HAL open science

Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables

Martin Lentschat, Patrice Buche, Luc Menut, Romane Guari, Mathieu Roche

▶ **To cite this version:**

Martin Lentschat, Patrice Buche, Luc Menut, Romane Guari, Mathieu Roche. Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables. Data in Brief, 2022, 41, pp.108000. 10.1016/j.dib.2022.108000 . hal-03610433

HAL Id: hal-03610433

https://hal.science/hal-03610433v1

Submitted on 16 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Data in Brief 41 (2022) 108000

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables☆

Martin Lentschat[a,b,*], Patrice Buche[a], Luc Menut[a], Romane Guari[a,b], Mathieu Roche[b]

[a] UMR IATE, Univ Montpellier, INRAE, Institut Agro, 2 place Pierre Viala, Montpellier 34060, France
[b] UMR TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, 500 Rue Jean François Breton, Montpellier 34090, France

ARTICLE INFO

Article history:
Received 12 November 2021
Revised 16 February 2022
Accepted 23 February 2022
Available online 2 March 2022

Keywords:
Table extraction
Natural language processing
Ontological and terminological resource
Food packaging
Permeability
Component
Quantity

ABSTRACT

This dataset is dedicated to text mining and is composed of partial n-Ary relation instances concerning food packaging composition and gas permeability. It was created from 31 tables derived from 10 English-language scientific articles in html format from several international journals hosted on the ScienceDirect website. This dataset includes two sets of data: manual table annotation results and automatic data extraction results. The tables were first annotated by one annotator and cross-curated by three different annotators. The annotation task aimed to identify all table data dealing with packaging permeability measurements and compositions. An Ontological and Terminological Resource (OTR) was used for the annotation process. The annotation guidelines were drawn up through a collective iterative approach involving the annotators, and they may be accessed alongside the data. This dataset of n-Ary relations can be used in natural language processing (NLP) approaches implemented in experimental fields, especially for n-Ary relation extraction research. It can also be useful for training or evaluation of

☆ Dataset: n-Ary relations on Permeability and Composition of food packaging

* Corresponding author.
E-mail addresses: martin.lentschat@umontpellier.fr (M. Lentschat), patrice.buche@inrae.fr (P. Buche).

https://doi.org/10.1016/j.dib.2022.108000
2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

methods for the extraction of experimental data from tables and text in scientific documents, especially in experimental domains such as food packaging.

© 2022 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications Table

Subject	Data Science: Data Mining and Statistical Analysis
Specific subject area	Food packaging permeability and component
Type of data	.csv Table,.py source codes and annotated.html files
How data were acquired	These data are the result of manual and automatic annotations. Manual annotation was performed by one annotator before being cross-curated by three different annotators. Automatic annotation reproduces the method described in [1,4]. Ten html files of scientific articles from different journals were manually downloaded from ScienceDirect and reprocessed. The pre-treatment programs are available within the dataset. The annotations were performed according to the annotation guidelines, which may also be accessed alongside the data (see Data accessibility). Only the 31 annotated tables extracted from the 10 articles are available with their respective captions and titles, while the full texts are not presented for copyright reasons.
Data format	analysed
Parameters for data collection	TRANSMAT [1–3] Ontological and Terminological Resource (OTR) was used to drive the data annotation. Articles to annotate were randomly selected on the @Web platform[1] among the 121 documents referenced with an html version available on ScienceDirect. The @Web articles were originally selected by domain experts based on their ability to answer specific Competency Questions (CQs): CQ1: What are the food packaging constituents and associated quantities in the packaging composition? CQ2: What are the O2/CO2/H2O permeability values and units associated with the different food packagings studied in the article? CQ3: What controlled parameter values and units are associated with the O2/CO2/H2O permeability measurements?
Description of data collection	The files were selected from the ScienceDirect website according to the competency questions. They were manually downloaded in html format. The tables were automatically preprocessed and manually verified to standardize the table layouts in order to be compatible with the automatic extraction process.
Data source location	The data are available on the CIRAD Dataverse portal. The data have beenwere manually collected by the UMR TETIS joint research unit (Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France) and manually and automatically annotated by the UMR IATE joint research unit (Univ. Montpellier, INRAE, Institut Agro, Montpellier, France).
Data accessibility	Repository name: TRANSMAT data tables Data identification number: 10.18167/DVN1/GCZBC9 Direct URL to data and annotation guidelines: https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/GCZBC9

[1] https://ico.iate.inra.fr/atWeb/.

Value of the Data

- This dataset enhances resources available for table extraction and NLP in specialized fields, especially in the new generation and bio-sourced food packaging domain.
- This dataset can be useful for computer and data scientists conducting NLP and data mining tasks.
- This dataset can be used in the evaluation or training of methods for different tasks: data table extraction, named entity extraction in a specialized domain, including terminological variations, measurement units, complex numerical values and relation extraction (binary or n-Ary).
- The dataset is scalable since all of the processing codes are available, thereby facilitating its extension with additional documents.

1. Data Description

The dataset [9] primarily consists of two data archives (.targz). The `manual annotation` and `automatic annotation` archives contain the annotated tables in html format, which are grouped in directories by document names. A result table (.csv) file is also present in each archive and represents the annotated data in terms of partial n-Ary relations. An annotation manual (.pdf) provides instructions for annotators and explains the choices. An archive (.targz) contains the python codes to make document modifications required for the annotation process and the result tables in the archives. Each result file contains the partial n-Ary relation instances present in a table (an example of a n-Ary relation instance row is given in Table 1). Each data entry is described as a partial n-Ary relation with a set of features:

Document the article title from which the data was extracted and annotated;
DOI the Digital Object Identifier of the article;
Relation the relation concept associated with the n-Ary relation instance;
Result_Argument the result argument instance associated with a given relation instance;
Arguments the other argument instances associated with the relation instance;
Table the table title;

Table 1
Example of a row in the `resTable.csv` file from the `manual annotation` archive.

Features	Values
Document	A method for the measurement of the oxygen permeability and the development of edible films to reduce the rate of oxidative reactions in fresh foods
DOI	https://doi.org/10.1016/S0308-8146(02)00485-5
Relation	O2_Permeability_Relation
Result_Argument	{'O2_Permeability': [['6.8 ± 0.410^{9}', '1E9 Gram by reciprocal day by reciprocal pascal by reciprocal meter'], 'OP 10 9 g d -1 Pa -1 m -1']}
Arguments	{'Thickness': [['1.86 ± 0.0010^{5}', 'Meter'], 'Thickness 10 5 m'], 'Method': ['', ''], 'Number_Of_Repetitions': ['', ''], 'Temperature': ['', ''], 'Relative_Humidity': ['', ''], 'Partial_Pressure_Difference': ['', ''], 'Packaging': ['', ''], 'Partial_Pressure': ['', '']}
table*	Table 2
Caption	The SA content, the thickness and the OP values of edible films at 25 °C and 0% RH
Segment	['Results and discussion', 'The effect of antioxidants on the OP of edible films containing SA']

Caption the textual content of the table caption;
Segment the names of the section and sub-section in which the table is present.

The annotation guideline (`data tables annotation rules.pdf`) presents the annotation schemea and the instructions given to the annotators. These instructions are summarized in the next section along with the choices made. The annotation schemea defines several tags to annotate the data in the tables. These tags are embedded in the html file and used both in the manual and automatic annotation processes.

2. Experimental Design, Materials and Methods

The dataset consists of manual and automatic annotations of the scientific article tables. All pre-processing codes used are available in the dataset. This dataset is complementary to the work presented in [6] focused on the annotation of argument instances in document texts. This previous dataset was used to evaluate approaches dealing with the extraction of argument instances from scientific documents [7,8]. The work presented here includes a subset of documents associated with this previous work. This data paper is not devoted to argument instance annotation in texts but rather to the annotation of partial n-Ary relation instances in the document tables. These n-Ary relations are composed of several argument instances that are gathered in a relation concept instance representing the information on a more complex level. This explains the high complexity of the annotation task outlined in this paper.

2.1. Data (input)

The dataset includes 31 tables from 10 articles which were manually downloaded in html format from different journals on the ScienceDirect website. We discarded all other textual content so as to comply with the article copyrights (not all papers are from open access journals). Two features, i.e. DOIs and document names, are also provided to enable recovery of the original documents.

TRANSMAT [2,3] Ontological and Terminological Resource (OTR) represents concepts in the food packaging domain and the relations between them. This OTR is structured in two parts, a core ontology and a domain ontology (see Fig. 1 and [1] for further details). The up-core ontology is the representation of the n-Ary relations structure, defined as a relation concept linked to the arguments composing the relation. The down-core ontology includes concepts specific to the experimental fields, while the concept is categorized as symbolic or quantitative which are associated with measurement units used in experimental fields.

A symbolic concept is typically expressed as a word or phrase. A quantitative concept is associated with numerical values and linked to unit concepts corresponding to the concerned quantity (e.g. $O2_Permeability \rightarrow cm^3/(m^2.day.MPa^{-1})$). The domain ontology contains the concepts related to the field at hands, i.e. matter transfer in the food packaging domain. Each symbolic or quantitative concept has a terminological component represented as a set of *labels* (*preferred* and *alternatives*). Thus, the OTR conceptualizes the domain of interest on the basis of relation concepts, each of which is defined by a set of *labels* and a signature, where arguments are symbolic or quantitative concepts. *Labels* associated with argument concepts of the n-Ary relations of interest drive the recognition and selection of the tokens in the documents.

The tables were first automatically pre-processed to normalize the table layouts. This step was subject to manual verification. The aim was to produce html files tailored to the automatic annotation process described in [1,4]. The character encoding was also normalized in utf-8. An example of a table selected for annotation is presented in Fig. 2.

Four n-Ary relations to be annotated were selected in the OTR: permeability relations (carbon dioxide, oxygen and water permeability) and the food packaging composition (impact factor component). The annotation task was first conducted by one annotator to obtain a prelimi-

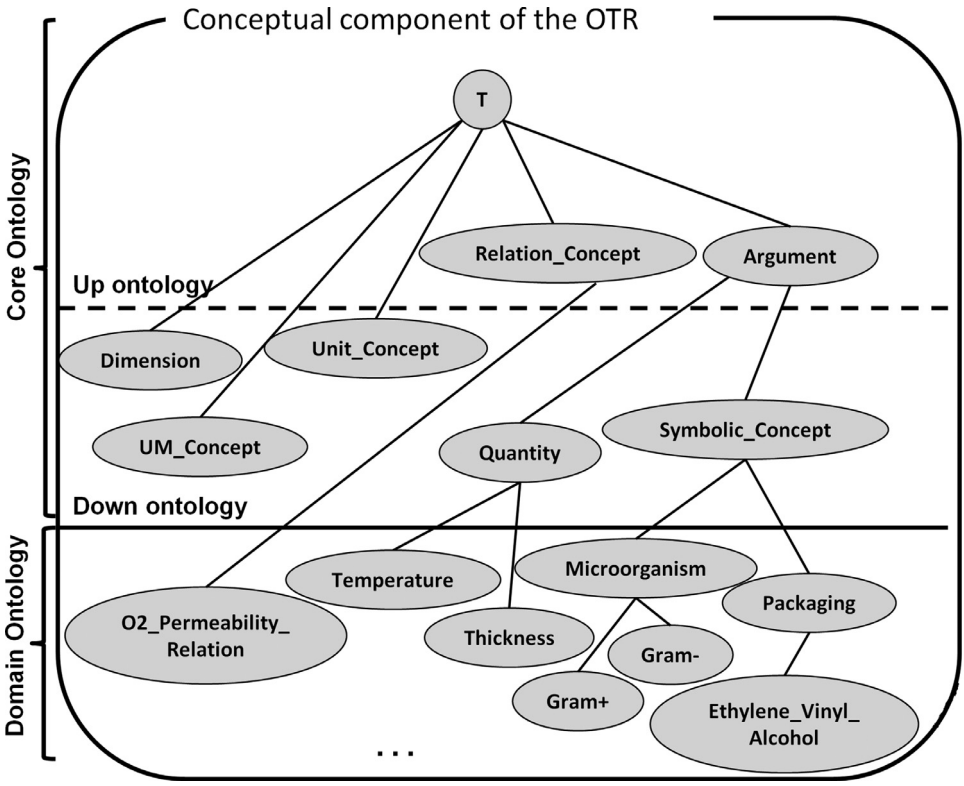

Fig. 1. Excerpt of the Transmat OTR structure.

Table 1 . Water vapour permeability (WVP) at 25 °C of PE films coated with chitosan compared to chitosan self standing films prepared with different casting solvents and plasticizers.

Sample	WVP × 10^{-13} (g/m s Pa) Δ RH 70%	WVP × 10^{-13} (g/m s Pa) Δ RH 45%	WVP × 10^{-13} (g/m s Pa) Δ RH 33%
PE	4.62 ± 0.73f	5.55 ± 0.23f	7.72 ± 2.58f
CS coated PE	12.37 ± 1.14f	6.67 ± 0.23f	7.88 ± 2.39f
PECSEinv	9.14 ± 1.09f	6.41 ± 3.28f	2.85 ± 0.34f
CSA	4161.31 ± 656.17a,b	2199.80 ± 1048.33d,e	25.71 ± 2.20f
CSE	4100.77 ± 588.88a,b	2884.37 ± 346.43b,c,d,e	38.71 ± 2.61f
CSAGLY	5410.08 ± 1543.67a	1905.39 ± 149.64e	26.14 ± 1.24f
CSEGLY	3481.46 ± 343.88b,c,d	2635.38 ± 414.28c,d,e	105.17 ± 6.57f

PE, polyethylene; CS coated PE, chitosan (CSE) coated polyethylene; PESCEinv, coating exposed to dry compartment; CSA, chitosan film prepared with aqueous acid solvent; CSE, chitosan film prepared with hydroalcoholic acid solvent; CSAGLY and CSEGLY, glycerol plasticized samples.

Different letters (a–f) indicate significant differences between formulations ($p < 0.05$).

Fig. 2. An example of a table selected for annotation. This table is extracted from [5]

```
<ri type="H2O␣Permeability_Relation" id="2">
  <ai type="Polyethylene" id="0"> </ai>
  <ai type="H2O␣Permeability" id="2"> </ai>
  <aii type="Temperature"> </aii>
  <aii type="Relative_Humidity"> </aii>
</ri>
<td class="align-left">
  <ai type="Polyethylene" id="0"> PE </ai>
</td>
<td class="align-char">
  <ai type="H2O␣Permeability" id="0"> 4.62 +- 0.73 </ai>
</td>
<td class="align-char">
  <ai type="H2O␣Permeability" id="1"> 5.55 +- 0.23 </ai>
</td>
<td class="align-char">
  <ai type="H2O␣Permeability" id="2"> 7.72 +- 2.58 </ai>
</td>
```

Fig. 3. An example of an annotated row.

nary annotation. This task involved the identification of all arguments and relation instances of the four targeted n-Ary relations. These relations consisted of different arguments that could be identified in the table rows and categorized as symbolic or quantitative. The final version of the table annotation was produced via an iterative process involving four annotators.

2.2. Manual Annotation

The table headers were annotated to distinguish numerical and symbolic columns. The tag *'qc'* stands for *'quantity concept'* and annotates the presence of a quantity concept. Otherwise, the tag *sc* (i.e. *symbolic concept*) is selected. These tags are supplemented by a *type* attribute used to specify the concept of the OTR represented in the table column. Selected OTR concepts must be n-Ary relation arguments such as Packaging or Thickness. The *qc* tags also specify the measure unit and a possible exponent for the numerical value, respectively with the *unit* and *exponent* attributes.

The n-Ary relations identified in the table were indicated by *rc* tags in the caption. Its *type* attribute specifies the OTR relation concept.

Rows were annotated to describe the n-Ary relation instances in the table. Each relation instance in a row is indicated with a *ri* tag. This tag has a *type* attribute, the previously identified relation concept, and a *id* attribute used to distinguish instances of different relations. Argument instances in the row are indicated by *ai* tags. This tag surrounds the text composing the instance (e.g. '*<ai type = "O2Permeability" id="0" >* 5.2 ± 0.2 *</ai>*'). It also presents *type* and *id* attributes to specify the argument concept and instance. A third tag category, i.e. *aii*, is used to indicate the presence of an implicit argument instance belonging to the relation instance. An implicit argument instance is not present in the table row but instead in its caption or header (e.g. a control parameter shared by all data in the table is only reported once in the caption instead of constituting an additional column). The *aii* tags also present *type* and *id* attributes. As the *aii* tags concern argument instances present in the title or caption of the table, they are not

Table 2
Global comparison of manual and automatic annotations.

tag	Manual	Automatic
sc	48	50
qc	172	170
rc	73	86
ri	828	925
ai	1273	1454
aii	717	∅
TOTAL	3111	2685

sc: symbolic concept; *qc*: quantitative concept; *rc*: relation concept; *ri*: relation instance; *ai*: argument instance; *aii*: implicit argument instance

the focus of automatic annotation. Fig. 3 is an example of manual annotation associated with the first row of the table in Fig. 2.

The manual annotation was then cross-curated by three different annotators as compared to the first one. The manual and automatic annotations are compared in Table 2.

2.3. Automatic annotation

The method and associated algorithms described in [1,4] were used to automatically annotate tables after the preprocessing steps. This method involves the use of an OTR. It was selected for its compliance with the different data formats present in the tables and its compatibility with the OTR used.

The automatic table annotation method involves a multi-step process similar to that of the manual annotation described above. Tags generated by automatic annotation correspond to those obtained by manual annotation. First, the table headers are automatically annotated to distinguish between numeric and symbolic columns based on the cell contents. Each column has the same type as the majority of its cells. A cell is numeric if it contains a measure unit of a majority of number tokens. Otherwise the cell is symbolic.

Then the OTR concept to be associated with each column is automatically identified using similarity scores in order to match ontology concepts with the recognized entities. The concept of a numerical column is determined by proximity measures between the measure units in the column and the measures units of quantitative concepts in the OTR. The concept of a symbolic column is determined using cosine similarity [12] of word vectors [10] between the column words and the OTR concept labels.

The relations in the table are then identified based on the proportion of quantitative and symbolic concepts that belong to each relation signature in the OTR. Selected relations have the highest proportion of arguments in the table. They are annotated in the table caption with *rc* tags associated with the *type* attribute to specify the associated OTR relation concept.

The automatic annotation algorithm finally annotates the table rows with argument and relation instances. Argument instances are identified based on the relation signatures, i.e. arguments that belong to the relations identified in the table. A specific concept for a symbolic argument instance is chosen in the set of sub-concepts of the argument based on term similarity. Quantitative argument instances are defined by the measure unit of the concept identified for the column. An *ai* tag indicates the argument instance in the cell. It presents an *id* attribute for single identification purposes.

Relation instances in rows are then annotated using one, or more, of the relations identified in the table according to the relation description in the OTR and the argument instances in the row. They are annotated with *ri* tags associated with the rows and containing a *type* attribute to indicate the relation concept in the OTR, and an *id* attribute to just identify relation instances. Argument instances composing the relation are also presented with *ai* tags, repeating the *ai* tags in the cells (see Fig. 3).

Table 3
Result of the automatic annotation method.

tag	recall	precision	F1-score
sc	.81	.38	.52
qc	.75	.62	.68
rc	.79	.67	.73
ri	.77	.69	.73
ai	.75	.66	.70

2.4. Results

The automatic annotation process was evaluated on a corpus of 31 tables. The overall automatic annotation results are compared to the manual annotations in Table 2. Tags automatically attributed to the columns (i.e. *sc* and *qc*), relation concepts (i.e. *rc*) and cells (i.e. *ri* and *ai*) were compared to those obtained during the manual annotation process.

Table 3 presents the an evaluation of the automatic annotation in terms of recall, precision and F1-score [11]. The results are consistent and comparable to those obtained in state-of-the-art studies using the same algorithm [1,4]. The precision on the *sc* tags precision is low (i.e..38). This is due to the numerous columns that containing a term associated with a symbolic concept, and thus identified as such. As these columns seldom contain argument instances associated with n-Ary relations, the error propagation to identify relation concept instances is low. Consequently, the low *sc* tag precision score is not a major issue. Errors in the *qc* tag annotation are mostly due to measure unit recognition. The *ri* tag annotations are satisfactory, especially since they mostly rely on the identification of *sc* and *qc* tags. Similarly, *ri* and *ai* tag annotations rely on correct annotation of previous tags, thus the results of both steps are close.

Ethics Statement

Out of scope.

Declaration of Competing Interest

The authors declare that they have no financial or personal interests that could influence the work reported in this paper.

CRediT Author Statement

Martin Lentschat: Resources, Investigation, Validation, Data curation, Writing – original draft, Visualization; **Patrice Buche:** Investigation, Data curation, Validation, Writing – review & editing, Supervision; **Luc Menut:** Investigation, Validation, Data curation; **Romane Guari:** Methodology, Software, Investigation, Resources, Data curation; **Mathieu Roche:** Writing – review & editing, Supervision.

Acknowledgments

The TEXT4LOD project has received funding from the IDEX/I-SITE MUSE[2] Univ. Montpellier (France).

[2] https://muse.edu.umontpellier.fr/en/muse-i-site/.

References

[1] P. Buche, J. Dibie-Barthélemy, L. Ibanescu, L. Soler, Fuzzy web data tables integration guided by an ontological and terminological resource, IEEE Trans. Knowl. Data Eng 25 (4) (2013) 805–819.

[2] V. Guillard, P. Buche, L. Menut, S. Dervaux, Matter Transfer Ontology, 2018, 10.15454/NK24ID

[3] V. Guillard, O. Couvert, V. Stahl, P. Buche, A. Hanin, J. Dibie-Barthélemy, Map-opt: A software for supporting decision-making in the field of modified atmosphere packaging of fresh non respiring foods, Packaging Res. De Gruyter 2 (1) (2017) 28–47.

[4] G. Hignette, P. Buche, J. Dibie-Barthélemy, O. Haemmerlé, Fuzzy annotation of web data tables driven by a domain ontology, in: European Semantic Web Conference, Springer, 2009, pp. 638–653.

[5] M. Kurek, M. Šćetar, A. Voilley, K. Galić, F. Debeaufort, Barrier properties of chitosan coated polyethylene, J. Membr. Sci. 403-404 (2012) 162–168, doi:10.1016/j.memsci.2012.02.037.

[6] M. Lentschat, P. Buche, J. Dibie-Barthelemy, L. Menut, M. Roche, Food packaging permeability and composition dataset dedicated to text-mining, Data in Brief 36 (2021) 107135, doi:10.1016/j.dib.2021.107135.

[7] M. Lentschat, P. Buche, J. Dibie-Barthelemy, M. Roche, Scipure: a new representation of textual data for entity identification from scientific publications, in: Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics, 2020, pp. 220–226.

[8] M. Lentschat, P. Buche, J. Dibie-Barthelemy, M. Roche, Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications, Int. J. Intell. Inf. Database Syst. 15 (1) (2022) 78–103, doi:10.1504/IJIIDS.2021.10042229.

[9] M. Lentschat, P. Buche, L. Menut, R. Guari, TRANSMAT tables data, 2021b, 10.18167/DVN1/GCZBC9

[10] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[11] C. van Rijsbergen, Information retrieval, 2nd edbutterworths, 1979,

[12] P. Sitikhu, K. Pahi, P. Thapa, S. Shakya, A comparison of semantic similarity methods for maximum human interpretability, in: 2019 artificial intelligence for transforming business and society (AITB), volume 1, IEEE, 2019, pp. 1–4.