data

MDPI

Data Descriptor

Conflicting Marks Archive Dataset: A Dataset of Conflicting Marks from the Brazilian Intellectual Property Office

Igor Bezerra Reis [1], Rafael Ângelo Santos Leite [2], Mateus Miranda Torres [2], Alcides Gonçalves da Silva Neto [2], Francisco José da Silva e Silva [1] and Ariel Soares Teles [1,3,*]

[1] Postgraduate Program in Computer Science, Federal University of Maranhão, São Luís 65085-580 , Brazil; igor.bezerra@lsdi.ufma.br (I.B.R.); fssilva@lsdi.ufma.br (F.J.d.S.e.S.)
[2] Federal Institute of Piauí, Floriano 64808-475, Brazil; rafaelangelo@ifpi.edu.br (R.Â.S.L.); mateusmmt00@gmail.com (M.M.T.); bmalcidesneto@gmail.com (A.G.d.S.N.)
[3] Federal Institute of Maranhão, Araioses 65570-000, Brazil
* Correspondence: ariel.teles@ifma.edu.br

Abstract: A registered trademark represents one of a company's most valuable intellectual assets, acting as a safeguard against possible reputational damage and financial losses resulting from infringements of this intellectual property. To be registered, a mark must be unique and distinctive in relation to other trademarks which are already registered. In this paper, we describe the CMAD, an acronym for Conflicting Marks Archive Dataset. This dataset has been meticulously organized into pairs of marks (Number of pairs = 18,355) involved in copyright infringement across word, figurative and mixed marks. Organizations sought to register these marks with the National Institute of Industrial Property (INPI) in Brazil, and had their applications denied after analysis by intellectual property specialists. The robustness of this dataset is ensured by the intrinsic similarity of the conflicting marks, since the decisions were made by INPI specialists. This characteristic provides a reliable basis for the development and testing of tools designed to analyze similarity between marks, thus contributing to the evolution of practices and computer-based solutions in the field of intellectual property.

Dataset: https://doi.org/10.5281/zenodo.10608109

Dataset License: CC-BY 4.0

Keywords: mark; trademark; brand; similarity; copyright; intellectual property

Citation: Reis, I.B.; Leite, R.Â.S.; Torres, M.M.; Neto, A.G.d.S.; Silva, F.J.d.S.e.; Teles, A.S. Conflicting Marks Archive Dataset: A Dataset of Conflicting Marks from the Brazilian Intellectual Property Office. *Data* **2024**, *9*, 33. https://doi.org/10.3390/data9020033

Academic Editor: Sharad Mehrotra

Received: 21 November 2023
Revised: 2 February 2024
Accepted: 5 February 2024
Published: 9 February 2024

1. Summary

Trademarks are instrumental in identifying and distinguishing goods and services in the global market [1]. This type of intellectual property, encompassing text, logos, sounds, colors, and even smells, are valuable assets for companies of all magnitudes. Consumer confidence and a trademark's reputation are intrinsically linked, highlighting the need for protection against unauthorized usage [2,3]. Also, it is essential to avoid copying any features of a previous trademark and exploiting any potential advantages when creating a new mark [4].

The World Intellectual Property Organization (WIPO) defines a trademark as a sign capable of distinguishing the goods or services of different companies [5]. The National Institute of Industrial Property (INPI), the Brazilian institute for intellectual property, defines a trademark as a distinctive sign whose main functions are to identify the origin of and distinguish goods or services from other identical, similar or related goods from different origins [4]. A trademark can take the form of a word mark (i.e., a sign consisting of one or more words), a figurative mark (i.e., a drawing, image, figure and/or symbol), a mixed mark (i.e., a word mark and a figurative mark), or a three-dimensional mark (i.e., a distinctive plastic form in itself) [4]. This understanding of shapes is also understood by the world's main intellectual property offices, such as the United States Patent and Trademark

Office (USPTO) [6], the Canadian Intellectual Property Office (CIPO) [7], and the China National Intellectual Property Administration (CNIPA) [8].

During a mark registration process, the initial search phase is essential. This phase requires the owner of the mark applying for registration to search for already registered trademarks that could potentially cause conflicts, i.e., when two marks are considered similar, which is a copyright infringement. Marks are considered conflicting when there is similarity in the nominative, phonetic, ideological, or visual aspects [4].

Figure 1 provides an example of similarity between marks that may lead to consumer conflict. The example displays a nominative similarity between the terms "DOMINUS PIZZAS E ESFIHAS ABERTAS" and "DOMINO'S PIZZA". As these terms have significant phonetic similarity, they can create conflict among consumers. The similarity in the pronunciation of the words is also phonetic, particularly in the initial syllables "DOMINUS" and "DOMINO'S".

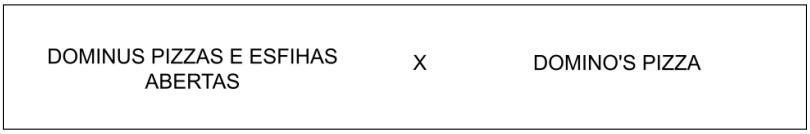


Figure 1. Example of nominative similarity.

Figure 2 displays an example of ideological similarity, in which, although the marks are visually presented differently, they evoke identical or similar ideas. In the example, "Café Brasileiro" translated to English language means "Brazilian Coffee" (i.e., text in the rejected mark), which could lead the target audience into a conflict or an inappropriate association.

Figure 2. Example of ideological similarity.

Figure 3 exemplifies a visual similarity, which presents elements which are graphically similar.

Figure 3. Example of visual similarity.

There is a continuous growth in the number of mark registrations worldwide, which reached around 10.9 million in 2018 [9]. The process of recognizing conflicting marks is performed manually by specialists, and it is a time-consuming and demanding task given the high number of applications for registration. The intellectual property offices of each country ensure trademark exclusivity, but the process is prone to human error and, in

severe cases, may result in the registration of similar marks. Such occurrences increase the complexity of resolving legal disputes between mark owners.

Therefore, law offices and intellectual property institutions require automated solutions to prevent new applications for mark registrations from conflicting with those marks already registered (i.e., trademarks). Such solutions will be able to avoid cases of litigation, which is a formal process of resolving legal disputes through the judicial system [10]. Automating the process of identifying conflicting marks through computer-based tools presents a promising solution [11]. Such tools can make the work of intellectual property officers more efficient [12], while simultaneously reducing costs in the trademark examination process. Also, they can improve decision making in the application for mark registration for companies and professionals dealing with trademark issues.

This study aimed to develop CMAD (acronym for ***C****onflicting* ***M****arks* ***A****rchive* ***D****ataset*), a dataset focused on litigation cases for use in trademark similarity experiments. The cases selected were those in which registration applications were rejected by the INPI to prevent conflicts between marks and reduce the possibility of litigation. The developed dataset has the potential to play a crucial role in developing tools for law offices and intellectual property officers globally. We believe that CMAD will be a fundamental instrument for boosting research and development in the area of intellectual property.

The remaining sections of this article are organized as follows. Section 2 discusses the related work. Section 3 describes the methods used to collect and organize the data. Next, Section 4 presents a detailed description of the dataset, while Section 5 discusses our work. Finally, Section 6 presents how to access and use the dataset.

2. Related Work

In the literature, the terms "trademark" and "logo" are distinct and have specific meanings. A trademark refers to a distinctive sign, such as a name, symbol, or design, which is used to identify a company's goods or services and distinguish them from those of other companies, and which are already legally protected by the competent authorities to prevent unauthorized use by third parties [4,5,7]. The term logo refers to a graphic element or symbol that represents a company or organization and is usually part of the brand's visual identity, but which is not yet legally protected [13].

The tasks known in the literature as "Logo Detection", "Image Retrieval", and "Trademark Similarity" are distinct from each other, but may be related depending on their use. Logo detection refers to the process of identifying the presence or location of specific logos in an image or video [14,15]. Image retrieval refers to the task of finding similar or relevant images in a dataset based on a query image provided by the user [16]. The idea is to retrieve images that share visual or semantic characteristics with the query image. Trademark similarity refers to the extent to which two trademarks are visually, phonetically or conceptually similar [4].

In the literature, there are different datasets originally developed for the logo detection and image retrieval tasks, in which some studies have adapted them to be applied to the trademark similarity task [1,11,17]. Next, we describe the datasets created for the tasks.

- BelgaLogos [18,19] was created for logo detection. It contains 10,000 manually annotated images of 26 logos. Each image is labeled for each logo, with 1 indicating when the logo is present in the image and 0 otherwise. The images in the dataset may contain one or several logos, or no logo. The test dataset contains the mark name present in the image, the file name, and coordinates of the pixels delimiting the logo present in the image;
- FlickrLogos-32 [20,21] contains images of 32 different logos and their labels. It was created and divided into three distinct sets, named P1, P2, and P3. The first set (P1) is intended for training Machine/Deep Learning (ML/DL) algorithms and has 10 images per class, which are logos in different perspectives. Sets P2 and P3, respectively, are used for validation and testing (or consultation), and contain 30 images per class, in which there is at least one instance of a logo;

- Logo-2K+ [22,23] was created for logo detection tasks, and has a total of 167,140 images. The images belong to marks divided into 10 classes (e.g., food, clothing, institutions, accessories), and subdivided into 2341 sub-categories representing each mark;
- LogoDet-3K [24,25] was created for logo detection and contains 3000 logo categories, with around 200,000 manually annotated logo objects and 158,652 images. The logo images are divided into nine categories (i.e., food, clothes, necessities, electronic, transportation, leisure, sports, medical, and others), and subdivided into 3000 sub-categories;
- LOGO-Net [26,27] is a large dataset of images for logo detection, including two sets with a total of 81,874 images: the "logos-18" set has a total of 16,043 logo objects in 8460 images, and the "logos-160" set has a total of 130,608 logo objects in 73,414 images. They were created using a web crawler (i.e., automated collection on the Internet [28]) on shopping mall websites, after which each image was manually annotated, thus delimiting the region of the logo;
- METU [29,30] is a dataset developed for image retrieval, and has 923,343 images of different types: logos with only text, only figures, and both images and text. It has two main sets: the query set and the test set. The query set contains 417 mark images manually labeled and grouped by similarity into 35 classes.

Table 1 presents a comparative analysis of related datasets with their tasks, number of samples, and data types. Related datasets are not specifically oriented towards the trademark similarity task, but rather towards logo detection and image retrieval tasks. The number of samples in the related datasets varies from less than 10 thousand to almost 1 million. The related datasets are composed of only images, without having any tabular data, descriptors, or metadata.

Table 1. Comparative analysis of related datasets and our developed dataset.

Dataset	Task	Number of Samples	Data Type
BelgaLogos [18,19]	Logo detection	10,000	Images
FlickrLogos-32 [20,21]	Logo detection	8240	Images
Logo-2K+ [22,23]	Logo detection	167,140	Images
LogoDet-3K [24,25]	Logo detection	158,652	Images
LOGO-NET [26,27]	Logo detection	81,874	Images
METU [29,30]	Image retrieval	930,328	Images
CMAD	Trademark similarity	Number of pairs = 18,355	Images and Tabular

Differently to the related datasets, CMAD was created specifically for the trademark similarity task. It was produced based on marks that have applied for registration at the INPI, but they have been rejected due to any conflict. These rejections have been carefully analyzed by INPI specialists. Therefore, trademark similarity in CMAD represents reliably labeled conflicts. Furthermore, to the best of our knowledge, CMAD is the first dataset that has cases of conflicting marks due to three types of similarity (i.e., nominative, ideological, and visual) including copyright infringements between word, figurative, and mixed marks. Importantly, word marks (i.e., in text form) may represent the majority of trademarks in intellectual property offices [31].

Also, different to BelgaLogos [18,19], FlickrLogos-32 [20,21], and LogoDet-3K [24,25], CMAD is composed of mark images and tabular data, which are organized to easily provide information related to the conflicting marks. CMAD can provide a solid basis for developing and testing tools designed to analyze similarity between marks, thus contributing to the evolution of practices and computer-based solutions in the field of intellectual property. Therefore, by choosing CMAD as a dataset for trademark similarity tasks, researchers

and law practitioners can benefit not only from the diversity and reliability of real-world trademark litigation cases, but also from its organizational and tabular structure.

3. Methods

The "Revista da Propriedade Industrial" (Industrial Property Magazine, in English language) is an official document published by the INPI on a weekly basis [32]. This magazine has several sections, one of which is specifically dedicated to trademark registration applications. The trademark section plays an important role in disseminating information related to trademark registration in Brazil, as it publishes information on registered trademarks and marks that have applied for registration. The publication of this information allows third parties to oppose the registration of a mark if they believe it may infringe their rights or cause conflict with an existing trademark. This ensures transparency in the registration process and allows third parties to monitor and evaluate ongoing processes.

During the mark registration process, the substantive examination stage verifies whether applications for mark registrations meet the legal conditions imposed by the INPI [33]. After initial screening, the mark is published in the Industrial Property Magazine for 60 days. After the 60-day period, applications that have not received any opposition to the registration proceed to the next phase of the process. Applications that receive opposition to registration by third parties (e.g., trademark owners, or law offices specialized in intellectual property acting on behalf of trademark owners), which have identified possible conflicts with their marks, are then evaluated by intellectual property specialists. These professionals must have: (1) delegation of competence; (2) received specific training for the opposition exam; and (3) high technical-professional qualifications in the field of trademark law [32]. After analyzing the oppositions, specialists have the authority, on behalf of the INPI, to reject the mark registration.

The methodology of this study involves analyzing and extracting data from magazines published by INPI. The flowchart depicted in Figure 4 describes the steps of the methodology.

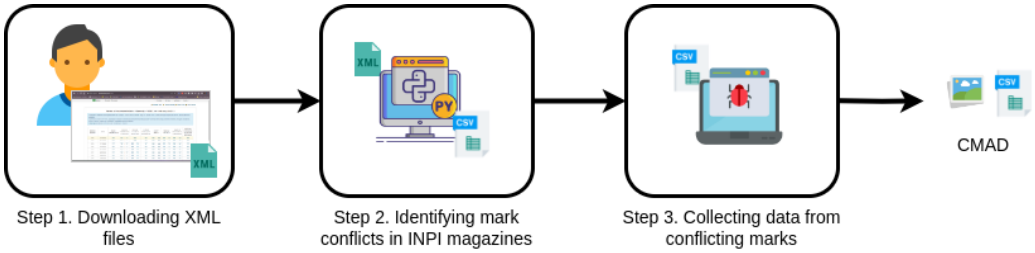


Figure 4. Methodology steps.

In Step 1, we manually download the magazines in XML format. This file is used as input to the algorithm that analyzes and extracts the applications. In the magazine, there is a section describing the applications that have been rejected and the reason, as can be seen in Figure 5.

For Step 2, we developed an algorithm written in Python programming language that receives the XML file (Figure 6) as input and identifies content related to mark applications (e.g., process number, complementary text) using regular expressions. This algorithm generates a CSV file that contains the conflicting marks.

In Step 3, each denied application is analyzed. In the complementary text of the application, we identify and extract the conflicting marks, and then start the crawler to collect their data from the INPI website. For this purpose, we created another algorithm in Python using the Selenium library [34] (see screen recording of the process in Supplementary File S1). The web crawler first collects information about the denied application (e.g., process number, name, presentation form, nature, Nice classification), as shown in Figure 7. When the mark presentation is figurative or mixed, the crawler also collects the mark image. The crawler algorithm then collects the same information for the conflicting trademark. If there is more than one conflicting trademark, data are collected

for all of them. At the end, a CSV file is generated containing the rejected mark for each conflicting trademark.

829220267 Indeferimento do pedido
Titular: EPI - EMPREENDIMENTO PATRIMONIAL INDUSTRIAL S/A [BR/PR]
Procurador: Manoel Paixao do Nascimento
NCL(9): 21
Especificação: ESPONJAS DE LIMPEZA PARA USO DOMÉSTICO, ESCOVAS PARA LOUÇAS, ESCOVAS PARA CALÇADOS, ESCOVAS PARA ESFREGAR, VASSOURAS. (DA CLASSE 21)

Detalhes do despacho: A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregistrável de acordo com o inciso XIX do Art. 124 da LPI: Processo 828906564 (BRILLMAX).

(**a**)

829220267 Rejection of application
Holder: EPI - EMPREENDIMENTO PATRIMONIAL INDUSTRIAL S/A [BR/PR]
Attorney: Manoel Paixao do Nascimento NCL(9):
21
Specification: HOUSEHOLD CLEANING SPOONS, DRAWING BRUSHES, SHOE BRUSHES, SCRUBBING BRUSHES, VASSOURGES. (FROM CLASS 21)

Dispatch details: The trademark reproduces or imitates the following third-party registrations and is therefore unregistrable in accordance with item XIX of Article 124 of the IPL: Process 828906564 (BRILLMAX).

(**b**)

Figure 5. Example of rejection: (**a**) originally written in PT-BR; (**b**) translated by the authors to English language.

```
<processo numero="927779820">
  <despachos>
    <despacho codigo="IPAS024" nome="Indeferimento do pedido">
      <texto-complementar>A marca reproduz ou imita os seguintes registros de terceiros, sendo,
      portanto, irregistrável de acordo com o inciso XIX do Art. 124 da LPI: Processo 918753252 (YUPY KIDS),
      Processo 921945086 (YUUUP! BABY) e Processo 830036970 (IUPII!!! BEBÊ).
      Art. 124 - Não são registráveis como marca: XIX - reprodução ou imitação, no todo ou em parte, ainda que com acréscimo,
      de marca alheia registrada, para distinguir ou certificar produto ou serviço idêntico, semelhante ou afim,
      suscetível de causar confusão ou associação com marca alheia;</texto-complementar>
    </despacho>
  </despachos>
  <titulares>
    <titular nome-razao-social="IZAURA MARIA VICTORINO DA SILVA" pais="BR" uf="PR"/>
  </titulares>
  <lista-classe-nice>
    <classe-nice codigo="35">
      <especificacao>Comércio (através de qualquer meio) de artigos do vestuário;Comércio (através de qualquer meio) de fitas e laços; </esp
      <status>Deferida</status>
    </classe-nice>
  </lista-classe-nice>
  <procurador>Danielle Juliá Lopes Brites</procurador>
</processo>
```

Figure 6. Example of an XML file containing a mark application.

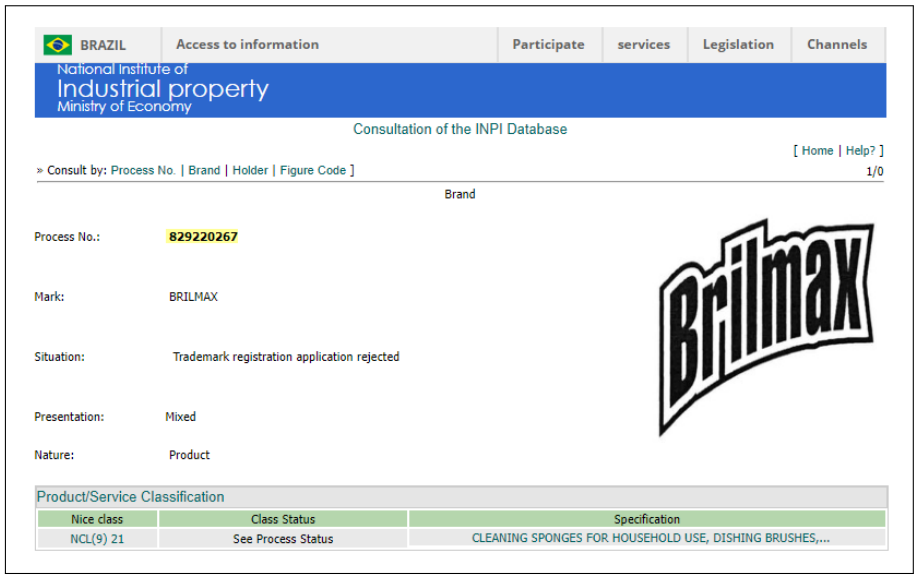


Figure 7. Example of collected data from a rejected mark.

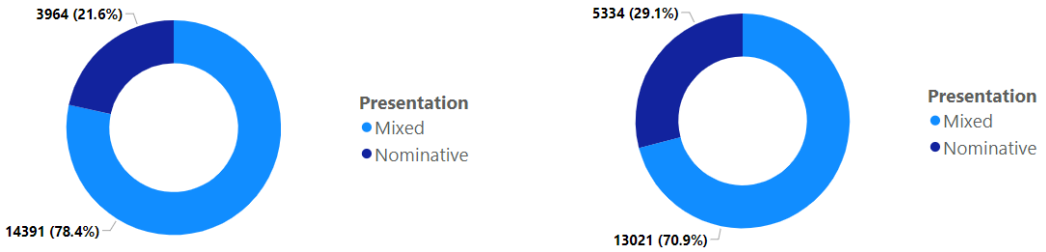
4. Data Description

CMAD contains 18,355 samples of mark conflicts, in which a mark applying for registration conflicted with a trademark. The dataset has a directory with images in PNG format, and a CSV file containing tabular data related to each sample. In the CSV file, each sample is structured in pairs and includes eight columns for each (i.e., the rejected mark and the trademark already registered in the INPI), a complementary text, and the magazine. For the columns that refer to the registered trademark, their headings are followed by the acronym TM (i.e., TradeMark) and, for the columns that refer to the rejected mark, their headings are followed by the acronym RM (i.e., Rejected Mark). The final two columns, complementary text and magazine, refer to the justification given for the opposition to register the trademark and the magazine number, respectively. Table 2 presents the columns of the dataset. Importantly, the CSV file was created in the Brazilian Portuguese language, with data from the INPI.

Table 2. Data description of the CMAD CSV file.

Field Name	Description	Type
Process number	Process number given to the registration application, which is used to uniquely identify the trademark in the dataset, as well as to access the image path if it has one (i.e., there are no images for word marks).	Numeric
Name	Mark name.	Text
Status	Mark status (e.g., rejected mark, registered, waiting for analysis).	Text
Presentation	Type of mark presentation (e.g., Nominative, Figurative and Mixed).	Text
Nature	Mark nature (e.g., goods, services).	Text
Nice classification	It is an international classification of goods and services, adopted globally and managed by WIPO. It is used to categorize marks in their area of application [35,36].	Text
Vienna classification	It is an international classification managed by WIPO to categorize graphic elements into figurative, mixed and three-dimensional marks. It helps to describe and specify visual elements during mark registration, so avoiding conflicts and providing clarity [37].	Text
Application date	Date the mark was applied for.	Date
Complementary text	Text describing the reasons why the application was denied.	Text
Magazine	Magazine publication number.	Text

The mixed form of presentation accounts for most of the records in the dataset. Of the rejected marks (Figure 8a), 14,391 (78.4%) are in the mixed form and 3964 (21.6%) in word form. While for trademarks (Figure 8b), 13,021 (70.9%) are in the mixed form, and 5334 (29.1%) in word form.



(**a**) (**b**)

Figure 8. Sample distribution of presentations for: (**a**) rejected mark applications, and (**b**) trademarks.

Figure 9 illustrates three samples of conflicting marks with pairs of mark images and, respectively, their entries in the CSV file.

(**a**)

Process number	Mark rejected	Status	Presentation	Process number	Trademark	Status	Presentation
924058803	My Bank$	Aguardando apresentação	Mista	918356849	MEUBANK	Registro de marca em vigor	Mista
925224022	AUTOBOX LUBRIFICANTES	Aguardando apresentação	Mista	908946481	AUTOBOX	Registro de marca em vigor	Mista
927878143	SOCORRO AUTO CHICO	Aguardando apresentação	Mista	912622962	AUTO SOCORRO Silva & Silva	Registro de marca em vigor	Mista

(**b**)

Figure 9. Three samples of conflicting marks in the CMAD: (**a**) three pairs of mark images, and (**b**) respective entries in the CSV file.

5. Discussion

5.1. CMAD Applications

CMAD is useful in various fields. In industry, it is particularly valuable for developing and testing software that assesses similarities between trademarks. Specifically, CMAD can aid in the development and improvement of systems designed to compare and identify similarities between trademarks.

In academia, the use of CMAD is highly beneficial in specific studies for analyzing trademark similarity, whether for the validation or training of ML/DL algorithms. Several studies [12,31,38,39] stand out for developing DL models that employ similar data pairs during training, thus assessing the similarity between marks. These studies exemplify the potential of CMAD to effectively contribute to the development of ML/DL models, particularly those based on Siamese Neural Networks [40]. By adopting CMAD as a data source, researchers can take the opportunity to explore a diverse range of mark conflict cases.

Although CMAD is focused on trademark similarity, it may have broader applicability. For instance, it can be used in the field of image retrieval. Different studies [17,41–43] utilize trademark datasets to implement advanced feature extraction algorithms, thus improving the efficiency and accuracy of trademark identification. The CMAD image set is an option for developing feature extraction algorithms aimed at the image retrieval task. Therefore,

our dataset can offer diversity and representativeness for the training and validation of such algorithms.

5.2. Ethical Considerations

CMAD only includes characteristics of marks, such as their name, classification, and image, and does not contain any personal information about their owners. These data are already publicly available on the INPI online platform. Therefore, there are no ethical or legal impediments to using this information for research and analysis. The use of CMAD ensures compliance with ethical and legal regulations by excluding sensitive personal data. The utilization of the dataset aligns with the INPI Open Data Plan [44,45], which establishes guidelines for the implementation and promotion of INPI data openness.

5.3. Strengths and Limitations

By focusing on mark conflict cases exclusively in Brazil, CMAD offers a unique and valuable perspective for analysis in the national context. This means that the data reflects the specific nuances of the Brazilian market, thus making CMAD particularly relevant for studies and applications focused on Brazil. Local legal and cultural aspects play a crucial role in shaping the mark. For example, the way marks are perceived and interpreted can be influenced by cultural factors such as language, symbolism, and social norms. Furthermore, Brazilian intellectual property laws, which govern the registration and protection of marks, can have significant implications for the existence and resolution of mark conflicts. Therefore, the use of CMAD data must take these aspects into account. Understanding these elements can help identify more precise trends, patterns, and insights contextualized to a national scenario.

5.4. Future Work

CMAD brings together mark conflicts published between August and October 2023. The web crawler algorithm developed in this study allows for the expansion of the dataset. By using it, a potential future task involves incorporating additional samples from the INPI database to broaden the scope of the CMAD dataset. Also, we plan to leverage CMAD to develop a DL-based multimodal method that will consider the nominative, ideological, and visual similarities of marks.

6. Usage Notes

The dataset provides an efficient organization of public information related to mark applications and trademarks already registered. Each figurative or mixed mark sample in the CSV file has a unique file name that is associated with the process number, thus facilitating access to its corresponding path in the folder. Word marks have no images. The format of the path to access a mark image in PNG format is structured as follows: `...directory_path/[process_number].png`. To read the .CSV file, a semicolon (;) should be used as a delimiter. When employed in the development of ML/DL solutions, the dataset must undergo a preparation process tailored to the targeted task, adhering to best practices [46,47].

Supplementary Materials: The following are available online at: https://www.mdpi.com/article/10.3390/data9020033/s1, Supplementary File S1: Screen recording.

Author Contributions: Conceptualization, I.B.R., A.S.T., R.Â.S.L., M.M.T. and A.G.d.S.N.; methodology, I.B.R., R.Â.S.L., F.J.d.S.e.S. and A.S.T.; data curation, I.B.R.; project administration, A.S.T. and R.Â.S.L.; software, I.B.R.; supervision, A.S.T.; Writing—original draft, I.B.R., A.S.T. and R.Â.S.L.; Writing—review and editing, I.B.R., F.J.d.S.e.S. and A.S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Coordination for the Improvement of Higher Education Personnel (CAPES) [Finance Code 001] and Brazilian National Council for Scientific and Technological Development (CNPq) [grant 308059/2022-0].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is publicly available on https://doi.org/10.5281/zenodo.10608109, accessed on 21 November 2023. The code that has been used in this study for data collection is available on https://github.com/igorbezerrar/cmad, accessed on 21 November 2023.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CIPO	Canadian Intellectual Property Office
CMAD	Conflicting Marks Archive Dataset
CNIPA	China National Intellectual Property Administration
CSV	Comma-separated values
DL	Deep Learning
INPI	Instituto Nacional da Propriedade Industrial
ML	Machine Learning
PNG	Portable Network Graphics
USPTO	United States Patent and Trademark
WIPO	World Intellectual Property Organization
XML	Extensible Markup Language

References

1. Vesnin, D.; Levshun, D.; Chechulin, A. Trademark Similarity Evaluation Using a Combination of ViT and Local Features. *Information* **2023**, *14*, 398. [CrossRef]
2. Trappey, C.; Trappey, A.; Lin, S. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. *Adv. Eng. Inform.* **2020**, *45*, 101120. [CrossRef]
3. Macías, W.; Cerviño, J. Trademark dilution and its practical effect on purchase decision. *Span. J. Mark. Esic* **2017**, *21*, 1–13. [CrossRef]
4. National Institute of Industrial Property. Manual de Marcas. Available online: http://manualdemarcas.inpi.gov.br/projects/manual/wiki/02_O_que_%C3%A9_marca#2-O-que-%C3%A9-marca (accessed on 24 October 2023).
5. World Intellectual Property Organization. Trademarks. Available online: https://www.wipo.int/trademarks/en/ (accessed on 24 October 2023).
6. United States Patent and Trademark Office. Trademark Decisions and Proceedings. Available online: https://developer.uspto.gov/tm-decisions/search/expungement (accessed on 24 October 2023).
7. Canadian Intellectual Property Office. What Are Trademarks? Available online: https://ised-isde.canada.ca/site/canadian-intellectual-property-office/en/what-intellectual-property/what-are-trademarks (accessed on 24 October 2023).
8. China National Intellectual Property Administration. Trademarks. Available online: https://english.cnipa.gov.cn/col/col2996/index.html (accessed on 24 October 2023).
9. WIPO. Indicadores Mundiais em Propriedade Intelectual: Pedidos de Patentes, Marcas e Desenhos Industriais Atingem Níveis Históricos em 2018. Available online: https://www.wipo.int/pressroom/pt/documents/pr_2019_838.pdf (accessed on 8 November 2023).
10. National Institute of Industrial Property Despachos Aplicáveis-Indeferimento. Available online: http://manualdemarcas.inpi.gov.br/projects/manual/wiki/5%C2%B719_Despachos_aplic%C3%A1veis#5194-Indeferimento (accessed on 19 November 2023).
11. Alshowaish, H.; Al-Ohali, Y.; Al-Nafjan, A. Trademark Image Similarity Detection Using Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 1752. [CrossRef]
12. Trappey, A.J.; Trappey, C.V.; Lin, S.C.C. Detecting Trademark Image Infringement Using Convolutional Neural Networks. *Adv. Transdiscipl. Eng.* **2020**, *10*, 477–486. [CrossRef]
13. Kesidis, A.; Karatzas, D. Logo and Trademark Recognition. In *Handbook of Document Image Processing and Recognition*; Doermann, D., Tombre, K., Eds.; Springer: London, UK, 2014; pp. 591–646. [CrossRef]
14. Bao, Y.; Li, H.; Fan, X.; Liu, R.; Jia, Q. Region-Based CNN for Logo Detection. In Proceedings of the International Conference on Internet Multimedia Computing and Service, Xi'an, China, 19–21 August 2016 ; pp. 319–322. [CrossRef]
15. Psyllos, A.P.; Anagnostopoulos, C.N.E.; Kayafas, E. Vehicle Logo Recognition Using a SIFT-Based Enhanced Matching Scheme. *IEEE Trans. Intell. Transp. Syst.* **2010**, *11*, 322–328. [CrossRef]
16. Smeulders, A.; Worring, M.; Santini, S.; Gupta, A.; Jain, R. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1349–1380. [CrossRef]

17. Lan, T.; Feng, X.; Xia, Z.; Pan, S.; Peng, J. Similar Trademark Image Retrieval Integrating LBP and Convolutional Neural Network. In Proceedings of the Image and Graphics, Shanghai, China, 13–15 September 2017; pp. 231–242.
18. Joly, A.; Buisson, O. Logo Retrieval with a Contrario Visual Query Expansion. In Proceedings of the 17th ACM International Conference on Multimedia, Beijing, China, 19–24 October 2009; pp. 581–584. [CrossRef]
19. Joly, A.; Buisson, O. BelgaLogos Dataset. Available online: https://www-sop.inria.fr/members/Alexis.Joly/BelgaLogos/BelgaLogos.html (accessed on 8 November 2023).
20. Romberg, S.; Pueyo, L.G.; Lienhart, R.; van Zwol, R. Scalable Logo Recognition in Real-World Images. In Proceedings of the ACM International Conference on Multimedia Retrieval 2011 (ICMR11), ACM, Trento, Italy, 18–20 April 2011.
21. Romberg, S.; Pueyo, L.G.; Lienhart, R.; van Zwol, R. FlickrLogos. Available online: https://www.uni-augsburg.de/en/fakultaet/fai/informatik/prof/mmc/research/datensatze/flickrlogos/ (accessed on 8 November 2023).
22. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Wang, H.; Jiang, S. Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 6194–6201. [CrossRef]
23. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Jiang, S. Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification. Available online: https://github.com/msn199959/Logo-2k-plus-Dataset (accessed on 8 November 2023).
24. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Jiang, S. Logodet-3k: A large-scale image dataset for logo detection. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2022**, *18*, 1–19. [CrossRef]
25. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Jiang, S. LogoDet-3K: A Large-Scale Image Dataset for Logo Detection. Available online: https://github.com/Wangjing1551/LogoDet-3K-Dataset (accessed on 8 November 2023).
26. Hoi, S.C.H.; Wu, X.; Liu, H.; Wu, Y.; Wang, H.; Xue, H.; Wu, Q. LOGO-Net: Large-scale Deep Logo Detection and Brand Recognition with Deep Region-based Convolutional Networks. *arXiv* **2015**, arXiv:1511.02462.
27. HOI, D.S. LOGO-Net Dataset. Available online: http://www.mysmu.edu.sg/faculty/chhoi/logonet/index.html (accessed on 20 November 2023).
28. Kumar, M.; Bhatia, R.; Rattan, D. A survey of Web Crawlers for Information Retrieval. *Wires Data Min. Knowl. Discov.* **2017**, *7*, e1218. [CrossRef]
29. Tursun, O.; Aker, C.; Kalkan, S. A Large-scale Dataset and Benchmark for Similar Trademark Retrieval. *arXiv* **2017**, arXiv:1701.05766.
30. Tursun, O.; Aker, C.; Kalkan, S. Metu Trademark Dataset. Available online: https://github.com/neouyghur/METU-TRADEMARK-DATASET (accessed on 8 November 2023).
31. Liu, Y.; Li, Q.; Sun, C.; Si, L. Similar Trademark Detection via Semantic, Phonetic and Visual Similarity Information. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; pp. 2025–2030. [CrossRef]
32. INPI. Revista da Propriedade Industrial. 2023. Available online: http://revistas.inpi.gov.br/rpi/ (accessed on 21 November 2023).
33. National Institute of Industrial Property (INPI). INPI Trademark Manual—Chapter 5: Substantive Examination, 2022. Available online: https://manualdemarcas.inpi.gov.br/projects/manual/wiki/05_Exame_substantivo (accessed on 21 November 2023).
34. Contributors, S. Selenium: Browser Automation. 2023. Available online: https://www.selenium.dev/ (accessed on 21 November 2023).
35. WIPO. Nice Classification—WIPO—World Intellectual Property Organization. Available online: https://www.wipo.int/classifications/nice/en/ (accessed on 24 January 2024).
36. INPI. Classification of products and services—INPI. Available online: https://www.gov.br/inpi/pt-br/servicos/marcas/classificacao-marcas/classificacao (accessed on 24 January 2024).
37. WIPO. Vienna Classification—WIPO. Available online: https://www.wipo.int/classifications/vienna/en/ (accessed on 24 January 2024).
38. Lan, T.; Feng, X.; Li, L.; Xia, Z. Similar Trademark Image Retrieval Based on Convolutional Neural Network and Constraint Theory. In Proceedings of the 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA), Xi'an, China, 7–10 November 2018; pp. 1–6. [CrossRef]
39. Trappey, A.J.; Trappey, C.V.; Shih, S. An intelligent content-based image retrieval methodology using transfer learning for digital IP protection. *Adv. Eng. Inform.* **2021**, *48*, 101291. [CrossRef]
40. Chicco, D. Siamese neural networks: An overview. *Artifi. Neural Netw.* **2021**, *2190*, 73–94. [CrossRef]
41. Velmurugan, K.; Baboo, L. Image Retrieval using Harris Corners and Histogram of Oriented Gradients. *Int. J. Comput. Appl.* **2011**, *24*, 6–10.
42. Cao, J.; Huang, Y.; Dai, Q.; Ling, W.K. Unsupervised Trademark Retrieval Method Based on Attention Mechanism. *Sensors* **2021**, *21*, 1894. [CrossRef] [PubMed]
43. Tursun, O.; Denman, S.; Sivapalan, S.; Sridharan, S.; Fookes, C.; Mau, S. Component-Based Attention for Large-Scale Trademark Retrieval. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 2350–2363. [CrossRef]
44. INPI. Portaria /INPI/PR Nº 70. Available online: https://www.gov.br/inpi/pt-br/acesso-a-informacao/dados-abertos/arquivos/documentos/diversos/plano-de-dados-abertos-bienio-2022-2024.pdf (accessed on 29 January 2024).
45. INPI. Dados Abertos. Available online: https://www.gov.br/inpi/pt-br/acesso-a-informacao/dados-abertos (accessed date 29 January 2024).

46. Fernandes, A.A.A.; Koehler, M.; Konstantinou, N.; Pankin, P.; Paton, N.W.; Sakellariou, R. Data Preparation: A Technological Perspective and Review. *SN Comput. Sci.* **2023**, *4*, 425. [CrossRef]
47. Aguinis, H.; Hill, N.S.; Bailey, J.R. Best Practices in Data Collection and Preparation: Recommendations for Reviewers, Editors, and Authors. *Organ. Res. Methods* **2021**, *24*, 678–693. [CrossRef]