




Article

A NLP Pipeline for the Automatic Extraction of a Complete Microorganism's Picture from Microbiological Notes

Sara Mora ¹, Jacopo Attene ¹, Roberta Gazzarata ^{2,3}, Daniele Roberto Giacobbe ^{4,5}, Bernd Blobel ^{6,7,8},
Giustino Parruti ⁹ and Mauro Giacomini ^{1,*}

¹ Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, 16145 Genoa, Italy

² Healthropy, Corso Italia 15/6, 17100 Savona, Italy

³ HL7 Europe, 1000 Brussels, Belgium

⁴ Infectious Diseases Unit, IRCCS San Martino Polyclinic Hospital, 16132 Genoa, Italy

⁵ Department of Health Sciences (DISSAL), University of Genoa, 16132 Genoa, Italy

⁶ Medical Faculty, University of Regensburg, 93053 Regensburg, Germany

⁷ eHealth Competence Center Bavaria, Deggendorf Institute of Technology, 94469 Deggendorf, Germany

⁸ First Medical Faculty, Charles University Prague, 12800 Prague, Czech Republic

⁹ Department of Infectious Diseases, AUSL Pescara, 65124 Pescara, Italy

* Correspondence: mauro.giacomini@unige.it

Abstract: The Italian “Istituto Superiore di Sanità” (ISS) identifies hospital-acquired infections (HAIs) as the most frequent and serious complications in healthcare. HAIs constitute a real health emergency and, therefore, require decisive action from both local and national health organizations. Information about the causative microorganisms of HAIs is obtained from the results of microbiological cultures of specimens collected from infected body sites, but microorganisms’ names are sometimes reported only in the notes field of the culture reports. The objective of our work was to build a NLP-based pipeline for the automatic information extraction from the notes of microbiological culture reports. We analyzed a sample composed of 499 texts of notes extracted from 1 month of anonymized laboratory referral. First, our system filtered texts in order to remove nonmeaningful sentences. Thereafter, it correctly extracted all the microorganisms’ names according to the expert’s labels and linked them to a set of very important metadata such as the translations into national/international vocabularies and standard definitions. As the major result of our pipeline, the system extracts a complete picture of the microorganism.

Keywords: hospital-acquired infections; international coding system; laboratory information systems; natural language processing; information extraction



Citation: Mora, S.; Attene, J.; Gazzarata, R.; Giacobbe, D.R.; Blobel, B.; Parruti, G.; Giacomini, M. A NLP Pipeline for the Automatic Extraction of a Complete Microorganism's Picture from Microbiological Notes. *J. Pers. Med.* **2022**, *12*, 1424. <https://doi.org/10.3390/jpm12091424>

Academic Editor: Norman R. Williams

Received: 10 July 2022

Accepted: 26 August 2022

Published: 31 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The recent COVID-19 pandemic highlighted even more the worrying and widespread increasing circulation of pathogenic microorganisms in hospitals, sheltering for elderly, and assisted residences. The Italian “Istituto Superiore di Sanità” (ISS) [1] identifies hospital-acquired infections (HAI) as the most frequent and serious complications of healthcare. A possible definition of HAI is “infections that first appear 48 h or more after hospital admission or no later than 30 days after discharge following inpatient care” [2]. HAIs constitute a real health emergency and require decisive action from both local and national health organizations. The main objective is to build stable and automatic systems dedicated to reporting and epidemiological surveillance. When a pathogenic microorganism responsible for HAI is identified, the information can allow targeted antibiotic treatment, as well as the prompt adoption of specific infection-control measures for multidrug-resistant (MDR) bacteria.

This information is obtained from microbiological cultures of specimens collected from infected body sites, and the outcomes are usually reported in the results of laboratory

exams [3,4]. MDR bacteria are defined as those organisms that show resistance to one or more agents in at least three antimicrobial categories during *in vitro* antimicrobial susceptibility tests [5,6].

MDR organisms are a world health problem causing about 33,000 deaths per year in Europe [7] and 35,000 in the United States [8]. It becomes even more serious when the infection affects critically ill patients, e.g., those admitted to intensive care units (ICUs) [9,10], because it is associated with increased mortality [11]. However, this is not only a human-related problem, because MDR bacteria are frequently found in the environment [12], especially in intensive farming both in agriculture (livestock and poultry) [13–15] and in aquaculture [16–18].

Although for more than 20 years modern laboratory information systems (LISs) [19,20] managed laboratory analyses, individual centers created their own vocabulary. This is mainly due to the fact that the development of information systems has been non-simultaneous and strongly localized. The resulting need to make results coming out of the single laboratories comparable led to the development of international coding systems and standards devised for the management of terminologies. An example of an international vocabulary applicable in this field is the Logical Observations Identifiers Names and Codes (LOINC). The Common Terminology Service Release 2 (CTS2) [21] is a standard dedicated to terminology management, whose specifications result from the collaboration between the Object Management Group (OMG) [22] and Health Level 7 (HL7) [23].

One of the fields where computerized systems faced many problems was the management of microbiology. This resulted from the high variability of the discipline and the strict link to the habits of individual laboratories (i.e., the coding system on which the nomenclature of bacteria relies, how sensitivity analyses are performed, etc.). Therefore, a contrast arose between the need for more variability and the mandatory use of LISs, imposed by national laws, whose structure in some specific cases may be considered too stiff. Therefore, because of the lack of ad hoc and appropriate fields for representing and managing microbiological-related information, the staff instead preferred exploiting clinical notes written as natural text. Accordingly, clinical notes became an important source of information for biomedical research, patient management, and care, while the necessity of manual inspection made their use expensive in terms of personnel effort and time. The limitations of data/information collection in LIS, on the one hand, and the advantage of concept representations using domain-specific languages instead of data/information representation, on the other hand, are discussed in more detail in Section 4.

This kind of problem can be addressed using artificial intelligence (AI) tools, especially with natural language processing (NLP). It is a branch of computer science that deals with the processing of natural human language by computers, studying the problems connected to the learning, understanding and automatic generation of human language both in written and in spoken form [24–26].

The objective of our work was to build a NLP-based pipeline for automatic information extraction from the notes of microbiological culture reports.

This could represent a first step toward the development of a system able to monitor antibiotic prescriptions at a hospital and territorial level in the Abruzzo Region [27].

This paper is an extension of work originally presented at the 18th International Conference on Wearable, Micro, and Nano Technologies for Personalized Health (pHealth 2021) titled “A NLP pipeline for the automatic extraction of microorganisms names from microbiological notes” [28]. The extended version addresses the problem of managing the national and international terminology systems linked to the project and of filtering clinical notes in order to exclude nonsignificant sentences.

2. Materials and Methods

The complete schema of the developed pipeline is presented in Figure 1, and it can be divided into four main sections (discussed in detail below). “Data preparation” prepares the inputs values, explaining how we build the vocabulary database, and it loads and

preprocesses the clinical notes. “Pattern recognition” takes as input the textual data, converting them into a numerical representation, and it builds three ML-based models that predict if the specific pattern is contained or not in a clinical note. “Pattern removal” takes as input the sentences identified by the model as containing the specific pattern and removes it. “Microorganism detection” takes as input both the filtered and the nonfiltered sentences and first performs the “Genus extension” process before extracting from the clinical notes meaningful information such as the microorganism name and if it belongs to the group of dangerous microorganisms.

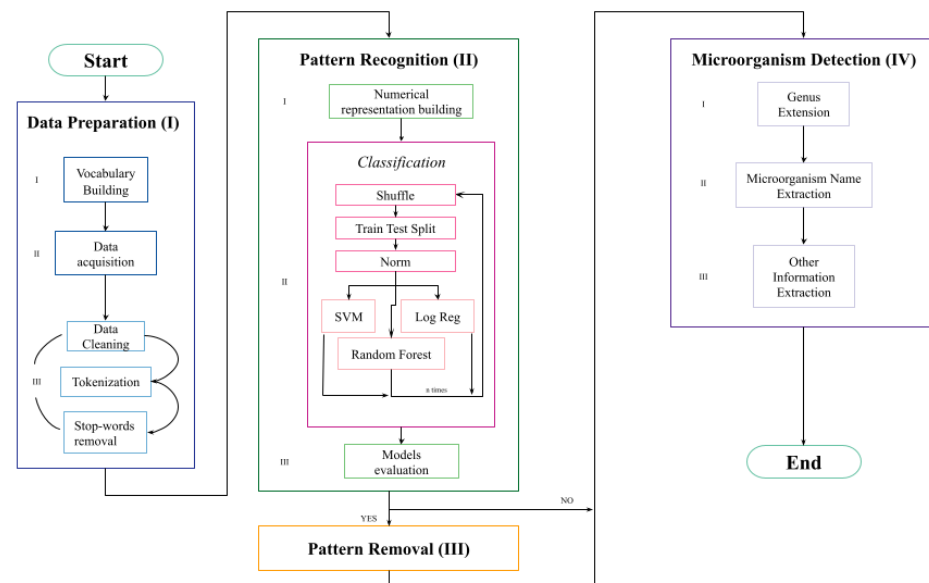


Figure 1. Complete schema of the pipeline. It can be divided into 4 main sections: data preparation, pattern recognition and removal, and microorganism detection.

2.1. Characteristics of the Sample

The collected sample was derived from the LIS of the main hospital of Pescara in Abruzzo Region and was obtained from clinical notes extracted from a 1 month collection of anonymized laboratory referral. It was composed of 499 texts from culture reports, classified as follows:

- 276 (55.3%) texts containing the name of a microorganism where an expert from the hospital confirmed its presence;
- 56 (11.2%) texts needing to be filtered because they contained a pattern that is not useful for our analysis and was, thus, removed. An example of a sentence belonging to that pattern is the following: “Integration of the preliminary report sent on . . . “. Indeed, we considered the use of synonyms, e.g., “provisional” instead of “preliminary”, and the presence of orthographic errors, e.g., missing letters. Therefore, we decided not to use regular expressions alone as first attempt.

We are waiting for the authorization from Abruzzo region to access the entire LIS system at a regional level to massively test the proposed system with notes produced by a wide range of persons.

2.2. Environment and Libraries

We completely developed the pipeline in Python language, and we used the Jupyter Notebook environment. The Python libraries used within this project were as follows:

Pandas: This is an open-source Python library containing data analysis and manipulation tools [29].

Pyodbc: This is an open-source module developed in Python that allows accessing databases through the ODBC (Open DataBase Connectivity).

Natural Language Toolkit (NLTK): This is a worldwide used library to perform text analysis in multiple languages; therefore, it is popular in both academia and research [30]. Some of the operations supported by the NLTK are tokenization, stemming, part of speech tagging (POS tagging), and disambiguation.

SpaCy: This is an open-source library for NLP in Python supporting different languages [31].

Re: This is a Python module that provides operations useful to work with regular expressions [32].

Scikit-learn: This is an open-source library that contains several machine learning algorithms, e.g., classification, regression, and clustering [33].

Seaborn, matplotlib: These are libraries used to produce graphics [34,35].

FuzzyWuzzy: This is a Python library used to manage the comparisons between strings. In detail, it is used to compute the distance between two strings with the same number of characters or not, taking into account the order of words and the allowed maximum frequency of a string. This comparison is based on the Levenshtein distance.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_i)} \end{cases} & \text{otherwise} \end{cases}$$

where i and j constitute the indices of the last character of the two substrings [36].

2.3. Data Preparation (I)

Vocabulary building (I.I): We built a vocabulary containing the names of microorganisms (bacteria, fungi, yeasts, and viruses) from the “National Healthcare Safety Network organism list”, including the current taxonomic subdivision which was proposed by Carl Woese in 1990. We mapped the microorganism’s genus and specie into three standard coding systems, at national and/or international level: Italian Clinical Microbiologists Association (AMCLI), Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), and National Healthcare Safety Network (NHSN). Together with the name of the microorganism, we retrieved other metadata, such as the microorganism’s definitions according to Medical Subject Headings (MSH) and the National Cancer Institute (NCI). We stored all the information in an SQL Server database, and we loaded them using the pyodbc tool.

Data acquisition (I.II): We used the pandas library to import data.

Data cleaning, tokenization, and stopword removal (I.III): We cleaned the clinical notes first by removing punctuation and substituting patterns that could be dates with the word “data”. Then, we divided them into minimal text units of analysis, which we called tokens. Then, we proceeded with stopword removal, but considering only words longer than one character in order to exclude from the analysis strings belonging to the class of prepositions, articles, and adverbs, while keeping single letters that could be the abbreviation of a genus name.

2.4. Pattern Recognition (II)

Once we loaded and cleaned data, we needed to convert text into a numerical representation that could be used as input for ML algorithms, and we adopted the bag of words technique. This choice was guided by the structure of the sentences that was fragmentary and did not respect any strict syntactic rules. Therefore, we preferred to use a context-free representation.

Numerical representation building (II.I): Bag of words (BoW) is a numerical representation of text that describes the occurrence of words within a document. It involves two main issues: a vocabulary of known words (or n-grams of characters) and a measure of their presence in the text. BoW representation does not keep any information about the structure or order of words in the document. The possibility to add grouped words (called n-grams of words) to the vocabulary allows capturing a little of the meaning from

the document. The resulting numerical representation was composed of both n-grams of characters and n-grams of words following the proportion of 70:30. We decided to select more features composed by n-grams of characters in order to deal with misspellings, abbreviations, and limited syntactic rules. We tested the model performance considering 10 possible total numbers of selected features from 10 to 100 with a step size of 10. We obtained the best performance with a total number of features equal to 90.

Classification (II.II): We used the aforementioned numerical representation to learn a supervised binary classifier to predict whether the observed pattern was present in the clinical note or not. Specifically, we compared the performances of three classifiers, as described below.

SVM is a supervised learning method that can be used to perform classification analysis on both linear and nonlinear data [37]. The main aim of SVM is to find a line or a hyperplane that maximizes the distance between the classes (support vectors) when placed between them [38]. If data are not linearly separable, then they can be transformed using a kernel function from a low-dimensional to a high-dimensional structure to make the data separable.

Logistic regression (LR) is a method of statistical analysis used to estimate the relationship between a dependent variable and at least one independent variable, minimizing the Euclidean distance between the true label and the model output. Specifically, for binary classification, the output variable is modeled by a sigmoid ranging between 0 and 1. We introduced model sparsity adding an L_1 penalty term [39].

Random forest (RF) is an ensemble of K decision tree classifiers created from a different bootstrap sample. The trees are built by sampling a random subset of the attributes at each internal node in the decision tree. The random sampling of the attributes reduces the correlation between the trees in the ensemble [40].

We split the dataset into a learning and a testing set with the proportion of 80:20. On the learning set, we performed the hyperparameter search through a tenfold cross-validation, which iteratively split the learning set into a training and validation set. They were respectively used to learn the model with all the possible combinations of hyperparameters and to evaluate the performances thereafter. Then, we learned the three models with the selected set of best hyperparameters, and we evaluated the model performances on the testing set. We repeated the classification 20 times, shuffling the data each time. In order to guarantee reproducibility of results, we set the random state equal to the loop index.

Models Evaluation (II.III): We evaluated the performance of the three ML models in terms of accuracy.

2.5. Pattern Removal (III)

Once the algorithm classified the clinical notes as “containing”/“not containing” the pattern, we used regular expressions to remove the uninformative pattern from the identified notes.

The schema of the regular expression was as follows: `\b[Ii]\w.+?\bdata\b`.

The elements of the expression are defined below.

`\b` asserts the position of a word boundary. In this case, we want the pattern beginning with ‘I’ (the first letter of the word ‘Integrazione’ (integration), which can be abbreviated and/or can be uppercase or lowercase in the notes).

`\w` matches any word character and ends with ‘data’ (the word that we substituted for all dates in the data cleaning phase).

`.` matches any character (e.g., letters, numbers, and punctuation) except for line terminators.

`+?` matches the previous token between one and unlimited times, the fewest times possible, but expanding as needed.

2.6. Microorganism Detection (IV)

Genus Extension (IV.I): We stored the microorganism names using the binomial nomenclature originating from the Linnaeus classification [41]. It is composed of two terms: the first is the genus name with the first letter capitalized; the second is the species name in lower case. Usually, after a microorganism’s name is written once in a text, then it can be substituted with its first capital letter, followed by a period, in subsequent mentions. However, considering the brevity of the clinical notes, a shared agreement is to always write the abbreviated form, despite the entire genus having not yet been introduced. This binomial nomenclature does not allow the use of an abbreviation composed of two letters for the genus. Nevertheless, even though the microorganisms should be written according to this strict rule, we decided to keep words composed of only one character and not to use a regular expression, because we considered that abbreviations may be not written correctly, e.g., by using abbreviations that are not followed by a period, or where uppercase letters are followed by a period but not followed by lowercase letters. Hence, we performed the extension of the microorganism genus. Specifically, we compared the “ $n + 1$ ” token with each species of the vocabulary. If the similarity index between the two tokens was greater than or equal to 98, then we checked the token “ n ”. If the token “ n ” began with the same letter of the genus of the species in position “ $n + 1$ ”, we substituted the token “ n ” with the genus name found in the vocabulary. The schema of the treatment is presented in Figure 2.

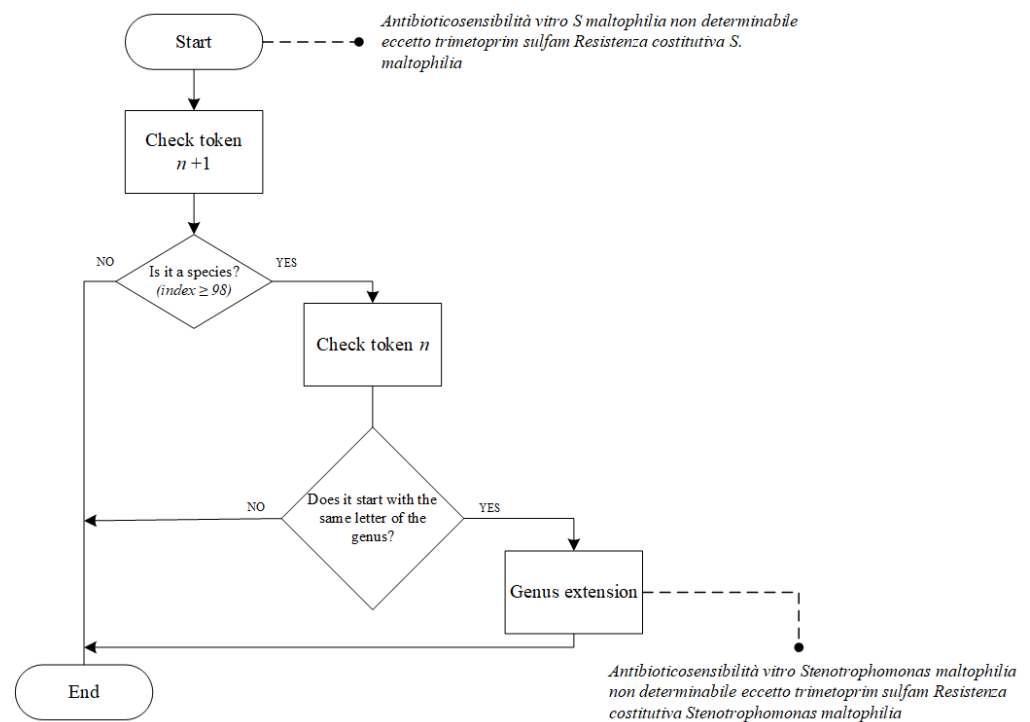


Figure 2. Genus extension decision flow. The figure also includes an example. In the upper part of the figure, there is a sentence already preprocessed but before the genus extension phase, whereas, in the lower part, we can see the extended version. First, the maltophilia species is identified, while “S” as the first letter of genus Stenotrophomonas is extended.

Microorganism name extraction (IV.II): Initially, we tried to carry out a lexical and morphological analysis, but the lack of morphological structure of the clinical notes resulted in no good results. Therefore, we extracted the microorganism name by comparing each token “ n ” of the preprocessed clinical notes and the genera in the vocabulary using the FuzzyWuzzy library. The complete workflow of the microorganism name extraction phase is shown in Figure 3.

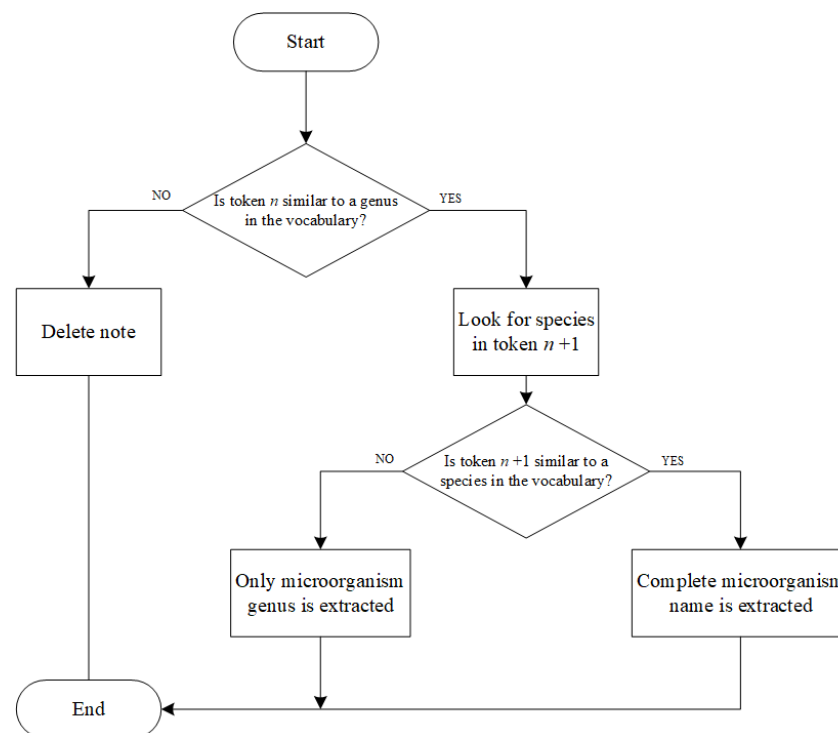


Figure 3. Microorganism name extraction decision flow. In contrast to the genus extension phase, in this process, the token “ n ” is first considered, and it is compared to the genera listed in the vocabulary. Only if a match is found (with a high similarity index) is the token “ $n + 1$ ” considered and compared to the list of species in the vocabulary that belong to the identified genus.

In particular, considering the genus extraction, we set the threshold of the similarity index at 75, while we set the threshold of the species index as 85 (as they were typically written correctly).

Other Information Extraction (IV.III): Together with the identification of the genus and species, in order to highlight microorganisms that could be potentially dangerous, we also searched the clinical notes for the keyword “alert”, which is an explicit indication of microbiologists regarding the danger of the identified microorganism. Similarly, but much less frequently, the “non-alert” bi-gram, with which the microbiologists indicate the harmlessness of the microorganism, may be present. To address both cases, we performed a search at the token level for the keyword “alert”. If identified at the n position, we checked if token $n - 1$ matched the negation “non”.

3. Results

3.1. Identification and Removal of a Specific Pattern

In the process of information extraction from the microbiological notes, it is useful to identify nonmeaningful sentences, e.g., “integration of the provisional report of . . .”. The lack of morphological structure in the sentences led us to use a count-based method to build a numerical representation of the clinical notes.

Figure 4 summarizes the mean values of accuracy obtained using the three classifiers over the 20 iterations per each total number of features, shuffling the data each time.

We obtained best results in terms of mean accuracy across classifiers (99.06%) with a total number of features equal to 90. The SVM classifier with a Gaussian kernel obtained a mean accuracy of 98.99%, logistic regression obtained a mean accuracy of 98.99%, and random forest obtained a mean accuracy of 99.19%.

This means that the pattern was correctly identified using all classifiers, and it can be securely removed from the specific clinical notes.

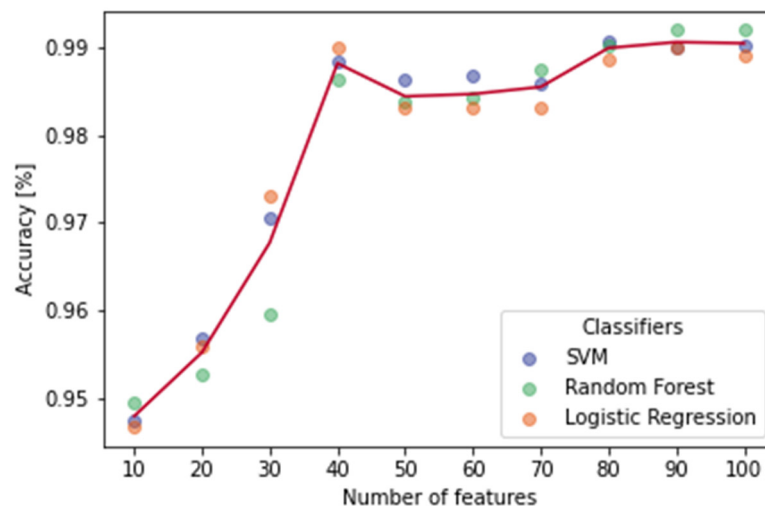


Figure 4. Mean accuracy performances of the three classifiers displayed for each value of the total number of features. Each data point is the mean value of 20 values obtained by shuffling the data.

3.2. Genus Extension

Our sample of clinical notes contained a total number of 107 abbreviated genera followed by their species. After the system elaboration process of the notes, all 107 genera were extended, and they completely matched with the expert indications.

3.3. Microorganism Detection

Our sample of clinical notes was composed of 499 texts, and 276 (55.3%) of them actually presented the name of a microorganism. We performed two tests.

First, we introduced the entire sample into the module for microorganism extraction. Then, we introduced only those notes that actually contained the name of the microorganisms.

The system correctly identified all microorganism names in both cases. In detail, it found 416 genera, and, as shown in Figure 5a, the majority of them (321) had a similarity index of 100. This was also a consequence of the genus extension process.

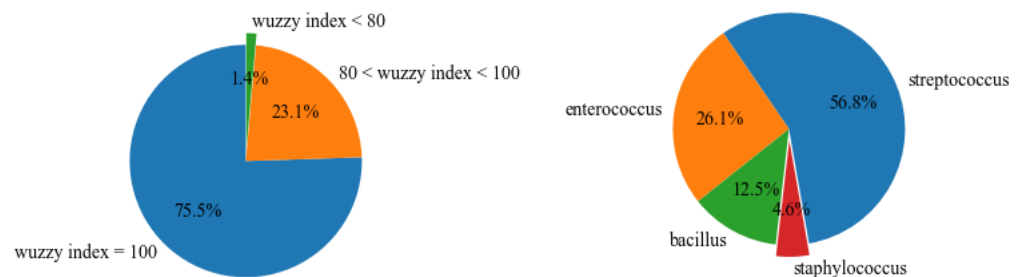


Figure 5. System performance for genus extraction. (a) Similarity index percentage distribution of genera. (b) Percentage distribution of genera found with low indices.

Figure 5b shows that ‘*Staphylococcus*’ was the genus name with the lowest score; in particular, it usually presented a very low similarity index, between 76 and 80, if a species was not specified. Indeed, we frequently found not only the strictly scientific term, but also the Italian term in the notes, because *Staphylococcus* is among the most widespread bacteria and is frequently mentioned in the common discourse.

This behavior affected the similarity index; in particular, *Staphylococcus* and ‘stafilococco/stafilococchi’ (Italian terms referring to the *Staphylococcus* genus) have 14 and 12 letters, respectively, within which only nine coincide, representing a Levenshtein distance of 5 (i.e., five changes are needed to transform the first word into the other). On the other hand, species never showed a Wuzzy index lower than 88.

Lastly, we introduced a weight, which was a decimal parameter ranging between 0 and 1. It could be associated with the genus–species couple or only with the genus, if present alone. As the same word (genus and/or species) could be associated with more than one genus/species, this process was necessary to highlight the maximum Wuzzy indices. An example of the system output is shown in Table 1; the similarity index of the two genera *Acinetobacter* and *Acetobacter* was 92, which is quite high. Thus, in order to identify the correct genus without any doubt, we compared the following token with all species of that specific genus present in the vocabulary. If a match was found (with a Wuzzy index over 98), then we assigned to that genus–species couple a weight parameter equal to 1, while the others received a weight of 0.

Table 1. Example of the system output returned when the input was “Gram negativi profilo proteomico riferibile ad *A baumannii*. Propensione di *A baumannii* alla pan-resistenza eccetto colistina (Microorganismo alert)”. The displayed columns correspond to the genus from the vocabulary, the specific word or character in text which the genus matches to, the genus Wuzzy similarity index, the species from the vocabulary, the word in text which the species matches to, the species Wuzzy similarity index, the clinical note divided into tokens, and the weight parameter.

Genus	Match Genus	Wuzzy Index Genus	Species	Match Species	Wuzzy Index Species	Clinical Notes	Weight
<i>Acinetobacter</i>	A	100.0	<i>baumannii</i>	<i>baumannii</i>	100.0	[94,'propensione','NaN','Acinetobacter','b ...	1
<i>Acetobacter</i>	A	92.0	NaN	NaN	NaN	[94,'propensione','NaN','Acinetobacter','b ...	0
<i>Aminobacter</i>	A	83.0	NaN	NaN	NaN	[94,'propensione','NaN','Acinetobacter','b ...	0
<i>Citrobacter</i>	A	83.0	NaN	NaN	NaN	[94,'propensione','NaN','Acinetobacter','b ...	0

Otherwise, if the clinical note did not contain any species, and the two genera that could correspond to the same word had an identical Wuzzy index, e.g., as a consequence of a spelling error, then the algorithm would assign to both genera an equal weight of 0.5.

3.4. Other Information Extraction

The whole sample included 48 clinical notes that contained the keyword “alert”. Our algorithm was able to correctly discriminate between the notes that contained the bi-gram “non-alert” ($N = 9$) and those that contained the keyword alone ($N = 39$).

4. Discussion

In general, the pattern recognition and the genus extension phases led to good results. The first achieved a mean accuracy value of 99.06% considering all three classifiers, while the second extracted all the names of microorganisms reported by the experts from the hospital. We should consider, however, that some ambiguities could be found during this second phase. Indeed, there are a few microorganisms with identical species and whose genera begin with the same letter. If one such case appears, then the system will duplicate the clinical note and it will extract both microorganisms, but both notes will be associated with a weight equal to 0.5. However, we should specify that, luckily, these kinds of ambiguities are quite rare. A well-known example is the *intermedius* species, which can belong to both the *Staphylococcus* genus and the *Streptococcus* genus. *Staphylococcus intermedius* is quite frequent in animals; however, it is reported as a human pathogen in very few cases, most of which are associated with exposure to animals, especially dogs. On the contrary, *Streptococcus intermedius* is one of the major causes of brain abscesses, but very few cases of this condition are documented annually in Italy, with an incidence that is less than 0.1% per year. Therefore, we can affirm that the probability that such ambiguity is present in the report notes of the microbiological laboratory is extremely rare.

The major result of our pipeline is that we can extract a wider picture of the microorganism, because each microorganism is stored together with other metadata in the build vocabulary, such as the definition according to MeSH and its translation into national

and international vocabularies. Furthermore, the pipeline also extracts the property of the microorganism under healthcare surveillance. Therefore, we can say that the system returns an object with its main characteristics. Once we accurately describe the microorganism, we can consider its identification in the clinical note as a trigger event of a series of messages and communications in accordance with the management policies of resistant microorganisms. Thus, it is possible to build a path to safeguard the patient and the community against the resistant microorganism [42]. The above-described system should be integrated in a multidisciplinary context. Correctly integrating objects from any viewpoint of the system in question requires its formal representation and management using the ISO 23903 Interoperability and Integration Reference Architecture [43]. ISO 23903 standardizes a model and framework for representing any type of system from the perspectives of the involved domains, its architectural composition/decomposition, and the related development process of implementable information and communications technology (ICT) solutions.

A limitation of the presented work is the low number of samples considered due to the fact that, to be delivered to researchers outside the laboratory, all these notes were checked individually and manually in order to avoid the illicit dissemination of personal data. In the near future, the correct structuring of electronic health records (which enables in constitutive law the reuse of clinical data for the purposes of scientific research) and greater awareness of the health risk that antibiotic-resistant bacteria constitute will result in a much higher number of notes to be analyzed. The more important methodological limitations of our project and ways to overcome them are discussed in the next section.

5. Future Work—Challenges and Solutions

Collecting, as well as storing and retrieving, representational objects of facts, systems, and processes is always a matter of the language and related grammar used to perform those actions. A simpler and more constrained language and ruling grammar, which is equivalent to the expressivity of languages/ontologies, facilitates processing of the outcome. However, highly expressive languages are less complete. This is a crucial challenge of knowledge representation in any business system including health and its special domains such as microbiology or infectious diseases.

Any business system can be represented using ICT ontologies. This holds for data stored in LIS databases, information models to represent the system's objects, and business process models representing the business processes needed to meet the intended business objectives. However, the justification of correctness and completeness of structure and behavior of the represented ecosystem can only be provided from the ecosystem's business view using the involved domains' ontologies. Justification of structure and behavior representation includes the representational components, their underlying concepts, their relations, and the related constraints. Figure 6 illustrates the related business system according to ISO 23903. The domains involved are clinical domains, managing patient care, and supporting facilities such as laboratories and microbiological departments to provide diagnostic services, as well as regulatory domains (policy domains) such as legal affairs, administration, security, and privacy management. Within the development process, the real-world system is then transformed into the different viewpoints of the intended ICT solution from the business process modeling (enterprise view) through the informational representation of all entities involved (information view) up to implementable artifacts (engineering view) and their management (technology view). The views in that order are represented by languages/grammars with increasing constraints and decreasing generative power, as well as decidability.

The technology view and engineering view are represented by data, while the computational view and information view are represented by information using related presentation styles such as programming languages or UML, respectively. The enterprise view represents the enterprise knowledge using business process modeling languages, and the business view represents the domain knowledge using domain ontologies (Figure 7). The

different levels in the model hierarchy allow for different actions necessary for designing and running the business process according to Krogstie [45].

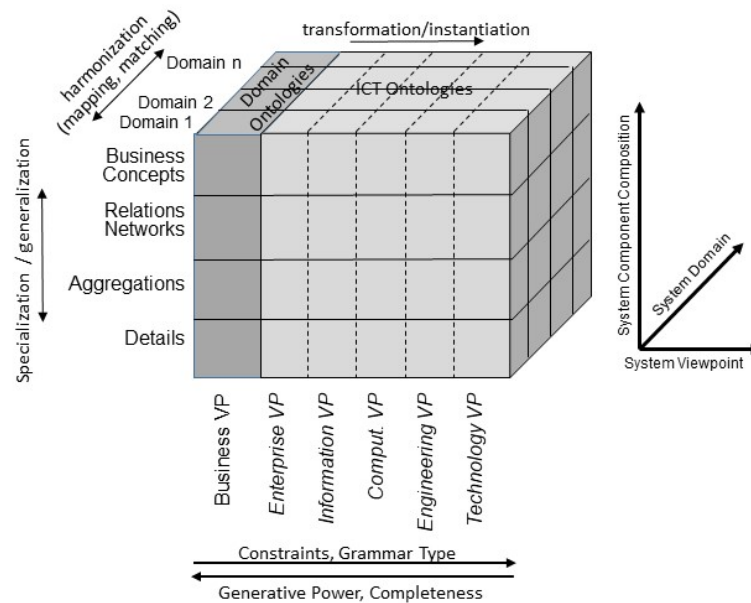


Figure 6. Generic business system representation according to ISO 23903, including the language/grammar characterization according to Chomsky [44].

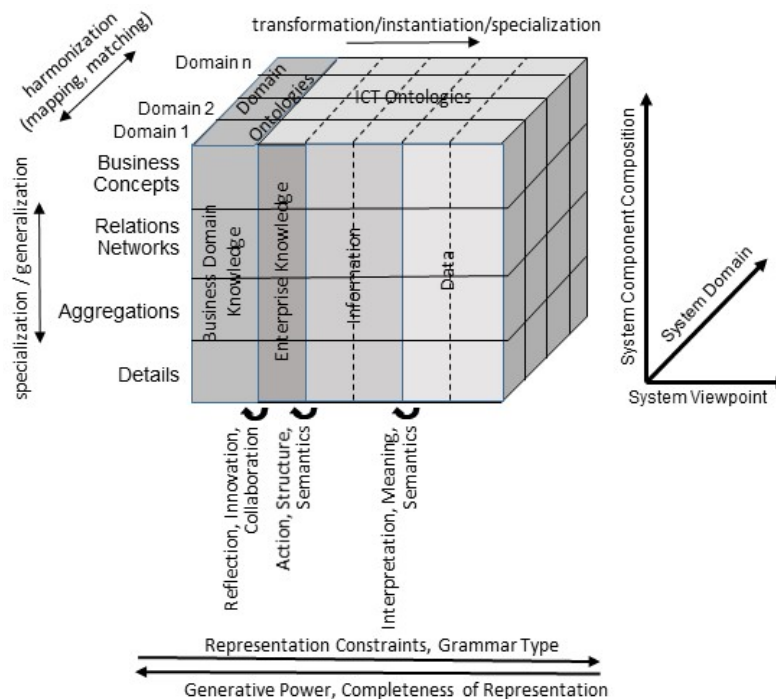


Figure 7. Generic business system representation according to ISO 23903 from the perspective of the knowledge pyramid after Aamodt and Nygard [46] and the model hierarchy after Krogstie [45].

For performing process-related actions, the enterprise view is necessary. For taking strategic and operational decisions, driving innovations, and enabling comprehensive collaboration, the representation of the business system in its comprehensive context using the ontologies of the directly and indirectly involved domains, guided by top-level ontologies according to ISO 21838 [47], is inevitable. In other words, the taxonomies used to analyze the business system must be replaced by ontologies, thereby not just considering

the domain knowledge, but also the knowledge space in question. This involves not only the naming of entities, but also the underlying concepts and comprehensive relations. More details on the challenges and solutions can be found, e.g., in [48] or [49], as well as in the introductory paper to this volume.

6. Conclusions

The main aim of this work was to develop an NLP pipeline to support the automatic extraction of the microorganisms' names and important information contained in microbiological notes of culture reports. We decided to preprocess the notes before the extraction process by removing not meaningful sentences, such as "integration of the provisional report of ...". We performed this task by applying machine learning methods to the numerical representation of the texts obtained using the bag of words technique. All the microorganisms present were extracted correctly; hence, the main goal was achieved. Lastly, considering that our vocabulary is based on international nomenclature standards, the presented pipeline can be applied to similar laboratory notes from other hospitals across the nation.

A next step of this project will be to also automatically extract from the same clinical notes the antibiogram data. In particular, key information that should be considered is the sensitivity of the specific microorganism to each single antibiotic tested. In the context of healthcare transformation toward pHealth or even 5P medicine (personalized, preventive, predictive, participative, and precision medicine), we have to advance the system further.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jpm12091424/s1>.

Author Contributions: Conceptualization, M.G., B.B. and R.G.; methodology, S.M.; software, J.A. and S.M.; validation, S.M.; formal analysis, S.M.; investigation, S.M.; resources, S.M.; data curation, G.P., J.A. and S.M.; writing—original draft preparation, S.M.; writing—review and editing, S.M., D.R.G., M.G. and B.B.; visualization, S.M.; supervision, M.G.; project administration, R.G.; funding acquisition, G.P. 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: Data used in this project are available as Supplementary Materials.

Acknowledgments: The authors want to thank the technical staff of the AUSL Pescara who contributed to the collection of the data used in the project.

Conflicts of Interest: Outside of the submitted work, D.R.G. reports investigator-initiated grants from Pfizer, Gilead Italia, and Shionogi, and speaker/advisory board fees from Pfizer and Tillotts Pharma. The other authors declare no conflict of interest.

References

1. Available online: <https://www.epicentro.iss.it/> (accessed on 25 August 2022).
2. Angela, R. Healthcare-associated infections: A public health problem. *Niger. Med. J. Niger. Med. Assoc.* **2012**, *53*, 59.
3. Huys, G.; Cnockaert, M.; Bartie, K.; Oanh, D.; Phuong, N.; Somsiri, T.; Chinabut, S.; Yussoff, F.; Shariff, M.; Giacomini, M.; et al. Intra- and interlaboratory performance of antibiotic disk-diffusion-susceptibility testing of bacterial control strains of relevance for monitoring aquaculture environments. *Dis. Aquat. Org.* **2005**, *66*, 197–204. [[CrossRef](#)] [[PubMed](#)]
4. Adamu, J.Y.; Raufu, A.I.; Chimaroke, F.C.; Ameh, J.A. Antimicrobial susceptibility testing of *Staphylococcus aureus* isolated from apparently healthy humans and animals in Maiduguri, Nigeria. *Int. J. Biomed. Health Sci.* **2021**, *6*, 4.
5. Magiorakos, A.-P.; Srinivasan, A.; Carey, R.B.; Carmeli, Y.; Falagas, M.E.; Giske, C.G.; Harbarth, S.; Hindler, J.F.; Kahlmeter, G.; Olsson-Liljequist, B.; et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: An international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* **2012**, *18*, 268–281. [[CrossRef](#)]
6. Basak, S.; Singh, P.; Rajurkar, M. Multidrug Resistant and Extensively Drug Resistant Bacteria: A Study. *J. Pathog.* **2016**, *2016*, 1–5. [[CrossRef](#)] [[PubMed](#)]

7. Organisation for Economic Co-operation and Development, European Centre for Disease Prevention and Control. *Antimicrobial Resistance—Tackling the Burden in the European Union—Briefing Note for EU/EEA Countries*; OECD: Paris, France, 2019. Available online: <https://www.oecd.org/health/health-systems/AMR-Tackling-the-Burden-in-the-EU-OECD-ECDC-Briefing-Note-2019.Pdf> (accessed on 25 August 2022).
8. Centers for Disease Control and Prevention. 2019. Available online: <https://www.cdc.gov/drugresistance/pdf/threats-report/2019-ar-threats-report-508.pdf> (accessed on 25 August 2022).
9. Timsit, J.-F.; Ruppé, E.; Barbier, F.; Tabah, A.; Bassetti, M. Bloodstream infections in critically ill patients: An expert statement. *Intensive Care Med.* **2020**, *46*, 266–284. [[CrossRef](#)]
10. Giacobbe, D.R.; Roberts, J.A.; Abdul-Aziz, M.H.; de Montmollin, E.; Timsit, J.F.; Bassetti, M. Treatment of ventilator-associated pneumonia due to carbapenem-resistant Gram-negative bacteria with novel agents: A contemporary, multidisciplinary ESGCIP perspective. *Expert Rev. Anti Infect.* **2022**, *20*, 963–979. [[CrossRef](#)]
11. Bassetti, M.; Poulakou, G.; Ruppe, E.; Bouza, E.; Van Hal, S.J.; Brink, A. Antimicrobial resistance in the next 30 years, humankind, bugs and drugs: A visionary approach. *Intensive Care Med.* **2017**, *43*, 1464–1475. [[CrossRef](#)]
12. Costa, D.; Martins, P.; Loureiro, L.; Matos, A.J.F. Transfer of multidrug-resistant bacteria between intermingled ecological niches: The interface between humans, animals and the environment. *Int. J. Environ. Res. Public Health* **2013**, *10*, 278–294. [[CrossRef](#)]
13. Saud, B.; Paudel, G.; Khichaju, S.; Bajracharya, D.; Dhungana, G.; Awasthi, M.S.; Shrestha, V. Multidrug-Resistant Bacteria from Raw Meat of Buffalo and Chicken, Nepal. *Vet. Med. Int.* **2019**, *2019*, 7960268. [[CrossRef](#)]
14. Rahman, M.; Husna, A.; Elshabrawy, H.A.; Alam, J.; Runa, N.Y.; Badruzzaman, A.T.M.; Banu, N.A.; Al Mamun, M.; Paul, B.; Das, S.; et al. Isolation and molecular characterization of multidrug-resistant *Escherichia coli* from chicken meat. *Sci. Rep.* **2020**, *10*, 21999. [[CrossRef](#)] [[PubMed](#)]
15. Jeżak, K.; Kozajda, A. Occurrence and spread of antibiotic-resistant bacteria on animal farms and in their vicinity in Poland and Ukraine—review. *Environ. Sci. Pollut. Res.* **2021**, *29*, 9533–9559. [[CrossRef](#)]
16. Huys, G.; Bartie, K.; Cnockaert, M.; Oanh, D.T.H.; Phuong, N.T.; Somsiri, T.; Chinabut, S.; Yusoff, F.M.; Shariff, M.; Giacomini, M.; et al. Biodiversity of chloramphenicol-resistant mesophilic heterotrophs from Southeast Asian aquaculture environments. *Res. Microbiol.* **2007**, *158*, 228–235. [[CrossRef](#)]
17. Pham, T.T.H.; Rossi, P.; Dinh, H.D.K.; Pham, N.T.A.; Tran, P.A.; Ho, T.T.K.M.; Dinh, Q.T.; De Alencastro, L.F. Analysis of antibiotic multi-resistant bacteria and resistance genes in the effluent of an intensive shrimp farm (Long An, Vietnam). *J. Environ. Manag.* **2018**, *214*, 149–156. [[CrossRef](#)] [[PubMed](#)]
18. Higuera-Llantén, S.; Vásquez-Ponce, F.; Barrientos-Espinoza, B.; Mardones, F.; Marshall, S.H.; Olivares-Pacheco, J. Extended antibiotic treatment in salmon farms select multiresistant gut bacteria with a high prevalence of antibiotic resistance genes. *PLoS ONE* **2018**, *13*, e0203641. [[CrossRef](#)] [[PubMed](#)]
19. Grimson, W.; Brender, J.; Grimson, J.; Groth, T.; Hermanson, B.; Yearworth, M.; Wade, V. Specifying an open clinical laboratory information system. *Comput. Methods Programs Biomed.* **1996**, *50*, 95–109. [[CrossRef](#)]
20. Aller, R.D. Software standards and the laboratory information system. *Am. J. Clin. Pathol.* **1996**, *105*, S48–S53.
21. Gazzarata, R.; Monteverde, M.E.; Vio, E.; Saccavini, C.; Gubian, L.; Borgo, I.; Giacomini, M. A Terminology Service Compliant to CTS2 to Manage Semantics within the Regional HIE. *Eur. J. Biomed. Inform.* **2017**, *13*. [[CrossRef](#)]
22. Available online: <https://www.omg.org/> (accessed on 25 August 2022).
23. Available online: <https://www.hl7.org/> (accessed on 25 August 2022).
24. Matheny, M.E.; Fitzhenry, F.; Speroff, T.; Hathaway, J.; Murff, H.J.; Brown, S.H.; Fielstein, E.M.; Dittus, R.S.; Elkin, P.L. Detection of blood culture bacterial contamination using natural language processing. *AMIA Annu. Symp. Proc. AMIA Symp.* **2009**, *2009*, 411–415.
25. Maganti, N.; Tan, H.; Niziol, L.M.; Amin, S.; Hou, A.; Singh, K.; Ballouz, D.; Woodward, M.A. Natural Language Processing to Quantify Microbial Keratitis Measurements. *Ophthalmology* **2019**, *126*, 1722–1724. [[CrossRef](#)]
26. Fu, S.; Wyles, C.C.; Osmon, D.R.; Carvour, M.L.; Sagheb, E.; Ramazanian, T.; Kremers, W.K.; Lewallen, D.G.; Berry, D.J.; Sohn, S.; et al. Automated Detection of Periprosthetic Joint Infections and Data Elements Using Natural Language Processing. *J. Arthroplast.* **2020**, *36*, 688–692. [[CrossRef](#)] [[PubMed](#)]
27. Gazzarata, R.; Monteverde, M.E.; Bonetto, M.; Savini, V.; Polilli, E.; Corridoni, S.; Costantini, A.; Santoleri, F.; Rapacchiale, G.; Palmieri, D.; et al. A SOA based solution for MDRO surveillance and improved antibiotic pre-prescription in the Abruzzo region. *pHealth* **2019**, *2019*, 49–54.
28. Mora, S.; Attene, J.; Gazzarata, R.; Parruti, G.; Giacomini, M. A NLP Pipeline for the Automatic Extraction of Microorganisms Names from Microbiological Notes. *Stud. Health Technol. Inform.* **2021**, 153–158. [[CrossRef](#)]
29. Pandas. Available online: <https://pandas.pydata.org/> (accessed on 25 August 2022).
30. Steven, B. NLTK: The natural language toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia, 17–21 July 2006.
31. Honnibal, M.; Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Appear* **2017**, *7*, 411–420.
32. Van Rossum, G. *The Python Library Reference, Release 3.8.2*; Python Software Foundation: Fredericksburg, Virginia, 2020.
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

34. Waskom, M.L. seaborn: Statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [[CrossRef](#)]
35. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [[CrossRef](#)]
36. Available online: <https://github.com/seatgeek/thefuzz> (accessed on 25 August 2022).
37. Ghosh, S.; Dasgupta, A.; Swetapadma, A. A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification. In Proceedings of the 2019 International Conference on Intelligent Sustainable Systems (ICISS), IEEE, Palladam, India, 21–22 February 2019; pp. 24–28. [[CrossRef](#)]
38. Chapelle, O.; Haffner, P.; Vapnik, V. Support vector machines for histogram-based image classification. *IEEE Trans. Neural Netw.* **1999**, *10*, 1055–1064. [[CrossRef](#)]
39. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: New York, NY, USA, 2002.
40. Leo, B. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
41. Linnaeus, C. *Systema naturae*. *Stockh. Holmiae (Laurentii Salvii)* **1758**, *10*, 551.
42. Mondain, V.; Secondo, G.; Guttman, R.; Ferrea, G.; Dusi, A.; Giacomini, M.; Courjon, J.; Pradier, C. A toolkit for the management of infection or colonization by extended-spectrum beta-lactamase producing Enterobacteriaceae in Italy: Implementation and outcome of a European project. *Eur. J. Clin. Microbiol.* **2018**, *37*, 987–992. [[CrossRef](#)]
43. *ISO 23903:2021*; Interoperability and Integration Reference Architecture–Model and Framework. ISO: Geneva, Switzerland, 2021.
44. Chomsky Hierarchy in Theory of Computation. Available online: <https://www.geeksforgeeks.org/chomsky-hierarchy-in-theory-of-computation/> (accessed on 25 August 2022).
45. Krogstie, J. Business Information Systems Utilizing the Future Internet. *Data Knowl. Eng.* **2011**, *90*, 1–18. [[CrossRef](#)]
46. Aamodt, A.; Nygård, M. Different roles and mutual dependencies of data, information, and knowledge—An AI perspective on their integration. *Data Knowl. Eng.* **1995**, *16*, 191–222. [[CrossRef](#)]
47. *ISO/IEC 21838:2021*; Information Technology–Top-Level Ontologies (TLO). ISO: Geneva, Switzerland, 2021.
48. Blobel, B.; Ruotsalainen, P.; Oemig, F. Why Interoperability at Data Level Is Not Sufficient for Enabling pHealth? In *pHealth*; IOS Press: Amsterdam, The Netherlands, 2020; Volume 273, pp. 3–20. [[CrossRef](#)]
49. Blobel, B.; Oemig, F.; Ruotsalainen, P.; Lopez, D.M. Transformation of Health and Social Care Systems—An Interdisciplinary Approach Toward a Foundational Architecture. *Front. Med.* **2022**, *9*, 802487. [[CrossRef](#)]