos de indización. https://nanopdf.com/download/metodos-para-el-analisis-de-documentos-de-terminacion-de-su_pdf
8. **Polo Bautista, L. R., y Martínez Acevedo, K. V. (2021).** Algoritmo para el análisis temático de documentos digitales. Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información, 35 (89), 13–31. https://doi.org/http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419
9. **Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., y Jatowt, A. (2018).** A Text Feature Based Automatic Keyword Extraction Method for Single Documents. En H. A. Pasi G., Piwowarski B., Azzopardi L. (Ed.), Advances in Information Retrieval. ECIR 2018. Lecture Notes in Computer Science (pp. 684–691). Springer International Publishing AG, part of Springer Nature. https://doi.org/10.1007/978-3-319-76941-7_63
10. **M. G. Thushara, T. Mownika y R. Mangamuru. (2019).** A Comparative Study on different Keyword Extraction Algorithms. Trabajo presentado en 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 969-973: 10.1109/ICCMC.2019.8819630.
11. **Li, Jinye. (2021).** A comparative study of keyword extraction algorithms for English texts. Journal of Intelligent Systems. 30. 808-815. 10.1515/jisys-2021-0040.
12. **M. Saef Ullah Miah, Junaida Sulaiman, Talha Bin Sarwar, Kamal Z. Zamli, Rajan Jose. (2021).** Study of Keyword Extraction Techniques for Electric Double-Layer Capacitor Domain Using Text Similarity Indexes: An Experimental Analysis, Complexity. p. 12. https://doi.org/10.1155/2021/8192320
13. **Grootendorst, M. (2020).** Keyword Extraction with BERT: a minimal method for extracting keywords and keyphrases. Zendo. https://doi.org/10.5281 / zenodo.4461265
14. **McKi**nney, W. (2010). Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a
15. **Reimers, N., y Gurevych, I. (2019).** Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. https://arxiv.org/abs/1908.10084