INTERNATIONAL DESIGN CONFERENCE – DESIGN 2024
https://doi.org/10.1017/pds.2024.199

Assessing text-image patent datasets with text-based metrics for engineering design applications

Marco Consoloni [1,2,✉], Vito Giordano [1,2] and Gualtiero Fantoni [1,2]

[1] University of Pisa, Italy, [2] Business Engineering for Data Science Lab (B4DS), Italy

✉ marco.consoloni@phd.unipi.it

Abstract

Images provide concise representations of design artifacts and emerge as the primary mode of communication among innovators, engineers, and designers. The advanced of Artificial Intelligence tools which integrates image and textual information can significantly support the Engineering Design process. In this paper we create 5 different datasets combining both images and text of patents and we develop a set of text-based metrics to assess the quality of text for multimodal applications. Finally, we discuss the challenges arising in the development of multimodal patent datasets.

Keywords: engineering design, product development, natural language processing (NLP), image data, patent analysis

1. Introduction

In recent years, the field of Artificial Intelligence (AI) has witnessed significant advancements in the realm of text-to-image AI. Modern text-to-image systems such as Midjourney, DALL-E 2, and Stable Diffusion take a textual description as input and generate images conveying concepts identifiable in the text. These systems are a subset of multimodal AI systems. In fact, unlike traditional methods that depend exclusively on a single type of data source (unimodal data), such as Natural Language Processing (NLP) which process only text, multimodal AI systems are developed to harness information from various types of data.

Integrating visual information with textual information to bridge the gap between visual perception and language understanding can significantly support the Engineering Design (ED) process. In fact, images provide synthetic representation of design artifact/process, and they emerge as the primary mode of communications among innovators, engineers, and designers throughout the ED phases (Jiang et al., 2022). The integration of patent's textual and visual information presents a promising opportunity to the ED process. This integration can significantly contribute to (1) supporting creative thinking and concept generation with design stimuli (Jiang et al., 2021); (2) developing a more systemic understanding of design artifacts (Jiang et al., 2022); (3) improving the performance of patent retrieval and prior art mapping tasks, thereby facilitating designers' access to a broader range of existing solutions which mitigate the risk of design fixation (Atherton et al., 2018).

In the context of patent documents, many international patent offices provide free access to their patent databases organizing in only one structured database a huge amount of technical information from around the globe. Researchers and practitioners are taking advantage of the rapid growth of these open-access patent databases by using NLP techniques as a tool to effectively extract textual information from patents (Chiarello et al., 2021). However, existing approaches have overlooked the potential of analysing patent images in conjunction with textual data to support engineer, and innovators in knowledge-intensive design activities (Jiang et al., 2022). In ED literature, there are different works

combining text and image. Lin et al., (2023) extract Subject-Action-Object structures from patent text and use them in combination with front-page patent images to search for similar patent. Jee et al., (2022) use keywords extracted from patent drawings depicting block diagrams to measure patent similarity. Pustu et al., (2022), develop a multimodal system which combines patent images and text extracted from images to enhance patent retrieval task. Vrochidis et al., (2010) combine keywords automatically extracted from sentences describing patent drawings in order to support patent retrieval task. Luo et al. (2023) combine, patent titles, abstract and bibliometric data with patent drawings to classify high-value patent documents. Jiang et al., (2022) develop a multimodal system to predict IPC-classes of patents based on patent titles, abstract and drawings. Vrochidis et al., (2012) combine figure descriptions manually associated to patent images and train detectors, which can identify global concepts in patent figures in order to classify patent documents according to those global concepts.
Most literature on the applications of multimodality in patents focuses on tasks of patent retrieval and patent classification. Moreover, all these studies focus on optimizing the performance of the proposed multimodal systems, without investigating the quality of data. When data is studied, it is mostly to look at the effect of data size on the system performance (Kucer et al., 2023), rather than the quality of the data (Rao et al., 2020). To the best of our knowledge no scientific approaches have been developed to assess the quality of image-text alignment within the context of multimodal patent datasets. Nonetheless, investigating data quality is essential to train multimodal models with less data and less computational resources (Giordano et al., 2023). Moreover, investigating data quality can improve the interpretability of the output generated by multimodal systems and Large Language Models (LLMs), thereby increasing their transparency (Arrieta et al., 2019). As opposed to background literature, this work proposes a first attempt to 1) explore different strategies to combine textual and visual information contained in patent documents, 2) assess the quality of different multimodal patent datasets using text-based metrics and 3) discuss potential implications of data quality in multimodal applications in the field of ED.

2. Methodology

The methodology we adopted is structured in three phases. First, we retrieved full text and images of patent documents from the patent database of the United States Patent and Trademark Office (USPTO). Second, we created five different multimodal patent datasets by aligning different textual information and patent images into (text, image) pairs. Finally, we defined and compute a set of text-based metrics across the generated multimodal patent datasets to highlight potential linguistic and content-related differences. Future studies will focus on exploring metrics for patent image analysis.

2.1. Data collection

We collected 6,715 utility patents from the archive dated 03.01.2023 of the USPTO Bulk Storage System. The archive contains for each patent: 1) the patent full text stored in an XML file and 2) the single-page patent images stored in TIF files. The XML files are structured in a hierarchical tree-like structure of XML nodes designed to store the patent full text. The TIF files include the patent front image, and the patent drawings. Patents without associated images were excluded since cannot be used to create a multimodal dataset, resulting in a final set of 6,013 patent documents.

2.2. Dataset creation

Patent documents consists of three primary types of data: 1) bibliometric information, 3) textual information and 3) visual information (or patent images).
Bibliometric information includes patent metadata such as applicants, inventors, assignee, publication date and International or Cooperative Patent Classifications (IPC/CPC).
Textual information includes the patent title, abstract, claims and description. The title provides a concise and descriptive name for the invention. The abstract provides a summary of the invention, highlighting the key features of the disclosure. The claims define the scope of the disclosure, enumerating the specific features/elements for which protection is sought. The description section typically contains two main sub-sections known as "Detail Description" and "Brief Description of the

Drawing". The former, provides detailed information about any specific embodiments of the invention, aiming to enable someone skilled in the relevant field to replicate the invention based on the provided information. The latter provides a brief explanation of each accompanying patent images. The description section typically includes two additional sections: the "background of the invention" and "cross-reference to related applications". These sections are excluded from our analysis since they are primarily focused on establishing connections with prior art and other patent documents, rather than providing information about patented inventions. Moreover, they do not explicitly reference to patent drawings and cannot be used to create a multimodal dataset.

Visual information consists of the front image and patent drawings. The former is the image situated on the cover page of patent documents and it is regarded as the representative image of the invention. The latter provides visual representations of the invention, such as technical drawings, block diagrams, and graphs, each distinguished by a unique ID.

A multimodal patent dataset consists of textual and visual information combined in (text, image) pairs. In creating such dataset, different combinations of the textual and visual information can be used, resulting in a diverse set of multimodal patent datasets. As shown in Figure 1, we adopted five different strategies of text-image alignment to create five different multimodal patent datasets.

The datasets A, B, and C are created using the invention title, the abstract, and the claims of patents. These textual elements are aligned with patent front images as they do not explicitly reference patent drawings, providing a natural synergy to be aligned with patent front images (Liu et al. 2020).

The datasets D and E are created using the "Brief Description of the Drawing " and the "Detailed description". These sections are structured in distinct paragraphs containing explicit reference to patent drawings. Therefore, we extract all the paragraphs from the "Brief Description of the Drawing" section and only those paragraphs from the "Detailed description" section which contain at least one figure reference. In this paper, a paragraph is a unit of text separated by line breaks. Using the XML nodes, we extract the unique figure references contained in the paragraphs of these section disregarding literal sub-levels and we associate each paragraph to the corresponding patent drawings. For instance, the paragraph *"FIG 4A and 4 B are front views of the tuft spike of FIG 2 shown adjacent a receptacle of the brush assembly of FIG 3, respectively."* (US11540621) refers to the patent drawings 4A, 4B, 2 and 3. Therefore, after dropping the literal sub-levels *"*A*"* and *"B"*, we associate this textual element with the patent drawings 4, 2 and 3, resulting in three distinct (text, image) pairs.

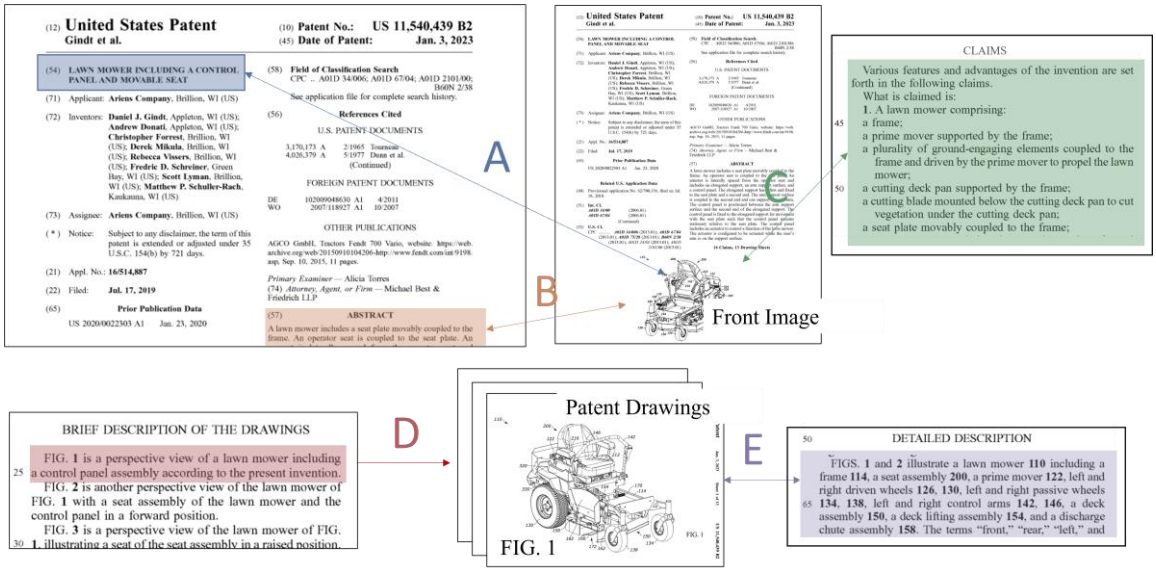

Figure 1. Text-image alignment strategies to create dataset A-E

2.3. Dataset analysis

We compute a set of text-based metrics to evaluate the linguistic (syntactic) and content-related (semantic) aspects of the textual elements of datasets A-E (refer to Figure 1). The goal of the metrics

is three-fold: 1) provide descriptive analytics of semantic and syntactic aspects of patents' textual elements, 2) identify textual elements rich in semantic information which can enhance the performance of multimodal models and 3) highlight structural differences among the multimodal datasets, which can impact the dataset's applicability.
Table 1 provides a descriptive summary of the metrics, detailing for each metric the designation name (Metric), what it measures (Objective) and the linguistic or content-related focus of the metric (Class).
The metric *"N. of words"* counts the total number of words in a textual element and it measures its brevity. The higher the metric, the longer the textual element is. To compute this metric, we consider as a word any alphanumeric character including underscores, hyphens and slashes. These three special characters are frequently used in word-tuples, such as *"multi_sensor", "pre-heating" and "AC/DC"*, which can be regarded as single terms.

Table 1. Descriptive summary of the metrics

#	Metric	Objective	Class
1	N. of words	text brevity	Linguistic
2	Avg. N. words per sent.	syntactic complexity	Linguistic
3	% of stopwords	portion of non-informative content	Linguistic
4	% of duplicated words	portion of repetitive content	Linguistic
5	% of components	focus on the architecture of patented devices	Content-related
6	% of functional verbs	focus on the functioning of patented devices	Content-related
7	Avg. TF-IDF	significance within the entire patent document	Content-related
8	N. of fig. references	degree of consistency in text-image alignment	Content-related

The metric *"Avg N. words per sent."* is computed as the average number of words per sentences within a given textual element. This metric aims to estimate the syntactic complexity of the textual elements. High values of this metric indicate long sentences, which can potentially compromise clarity and readability. Sentences are split on periods while avoiding splitting on decimal numbers with a regex pattern. To avoid sentence splitting errors, before preforming sentence splitting, we removed periods from figure references (e.g., FIG., FIGS.) and from common acronyms (e.g., "et al.", "i.e."). Instead of conventional sentence splitting tools offered by NLTK and spaCy libraries, a regex pattern is utilized since these commercial tools are not specifically designed for patent documents, leading to sentence splitting errors when applied to the complex technical-juridical jargon of patents.
The metric *"% of stopwords"* calculates the percentage of stop words in a textual element, aiming to estimate the portion of non-informative words. A higher value for this metric indicates a greater presence of words that do not significantly contribute to the core technical meaning of the textual element. We used the comprehensive stop word list developed for patent documents by Sarica and Luo (2020). The list consists of 365 terms drawn from the NLTK English stop words (179), the USPTO stop words (99) and the custom technical stop words developed by the authors (87).
The metric *"% of duplicated words"* measures the percentage of duplicated words in a textual element, aiming to estimate the portion of repetitive content. For instance, the abstract *" Various embodiments of exchanges are described. Methods and other embodiments are also described."* (US11541153) contains 3 words ("embodiments", "are", "described") duplicated 2 times, and it consists of 13 words. Therefore, the metric value is calculated as (3*2) / 13 = 46.15%.
The metric *"% of components"* is computed by dividing the number of components mentioned in a textual element by its total number of words. It aims to estimate the extent to which a textual element describes the architecture of patented devices. When a textual element has a higher percentage of components, it is likely to delve into structural aspects of a device, leading to an elevated value for this metric. For instance, the text *"Referring to FIG 101-102 and 105-106, a tibial anchor guide 506 may include a housing 508, a shaft 510, and a pin 512."* (US11540928) contains 4 components ("tibial anchor guide", " housing " " shaft", " pin"), and it consists of 23 words. Therefore, the metric value is calculated as 5 / 23 = 21.7%. To map component names in the textual elements we use component

reference numbers, such as 506, 508, 510 and 512 in the previous example, which are explicitly designated in patent documents to ensure the unique identification of components.
The metric *"% of functional verbs"* is computed by dividing the N. of functional verbs mentioned in the textual element by its total N. of words. To map the functional verbs, we developed a list of 6,940 verb lemmas extracted from the TechNet list of words developed by (Sarica et al., 2020). The functional verb lemmas of the list are identified in the text using the spaCy POS-tagging. This metric aims to estimate the extent to which the text describes the functioning of patented devices. Essentially, a higher usage of functional verbs in a text indicates a more detailed description of the device's dynamics, resulting in an elevated value for this metric.
The metric *"Avg. TF-IDF"* is computed as the Average Term Frequency-Inverse Document Frequency (TF-IDF) of the words contained in a textual element. The TF-IDF scores for individual words are computed against the entire corpus of the patent documents' full-text. A high value of this metric suggests high relevance of a textual element's content in comparison to the content of the associated patent document.
The metric *"N. of fig. reference"* count the number of unique figure references contained in a textual element. The metric disregards literals sub-levels, which can be occasionally used in patent documents to denote multiple figures. For example, in the text, *"FIG 5A-11B are graphs and pressure loads…"* (US11540588), the literal sub-levels *"A"* and *"B"* are dropped, resulting in the figure reference *"5A-11B"* being simplified to *"5-11"*. Consequently, the metric is assigned the value of 7 as the textual element refers to figures 5, 6, 7, 8, 9, 10 and 11. The metric aims to estimate the level of ambiguity of a textual element used in a (text, image) pair. A high value for this metric indicates that a textual element refers to multiple patent drawings, resulting in a greater difficulty in establishing a one-to-one association between the text and the images.

3. Results and discussions

3.1. Dataset overview

Table 2 provides a descriptive summary of the key attributes pertaining to datasets A-E, detailing for each dataset the designation name of the dataset (Dataset), the number of textual elements (N. Text), the number of images (N. Imgs), the number of text-image pairs (N. Pairs), the number of sentences within the textual elements of the dataset (N. Sentences), the total number of words of the entire dataset (N. Words) and the tallied number of unique words (N. Unq. Words). The metrics "N. Text", "N. Imgs" and "N. Pairs" provide an estimation of the size of the datasets, whereas the metrics "N. Sentences", "N. Words " and "N. Unique Words" provide high-level information regarding the textual structure of the datasets. Table 2 clearly shows that the type of text-image alignment significantly influences the size of the dataset, which is notably smaller for datasets A-C in comparison to datasets D-E. This is particularly relevant for multimodal dataset, given the swift escalation in storage and processing cost for images. Moreover, Table 2 indicates that the textual elements present very different granularity (N. Sentences), linguistic richness (N. Words) and linguistic variety (N. Unq. Words). Researchers and practitioners can use this information to make informed decisions about dataset suitability for specific scientific and practical investigations.

Table 2. Descriptive summary of the multimodal patent datasets

#	Dataset	N. Text	N. Imgs	N. Pairs	N. Sentences	N. Words	N. Unq. Words
A	Title + Front Image	6,013	6,013	6,013	6,013	48,557	6,350
B	Abstract + Front Image	6,013	6,013	6,013	21,125	695,053	15,638
C	Claims + Front Image	6,013	6,013	6,013	100,371	7,080,247	27,923
D	Brief. Desc. of Drawgs. Parghs. + Patent Drawgs.	77,882	77,567	98,177	80,456	1,756,302	19,737
E	Detailed Desc. Parghs. + Patent Drawgs.	141,617	77,567	239,047	599,928	17,829,100	81,784

3.2. Qualitative discussion of the metrics

We analyse the metrics' distributions across datasets A-E. Figure 2 shows the box plots of the metrics 1-8 for the dataset A-E. Based on Figure 2, we provide the following examples to show how the metrics can effectively capture linguistic and content-related aspects of the textual elements under analysis.

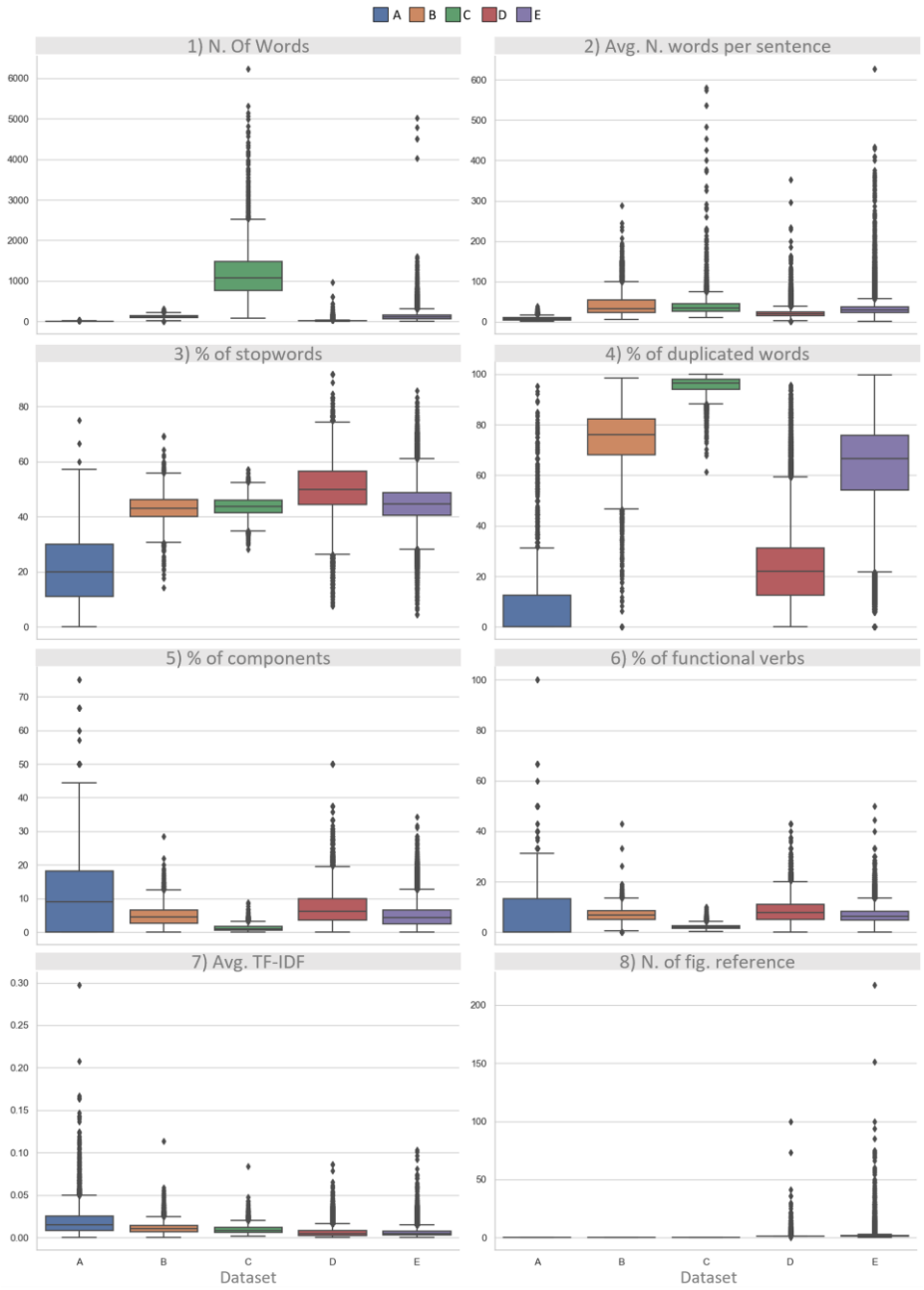

Figure 2. Box plots of the metrics 1-8 for dataset A-E

As for the metric *"N. of words"*, it provides quantitative estimation of the brevity of a textual element as in the case of the following patent titles: *"Electric glue gun"* (US11541417) and *"Glue applying mechanism of edge banding machine for applying glue to workpiece having oblique surface and edge banding machine using the glue applying mechanism"* (US11541415). Despite both titles describing a glue application system, they have different lengths, with 3 words for the former while 24 words for the latter. Notably, the first patent title is more concise, while the second one provides a more detailed description.

Concerning the metrics *"% of stopwords"* and *"% of duplicates"*, they serve as indicators of the portion of non-informative and repetitive content of a textual element, thereby reflecting its semantic richness. For example, high values of these metrics are observed in abstracts such as *"Various embodiments of exchanges are described. Methods and other embodiments are also described."* (US1154479) which presents 69. 23% of stopwords and 46.15% of duplicates and, similarly, *"The present invention relates to a door system for a refrigeration device and a refrigeration device with such a door system."* (US11543171) which exhibits 57.14% of stopwords and 57.14% of duplicates. Notably, both abstracts contained low level of technical content and limited substantive information, offering low "value" for further multimodal analyses. On the contrary, low values of these metrics can be observed in abstract such as *"A wellbore casing self-tightening tubular gripping device." (US11542761)* which present 14.29% of stopwords and 0% duplicates and, *"Superhydrophobic coatings to reduce deposit formation of diesel exhaust fluid (DEF) within selective catalytic reduction (SCR) systems."* (US11541380) which exhibits 17.65% of stopwords and 0% of duplicates. Both these abstracts contain high level of technical content and can be high-quality candidates for a multimodal patent dataset.

As for the metric *"Avg N. words per sentence"*, it provides a proxy of the syntactic complexity of a textual element, as shown by the comparison of the following two abstracts. The first one*:* "*A support module for a platform comprises a body and a lower surface. The body defines an opening configured to receive a pallet support. The lower surface is configured to abut a top deck of the platform."* (US11542062) is composed by 37 words and it features three concise sentences with an average of 12.67 number of words per sentence. Conversely, the abstract: *"Systems and methods that enable a user, such as a player, to manage persistent data, such as game state data, remote from any electronic gaming machine, such as via a mobile device executing a mobile device."* (US11544995) is composed by a single sentence of 37 words with an average of 37 words per sentence. This discrepancy in the syntactic structure underscores the diverse linguistic complexity exhibited by these abstracts.

Regarding the metrics *"% of components"* and *"% of functional verbs"*, they serve as indicators of a textual element's focus on describing structural/architectural and dynamic/behavioural aspects of patented devices, respectively. A higher metric value corresponds to a greater focus on the respective aspect. For instance, the paragraph from the description section *"As illustrated in FIG 19, server 20 includes communication unit 201, controller 202, and storage 203."* (US11546429) has 25% of components and 6.25% of functional verbs. Notably, it lists some of the device components without describing their operational details. On the contrary, the paragraph *"FIG 27 depicts typical operations performed when an S0 error occurs."* (US11544132) have 0% of components and 28% of functional verb*s.* In fact, it briefly describes the system's functioning in response to a specific error occurrence.

Concerning the metric *"Avg. TF-IDF"*, it provides a synthetic indicator of the significance of a textual element within the overall content of the corresponding patent document, as shown by the following two paragraphs. The former, *"FIG 51 is a detailed view of a main part of FIG 49."* (US11543095) yields an average TF-IDF score of 0.000 and it lacks meaningful technical information. Conversely, *"FIG 3, represents conditions for the isomerization of Δ9-THC to Δ10-THC."* (US11542243*)* has an average TF-IFD score 0.161, referring to the chemical process of isomerization and THC molecules, which are understandably crucial aspects for the patent's subject matter.

As for the metric *"N. of fig. reference"*, it captures the level of ambiguity in the alignment between text and images. As a matter of fact, text within patent documents contain either single or multiple figure references. For instance, in *"FIG 20A and 20B are front and back views of the bladder and enclosure of FIG 1-14"* (US11540615) reference is made to 15 distinct drawings, whereas *"FIG 3 is sectional view of the lift detection arrangement in a normal position."* (US11540440) refers to one specific drawing. This highlight that different degree of consistency in text-image alignment can be possible in a multimodal patent dataset.

The provided examples are meant to offer qualitative evidence that textual elements convey technical content heterogeneously from a linguistic and content-related perspective. This heterogeneity might significantly influence the performance of multimodal applications. For instance, our metrics can be leveraged to create stratified dataset of different "quality" to test the performance of multimodal models with different input data. Moreover, the metrics "% of components" and "% of functional verbs" can be exploited to create two different multimodal datasets with different scope. The first one

focuses on components, which can be used to automatically generate bill of materials which incorporate spatial relationships among components extracted form patent images. The second dataset, focuses on functional interfaces among components for use in abstract mapping of a device's functioning, providing beneficial for the divergent activities of the engineering design process (e.g. concept generation).
To validate our qualitative analysis, we analysed the correlation among the metrics to unveil potential positive and negative relationships within the metric set. Correlation matrix reveals that max correlation values is 0.504 and the average correlation value is -0.022, indicating the absence of robust relationships within the set of metrics. Hence, we argue that the metrics that we have defined are non-overlapping and distinctly capture heterogeneous facets of the semantic content expressed by the textual elements of dataset A-E.

3.3. Quantitative discussion of the metrics

We provide a comparative analysis of the five multimodal patent datasets by analysing the distributions of the metrics 1-8 across datasets A-E, as shown in Figure 2. Moreover, we perform the "Mann-Whitney U" statistic test to test whether the metrics' distributions are statistically different across datasets A-E.
In Figure 2, it is evident that dataset C exhibits the highest mean and standard deviation values for the metric "Number of Words," amounting to 1,177.490 and 599,879, respectively. This observation suggests that patent claims within dataset C consist of a greater number of words and display significant variations in length when compared to those textual elements found in the other datasets.
Observing Figure 2**,** dataset D has the highest mean values for the metric *"%. of stopwords"* (50.205), indicating a prevalence of non-informative content compared to the other datasets. However, dataset D yields the second-highest values for the metric *"% of components"* (6.751) and the highest values for the metric *"% of functional verbs"* (8.064), suggesting the presence of valuable technical content. Although this may seem a contradiction at the first sight, the textual elements within dataset D, such as *"FIG 4 shows side views for embodiments of the hybrid frame sleeve case."* (US11540602) and *"FIGS 2A-2C illustrate some representations of a body-surface-mountable fiducial patch in accordance with some embodiments of the invention."* (US11540767), share a common syntactical structure which explains the anomaly. This structure comprises: 1) reference keywords like "FIG" or "FIGS" in combination with general keywords/expression such as "embodiments", "invention" "in accordance with"; 2) a visual verb such as "show" and "illustrate" and 3) one or more component names such as "hybrid frame sleeve case" and " body-surface-mountable fiducial patch ". Hence, paragraphs of dataset D have understandably high values for the metric *"%. of stopwords"* due to the use of stop words keywords (1) and component names (2) in a concise descriptive text as demonstrated by the low mean values for the metrics *"N. of words"* (22.547) and *"Avg. N. words per sentence"* (21.896). Moreover, they yield high values for the metrics *"% of functional verbs"* because visual verbs (2) are erroneously mapped as functional verbs in succinct descriptive texts. This shows a limitation of our approach in mapping functional verbs with a keyword-based approach.
Moreover, Figure 2 shows that dataset C has a distribution of the metric *"% of duplicated words"* skewed towards higher values when compared to the other datasets. This skewness is attributed to the abundance of legal expressions commonly found in patent claims, such as *"not limited to"*, *and "according to claim"* which increase the number of duplicated words within the dataset.
Furthermore, Figure 2 clearly shows distinctive characteristics for the dataset A, which involves patent titles as textual elements. This dataset exhibits the highest *"Avg. TF-IDF"* (0.020), the highest *"% of components"* (11.157), the second highest *"% of functional verbs"* (7.858), as well as the lowest values for *"%. of stopwords"(20.049)* and *"% of duplicated words"* (8.075). These results suggest that dataset A contains substantial technical information in comparison to the other datasets. However, dataset A exhibits the lowest values for both the metrics *"N. of words"* (8.075) and *"Avg. N. words per sentence"* (8.075) indicating that, on average, patent tiles are composed by 8 words. Therefore, using patent titles in multimodal datasets as figure descriptors may have potential limitation in providing sufficiently detailed information regarding patent drawings. Hence, we argue that different textual elements may

have different levels of abstraction in describing patent images, therefore different applications in the context of multimodality.
From Figure 2, the distributions of the metric *"N. of fig. reference"* shows that datasets A-C lack explicit figure references hindering direct alignment of these patent sections with patent drawings, and, thereby, rendering them more suitable to be directly aligned with patent front images. On the contrary, the textual elements of datasets D-E do include reference to patent drawings. However, dataset D has both a lower mean (1.261) and a lower standard deviation (0.944) comparing to dataset E (1.688; 2.190). This suggests that in the paragraphs of dataset D reference is generally made to only one drawing, whereas the paragraphs from the description section make use of multiple figure references, resulting, in more ambiguous text-image alignment.
Figure 2 facilitated the identification of 1) statistical macro-level differences among the datasets and 2) potential insights and problems about the textual elements of the multimodal datasets. However, an in-depth assessment of the overall quality and suitability of the datasets for different multimodal applications necessitates rigorous experimentation in training and testing multimodal models.
To validate our quantitative analysis, we test whether the distributions of the metrics 1-8 are statistically different within the datasets A-E. Since the metrics distributions are not normal (after the Shapiro-Wilkinson test) we run the non-parametric statistic test of "Mann-Whitney U" with a confidence level of 0.05. The result shows a p-vale lower than 0.05 for each test, meaning that the distributions of the metric 1-8 are statistically different within the datasets A-E.

4. Conclusions and future developments

In Section 3, we observed certain technical challenges that arise when constructing multimodal datasets using patent documents. One limitation we encountered is evidenced by the *"% of functional verbs"* metric and it is associated with our approach to mapping functional verbs through a keyword-based approach. This approach has its shortcomings, in disambiguating the usage of these verbs in both functional and non-functional contexts. To address this issue, it may be essential to explore the integration of contextual embeddings. Additionally, we faced difficulties when calculating the metric *"N. of fig. reference*", especially when patent images are referenced in text with the same numerical identifier but differentiated by literal sub-level (e.g., "FIG 5A-5C"). In such cases, our algorithm struggled to discern that a textual element is referencing two distinct images. Moreover, current work has focused exclusively on the analysis of functions and components found within textual data. However, there are other ED entities that can be extracted within patent documents, such as failures, problems, users, and technical features (Giordano et al., 2023). Starting from these limitations we provide other challenges to be faced in creating text-image alignment. Table 3 classifies the main challenges related to the pre-processing, computational and storage issues.

Table 3. Challenges for creating multimodal patent datasets

	Text	Image	Alignment
Pre-processing	Multiple figure reference in consolidated expressions	Multiple images on a single patent drawing page	No standard practices to manage the text-image alignment
Computational	High processing-time of sentence splitting	High processing-time of image segmentation techniques	High processing-time of OCR techniques to extract figure reference
Storage	Demanding storage requirements for patent full text	Demanding storage requirements for patent images	No formatting standard to store the alignment between image-text

In Table 3 we observe that high processing time to split textual data and segment images hinders the scalability of multimodal applications. Moreover, both patent images and full-text patent require demanding storage capabilities, and, for the moment, no formatting standards exist to store the data.
This study advances our understanding on how to create a multimodal patent dataset shedding light on their potentials and limitations for downstream ED applications. It is a preliminary work to encourage

the creation of high-quality patent datasets for multimodal applications in the field of ED. Such efforts stand to benefit patent analysts, innovators, and AI researchers to make more informed decisions about which datasets to use as training data for multimodal models and how multimodal patent data can support the ED.

Our work compares different multimodal datasets solely relying on text-based metrics, with no metrics established to analyse patent images. For future studies, we plan to develop image-based metrics for evaluating the quality of patent images. In fact, patent images contain component reference numbers and other visual elements such as directional arrows, axes, symbols, and indications of geometric dimensions. These elements can be used to 1) measure the semantic content expressed by images and 2) measure the consistency of text-image alignment by matching these elements in text. Moreover, based on text-based and image-based metrics we are going to release different multimodal patent datasets to foster multimodal applications such as 1) text-to-image and image-to-text applied for patent data and 2) visual grounding which consists in establishing the correspondence between textual references and specific visual regions/objects in patent drawings.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): "Concepts, taxonomies, opportunities and challenges toward responsible AI". *Information fusion*, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Atherton, M., Jiang, P., Harrison, D., & Malizia, A. (2018). "Design for invention: annotation of functional geometry interaction for representing novel working principles." *Research in Engineering Design*, 29, 245-262. https://doi.org/10.1007/s00163-017-0267-2

Chiarello, F., Belingheri, P., & Fantoni, G. (2021). "Data science for engineering design: State of the art and future directions." *Computers in Industry*, 129, 103447. https://doi.org/10.1016/j.compind.2021.103447

Giordano, V., Puccetti, G., Chiarello, F., Pavanello, T., & Fantoni, G. (2023). "Unveiling the inventive process from patents by extracting problems, solutions and advantages with natural language processing." *Expert Systems with Applications*, 229, 120499. https://doi.org/10.1016/j.eswa.2023.120499

Jee, J., Park, S., & Lee, S. (2022). "Potential of patent image data as technology intelligence source." *Journal of Informetrics*, 16(2), 101263. https://doi.org/10.1016/j.joi.2022.101263

Jiang, S., Luo, J., Ruiz-Pava, G., Hu, J., & Magee, C. L. (2021). "Deriving design feature vectors for patent images using convolutional neural networks." *Journal of Mechanical Design*, 143(6). https://doi.org/10.1115/1.4049214

Jiang, S., Sarica, S., Song, B., Hu, J., & Luo, J. (2022). "Patent data for engineering design: A critical review and future directions." *Journal of Computing and Information Science in Engineering*, 22(6), 060902. https://doi.org/10.1115/1.4054802

Kucer, M., Oyen, D., Castorena, J., & Wu, J. (2022). "DeepPatent: Large scale patent drawing recognition and retrieval." *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2309-2318).

Lin, W., Yu, W., & Xiao, R. (2023). "Measuring Patent Similarity Based on Text Mining and Image Recognition." *Systems*, 11(6), 294. https://doi.org/10.3390/systems11060294

Pustu-Iren, K., Bruns, G., & Ewerth, R. (2021). "A multimodal approach for semantic patent image retrieval." *In PatentSemTech 2021-Patent Text Mining and Semantic Technologies*, July 15th 2021, online (Vol. 2909). Aachen, Germany: RWTH Aachen. https://doi.org/10.34657/6842

Rao, R., Rao, S., Nouri, E., Dey, D., Celikyilmaz, A., & Dolan, B. (2020). "Quality and relevance metrics for selection of multimodal pretraining data." *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 956-957). https://doi.org/10.1109/CVPRW50498.2020.00486

Sarica, S., & Luo, J. (2021). "Stopwords in technical language processing." *Plos one*, 16(8), e0254937. https://doi.org/10.1371/journal.pone.0254937

Sarica, S., Luo, J., & Wood, K. L. (2020). "TechNet: Technology semantic network based on patent data." *Expert Systems with Applications*, 142, 112995. https://doi.org/10.1016/j.eswa.2019.112995

Vrochidis, S., Moumtzidou, A., & Kompatsiaris, I. (2012). "Concept-based patent image retrieval." *World Patent Information*, 34(4), 292-303. http://dx.doi.org/10.1016/j.wpi.2012.07.002

Vrochidis, S., Papadopoulos, S., Moumtzidou, A., Sidiropoulos, P., Pianta, E., & Kompatsiaris, I. (2010). "Towards content-based patent image retrieval: A framework perspective." *World Patent Information*, 32(2), 94-106. https://doi.org/10.1016/j.wpi.2009.05.010