applied sciences

Article

SPUCL (Scientific Publication Classifier): A Human-Readable Labelling System for Scientific Publications

Noemi Scarpato [1,*], Alessandra Pieroni [2] and Michela Montorsi [1]

[1] Departement of Human Sciences and Promotion of the Quality of Life, San Raffaele Roma Open University, 00166 Rome, Italy; michela.montorsi@uniroma5.it
[2] Agency for Digital Italy (AgID), Smart Cities Service Via Liszt 21, 00144 Rome, Italy; alessandra.pieroni@agid.gov.it
* Correspondence: noemi.scarpato@uniroma5.it

Abstract: To assess critically the scientific literature is a very challenging task; in general it requires analysing a lot of documents to define the state-of-the-art of a research field and classifying them. The documents classifier systems have tried to address this problem by different techniques such as probabilistic, machine learning and neural networks models. One of the most popular document classification approaches is the LDA (Latent Dirichlet Allocation), a probabilistic topic model. One of the main issues of the LDA approach is that the retrieved topics are a collection of terms with their probabilities and it does not have a human-readable form. This paper defines an approach to make LDA topics comprehensible for humans by the exploitation of the Word2Vec approach.

Keywords: document classification; human-readable labelling; Word2Vec

Citation: Scarpato, N.; Pieroni, A.; Montorsi, M. SPUCL (Scientific Publication Classifier): A Human-Readable Labelling System for Scientific Publications. *Appl. Sci.* **2021**, *11*, 9154. https://doi.org/10.3390/app11199154

Academic Editors: Teen-Hang Meen and Chun-Yen Chang

Received: 23 July 2021
Accepted: 28 September 2021
Published: 1 October 2021

1. Introduction

Usually, for each research field it is possible to retrieve a very large collection of related papers that define the state-of-art of an argument, often the relationship between the searched argument and the retrieved papers have a different degree of correlation dependent on the content of the retrieved articles.

Organising the retrieved collection of papers in relation of theirs content and degree of correlation is not a trivial task.

In literature, many different approaches to document classification have been defined [1]. Some of these approaches require a training phase (supervised approaches) and need a set of annotated documents to train the system [2]; some others instead do not require the training phase and attempt to classify documents in an unsupervised way (unsupervised approaches) [3].

Although several unsupervised systems have been defined to classify different types of documents, all these systems are unable to assign a meaningful label to these documents.

The traditional unsupervised approaches, like Latent Dirichlet Allocation (LDA from now) [4], realise a document clusterization and extract a list of topics composed by a list of more relevant (frequency-based approach) terms and their related weights.

The supervised classification methods, instead, make possible the identification of a predetermined class or label related to a document but need to be trained by a set of labelled documents.

The key idea of this work is to adopt a hybrid strategy to achieve best performance in document classification exploiting both supervised, semi-supervised and unsupervised techniques in the same approach.

To reach this aim we have defined and implemented the Scientific Publication Classifier (SPUCL from now) system, a semi-supervised document classifier able to assign a human-readable label to classified documents. SPUCL aims to assess critically the scientific literature.

We tested the SPUCL system by adopting it in a practical research task: classifying papers reporting application of analytical DNA methods to the authenticity testing of complex food products.

Our approach performs unsupervised pre-training by using LDA, followed by a fine-tuning step realised through Word2Vec and a labelling step performed with the application of the cosine similarity metric.

SPUCL exploits the powerful of Neural Networks that have found new interest both for computing power of current electronic systems and for new technologies [5–7].

The following sections will be described the background of document classification from statistical methods to deep learning techniques, the architecture of SPUCL classifier. Following we will present an application of SPUCL system, in this esperiment SPUCL classifies a set of 119 scientific papers in the food analytic research field, the discussion of the achieved results and the conclusions.

1.1. Background

Recently, many attempts have been provided to realise collaborative environment for information management via artificial intelligent technologies [8]; these attempts should be applied in very different fields such as Cultural Heritage [9], Medicine [10–12], IoT [13,14], Internet of Vehicles [15,16], Cyber Security [17] and Smart Cities [18].

One of the main tasks in information management is document management, which makes it possible to extract relevant information from a set of documents, in particular scientific papers information management (see Section 1.1.1) makes it possible to extract relevant information from a set of scientific papers.

In literature, many approaches for documents management such as clustering, classification, Topic Modelling and the most recently neural networks approaches (see Section 1.1.1) have been defined.

All of these approaches are artificial intelligence tasks and in particular they are Natural Language Processing tasks (NLP from now).

The document clustering is an unsupervised model able to analyse a set of documents and to divide them into a set of clusters composed by similar documents (see Section 1.1.2). The document classification is a supervised method able to assign to each document a predefined class (see Section 1.1.3). The Topic Modelling is an approach able to define topics related to the analysed documents (see Section 1.1.4). The Word Embedding is a neural network approach based on a vector representation with variable length and it is able to capture the semantic and syntactic correlation between the words contained in a document (see Section 1.1.5).

In our work we aim to address the goal of human-readable labelling of scientific papers exploiting the above mentioned approaches. In the following, further details about scientific paper information management and relative approaches will be provided.

1.1.1. Scientific Papers Information Management

Scientific papers information management is a challenging task and it can be divided into many different sub-tasks such as classification, clustering, labelling and summarization.

There are many different research fields; among these, the biomedical field is one of those characterised by the highest number of articles published by topic and which has the highest number of citations and cross-references and for these reasons the document management of this kind of documents is a very challenging task. In this paper we present an experiment based on a corpus composed by a set of scientific papers related to DNA methods for food integrity analysis. Food safety and food quality have increasingly come to the forefront of consumer concerns, industry strategies and government policy initiatives.

In this context DNA technologies represent a useful tool in food inspection and regulation.

In fact, molecular markers are of extreme importance to traceability since they allow the outcome of a given item to be monitored at each stage in the food chain.

The aims of the considered experiment have been to understand the current state-of-the-art of know-how and methodologies.

Detection of fraud, especially in meat products, is important not only for economic problem but also for health, religious and ethical reasons.

In recent decades, the main DNA methods are based on the detection of species-specific DNA sequences. Most of these methods rely on the polymerase chain reaction (PCR) technique for its specificity, sensitivity, simplicity and rapidity, allowing the identification of species of origin even in complex and processed foods.

In our experiment (see Section 2.4) we aim to classify papers in relation to DNA methods and food analysed.

In literature, many works have been provided about this topic, for instance in [19] authors provide a method for labelling hierarchical clusters of scientific articles; differently from our approach authors define a label for a cluster of documents and not for a single paper. Furthermore, the label is defined as a set of terms extracted from the cluster and not as a complete sentence.

In [20] authors provide a large manually annotated corpus for scientific papers and a summarization method. The objective of this article is the summarization of the scientific papers; to achieve this aim the authors exploit both abstract text and references.

In [21] authors provide a classification approach for biomedical literature; in [22] authors provide an approach based on feature selection for medical full text classification.

In Table 1 the comparison between these approaches and SPUCL is shown.

Table 1. Comparison between SPUCL and the other approaches to the document classification.

			Approaches		
Feature	**Peganova** [19]	**Yasunaga** [20]	**Simon** [21]	**Goncalves** [22]	**SPUCL**
NLP task	Clustering, Labelling	Summariza-tion	Classification	Classification	Classification, Labelling, Topic Modelling
Input	Full text	Abstract text, References	Abstract text	Full text, List of features	Abstract text, Full text
Output	Labelled clusters list	Summaries list	Ranked paper list	Ranked paper list	Labelled paper list with human readable labels

1.1.2. Document Clusterization

Document Clusterization is a NLP) task able to divide documents into homogeneous groups based on meaning. The methodologies for document clusterization do not require a training phase and for this reason they are called unsupervised.

There are many techniques for document clustering [23]; the most popular is the K-means algorithm that was first published in 1955. The K-means algorithm uses a vector representation of the documents, in which each document is represented as a vector of features. The vector is composed by the words contained in the document. To improve the performance of the algorithm it is necessary to carry out some preprocessing such as the removal of stop-words. In addition, the weight of each word can be calculated using the formula TF–IDF (term frequency–inverse document frequency) and then the vector will be composed of the words in the document and weighed for their frequency.

1.1.3. Document Classification

The document classification is a task of the NLP that has the purpose of associating to each document analysed one or more classes, previously defined.

The document classification is a supervised method, unlike the clustering algorithms mentioned in the previous section, which means that the classification algorithm must be trained on a corpus of documents already classified.

In the literature there are several approaches of document classification such as Bayesian classifier [24], Decision Tree [25], K-nearest neighbor(KNN) [26], Support Vector Machines (SVMs) [27,28] and Neural Networks [29–31].

1.1.4. Topic Modelling

As mentioned above, LDA [4] is one of the most popular Topic Modelling approaches. Topic Modelling is an approach to analyse a corpus of unlabelled documents and define a set of topics related to these documents, in order to define a set of clusters of similar documents.

LDA (Latent Dirichlet Allocation) [4] is a Topic Modelling approach that takes a collection of unlabelled documents and attempts to find the topics in this collection.

In Figure 1 it is shown the LDA Model architecture.

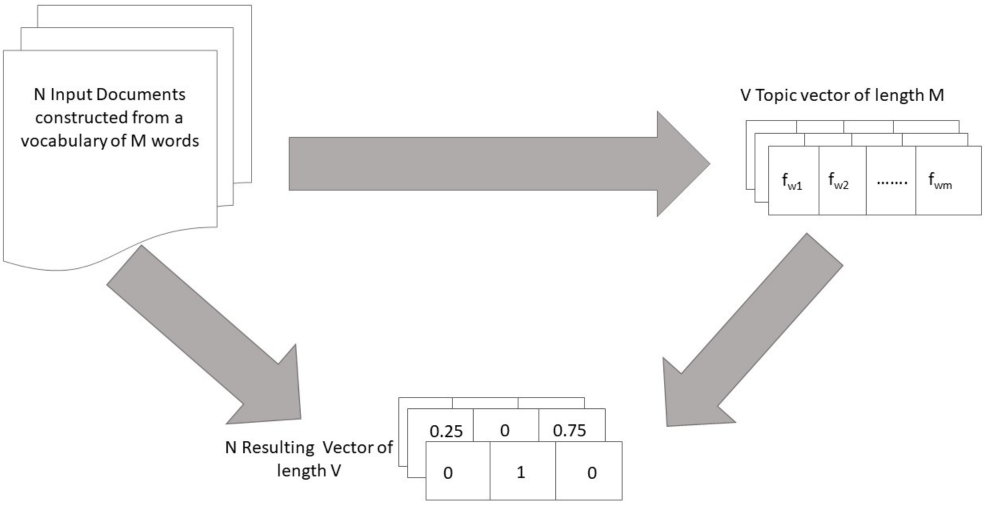

N Input Documents constructed from a vocabulary of M words
V Topic vector of length M
f_w1
f_w2
....
f_wm
N Resulting Vector of length V
0.25
0
0.75
0
1
0

Figure 1. LDA Model architecture.

In LDA each document is represented as a bag-of-word, i.e., a fixed-length vector with length equal to the vocabulary size. Each dimension of this vector corresponds to the count or occurrence of a word in a document. In order to enhance the performance of LDA some preprocessing is needed, such as stemming, lemmatization and stop-words identification; in this way the bag-of-words will be composed only by relevant terms.

The adoption of a bag-of-words model makes LDA prone to the same shortcomings of the other unsupervised models, i.e., not taking care of structure of the document or local interaction between words. In LDA it is assumed that word usage is correlated with topic occurrence, as documents on similar topics tend to use a similar sub-vocabulary; the resulting clusters of documents can be interpreted as discussing different 'topics'. The final result of a Topic Modelling performed by the LDA approach is a set of topics and related documents; for each document there is a vector of length equal to number of topics, containing the probability of relationship with respect to the topic. In the model settings it is possible to define the number of desired topics; in addition, it is necessary to manually define a human-readable label for each topic vector.

1.1.5. Word Embedding

Word Embedding [32] is a word representation based on a fixed-length vectors representation that can represent how words are semantically related to each other. The Word Embedding vectors do not contain all the words of the vocabulary coming from the starting corpus but only those that are semantically correlated with the represented word. In a Word Embedding representation two similar vectors, that are close in the vectorial space built from the training corpus, are the representation of two words semantically correlated. All the approaches based on Word Embedding in fact are based on the distributional hypothesis. Figure 2 shows an example of Word Embedding representation.

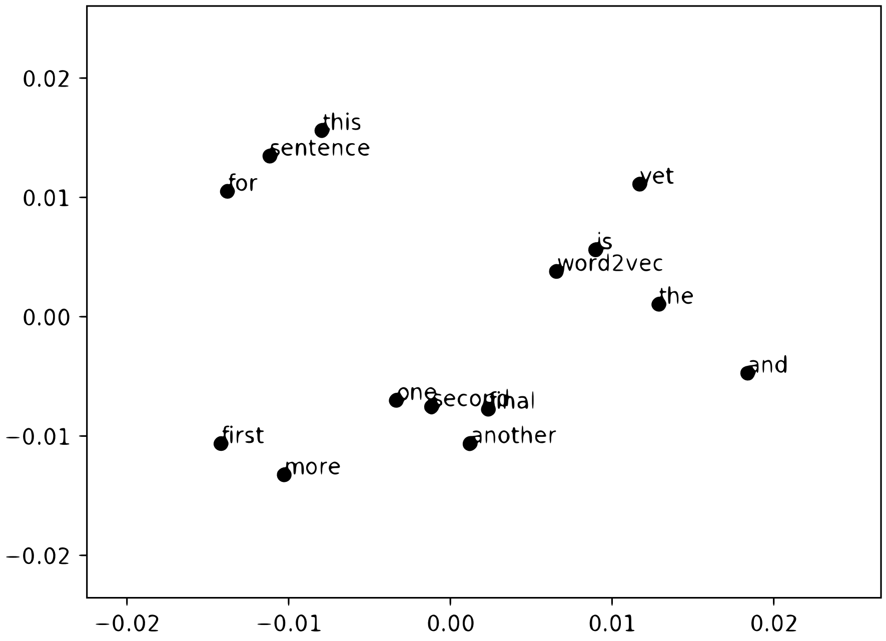


Figure 2. Word Embedding: the dots represent the word vectors contained in the word vector space related to a simple sentence.

Recently many Word Embedding models have been provided such as Glove [33], Elmo [34] and Word2Vec [35].

The Word2Vec neural network is one of the most popular Word Embedding models, developed by Mikolov et al. [35].

In this work the Word2Vec model has been adopted in consideration of its easy realisation and its ability to identify the context of a word in a document, its semantic and syntactic similarity with other words and its relations with them.

Word2Vec can be implemented through two different approaches: Skip Gram and Common Bag Of Words (CBOW) were both realised with the use of a neural network.

Skip Gram takes as input the vector representation of the word in question and tries to predict its context; usually there is only one hidden level for this configuration. CBOW instead takes as input the vectors of the words that represent a context and tries to predict a target word related to that context; this is shown in Figure 3.

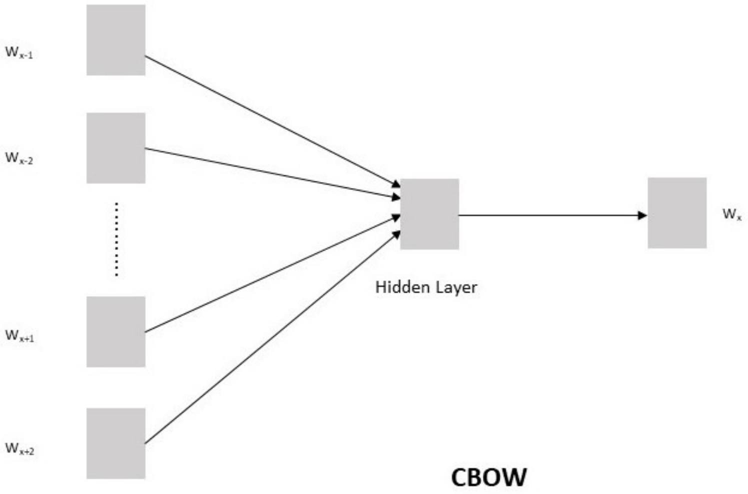
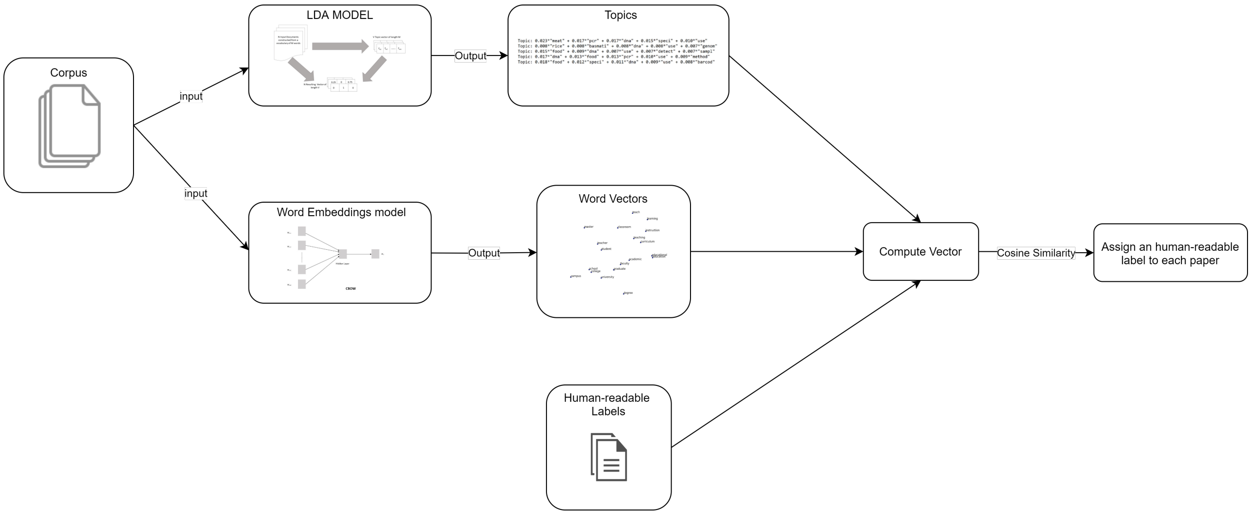


Figure 3. CBOW Model.

2. Materials and Methods

The core idea behind SPUCL is to achieve one of the main issues of the unsupervised topic modelling methods (i.e., LDA): the readability of the retrieved topics.

Indeed SPUCL is able to assign a human readable label to each topics, exploiting the artificial neural networks through a combination of the LDA and Word2Vec technologies. Since lda-based templates cannot provide readable labels, we decided to use the power of Word Embedding Artificial Neural Network (ANN from now) to identify the right label for each topic. As shown in Figure 4 in the workflow of our system, we first get all the abstracts of the papers contained in the training corpus and train the LDA model on these. We decided to exploit only the text contained in the abstracts in the LDA task, because in the abstract of a scientific paper are contained the main concepts (and related terms) of the entire article, in this way we can obtain a better topic distribution.



```
Topic: 0.023*"meat" + 0.017*"pcr" + 0.017*"dna" + 0.015*"speci" + 0.010*"use"
Topic: 0.008*"rice" + 0.008*"basmati" + 0.008*"dna" + 0.008*"use" + 0.007*"genom"
Topic: 0.015*"food" + 0.009*"dna" + 0.007*"use" + 0.007*"detect" + 0.007*"sampl"
Topic: 0.017*"dna" + 0.013*"food" + 0.013*"pcr" + 0.010*"use" + 0.009*"method"
Topic: 0.018*"food" + 0.012*"speci" + 0.011*"dna" + 0.009*"use" + 0.008*"barcod"
```

Figure 4. WORKFLOW of SPUCL.

In NLP systems the preprocessing phase is a crucial aspect of the workflow, indeed an efficient preprocessing strategy can improve the overall quality of the system [36]. Second, we compute the Word Embedding model on all corpus documents in their extended version, differently from the previous task we exploit the entire text because the performances of Word2Vec ANN can improve with a large number of examples (i.e., words). In addition to the text of all papers we exploit a pre-trained corpus provided by Gensim library in order to enlarge the number of computed vectors. The output of LDA, i.e., the word distribution over the topics, is then employed as input for the Word2Vec model trained in the previous step to compute the embedding vectors for all words in the topics.

We also compute the embedding vectors for the words that make up the human-readable labels.

Finally, we evaluate the cosine similarity between these two sets of vectors and assign the right label to each paper.

In order to realise the SPUCL system we have defined an architecture composed by two phases: preprocessing and core application (see Figure 5); further, we added a hyperparameters optimisation phase to enhance the performances of the system.

The first phase of the architecture aims to prepare the corpus. The core application provides the execution of the main tasks of the system. The hyperparameters optimisation is able to set the best configuration of the parameters of the system.

In the next sections we describe in detail our approach.

2.1. Preprocessing

The aim of this phase is to improve the quality of the analysed corpus; in order to achieve this objective we have decided to implement three main tasks: tokenization, stop-words exclusion and stemming.

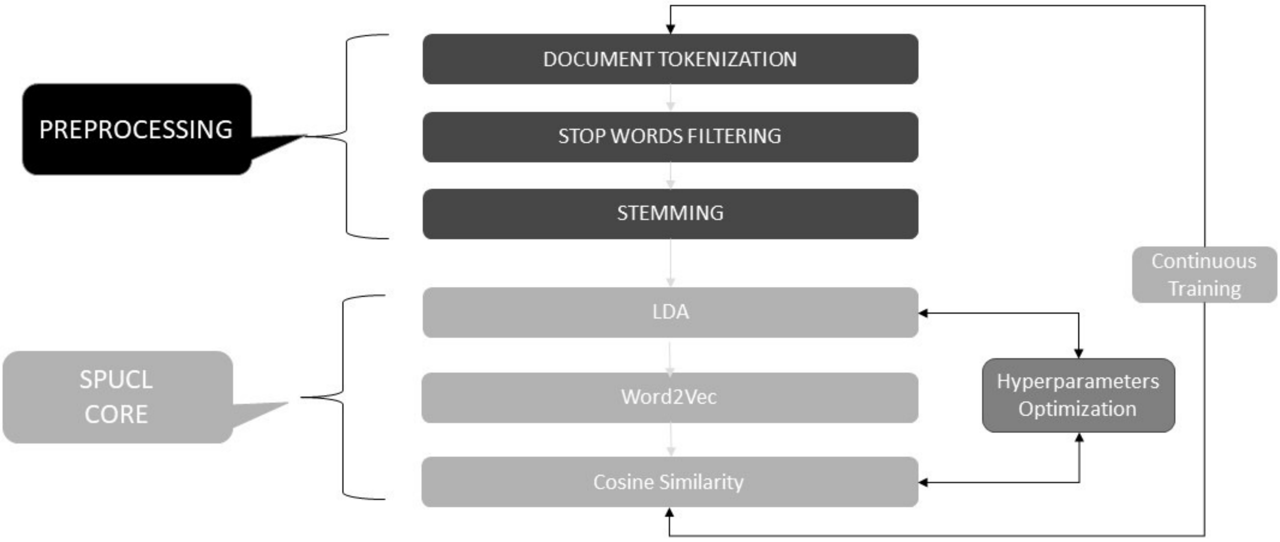


Figure 5. SPUCL Architecture.

The tokenization phase consists of dividing the text into single words or tokens. This operation is necessary in order to execute both the Topic Modelling algorithm and the Word Embedding algorithm.

The stop-words identification and exclusion phase consists in the identification of the tokens that represent stop-words such as articles and prepositions. First of all, we consider the predefined set of stop-words present in nltk (https://www.nltk.org/, accessed on 20 July 2021) library of python. Moreover, we have added an additional set of stop-words: “et al., Introduction, Conclusions, author, figure, table, paragraph, section, keywords, acknowledgments and references”, that are closely related to the specific field of the classification of scientific paper.

These words are often present in scientific papers and are not significant in order to understand the meaning of the papers. By adding these words in the stop-words set we improve the performance of our system because they do not appear in the topics retrieved by lda even though they are very frequent in scientific papers.

In order to reduce the space of represented words we have adopted the stemming; this process reduces the flexed form of a word to its root form, called the theme.

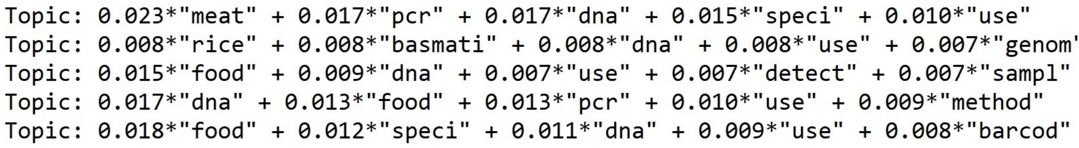
If the space of words is reduced, each of them has a greater meaning than the document analysed and the search of the topics is more accurate. Moreover, when we define the label related to the document, the evaluation of the cosine similarity is more accurate. At the end of the preprocessing phase the corpus is optimised to be processed by SPUCL.

This step is very important to get better performance from the SPUCL system and for this reason we allow the user to customise the stop-words set to further improve the system output.

2.2. SPUCL Core

This phase is the core of the application and consists of three steps: the execution of the LDA, the training of the Word2Vec neural network and the calculation of the cosine similarity. First of all, SPUCL system divides the documents into self contained groups using an implementation of LDA Algorithm realised in python language trough the Gensim library.

Trough the LDA algorithm, SPUCL defines five topics and assigns to each documents the topic with the highest probability (see Figure 6).



```
Topic: 0.023*"meat" + 0.017*"pcr" + 0.017*"dna" + 0.015*"speci" + 0.010*"use"
Topic: 0.008*"rice" + 0.008*"basmati" + 0.008*"dna" + 0.008*"use" + 0.007*"genom"
Topic: 0.015*"food" + 0.009*"dna" + 0.007*"use" + 0.007*"detect" + 0.007*"sampl"
Topic: 0.017*"dna" + 0.013*"food" + 0.013*"pcr" + 0.010*"use" + 0.009*"method"
Topic: 0.018*"food" + 0.012*"speci" + 0.011*"dna" + 0.009*"use" + 0.008*"barcod"
```

Figure 6. Topics.

Second, the SPUCL system provides training words to Embedding vector space through the Word2Vec ANN. To realise SPUCL, the topics vectors retrieved with LDA [4] are used to train a Word2Vec [35] Neural Network in the CBOW version (see Figure 3).

In SPUCL architecture the CBOW implementation of Word2Vec is adopted because it is faster and has better representations for more frequent words.

The designed LDA algorithm identifies five topics; this number has been identified as the best one to correctly divide the corpus into homogeneous sub-sets of papers, see Section 2.3.

At the end of the execution of the LDA algorithm thus configured, a list of topics associated with each of the documents was obtained.

It was decided to use the first three words of each topic as input of the neural network Word2Vec.

The first phase of using the word2Vec neural network consists in training through the output of the LDA algorithm.

The words present in the topics related to each document are then used as input for the neural network Word2Vec in order to obtain a vector that represents the label of the considered document.

After having trained the network with the topics, five human-readable labels have been defined formed by sentences of complete sense such as: "PCR method applied to the meat" and they have been used as input of the neural network to obtain a vector representing the label.

Finally, the similarity between the vector representing the document and the vectors of the predefined labels was calculated for each document by using the cosine similarity and the label with a greater similarity was assigned to the document.

The cosine similarity is a metric able to measure the similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Given two vectors, A and B, the cosine similarity, $\cos(\Theta)$, is represented using a dot product and magnitude as:

$$\cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} \tag{1}$$

The cosine similarity (1) should be close to 1, i.e., the angle between two considered vectors should be close to 0; we adopt the cosine similarity to evaluate the similarity between topics related to a document and a meaningful label.

SPUCL provides a continuous training, at the insertion of a new document in the corpus the related topic or topics are identified and this new input is used to re-train the neural network Word2Vec previously trained; furthermore, the human-readable label more similar to the new document inserted is calculated.

2.3. Hyperparameters Optimisation

The Hyperparameters Optimisation phase provides a recursive mechanism to define the better setting for the LDA parameters.

Traditionally, the measure employed to evaluate a Topic Model and define the better number of topics is the topic coherence [37].

Instead, to define our Hyperparameters Optimisation algorithm we have defined a mechanism based on a recursive method that tries to change recursively the number of topics and the number of words for each topic. In our Hyperparameters Optimisation algorithm we compute the performances achieved from the SPUCL system in its different configurations.

We decided to calculate the overall performances of the SPUCL system instead to the coherence in order to determine the number of topics.
This optimisation approach allowed us to improve the overall performances of the SPUCL systems and not only of the LDA algorithm.

In this version of the system we performed ten cycles with different numbers of topics (from one to ten topics); for each of them we performed five cycles with different numbers of words per topic (from five to ten words for each topic). In total, we have tried fifty different combinations for LDA.

We have evaluated the performance of the system in its different configurations and we have identified the better combination in five topics and five words (see Section 3).

2.4. Food Analytic Scientific Paper Classification Experiment

As mentioned above, SPUCL aims to classify scientific papers, in particular we have designed and developed SPUCL in the context of FOOD INTEGRITY Project to provide researchers an instrument to analyse the state of art of the DNA methods in food inspection.

We have exploited SPUCL to provide an easy and fast method of labelling the analysed papers. We called this task food analytic experiment.

To realise the food analytic experiment, a bibliographic research was carried out by keywords into PubMed and EBSCO libraries; through this research a set of articles composed by 119 different articles, published in the last 20 years and related to the DNA methods, has been selected.

First of all, domain experts have classified the papers per method and food type as shown in Tables 2 and 3; in these tables are reported the number of papers related to each method and product; note that each paper can be related to many different methods and/or products.

Table 2. Papers classified per method.

Methods	Number of Papers
real time PCR	78
multiplex PCR	54
PCR RFLP	46
duplex PCR	36
Single Nucleotide Polymorphism	24
DNA barcode	21
Next Generation Sequencing	12
Microsatellites	9
Microarray	4
LAMP	3
CAPS	1

Table 3. Papers classified per food product.

Food Products	Number of Papers
chicken	26
cattle	21
pig	21
horse	19
turkey	19
sheep	18
buffalo	14
goat	13
donkey	10
rat	6
rabbit	5
mouse	5
olive oil	13
rice	12
tomato	8
black pepper	6
celery	3
onion	3
grape	2
carrot	1
sunflower	1
sage	1

Second, we asked domain experts to summarise the table in a set of human-readable labels; These will be used in food analytic experiment. The domain experts have defined the following human-readable labels:

1. "PCR methods applied to meat";
2. "PCR methods applied to vegetables";
3. "PCR methods applied to sea food";
4. "DNA barcode applied to meat";
5. "DNA barcode applied to vegetables";
6. "DNA barcode applied to the sea food";
7. "Next Generation Sequencing applied to meat";
8. "Next Generation Sequencing applied to vegetables";
9. "Next Generation Sequencing applied to sea food";
10. "Microarray applied to meat";
11. "Microarray applied to vegetables";
12. "Microarray applied to sea food";

Then, we divided the dataset into training set (70%) and test set (30%).

Finally, we trained SPUCL System with the training set to calculate the most similar human-readable label for each article. In Section 3 we will describe the results achieved in the food analytic experiment, calculated on the test set.

3. Results

We ran food analytic experiment as described in the previous section.

Its results are summarised in the next subsection. Then we explored the obtained data in the discussion Section 4.

Food Analytic Experiment Results

As mentioned above SPUCL provides a a human-readable labelling system for scientific publications. Each paper analysed by SPUCL is classified according to the label automatically assigned by the system.

In order to evaluate SPUCL system we adopted the following metrics:

Precision

$$\frac{TruePositive}{TruePositive + FalsePositive} \tag{2}$$

Recall

$$\frac{TruePositive}{TruePositive + FalseNegative} \tag{3}$$

and F-Measure

$$2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

that are traditionally employed to evaluate the document classification task.

The F-Measure (2) score calculates the harmonic mean of Precision and Recall, giving each the same weighting. It allows one to evaluate a model taking both the Recall and Precision into account using a single score, so it is useful to evaluate the overall performance of the SPUCL model.

The first step of the evaluation process consists in assigning to each document a label chosen from the list in Section 2.4.

To perform this task we asked domain experts to assign manually to each of selected papers (Section 2.4) one of the predefined human-readable label, in order to create a manually annotated set of papers. Experts classified papers as shown in Table 4

Table 4. Papers classified per Human-Readable Label.

Label	Number of Papers
"PCR methods applied to meat"	35
"PCR methods applied to vegetables"	29
"PCR methods applied to sea food"	17
"DNA barcode applied to meat"	10
"DNA barcode applied to vegetables"	9
"DNA barcode applied to the sea food'	3
"Next Generation Sequencing applied to meat"	8
"Next Generation Sequencing applied to vegetables"	1
"Next Generation Sequencing applied to sea food"	3
"Microarray applied to meat"	2
"Microarray applied to vegetables"	1
"Microarray applied to sea food"	1

In the second step of the evaluation process, we computed the Precision, the Recall and the F-Measure.

First we ran SPUCL system in order to automatically assign a label to each paper of this pre-annotated set of papers.

Second we calculated these values for each of the selected labels and finally we calculated the average performances of the system.

The performance of the SPUCL system is shown in Table 5; in this table we reported the performance related to each label and the average performance of the SPUCL system.

Table 5. Performance of SPUCL.

Label	Precision	Recall	F-Measure
"PCR methods applied to meat"	0.7	0.8	0.75
"PCR methods applied to vegetables"	0.62	0.83	0.71
"PCR methods applied to sea food"	0.67	0.82	0.74
"DNA barcode applied to meat"	0.57	0.8	0.7
"DNA barcode applied to vegetables"	0.58	0.78	0.67
"DNA barcode applied to the sea food"	0.2	0.67	0.31
"Next Generation Sequencing applied to meat"	0.5	0.88	0.64
"Next Generation Sequencing applied to vegetables"	0.1	1	0.18
"Next Generation Sequencing applied to sea food"	0.5	1	0.67
"Microarray applied to meat"	0.5	1	0.67
"Microarray applied to vegetables"	0	0	0
"Microarray applied to sea food"	0.17	1	0.29
Average	0.43	0.8	0.53

4. Discussion

The SPUCL system provides a Scientific Publication Classifier based on a human-readable labelling algorithm.

In order to evaluate our system we adopted the Precision, Recall and f-measure metrics. As shown in Table 5 the average value of F-Measure of the SPUCL system is 0.53. We can observe that if we exclude from the model the labels with a very low number of associated papers (number of papers <=3), the average F-measure increases to 0.65; the same can be observed in relation to the average Precision value that increases to 0.61 and in relation to the average Recall that increases to 0.82.

The issue of imbalanced dataset in machine learning is a challenging task. In order to achieve this issue we will be implementing the SMOTE (Synthetic Minority Oversampling Technique) [38] technique in the next version of SPUCL system. This will allow us to oversample the minority classes and to improve the performances also in these classes.

As mentioned above, there are many approaches to the scientific papers information management (see Section 1.1.1). Due to the differences between these methods, each of them has been evaluated with different approaches. In Table 6 we provide a comparison between the evaluation methods of SPUCL and similar approaches.

Table 6. Comparison between SPUCL and the other approaches' evaluation methods.

	Approaches				
Metrics	Peganova [19]	Yasunaga [20]	Simon [21]	Goncalves [22]	SPUCL
Precision, Recall, F-Measure		x		x	x
PathRatio, Attempts, Jumps	x				
Human evaluation		x			
AUC			x		

5. Conclusions

In this work, SPUCL was presented, a semi-supervised document classification system capable of automatically assigning a human-readable label to the documents analysed.

The SPUCL system was tested through the analysis of a set of papers related to the state of the art of DNA methods in food inspection and regulation selected by domain experts.

The exploitation of SPUCL to analyse the state of the art of a given scientific topic has shown the possibility to perform a semi-supervised classification of documents and an unsupervised labelling task able to provide as output a label that is semantically relevant and understandable for the humans.

This approach can easily be applied to different research areas to facilitate the analysis of the state of the art.

The SPUCL system currently requires: a set of articles related to a research field and a set of human readable labels.

Although it is not necessary to read all the articles, it is required to recognise them as relating to the selected research field.

In Section 2.4, papers were selected through a bibliographic search. In addition, we asked the domain experts to define the human-readable labels and to assign to each paper one of these labels.

This can be seen as a limitation of the SPUCL system, and to overcome this limitation we intend to provide incremental training of the SPUCL system.

The incremental training is an advantage of using neural networks.

Some of the concepts of a given field of research learned earlier can be reused in other fields where the number of relevant papers is smaller.

When the system is properly trained, it will be possible to ask the experts only for the set of papers to be examined and the human-readable labels.

Author Contributions: Conceptualisation, N.S.; Methodology, N.S. and A.P.; Validation, M.M.; Investigation, N.S.; Resources, N.S.; Data curation, N.S. and M.M.; Writing—Original draft, N.S. and M.M.; Writing—Review and editing, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was partially supported by POR Campania FESR 2014–2020 RARE.PLAT.NET (CUP B63D18000380007) Grant from the European Social Fund. The funding sources had no involvement in the study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

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

References

1. Alić, B.; Gurbeta, L.; Badnjević, A. Machine learning techniques for classification of diabetes and cardiovascular diseases. In Proceedings of the 2017 6th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, 11–15 June 2017; pp. 1–4. [CrossRef]
2. Sueno, H.T.; Gerardo, B.D.; Medina, R.P. Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 3937. [CrossRef]
3. Afzal, M.Z.; Capobianco, S.; Malik, M.I.; Marinai, S.; Breuel, T.M.; Dengel, A.; Liwicki, M. Deepdocclassifier: Document classification with deep Convolutional Neural Network. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1111–1115. [CrossRef]
4. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
5. Acciarito, S.; Cardarilli, G.C.; Cristini, A.; Nunzio, L.D.; Fazzolari, R.; Khanal, G.M.; Re, M.; Susi, G. Hardware design of LIF with Latency neuron model with memristive STDP synapses. *Integr. VLSI J.* **2017**, *59*, 81–89. [CrossRef]
6. Acciarito, S.; Cristini, A.; Di Nunzio, L.; Khanal, G.M.; Susi, G. An a VLSI Driving Circuit for Memristor-Based STDP. In Proceedings of the 12th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME), Lisbon, Portugal, 27–30 June 2016.
7. Khanal, G.M.; Acciarito, S.; Cardarilli, G.C.; Chakraborty, A.; Di Nunzio, L.; Fazzolari, R.; Cristini, A.; Re, M.; Susi, G. Synaptic behaviour in ZnO-rGO composites thin film memristor. *Electron. Lett.* **2017**, *53*, 296–298. [CrossRef]
8. Pazienza, M.; Scarpato, N.; Stellato, A.; Turbati, A. Semantic Turkey: A browser-integrated environment for knowledge acquisition and management. *Semant. Web* **2012**, *3*, 279–292. [CrossRef]
9. Accardi, A.R.D.; Chiarenza, S. Digital museums of the imagined architecture: An integrated approach. *Disegnarecon* **2016**, *9*, 15-1.
10. Guadagni, F.; Zanzotto, F.M.; Scarpato, N.; Rullo, A.; Riondino, S.; Ferroni, P.; Roselli, M. RISK: A random optimization interactive system based on kernel learning for predicting breast cancer disease progression. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5th International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2017, Granada, 2017*; Springer: Cham, Switzerland, 2017; pp. 189–196.
11. Ferroni, P.; Zanzotto, F.M.; Scarpato, N.; Riondino, S.; Nanni, U.; Roselli, M.; Guadagni, F. Risk Assessment for Venous Thromboembolism in Chemotherapy-Treated Ambulatory Cancer Patients. *Med. Decis. Mak.* **2016**, *37*, 234–242. [CrossRef]

12. Ferroni, P.; Zanzotto, F.M.; Scarpato, N.; Riondino, S.; Guadagni, F.; Roselli, M. Validation of a machine learning approach for venous thromboembolism risk prediction in oncology. *Dis. Mark.* **2017**, *2017*, 1–7. [CrossRef]
13. Scarpato, N.; Pieroni, A.; Di Nunzio, L.; Fallucchi, F. E-health-IoT universe: A review. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2017**, *7*, 2328–2336. [CrossRef]
14. Guadagni, F.; Scarpato, N.; Patrizia, F.; D'Ottavi, G.; Boavida, F.; Roselli, M.; Garrisi, G.; Lisi, A. Personal and Sensitive Data in the e-Health-IoT Universe. In Proceedings of the 2nd International Summit on Internet of Things, IoT 360° 2015, Rome, Italy, 27–29 October 2016; Volume 170, pp. 504–514.
15. Pieroni, A.; Scarpato, N.; Brilli, M. Industry 4.0 Revolution in Autonomous and Connected Vehicle A non-conventional approach to manage Big Data. *J. Theor. Appl. Inf.* **2018**, *96*, 10–18.
16. Pieroni, A.; Scarpato, N.; Brilli, M. Performance study in autonomous and connected vehicles a industry 4.0 issue. *J. Theor. Appl. Inf. Technol.* **2018**, *96*, 984–994.
17. Cilia, N.; Scarpato, N.; Romano, M. A Semantic Approach to Reachability Matrix Computation; In Proceedings of the 10th Conference on Semantic Technology for Intelligence, Defense, and Security, STIDS 2015, Fairfax VA, USA, 18–20 November 2015; Volume 1523, pp. 91–94.
18. Pieroni, A.; Scarpato, N.; Di Nunzio, L.; Fallucchi, F.; Raso, M. Smarter City: Smart energy grid based on Blockchain technology. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2018**, *8*, 298–306. [CrossRef]
19. Peganova, I.; Rebrova, A.; Nedumov, Y. Labelling Hierarchical Clusters of Scientific Articles. In Proceedings of the 2019 Ivannikov Memorial Workshop, IVMEM 2019 Velikiy Novgorod, Russia, 2019 IEEE Computer Society, Velikiy Novgorod, Russia, 13–14 September 2019; pp. 26–32. [CrossRef]
20. Yasunaga, M.; Kasai, J.; Zhang, R.; Fabbri, A.R.; Li, I.; Friedman, D.; Radev, D.R. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 7386–7393. [CrossRef]
21. Simon, C.; Davidsen, K.; Hansen, C.; Seymour, E.; Barnkob, M.B.; Olsen, L.R. BioReader: A text mining tool for performing classification of biomedical literature. *BMC Bioinform.* **2019**, *19*, 165–170. [CrossRef]
22. Gonçalves, C.A.; Iglesias, E.L.; Borrajo, L.; Camacho, R.; Vieira, A.S.; Gonçalves, C.T. Comparative Study of Feature Selection Methods for Medical Full Text Classification Carlos. In *Bioinformatics and Biomedical Engineering. IWBBIO 2019. Lecture Notes in Computer Science*; Rojas, I., Valenzuela, O., Rojas, F., Ortuño, F., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 11466, pp. 514–523. [CrossRef]
23. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
24. Tang, B.; He, H.; Baggenstoss, P.M.; Kay, S. A Bayesian Classification Approach Using Class-Specific Features for Text Categorization. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1602–1606. [CrossRef]
25. Rokach, L.; Maimon, O. *Data Mining with Decision Trees: Theory and Applications*; World Scientific: Stevens Point, WI, USA, 2008; p. 244.
26. Han, E.H.; Karypis, G.; Kumar, V. *Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 53–65.
27. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000; p. 189.
28. Manevitz, L.M.; Yousef, M. One-Class SVMs for Document Classification. *J. Mach. Learn. Res.* **2001**, *2*, 139–154.
29. Zhou, C.; Sun, C.; Liu, Z.; Lau, F.C.M. A C-LSTM Neural Network for Text Classification. *arXiv* **2015**, arXiv:1511.08630.
30. Zhang, X.; Zhao, J.; LeCun, Y. Character-level Convolutional Networks for Text Classification. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 649–657.
31. Li, J.; Luong, M.T.; Jurafsky, D. A Hierarchical Neural Autoencoder for Paragraphs and Documents. *arXiv* **2015**, arXiv:1506.01057.
32. Turian, J.; Ratinov, L.; Bengio, Y. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10), Uppsala, Sweden, 11–16 July 2010; pp. 384–394.
33. Pennington, J.; Socher, R.; Manning, C.D. *GloVe: Global Vectors for Word Representation*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543.
34. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. *Deep Contextualized Word Representations*; Association for Computational Linguistics: New Orleans, Louisiana, 2018.
35. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
36. Uysal, A.K.; Gunal, S. The impact of preprocessing on text classification. *Inf. Process. Manag.* **2014**, *50*, 104–112. [CrossRef]
37. Douven, I.; Meijs, W. Measuring coherence. *Synthese* **2007**, *156*, 405–425. [CrossRef]
38. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]