ISSN 1678-765X doi 10.20396/rdbci.v21i00.8671396/30919

Article

RDBCI
Revista Digital de Biblioteconomia e Ciência da Informação
Digital Journal of Library and Information Science

Information architecture applied on natural language processing: a proposal

Information Science contributions on data pre-processing for training and learning of artificial neural networks

George Hideyuki Kuroki Júnior [1] Claudio Gottschalg-Duque [2]

ABSTRACT

Introduction: Natural Language Processing through artificial neural networks has gaps that can be addressed by Information Science through Information Architecture. Objective: To present Information Science contributions on Knowledge Organization applied to artificial neural networks training methods, positioning it as an active body of knowledge in artificial intelligence problems. Methodology: A three-leveled analysis path (metaphysical, scientific, and technological) is adopted to guide and ground the study. On metaphysical level, current development stage of natural language processing techniques is verified and analyzed. On scientific findings, a five-step procedure is proposed which aims to design, analyze, and prepare information spaces for artificial neural networks training and learning methods, fulfilling gaps identified by authors focused on Computer Science implementations. On technological implementation, the five-step procedure is applied to 3 datasets formed by texts from 16 scientific knowledge areas, as an evaluation basis. Results: Results obtained through pre-processed data and raw data where compared, showing great potential in developing a structured method of Multimodal Information Architecture that provide instruments able to organize data used as test and learning samples in artificial neural networks. Conclusion: This method could place Information Science as a producer of data pre-processing solutions, replacing its current role as consumer of prefabricated solutions made by Computer Science.

KEYWORDS

Information Science. Information architecture. Information treatment. Artificial Intelligence. Natural language processing.

Author's correspondence

[1] Universidade de Brasília
Brasília, DF - Brazil
e-mail: kurokijr@gmail.com

[2] Universidade de Brasília
Brasília, DF - Brazil
e-mail: klaussherzog@gmail.com

Uma proposta de arquitetura da informação aplicada ao processamento de linguagem natural:

contribuições da Ciência da Informação no pré-processamento de dados para treinamento e aprendizado de redes neurais artificiais

RESUMO

Introdução: O processamento de linguagem natural em redes neurais artificias possui lacunas passíveis de tratamento por parte da Ciência da Informação, utilizando-se de Arquitetura da Informação. Objetivo:

Propor contribuições da Ciência da Informação na Organização do Conhecimento para treinamento de redes neurais artificiais utilizando Arquitetura da Informação Multimodal, posicionando-a como área do conhecimento atuante em problemas de inteligência artificial. Metodologia: Adaptando um percurso de três níveis de análise (metafísico, científico e tecnológico), verifica o atual estágio de desenvolvimento de técnicas de processamento de linguagem natural (metafísico); utiliza definições de Arquitetura da Informação Multimodal propondo um procedimento de cinco passos para delineamento, análise e transformação do espaço informacional a ser utilizado em métodos de treinamento e aprendizagem de redes neurais, complementando lacunas identificadas por autores voltados a implementações da Ciência da Computação (científico); verifica a aplicabilidade da proposta em 3 conjuntos de dados advindos de 16 áreas do conhecimento como base de avaliação (tecnológico). Resultados: Os resultados obtidos nas situações com pré-tratamento e sem pré-tratamento foram comparados observando-se potencial para desenvolvimento de um método estruturado de Arquitetura da Informação Multimodal que forneça instrumentos para a organização do pré-processamento de dados a serem utilizados como massa de teste e aprendizado em redes neurais artificiais, em particular, no processamento de linguagem natural. Conclusão: Este método posicionaria a Ciência da Informação como atuante e produtora de soluções de pré-processamento de dados, sobrepondo o papel atual de mera consumidora de soluções pré-fabricadas pela Ciência da Computação.

PALAVRAS-CHAVE
Ciência da Informação. Arquitetura de informação. Tratamento da informação. Inteligência Artificial. Processamento de linguagem natural.

CRediT

- Recognitions: Not applicable.
- Funding: Not applicable.
- Conflicts of interest: Authors certify that they have no commercial or associative interest that represents a conflict of interest in relation to the manuscript.
- Ethical approval: Not applicable.
- Availability of data and material: The data have industrial property confidentiality.
- Authors' contributions: Conceptualization, Data Curation, Formal Analysis, Research, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing, Writing - review & editing: KUROKI JÚNIOR, G.H.; Supervision, Validation: DUKE, C. G.

JITA: BK. Information architecture.

Article submitted to the similarity system

Submitted: 03/11/2022 – **Accepted**: 16/12/2022 – **Published**: 08/02/2023

Editor:Gildenir Carolino Santos

1 INTRODUCTION

The increasing use of artificial intelligence models in everyday activities of classification and treatment of information places a new prism of observation to the question raised by Hjørland (2008). According to the author, Knowledge Organization as a study area would have Information Science and Librarianship as its central pieces, however, being seriously challenged by Computer Science.

At the time when this statement was made, a proposal of architecture and implementation of artificial neural networks developed by Hinton, Osindero and Teh (2006) made it possible to overcome a historical obstacle faced by Computing. Until then, the construction of artificial neural networks suffered from a lack of depth in their implementations: notoriously, the human brain, the basis for the development of intelligence models, has several layers of analysis, which allows the treatment of problems with greater complexity. With the advent of the proposal in question, the number of processing layers exceeded the limit of two or three.

The passing of this computational limitation gave rise to a great variety of technological implementations, giving rise to innumerable architectural designs of neural networks that apply multiple mathematical algorithms to obtain a measure of intelligence through pattern checking.

Although there have been advances in computer science, a criticism made by Hjørland is still open to discussion:

> There are many separate communities working with different technologies, but very little research into their basic assumptions and merits and weak sides. The problem is not just to formulate a theory, but to discover theoretical assumptions in different practices, to formulate these assumptions as clearly as possible, to

A point in common to all Computer Science initiatives is their dependence on a significant amount of data and/or records to obtain patterns to be seen. However, obtaining this data is not always possible, particularly in problems that require specialized knowledge, for example, the classification of technical-scientific texts, highly linked to the vocabulary of the area in question.

Historically, Computer Science was primarily concerned with the treatment of the complexity of a neural network model in the face of the data to be analyzed as it grows exponentially, what Bellman (1954) called the *problem of data dimensionality*. In the absence of larger registries, textual data enrichment techniques stick to everyday contexts of use, still making use of the wide range of information available in other common domains.

This paper positions the Multimodal Information Architecture (MIA) as a first contribution of Information Science, in the form of a theoretical counterpart of data pre-processing for further application in Artificial Intelligence (AI) models, more specifically in Natural Language Processing (NLP), in text classification problems in specific knowledge domains.

2 METHODOLOGICAL PROCEDURES

To analyze in a structured way the impacts of MIA application on NLP problems, we propose the use of the methodological path for the construction of a Worldview (M^3) created by Van Gigch and Moigne (1989).

This proposal considers the construction of knowledge along three stages that are closely related: a metaphysical level, prior to the formalization of the object of knowledge; a

level of the object of knowledge itself; and a level of application of the knowledge constructed. In this sense, this paper will adapt this method as follows:

(a) At the metaphysical level: find the fundamental issues of the current stage of NLP and fundamental issues of Multimodal Information Architecture;
b) At the knowledge object level: propose ways of applying MIA to NLP problems;
c) At the knowledge application level: to generate MIA products for implementation in NLP.

After going through the three levels of the World Vision adopted, we will have a set of knowledge, techniques and products that can be validated and verified for their adherence to the problem addressed, through a comparison of results obtained in simulations of artificial neural networks based on a set of data not treated by MIA and the same set of data treated by MIA.

3 DEEP LEARNING: APPLICATIONS, DEVELOPMENT, AND CHALLENGES IN NATURAL LANGUAGE PROCESSING

The fundamental dictates for the construction of artificial neural networks were sedimented throughout the 60s to the 90s. With the entry of the 2000s and the proposal by Hinton, Osindero, and Teh (2006), a new range of implementations began to take advantage of the depth of layers of analysis, giving rise to the term Deep Learning.

Wason (2018) performs a survey on the use of the discoveries made by Hinton, Osindero, and Teh (2006), verifying their broad use in a varied range of domains such as, for example, voice recognition independent of the sound source; recurrent neural networks; handwriting recognition; deep belief networks; auto-encoders; acoustic modeling; class feature detectors; handwriting synthesis; language modeling; model improvement and development among others. He concludes that three major challenges still linger in most AI applications:

(a) Data volume: the mass of data needed to achieve satisfactory learning would be of the nature of ten times the number of parameters (neurons) of the designed network;
b) Overfitting phenomenon: the larger the size of the net, in terms of the number of parameters, the greater the probability that learning will be oversized, resulting in a low generalization ability (small changes in the input objects result in an unsatisfactory result);
c) Fragile nature: neural networks tend to be specialized, so that when trained on one task, their performance on another task is extremely poor.

From the junction of the first two challenges cited by Wason (2018), one still finds problem previously mapped by Bellman (1954), also addressed by Arel, Rose, and Kanowski (2010) called the data dimensionality problem, where the learning complexity grows exponentially over the linear increase in the number of data dimensions.

According to Minaee (2021), the most recent attempts to obtain best results in NLP are based on Transformers and Pre-Trained Models - MPT. Since the first implementations of neural networks for NLP, such as Convolutional Networks, Recurrent Networks, and LSTM (Long Short-Term Memories) Networks, the difficulty in capturing the relationships between words within a sentence has been perceived. With the advent of Attention Mechanism-based models first proposed by Bahdanau, Cho, and Bengio (2015), neural networks began to treat various objects in a grouped manner. Based on this advance, Vaswani *et al.* (2017) proposed a

new architecture called Transformers, which brought two relevant innovations: assignment of an attention score that evaluates the influence of one word on another and improvement in parallelization methods, reducing training time. As of 2018 a growth in Transformer-based MPTs is seen, endowed with denser architectures and pre-trained on large volumes of textual data which jointly entails better contextualization of words and sentences. Qiu *et al.* (2020) conducted a survey on the most commonly used TPMs, classifying them by four categories:

(a) representation type: way of standing for the language, aiming at identifying implicit linguistic rules and common-sense knowledge that are not explicit in textual data;
b) Architectural model: how contexts are captured, whether sequentially (word after word) or non-sequentially (using a pre-defined syntactic or semantic structure);
c) Type of pre-training task: goal looked for during training. In supervised learning, one seeks a function capable of mapping input and output pairs; in unsupervised learning, one looks to obtain intrinsic knowledge from unclassified data; in self-supervised learning, there is a combination of the previous types, where the training method is based on supervised learning, but the data classification is generated automatically.
d) Extensions to the model: MPTs generally aim at universal representations of a language for generic applications. For specific applications, further enrichment of the model is desirable as multi-language, multimodal, or domain- or task-specific.

Qiu *et al.* (2020) also divide TPMs into two generations according to their goals. The first generation seeks good word mapping models, obtaining hierarchical word classification over a language model. They are context independent. Word2vec by Mikolov *et al.* (2013a), GloVe by Pennington, Socher, and Manning (2014) as well as CBow and Continuous Skip-Gram by Mikolov *et al.* (2013b) are examples. The second generation seeks to produce word vectors at the sentence level, considering the context in which the words are found. CoVe by McCann *et al.* (2017), ELMo by Peters *et al.* (2018), OpenAI GPT by Radford *et al.* (2018), and BERT by Devlin *et al.* (2019) are examples.

Given breadth of available models, Minaee *et al.* (2021) propose a five-step procedure for choosing an NLP neural network:

a) Selection of the MPT;
b) Adaptation to the problem domain;
c) Insertion of a layer adapted to the task;
d) Adjustment of weights to the task;
e) Compression of the model.

After analyzing more than 150 NLP-oriented models using more than 40 data sets, the authors conclude that no matter how much progress has been made, some issues remain challenging:

(a) lack of data for more complex tasks: although the amount of data collected over the years is expressive, tasks such as questions and answers with multi-step reasoning, text classification for documents with multiple languages, and text classification for long documents;
b) Common-sense knowledge models: the lack of models with common-sense knowledge limits the ability of neural networks to analyze, such as answering questions about the real world or dealing with incompleteness of information;
c) Memory-efficient models: Most modern models require large amounts of memory, which leads to the need for compression;

d) Low-effort learning: most Deep Learning models are trained through supervised learning. In practice, collecting and classifying data for a new domain is a complex and challenging task.

The advances of natural language processing tools are remarkable, both in diversity of implementations and spectrum of treatments undertaken, however, the representation of specific knowledge (treated to some extent by Minaee *et al.* (2021) as common-sense knowledge) still represents a challenge to be better addressed.

4 MULTIMODAL INFORMATION ARCHITECTURE: CONTRIBUTIONS TO THE DEVELOPMENT OF NLP

According to Kuroki Junior (2018), Multimodal Information Architecture - MIA - is defined as the construction and distinction of Architectural Worlds, through the assumption of Relational Models grouped by space-time contexts of correlated or uncorrelated Information States.

For the author, the term would be closely linked to Information Science, by its willingness to act in what Hjørland (2008) referred to as the strict sense of Knowledge Organization: description, indexing and classification of documents. An imposition of Order (by architecture) for both streams of concepts of Information, defined by Capurro and Hjørland (2007): an objective one, treating it as a thing (number of bits, for example) and a subjective one, which would depend on the interpretation of a cognitive agent. In both cases, say the authors, Information Science would focus on the phenomena of relevance and interpretation as basic aspects of the concept of information.

Kuroki Junior's (2018) proposal extends the traditional concept of Information Architecture by adding the concept of Mode given by Kress and Van Leeuween (2001) and Kress (2009), as any socially and culturally shaped resource for constructing meanings. For the authors, any Mode, including language (in the conception of written and spoken language and its possibilities) has both limitations and potentialities.

The expression of meanings and the consistency of a relational model among the various groupings formed must be marked by some measure of orderliness. The problem lies in situations in which the same premise may be considered true in one context, but false in another, and yet both contexts must coexist in the same informational model. In a simple and reduced illustrative way, the same NLP neural network should assume that a specific term (e.g., "system") has both positive and negative impact at the same time. Here is the issue with multimodality in information architectures: the cost of modeling every case in which all assumptions are true in all possible configurations exceeds the benefits found by this extreme individualization and granularization of problems. This is the Ockham's Razor dilemma also adopted by Kuroki Junior's MIA (2018), through two principles that keep intimate relation between them: economy and relevance. "Pluralitas non est ponenda sine necessitae" (plurality should not be put without necessity, represented by relevance) and "Frustra fit per plura quod potest fieri per pauciora" (it is fruitless to do with more what one can do with less, represented by economy).

In addressing the issue, Kuroki Junior (2018) uses modal logical structures, based on possibility and necessity operators according to Carnielli and Pizzi (2008) and Portner (2009). A proposition is possible if it is true in some configuration of a domain. A proposition is necessary if it is true in all configurations of a domain.

Information Science, through MIA, would act in the Knowledge Organization in the strict sense of Hjørland (2008), producing views or groupings of data that can more effectively express a domain or an information context to facilitate pattern recognition through neural

networks. In MIA, an architectural world is a context of relationships between subjects and objects, i.e., a set of semantic domains that can be shaped in multiple ways from the same set of subjects and objects.

The following items detail the five-step procedure proposed to obtain a new informational domain configuration aimed at NLP, as a phase prior to the data preprocessing performed in the development of artificial neural networks.

4.1 Identifying context entities

For NLP and MPTs, a context can be seen only as a group of texts grouped by linguistic, semantic, factual, common sense or any other similarity. This is not true for MIA. A context becomes an architectural space only when a subject's viewpoint of at least one object is considered. In contrast, multiple subjects may differently classify an object, just as a given sequence of text may express different meanings in different contexts. NLP neural networks aim to overcome this barrier by means of data volume which, according to Minaee *et al.* (2021), is restricted for more complex tasks. In this sense, the first intervention of MIA aims to define the subjects and objects of a context, where:

(a) SUBJECT is an entity endowed with the ability to produce and manipulate information;
b) OBJECT is an entity with signification potential, endowed with attributes that can be interpreted by subjects in a common way;
c) A CORRELATION occurs when a subject transforms an object by means of DEFINITION, COMPARISON, FUSION OR DECOMPOSITION and the product of this operation is accepted within the body of knowledge shared by the subjects that compose the context.

Once the subjects who figure in a context are defined, the way in which they manipulate objects determines the configuration of the observed moment. Kuroki Junior (2018) makes these various moments explicit by calling correlation the fundamental unit of connection between subjects and objects. Even if different subjects agree that a set of characteristics define an object, their correlations are distinct, subject to intrinsic differences not observable at the moment of analysis.

4.2 Identifying correlations between entities

For Kuroki Junior. (2018), relations connect instances of a context or the contexts themselves, and a correlation is a specific type of relation. A correlation is formed between a subject and an object in a given context. In an NLP approach in light of MIA, the proposed fundamental correlations are four:

(a) DEFINITION is a correlation performed by a subject that transforms the state of a being in a context to object, opening the possibility of adding other beings as attributes.
b) COMPARISON is only applicable to objects defined by a subject. Any level of comparison is done by analyzing the attributes assigned to different objects.
c) FUSION is the joining of two objects to form a third.
d) DECOMPOSITION is the operation opposite to fusion, where an object gives rise to two distinct objects.

Through these operations the impressions of the subjects acting in a context are collected regarding the characteristics of a group of objects called attributes. It is important to emphasize that only by adopting modal logic models can MIA treat the different *Modes* in which entities aggregate in different ways. For example, in a given *Technology Mode* the entity "system" would be an object with attributes [information, development, language], while in a *Policy Mode*, the same entity "system" would only be an attribute of the object "government". To accommodate both *Modes*, the informational context must be subdivided into smaller units, and then these units must have their relations (not correlations, which refer to subjects and objects) identified.

4.3 Domain distinction

In MIA applied to NLP, a DOMAIN is a group of object attributes that can be identified in a common way by different subjects through similar correlations. In this way, subjects and objects can make up several domains. Three possible ways of establishing domains are:

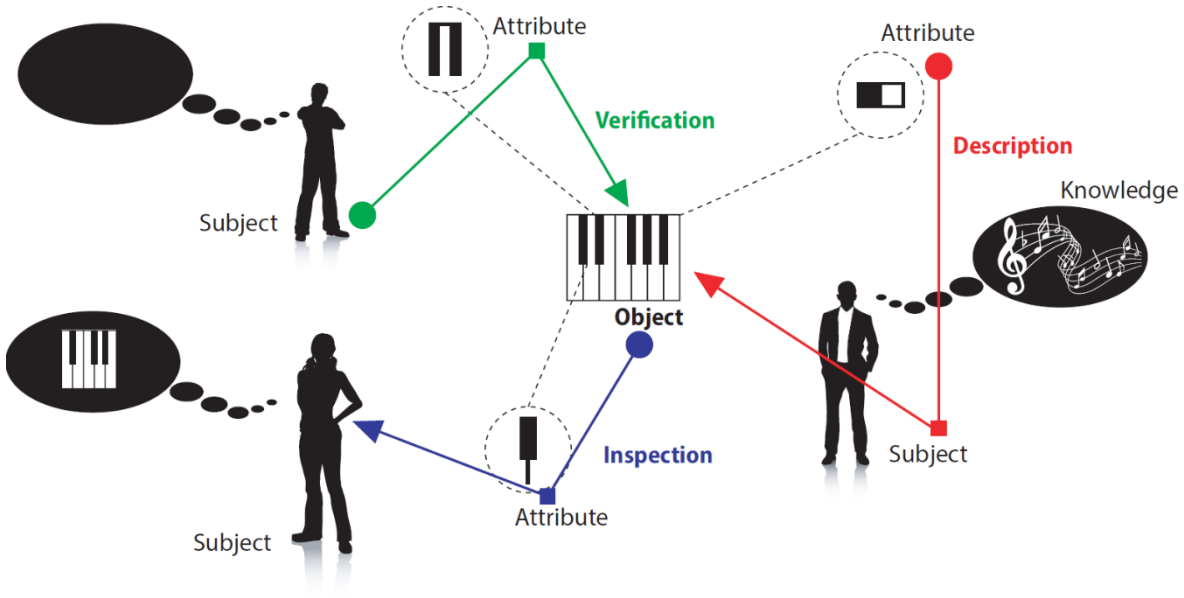
(a) Description: starting from a set of potential attributes, one checks their semantic reception by subjects to then find such attributes in certain objects, grouping them together;

b) Inspection: analyzing a set of objects, grouping them by common attributes and verifying the common recognition in a certain group of subjects;

c) Verification: by inquiring a certain group of subjects, one identifies attributes perceived in a common way by the individuals in this group and groups objects that contain these attributes.

A graphic representation of these forms can be seen in Figure 1 below

Figure 1. Ways of establishing and distinguishing domains



Source: Produced by the authors in May 2022.

The subdivisions from this step complete the information cycle of MIA: the scope of *"information states*" (from items 4.1 Identification of context entities and 4.2 *Identification of*

correlations between entities) "correlated or uncorrelated" (from item 4.3 Domain distinction) is defined.

4.4 Proposition of relations between domains

The first three operations aim at identifying entities, correlations and domains of a model addressing the informational set to be treated. The relationships between these domains give the architectural character of the proposal, in the sense of an imposition of economy and order. In order for an MIA to somehow impact a context or even a domain, some change in this informational space must be performed. This happens through relationships between domains.

For Kuroki Jr. (2018), relations are endowed with rules that constrain them. The primary definition of MIA uses modal logic to express relations. Three basic domain manipulation relations are proposed to change this domain or produce a new one:

(a) Identity: an identity relation is obtained when all objects in one domain can be found in another domain. It corresponds to the modal operator of necessity;
b) Proximity: a proximity relation is identified when part of the objects from one domain can be found in another domain. It corresponds to the possibility modal operator;
c) Incidental: Incidental relations are not always perceptible, with a certain degree of randomness in their incidences. The simplest way to define them would be as a second order relation.

As for the extent of the relations, the author uses logical modal structures cited by Carnielli and Pizzi (2008):

(a) Reflexive: a reflexive structure is identified when a proposed relation is applicable from a domain to itself;
b) Serial: a serial structure is identified when a proposed relation is applicable from one domain to another;
c) Symmetric: a symmetric structure is identified when a proposed relationship is mutually applicable between two domains;
d) Transitive: a transitive structure is identified when, assuming three domains [A, B, C], if A has the proposed relation with B, and B has the proposed relation with C, then A has the proposed relation with C;
e) Euclidean: a Euclidean structure is identified when a proposed relation is reflexive, symmetric, and transitive.

From the combination of type and extent we get the complete classification of a relation. For example: the relations between the domains A = {1,3,4}; B = {1, 3, 5} and C = {1,2,3,4,5} would be and serial identity of A to C and B to C; and symmetric proximity between A, B and C.

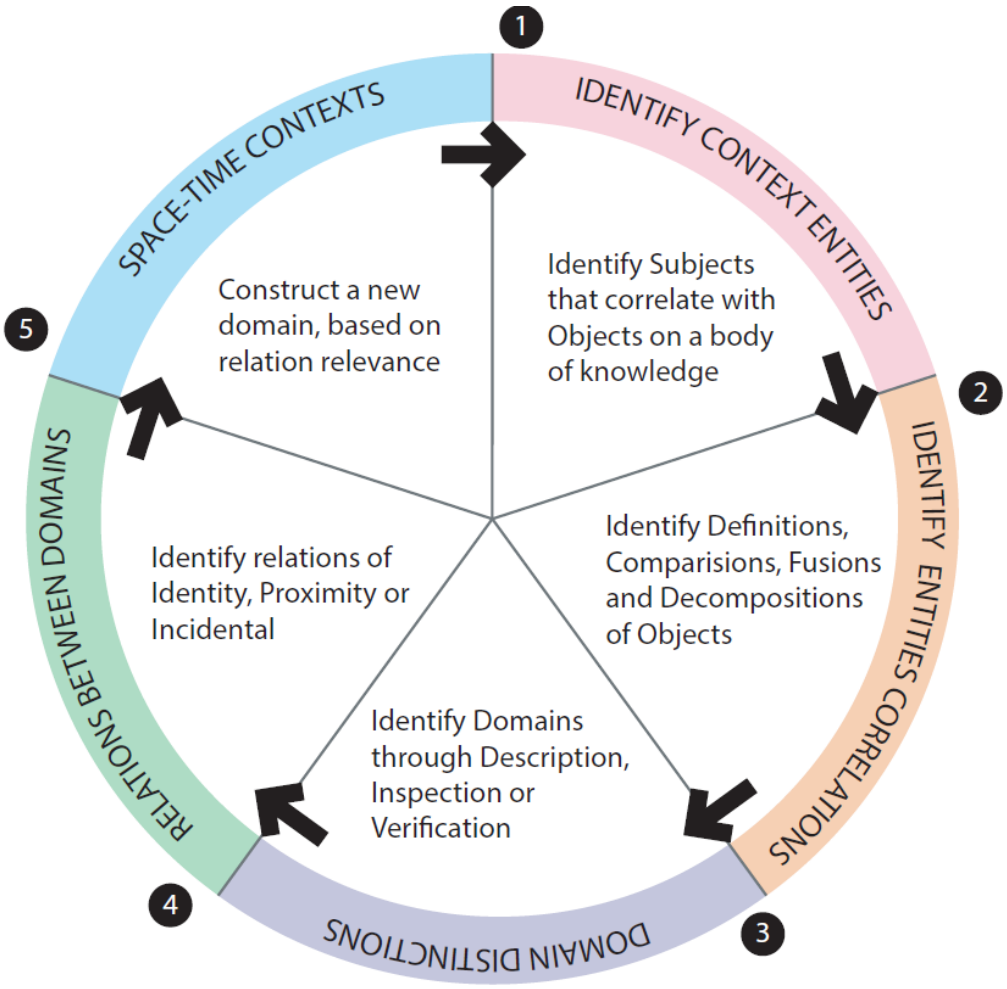
4.5 Grouping by space-time contexts

Applying all possible regulation to a domain or set of domains is not the goal of MIA. A measure of economy of relations must be taken into consideration, otherwise any configuration would tend to map objective reality as closely as possible. For Kuroki Junior (2018), space-time distinctions can be identified through deontic structures, which express a logic of obligations and permissions. These are distinguished from epistemic structures, which

deal with knowledge. The main difference lies in the impossibility for deontic structures to assume an immutable truth: they only consider the possibility of an occurrence. A simple example cited by Portner (2009) would be the moral rule "do not murder". Even though this is listed as necessary (it must exist in all possible contexts), murder still occurs.

All the rules listed so far address spatial issues of an information architecture: how comprehensive a model is with respect to the relationships, objects, and attributes it considers. The temporal issue becomes, in fact, a limiting factor for any static model, which leads to the need for a cyclic model, as per Figure 2.

Figure 2. MIA building cycle



Source: Produced by the authors in May 2022.

5 IMPLEMENTING A MULTIMODAL INFORMATION ARCHITECTURE

Following the proposed methodological path, an application of MIA to an NLP problem is suggested as an example. The selected situation refers to text classification. The difficulty lies both in the absence of sufficient data for learning and in the semantic scope of this data. In summary, it is a positive or negative trend analysis of a set of texts according to a legislation of incentives for research, development, and innovation. Each year, more than 10,000 texts are submitted, which can be classified into 16 categories of knowledge: Agribusiness, Food, Consumer Goods, Civil Construction, Pharmaceutical, Metallurgy, Mining, Furniture, Others, Paper and Pulp, Textile, Petrochemical, Mechanics and

Transportation, Electro-electronics, ICT and Telecommunications. So far, only the data from 2014 and 2015 have been analyzed and classified as "approved" or "failed".

5.1 Application of NLP to a data set not managed by MIA

The texts classified in the years 2014 and 2015 were submitted to training, validation, and testing in a neural network for text classification. The BERTimbau model by Souza, Nogueira, and Lotufo (2020) was used for this task, trained using the brWaC corpus by Filho *et al.* (2018), which has 3.5 million documents and 2.68 billion tokens. The model used separates the data into three parts: Training, Validation, and Test. For each set, two variables are observed. Loss represents the difference between the expected results and the results obtained by the machine. It is through the loss that one obtains the adjustments of the neural network's weights, which enables the advancement of learning throughout the experiment. Lower loss values indicate better learning of the network. Accuracy (acc) represents the percentage of correct answers obtained in each step of the experiment. This variable expresses how assertive the model is based on the data presented. To isolate the products of MIA from any interference from computer science techniques (enriching the database, changing the learning algorithm, increasing the scope of analysis), no improvement procedure will be applied to either the environment or the original data set, which ensures that any result is linked solely and exclusively to MIA.

Ten experiments with 20 training cycles were performed for the 2014, 2015, and 2014 and 2015 data together, yielding the following average results presented in Table 1:

Table 1. Average results of the experiments performed with untreated data

Variable	2014	2015	2014 e 2015
Training Loss	0,7087808	0,5627345	0,6463273
Training Accuracy	53,55%	76,38%	63,39%
Validation loss	0,6949488	0,5708822	0,6765008
Validation Accuracy	54,52%	74,14%	59,08%
Test loss	0,7416452	0,4740491	0,6142412
Accuracy in Testing	54,79%	77,57%	58,22%

Source: Produced by the authors in August 2022

There is a noticeable difference in the results from the 2014 and 2015 data, with the latter showing more assertive values. The percentage difference in test accuracy reaches 21.78% between the years separately. When joining both sets, the accuracy tends toward results closer to 2014, representing a decrease from the best result (2015) of 15.52%.

Starting from the same datasets provided, the objectives to be achieved through MIA-based data pre-processing will be:

(a) to find domain grouping configurations that increase the accuracy of the NLP algorithm without technical-computational interventions (based on source code changes);

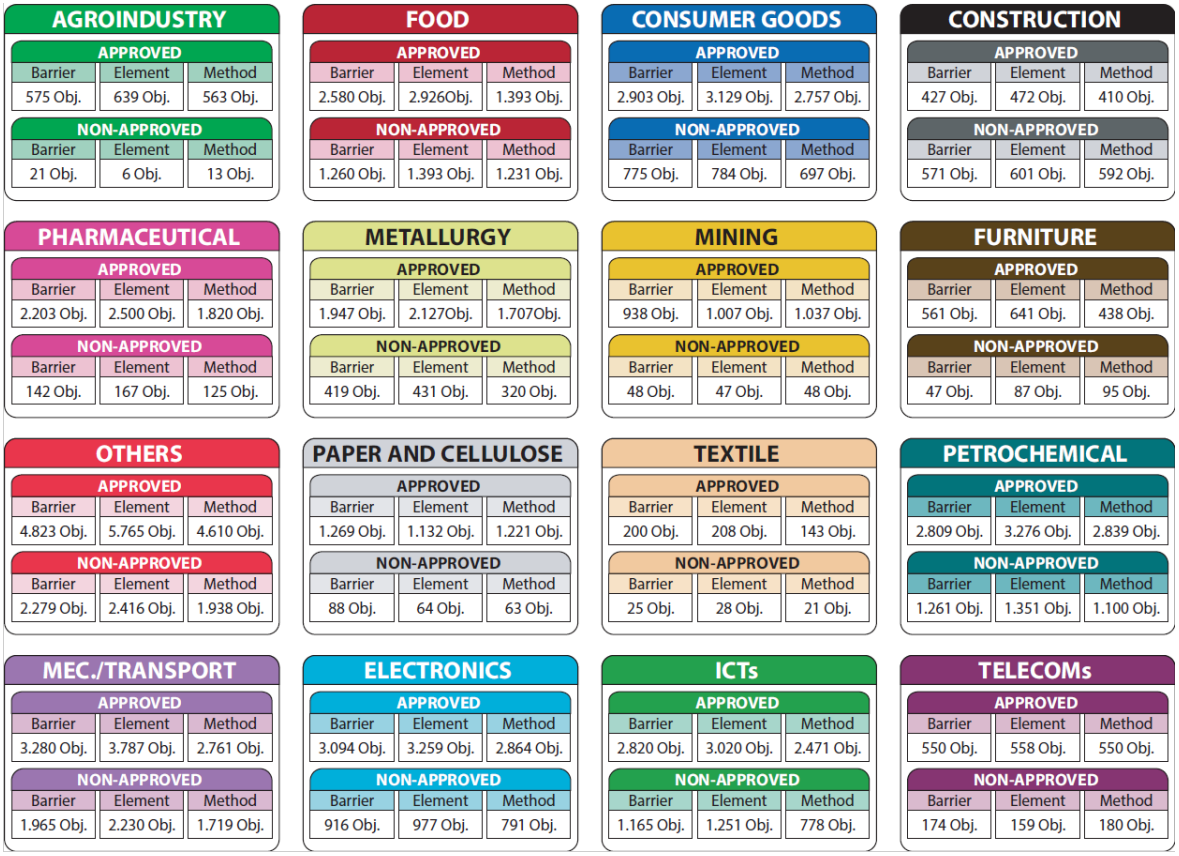
b) Identify domains that present data with higher or lower learning extraction potential.

5.1 Step 1: Identify context entities

The first step in transforming the informational environment in question is the identification of entities from each original context. The active subjects in the initial configuration analyze submitted texts in 16 knowledge areas. As the classification of these is given by means of several individuals (natural persons), applying the MIA of Kuroki Junior (2018), the set of knowledge expressed in each area can be considered a subject, obtaining, therefore, 16 subjects.

Reflexively, the corpus of objects is also defined by this distinction of subjects, given that there is a semantic agreement between the people who analyze the texts in each area (they are experts. The difference lies in the fact that each knowledge area has two binary value partitions - Approved or Not Approved - endowed with 3 semantic groupings - Innovative Element, Technological Barrier, and Method - totaling 96 semantic contexts. In this sense, given that objects are expressed through attributes, only nouns are eligible as entities, given their ability to absorb attributes through other semantic terms that modify them. Figure 3 shows the numbers obtained by context for the year 2015.

Figure 3. Objects identified by context - Base year 2015



Source: Produced by the authors in August 2022

5.2 Step 2: Identify correlations between entities

The second phase in producing an MIA is the identification of correlations between subjects and objects in the domain. In this sense, a technique called Inverse Document Frequency (IDF), originally proposed by Jones (1973), was used. This is a logarithmic measure of the relevance of a term in a set of documents: the lower the incidence of a given word in a text, the higher the probability of its relevance. The selection of entities in the model must

guarantee the maintenance of the relevance relation of the potential entity in the original untreated context. In this sense, 5 analysis steps are proposed:

a) Calculating the FID of each entity before each of the 96 semantic domains;
b) Obtaining the average FID of each entity considering the sum of the values of the 96 semantic domains;
c) Selecting the entities whose FID mean (calculated in the previous step) is greater than the standard deviation considering all the FID means;
d) Identification of objects by means of Definition, Comparison, Fusion, and Decomposition.

For the year 2015, 21,142 potential entities were identified. By applying the sequencing of steps "a", "b" and "c", this number decreases to 513. Among the potential entities, the semantic sets [method, methodology], [manufacturing, production], [needed, necessity], [productive, productivity], [end, final, result], [system, software] were identified, which present potential similarity. The attributes of such pairs were analyzed by means of comparison, in order to verify the need for defining two terms or merging them into one term. The results of the potential relationships in question were:

- Semantic set [method, methodology]: percentage of similarity between attributes of 1.69%. Definition relationship;
- Semantic set [manufacturing, production]: percent of similarity between attributes of 1.85%. Definition relationship;
- Semantic set [need, necessity]: percent of similarity between attributes of 5.26%. Definition relationship;
- Semantic set [productive, productivity]: percent of similarity between attributes of 8.47%. Definition relationship;
- Semantic set [end, final, result]: percent of similarity between attributes of 4.81%. Definition relationship;
- Semantic set [system, software]: percent of similarity between attributes of 1.96%. Definition relationship;

Thus, the 513 potential entities obtained through the four selection steps are recognized and correlated as domain objects.

5.3 Step 3: distinguish domains

Once the 16 subjects and 513 objects acting in the original domain have been identified, we proceed to change the configuration of this informational space by means of description, inspection, or verification. Given that the route to obtain this configuration started from the analysis of a set of texts by natural persons, the verification procedure becomes the most assertive choice for the distinction of the domains. The procedure is endowed with 3 steps:

(a) inquire a group of subjects;
b) Identify common attributes;
c) Groupings of objects that possess such attributes.

The first step was performed prior to the application of MIA, when the texts were analyzed by natural persons, that is, it was performed when the original data set was obtained, classified by knowledge area and approval/disapproval of each text individually. The second step was

performed in the item Identification of correlations between entities, where the 513 objects recognized by the 16 subjects of the initial context were obtained. For the third step, four procedures were performed

a) Calculation of the relevance of the objects for each of the 16 subjects: each area has two merit ratings (approved or disapproved) for three semantic contexts (Innovative Element, Technological Barrier, and Method), totaling six analysis parameters. The FID values of each object in the six analysis parameters are summed, obtaining the relevance value of the object for each of the 16 subjects. This value represents how relevant each object is to the subjects;
b) Index of subject adherence to the environment: equipped with the relevance value of the objects, the sum of these values represents how adherent the scope of knowledge of the subject is to the analyzed context;
c) Obtaining the dispersion index of the informational context: by calculating the standard deviation of the adherence indexes calculated in the previous procedure, it is possible to verify how uniform the informational environment is.
d) Designing domains based on the index of dispersion of the informational environment: the greater the dispersion index, the greater the quantity of clusters, observing the need for compensation between the adherence indices of the subjects to the environment.

The dispersion index calculated based on step "c" for the year 2015 had been 562.38, which divides the spectrum of values in Table 7 into 4 ranges:

- 0 to 562.38: composed of the subjects Metallurgy, Pharmaceuticals, Pulp and Paper, Mining, Furniture, Construction, Agribusiness, Telecommunications and Textile;
- 562.39 to 1,124.76: composed of the subjects Petrochemicals, Consumer Goods, ICTs, Food, and Electrical and Electronics;
- 1,124.77 to 1,687.14: composed of the subjects Mechanics and Transportation;
- 1,687.15 to 2,249.52: composed of the subject Others.

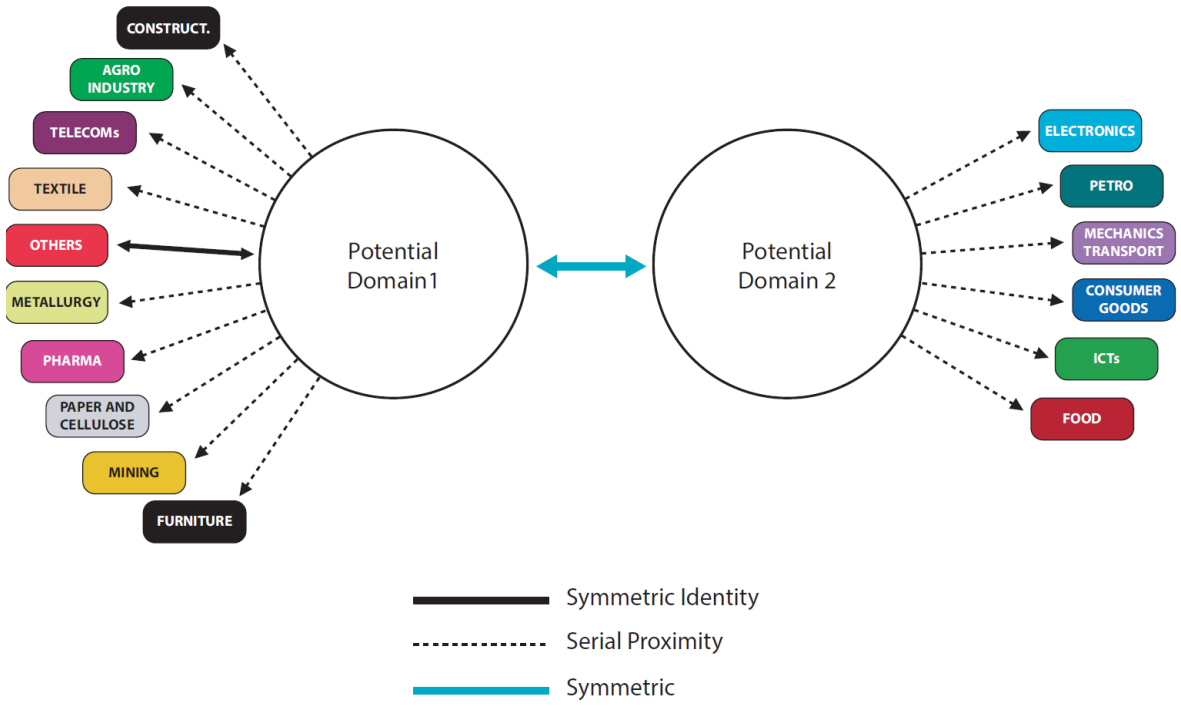
The lowest level of distinction/aggregation possible in the aforementioned informational context, defending the totality of the 16 subjects, is the division into two domains. Such division should consider a balance in the subject's rate of adherence to the informational context. In this sense, the groupings [1, 4] and [2, 3] present themselves as the most balanced, giving rise to:

- Potential domain 1, composed of the subjects Metallurgy, Pharmaceuticals, Pulp and Paper, Mining, Furniture, Civil Construction, Agribusiness, Telecommunications, Textile, and Other;
- Potential Domain 2, composed of the subjects Petrochemicals, Consumer Goods, ICTs, Food, Electrical and Electronic Products, and Mechanics and Transportation.

5.4 Step 4: Identify Relationships Between Domains

Given the two potential domains found in the previous step, we move on to establishing relationships between the knowledge areas and these domains, as well as between the domains themselves. In this sense, Figure 4 demonstrates the identity and proximity relationships that gave rise to the potential domains, as well as the extent of the relationships between these domains.

Figure 4. Relations between knowledge areas and potential domains - Base year 2015



Source: Produced by the authors in August 2022

It can be observed that, in its formation, only Potential Domain 1 has a Symmetric Identity relationship, since the knowledge area "Others" is the only one that has all the objects present in the domain. All relations identified for the formation of potential domains 1 and 2 are reflexive, since this operation starts from the identification of common objects, which necessarily requires the verification of the existence of this object in the domain itself, and only then proceed to verify the existence of said object in another domain.

Regarding the relations between the potential domains, there is a single symmetric relation [1,2], since all objects can be found in any possible configuration of both domains, which demonstrates that both coexist independently being micro-organizations of the original informational context.

5.5 Step 5: grouping by space-time contexts

As described in item 4.2 Identification of correlations between entities, the 2015 base year data was used to design the domain distribution obtained in item 4.3 Domain Distinction. In order to verify the temporal extension of the change in the proposed architecture over the years, the MIA cycle shown in Figure 3 was performed, along with the procedures described in items 4.1 to 4.4 for the base year 2014, obtaining a distinct configuration of domains.

For the step of identifying correlations between entities, the number of potential entities becomes 480 in 2014, to the detriment of the 513 obtained in 2015. The dispersion index of the informational context for 2014 was 798.84. Such a change resulted in a slightly different aggregation of subjects from the year 2015:

- 0 to 798.84: composed of the subjects Metallurgy, Pharmaceuticals, Pulp and Paper, Mining, Furniture, Construction, Agribusiness, Telecommunications, and Textile;

- 798.85 to 1,597.68: composed of the subjects Petrochemicals, Consumer Goods, ICTs, Food, and Electronics;
- 2,396.53 to 3,195.37: composed of the subjects Mechanics and Transportation and Others;

The three most significant changes are: the separation of the subjects Mechanics and Transportation and Others into two distinct ranges; reclassification of the subject Information and Communications Technology to the range below the context dispersion index; and the reordering of the aggregation ranges. Although the changes are apparently negligible, one has to consider the balance between the subjects' clustering indices. In this sense, 3 potential domains are proposed for the year 2014:

- Potential domain 3, composed of the subject Mechanics and Transportation and of part of the subjects that make up the first aggregation band of the original context for the year 2014, namely: Agribusiness, Furniture, Pulp and Paper, Pharmaceuticals, and ICTs;
- Potential domain 4, composed of the subject Others and the remaining part of the subjects that make up the first aggregation range of the original context for the year 2014, namely: Textile, Telecommunications, Construction, Mining and Metallurgy;
- Potential domain 5, composed of the totality of the subjects that make up the second aggregation band, namely: Chemicals and Petrochemicals, Consumer Goods, Electrical and Electronics, and Food.

We verify the high sensitivity of the problem to spatial-temporal separation: an MIA used in one year cannot be taken, at first, as applicable to a new temporal context. This premise is confirmed when the 2014 and 2015 data are analyzed together. The number of potential entities identified is 1,192. The dispersion index of the informational context has risen to 10,243.65, creating 3 different domains from those previously identified:

- Potential domain 6, composed of the subjects Mechanics and Transportation, Telecommunications, Civil Construction, Pulp and Paper, Pharmaceuticals, and Metallurgy;
- Potential domain 7, composed of the subjects Others, Textiles, Agroindustry, Furniture, Mining and Consumer Goods;
- Potential domain 8, composed of the Chemicals and Petrochemicals, Food, ICTs, and Electronics subjects.

6 APPLYING NLP WITH MIA PRE-TREATED DATA

Once the impossibility of producing a predictive model for the selected problem based on the indistinct selection of data has been identified and, equipped with the MIA products obtained through the steps of identifying context entities until the grouping by space-time contexts, we will proceed to validate the model obtained. For such an intent, the 2014 and 2015 data were split and concatenated according to the potential domains built and trained for 10 times, keeping the training conditions described in item 5.1 Application of NLP in a dataset not treated by MIA. The results obtained are presented in Table 2.

Table 2. Average results of experiments performed with MIA-treated data

Potential domain	Loss in training	Training Accuracy	Loss in Validation	Validation Accuracy	Loss under test	Accuracy in Testing
Potential Domain 1 (2015)	0,5296761	78,43%	0,5286451	80,19%	0,4946408	84,88%
Potential Domain 2 (2015)	0,5505137	75,77%	0,5717295	72,78%	0,5767502	72,65%
Potential domain 3 (2014)	0,7006512	55,88%	0,6701859	58,00%	0,6577451	58,70%
Potential domain 4 (2014)	0,7183891	55,41%	0,7043763	54,85%	0,6313299	54,98%
Potential domain 5 (2014)	0,7111632	51,85%	0,6945764	52,65%	0,7277233	52,80%
Potential domain 6 (2014 and 2015)	0,6629338	63,30%	0,6571880	63,40%	0,5833146	63,94%
Field Potential Domain 7 (2014 and 2015)	0,6799421	59,75%	0,6634957	56,19%	0,6887856	55,15%
Potential domain 8 (2014 and 2015)	0,6265331	67,11%	0,6602471	63,38%	0,6573988	61,58%

Source: Produced by the authors in October 2022

7 DISCUSSION OF RESULTS

There is variation in the loss and accuracy values after the treatment of the original information set and its separation into relevance domains. Some domains present an improvement in prediction accuracy, while others present a worsening in prediction accuracy.

The year 2015, used as the basis for explaining the procedures proposed in item 5, had its dataset divided into 2 potential domains. The initial results presented themselves as the most assertive in the untreated context. Table 3 shows the comparison between the average loss and accuracy values for the year dataset.

Table 3. Average results of the experiments performed with untreated data

Variable	2015	2015 – Domain 1	2015 – Domain 2
Training Loss	0,5627345	0,5296761	0,5505137
Training Accuracy	76,86%	78,43%	75,77%
Validation loss	0,5708822	0,5286451	0,5717295
Validation Accuracy	74,14%	80,19%	72,78%
Test loss	0,4740491	0,4946408	0,5767502
Accuracy in Testing	77,57%	84,88%	72,65%

Source: Produced by the authors in August 2022

It can be seen that the potential domain 1 presented a gain of 7.31% in test accuracy in opposition to the loss of 4.92% pointed out for the potential domain 2. The learning potential follows the same trends in both domains, pointing out that there is an improvement in the performance of the NLP network when using the subset of domain 1 data and a worsening for

domain 2. Starting from the same dataset, the MIA pre-processing identified subdivisions that have higher and lower learning extraction capability, demonstrated through the variation in accuracy and loss in the two sets.

Following MIA validation, the temporal issue was addressed by running experiments based on pre-treatment of data from the year 2014, as well as joining data from 2014 and 2015. Table 4 presents the comparison of results for the year 2014.

Table 4. Comparison of results - Base year 2014

Variable	2014	2014 – Domain 3	2014 – Domain 4	2014 – Domain 5
Training Loss	0,7087808	0,7006512	0,7183891	0,7111632
Training Accuracy	53,55%	55,88%	55,41%	51,85%
Validation loss	0,6949488	0,6701859	0,7043763	0,6945764
Validation Accuracy	54,52%	58,00%	54,85%	52,65%
Test loss	0,7416452	0,6577451	0,6313299	0,7277233
Accuracy in Testing	54,79%	58,70%	54,98%	52,80%

Source: Produced by the authors in October 2022

Throughout the MIA construction procedure for the year 2014, a reduction in the number of potential entities is observed compared to the year 2015 (513 to 480) and an increase in the dispersion index of the informational context (from 562.38 to 798.84). Such figures lead to the following considerations that guide the analysis:

(a) subjects who acted in the 2014 informational context recognized fewer entities as relevant objects, with a large variation in their environment adherence indices, i.e., there are subjects who have a high context adherence (the objects he recognizes are mostly in the relevant informational context), and others who have a low context adherence (their recognized objects are mostly not in the relevant informational context).

b) The relevant informational context to be treated had been more dispersed, requiring more subdivisions of the original context, going from 2 domains to 3.

Domain 3 showed a 4.01% improvement in test accuracy levels, a smaller gain than that recorded for domain 1 in 2015. Domain 4 remained virtually unchanged from the original context, with a slight improvement of 0.19% in test accuracy. Domain 5, in turn, showed a 1.99% decrease in test accuracy levels.

This reflects the high dispersion of data and the low adherence of the subjects to the relevant context: the data set with the best predisposition to learning becomes smaller and, even so, with a negligible gain.

Another situation analyzed in the experiments was the pooling of data from 2014 and 2015. The comparison between the results without pretreatment and with pretreatment is shown in Table 5.

Table 5. Comparison of results - Base years 2014 and 2015 combined

Variable	2014/2015	2014/2015 – Domain 6	2014/2015 – Domain 7	2014/2015 – Domain 8
Training Loss	0,6463273	0,6629338	0,6799421	0,6265331
Training Accuracy	63,39%	63,30%	59,75%	67,11%

Validation loss	0,6765008	0,6571880	0,6634957	0,6602471
Validation Accuracy	59,08%	63,40%	56,19%	63,38%
Test loss	0,6142412	0,5833146	0,6887856	0,6573988
Accuracy in Testing	58,22%	63,94%	55,15%	61,58%

Source: Produced by the authors in October 2022

As presented in 5.5 Step 5: grouping by space-time contexts, the number of potential entities identified was 1,192, however, the dispersion index of the informational context rose to 10,243.65. Again, we have a mismatch between how adherent the subjects' knowledge is to the relevant informational context. The gain in test accuracy was observed in domains 6 (5.72%) and 8 (3.36%) while for domain 7 a decrease of 3.07% was observed.

Of the 8 proposed domains, taking the test accuracy results as an analysis parameter, 4 (four) showed a gain, 3 (three) showed a loss, and 1 (one) maintained the previous levels, with a small increase. Based on this analysis, it is possible to identify the data sets that have more and less potential for learning extraction.

Table 6. Analysis of learning potential by knowledge area

Knowledge area	2014	2015	2014/2015	Potential
Agribusiness	1	1	-1	1
Foodstuffs	-1	-1	1	-1
Consumer goods	-1	-1	-1	-3
Civil construction	0	1	1	2
Electro-electronics	-1	-1	1	-1
Pharmaceutical	1	1	1	3
Mechanics and Transportation	1	-1	1	1
Metallurgical	0	1	1	2
Mining	0	1	-1	0
Furniture	1	1	-1	1
Paper and Cellulose	1	1	1	3
Chemical and Petrochemical	-1	-1	1	-1
TICs	1	-1	1	1
Telecommunications	0	1	1	2
Textile	0	1	-1	0
Others	0	1	-1	0

Source: Produced by the authors in October 2022

8 CONCLUSION

Through this article, we aimed to position Information Science as an integral part of the process of building artificial intelligence, figuring as a discipline prior to the formalization of neural network algorithms. The pre-processing of data provided by MIA can contribute to

increase the accuracy of predictions by simply rearranging the data provided, that is, by imposing a sense of dynamic organization according to the space-time treated.

In section 3 it was identified that the current stage of development of NLP provides a diverse range of algorithmic implementations, however, the most used training techniques (such as supervised learning) still require large volumes of classified data and improvements in specific knowledge or common-sense models (focused on questions about the real world) and with incomplete information.

In section 4, MIA, and its treatment of Modes of meaning expression were presented, following Kress and Van Leeuween (2001) and Kress (2009); through modal logical structures, according to Carnielli and Pizzi (2008) and Portner (2009). By combining the two schools of thought, it becomes possible to manage different semantics in the same informational context, a very common problem in NLP tasks. The MIA approach is based, among other principles, on economy and relevance to provide the best possible informational configuration. It uses a 5-step procedure to identify subjects and their correlations with objects, as well as the domains to which subjects and objects belong and the relations between these domains.

In section 5 the MIA product construction procedure is applied to a real problem of classifying texts coming from 16 knowledge areas. Eight subdomains were designed without any change in the original amount of data. Using a widely used NLP algorithm for the Brazilian Portuguese language, the results obtained from data treated by MIA were compared to those obtained without such treatment.

Although the observed values were numerically discrete from the point of view of prediction accuracy, there is room for improvement in most of the distinguished domains. Considering that no data enrichment procedure or improvement of the linguistic model was performed, it is plausible to conclude that MIA, by itself, indicated the best possible grouping of data in each temporal moment, based only on the records initially presented.

Finally, in this paper, the choice of the FID technique initially proposed by Jones (1973) to obtain correlations between subjects and objects in item 5.2 Step 2: identify correlations between entities, does not bind MIA to its use, and can be replaced by any other technique that provides a measure of object relevance for each subject. Investigation of other methods of obtaining such a level of relevance is encouraged.

REFERENCES

AREL, I; ROSE, D. C.; KARNOWSKI, T. P. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. **IEEE computational intelligence magazine**, *[S.l.]* v. 5, n. 4, p. 13-18, 2010. DOI: 10.1109/MCI.2010.938364. Access on: 9 Jan. 2023.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 3, 2015, San Diego, CA. **Anals** [...]. San Diego, CA, 2015. DOI: 10.48550/arXiv.1409.0473. Access on: 9 Jan. 2023.

BELLMAN, R. The theory of dynamic programming. **Bulletin of the American Mathematical Society**, Providence, RI, v. 60, n. 6, p. 503-515, 1954. DOI: 10.1090/S0002-9904-1954-09848-8. Access on: 9 Jan. 2023.

CAPURRO, R.; HJORLAND, B. O conceito de informação. **Perspectivas em ciência da informação**, Belo Horizonte, v. 12, n. 1, p. 148–207, 2007. DOI: 10.1590/S1413-99362007000100012. Access on: 9 Jan. 2023.

CARNIELLI, W.; PIZZI, C. **Modalities and multimodalities**. Springer Science & Business Media, 2008. 304p.

JONES, K. S. Index term weighting. **Information storage and retrieval**, Cambridge, UK, v. 9, n. 11, p. 619-633, 1973. DOI: 10.1016/0020-0271(73)90043-0. Access on: 9 Jan. 2023.

HJØRLAND, B. What is knowledge organization (ko)? Knowledge organization. **International journal devoted to concept theory, classification, indexing and knowledge representation**, *[S.l.]*, ERGON-Verlag GmbH, 2008. Available at: http://bit.ly/3vQG7Ry. Access on: 9 Jan. 2023.

HINTON, G. E.; OSINDERO, S.; TEH, Y. A fast learning algorithm for deep belief nets. **Neural computation**, Boston, MA, v. 18, n. 7, p. 1527-1554, 2006. DOI: 10.1162/neco.2006.18.7.1527. Access on: 9 Jan. 2023.

KRESS, G. What is mode? *In*: Jewitt, C. (ed.). **The Routledge Handbook of Multimodal Analysis**. London, UK, Routledge, 2009. 340 p.

KRESS, G.; VAN LEEUWEN, T. **Multimodal discourse**: The modes and media of contemporary communication. London: Hodder Arnold Publication, 2001. 142 p.

KUROKI JÚNIOR, G. H. **Sobre uma arquitetura da informação multimodal**: reflexões sobre uma proposta epistemológica. Dissertação (Mestrado) — Universidade de Brasília, 2018. DOI: 10.26512/2018.02.D.31920. Access on: 9 Jan. 2023.

DEVLIN, J.D. *et al.* Pre-training of deep bidirectional transformers for language understanding**.** *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, Minneapolis, MI. **Proceedings** […]. Minneapolis, MI, 2019. DOI: 10.48550/arXiv.1810.04805. Access on: 9 Jan. 2023.

MCCANN, B. *et al.* Learned in translation: Contextualized word vectors. *In:* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30, Long Beach, CA, **Proceedings** […]. Boston, MA, 2017. DOI: 10.48550/arXiv.1708.00107. Access on: 9 Jan. 2023.

MINAEE, S. *et al.* Deep learning-based text classification: a comprehensive review. **ACM Computing Surveys (CSUR)**, *[S.l.]*, v. 54, n. 3, p. 1-40, 2021. DOI: 10.1145/3439726. Acessado em: 9 Jan. 2023.

MIKOLOV, T. *et al.* Efficient estimation of word representations in vector space. *[S.l.]*, **arXiv preprint**, arXiv:1301.3781. DOI: 10.48550/arXiv.1301.3781, 2013a. Acessado em: 9 Jan. 2023.

MIKOLOV, T. *et al.* Distributed representations of words and phrases and their compositionality. *In:* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2013), Lake Tahoe, NV, **Proceedings** […]. Boston, MA, 2013b. v. 26. DOI: 10.48550/arXiv.1310.4546. Access on: 9 Jan. 2023.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. *In:* CONFERENCE ON EMPIRICAL METHODS IN NATURAL

LANGUAGE PROCESSING (EMNLP), Doha, Qatar. **Proceedings** […]. DOI: 10.3115/v1/D14-1162, 2014. Access on: 9 Jan. 2023.

PETERS, M. E. *et al.* Deep contextualized word representations. *In:* NAACL. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 1 (long papers), New Orleans, Louisiana, 2018. **Proceedings** […]. DOI: 10.18653/v1/N18-1202 2018. Access on: 9 Jan. 2023.

PORTNER, P. **Modality**. London: Oxford University Press, 2009. 320 p.

QIU, X. *et al.* Pre-trained models for natural language processing: A survey. **Science China Technological Sciences**, *[S.l.]*, v. 63, n. 10, p. 1872-1897, 2020. DOI: 10.1007/s11431-020-1647-3. Access on: 9 Jan. 2023.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. **OpenAI**, *[S.l.]*. v. 4, n. 19, 2018. Available at: http://bit.ly/3Xhzao1. Access on: 9 Jan. 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. *In:* CERRI, R., PRATI, R.C. (ed.) **Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science**, *[S.l.]*, v. 12319. Springer, Cham. p. 403-417. DOI: 10.1007/978-3-030-61377-8_28. Access on: 9 Jan. 2023.

VAN GIGCH, J. P.; MOIGNE, J. L. A paradigmatic approach to the discipline of information systems. **Behavioral Science**, *[S.l.]*, v. 34, n. 2, p. 128-147, 1989. DOI: 10.1002/bs.3830340203. Access on: 9 Jan. 2023.

VASWANI, A. *et al.* Attention is all you need. **Advances in neural information processing systems**, Long Beach, CA, v. 30, p.1-15, 2017. DOI: 10.48550/arXiv.1706.03762. Available at: https://arXiv:1706.03762. Access on: 9 Jan. 2023.

WAGNER FILHO, J. A. *et al.* The brWaC corpus: a new open resource for Brazilian Portuguese. *In:* Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), Miyazaki, Japan. **Proceedings** […]. Miyazaki, Japan: ELRA, 2018. Available at: http://bit.ly/3CBRzUR. Access on: 9 Jan. 2023.

WASON, R. Deep learning: Evolution and expansion. **Cognitive Systems Research**, *[S.l.]*, v. 52, p. 701-708, 2018. DOI: 10.1016/j.cogsys.2018.08.023. Access on: 9 Jan. 2023.