A NLP pipeline for the automatic extraction of a complete microorganism's picture from microbiological notes

Sara MORA [a,1], Jacopo ATTENE [a], Roberta GAZZARATA [b], Daniele Roberto GIACOBBE [c], Bernd Blobel [d], Giustino PARRUTI [e] and Mauro GIACOMINI [a]

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

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

c Infectious Disease Clinic, IRCCS Policlinic San Martino Hospital, 16132 Genoa, Italy

d Medical Faculty, University of Regensburg, Regensburg, Germany

e Department of Infectious Diseases, AUSL Pescara, Pescara, Italy

Abstract. The *"Istituto Superiore di Sanita`"* (ISS) identifies the *hospital-acquired infections* (HAIs) as the most frequent and serious complications in health care. HAIs constitute a real health emergency and therefore require decisive action from both local and national health organizations. Information about the presence of HAIs is obtained from microbiological cultures of specimens collected from infected body sites, but the outcomes are usually reported in the notes field of the laboratory exams' results. The objective of our work is to build an NLP-based pipeline for the automatic information extraction from the notes of microbiological culture reports. We analyzed a sample composed of 499 texts extracted from 1 month of anonymized laboratory referral. First, our system filtered texts in order to remove non-meaningful 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

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 *"Istituto Superiore di Sanita`"* (ISS) [1] identifies the *hospital-acquired infections* (HAI) as the most frequent and serious complications of health care. A possible definition of HAI is *"infections that first appear 48 hours or more after hospital admission or no later than 30 days after discharge following in patient 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 the reporting and epidemiological surveillance. When a pathogenic microorganism, multi-resistant or not, responsible for HAI, is identified, they allow the prompt adoption of specific control measures. 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]. *Multi-drugs resistant bacteria* (MDR) are defined as those organisms that in *in-vitro* antimicrobial susceptibility tests show resistance to one or more agents in at least three antimicrobial categories [5,6].

MDR organisms are a world health problem that causes about 25 000 deaths per year in Europe [7] and 23 000 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 an 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) [13,14,15] and sea farming [16,17,18].

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 information systems development has been non-simultaneous and strongly localized. The resulting needs of making 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 the 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., which coding system the nomenclature of bacteria relies on, 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 show as too stiff. Therefore and because of the lack of ad hoc and appropriate fields for representing and managing microbiological-related information, the staff preferred over years exploiting clinical notes written as natural text instead. That way, clinical notes became an important source of information, both for biomedical research, patient management and care, but 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, will be discussed in more detail in Section 4.

This kind of problem can be addressed with *Artificial Intelligence* (AI) tools and 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 spoken form [24,25,26].

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

It could represent a first step towards 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 & 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 not significant sentences.

2. Materials & Methods

2.1. Characteristics of the sample

The collected sample derives from the main hospital of Pescara in Abruzzo Region and is obtained from clinical notes extracted from a 1-month collection of anonymized laboratory referral. It is composed of 499 texts from culture reports, among them:

- 276 *(55.3%)* contain the name of a microorganism and an expert from the hospital confirmed its presence.
- 56 *(11.2%)* should be filtered because they contain a pattern that is not useful for our analysis so we decided to remove it. An example of a sentence belonging to that pattern is: *"Integration of the preliminary report sent on ..."*, but it is only one of the possibilities. Indeed, we should consider 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 only use regular expressions.

We expect to be able to acquire more data in the near future.

2.2. Environment & Libraries

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

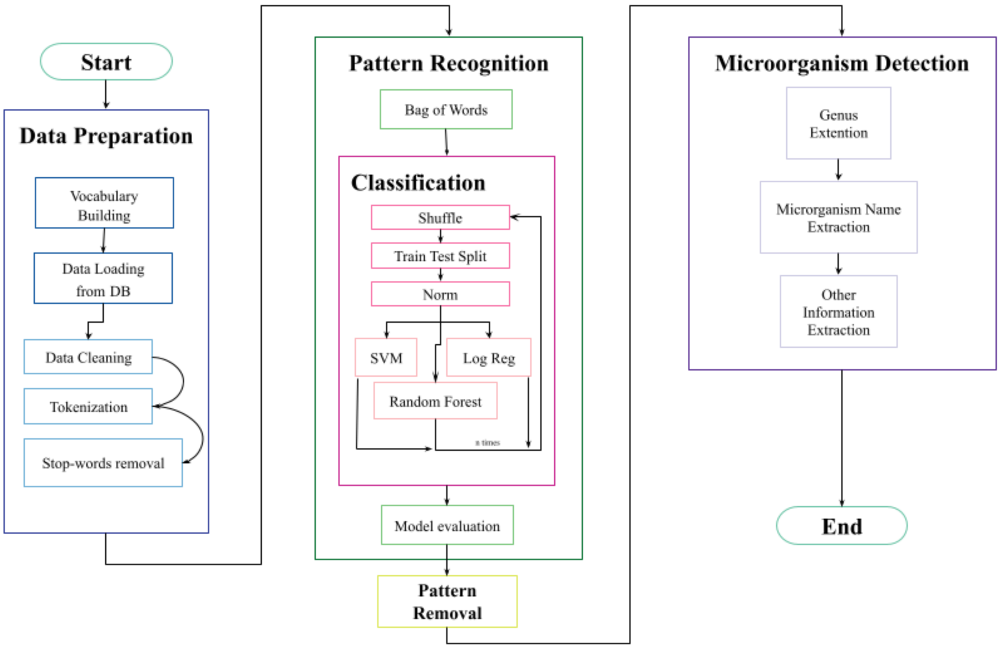


Figure 1. *Complete schema of the pipeline. It can be divided into 4 main sections: data preparation, pattern recognition and removal, microorganism detection.*

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

2. **Pyodbc:** This is an open source module developed in Python that allows accessing databases through the ODBC *(Open DataBase Connectivity)*.
3. ***Natural Language Toolkit* (NLTK):** This is a worldwide used library to perform text analysis in multiple languages, therefore it is popular both in academia and for research [30]. Some of the operations supported by nltk are: tokenization, stemming, *part of speech tagging* (POS tagging) and disambiguation.
4. **SpaCy:** This is an open source library for NLP in Python supporting different languages [31].
5. **Re:** This is a Python module that provides operations useful to work with regular expressions [32].
6. **Scikit-learn:** This is an open source library that contains several *machine learning* (ML) algorithms, e.g. classification, regression, clustering, etc [33].
7. **Seaborn, matplotlib:** These are libraries used to produce graphics [34,35].
8. **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 words' order and the allowed maximum frequency of a string. This comparison is based on the *Levenshtein distance*:

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

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

2.3. Data Preparation

Vocabulary building: We built a vocabulary containing the names of microorganisms (bacteria, fungi, yeasts, viruses) from the "**National Healthcare Safety Network** organism list", including the current taxonomic subdivision which was proposed by Carl Woase in 1990. We mapped the microorganism's genus and specie into 4 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: We used the pandas library to import data.

Data cleaning, tokenization, stopwords removal: 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, that we call tokens. Then we proceeded with stopwords removal, but considering only words longer than 1 character in order to exclude from the analysis strings belonging to the class of prepositions, articles and adverbs, but keeping single letters that could be the abbreviation of a genus' name.

2.4. Pattern Recognition

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.
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 to capture a little of 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 step 10. We obtained the best performance with a total number of features equal to 90.

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:

SVM is a supervised learning method that can be used to perform classification analysis on both linear and non-linear 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 dimension 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 a 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 hyperparameters search through a ten-fold cross-validation, which iteratively split the learning set into 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 twenty times, shuffling the data each time. In order to guarantee reproducibility of results, we set the random state equal to the loop index.

2.5. Pattern Removal

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

The schema of the regular expression is: **\b[Ii]\w.+?\bdata\b**

where:

\b asserts position of a word boundary. In this case we want that: the pattern begins with ‘I’ (the first letter of the word ‘Integrazione’ (integration) which can be abbreviated and/or it can be capital or lowercase in the notes); **\w** matches any word character and ends with ‘data’ (the word that we substituted to the all dates in the data cleaning phase).

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

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

2.6. Genus Extension

We stored the microorganism names using the binomial nomenclature which originates from the Linnaeus classification [41]. It is composed of two terms: first the name of the genus with the first letter capitalized, second the name of the species in lower case. Usually, after a microorganism’s name is introduced once in a text, the genus can then be abbreviated to the first letter (followed by a fullstop) in subsequent mentions. However, considering the shortness of the clinical notes, a shared agreement is to always use the abbreviated form, even though the entire genus has not been introduced yet. This binomial nomenclature does not allow the use of a two-letter abbreviation for the genus. Anyway, even if 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 could be spelled incorrectly, e.g., by using abbreviations not followed by the fullstop, or letters followed by the fullstop without using lower-case letters. Then we performed the extension of the microorganism genus. In detail, we compared the “n+1” token with each species of the vocabulary. If the two tokens were very close (similarity index 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” by the genus name found in the vocabulary.

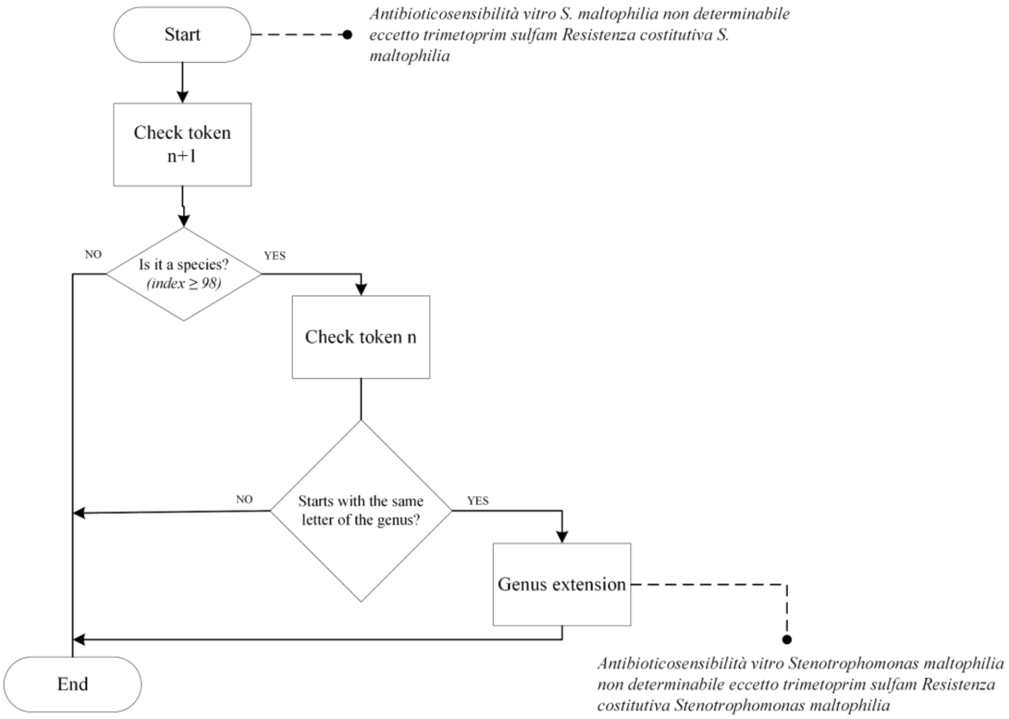

Start
Antibioticosensibilità vitro S. maltophilia non determinabile eccetto trimetoprim sulfam Resistenza costitutiva S. maltophilia
Check token n+1
NO
Is it a species? (index ≥ 98)
YES
Check token n
NO
Starts with the same letter of the genus?
YES
Genus extension
End
Antibioticosensibilità vitro Stenotrophomonas maltophilia non determinabile eccetto trimetoprim sulfam Resistenza costitutiva Stenotrophomonas maltophilia

Figure 2: Genus extension decision flow. *The figure also includes an example:. In the upper part of the picture, there is a sentence already preprocessed before the genus extension phase, while in the lower part we can see the extended version. First, maltophilia species is identified, and then, "S" as the first letter of genus Stenotrophomonas has been extended.*

2.7. Information Extraction

Initially, we tried to carry out a lexical and morphological analysis, but the lack of morphological structure of the reports did not produce any good results. Therefore, we extracted the microorganism name by comparing each token found within the document and the vocabulary using the FuzzyWuzzy library. In particular, considering the extraction of the genus we set the threshold on the similarity index at 75 and the threshold for the species at 85 (as usually they were written correctly).

Together with the identification of genus and species in order to highlight microorganisms that could be potentially dangerous, we searched into the clinical notes also for the keyword "alert", which is an explicit indication of microbiologists about the danger of the identified microorganism. Similarly, but much less frequent, 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 research at token level of the keyword "alert". If identified at the n position, we checked if token n-1 matched the negation "not".

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 non meaningful sentences, i.e., *"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.

Fig. 3 summarizes the mean values of the results obtained by 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 showed a mean accuracy of 98.99% and Random Forest showed a mean accuracy of 99.19%.

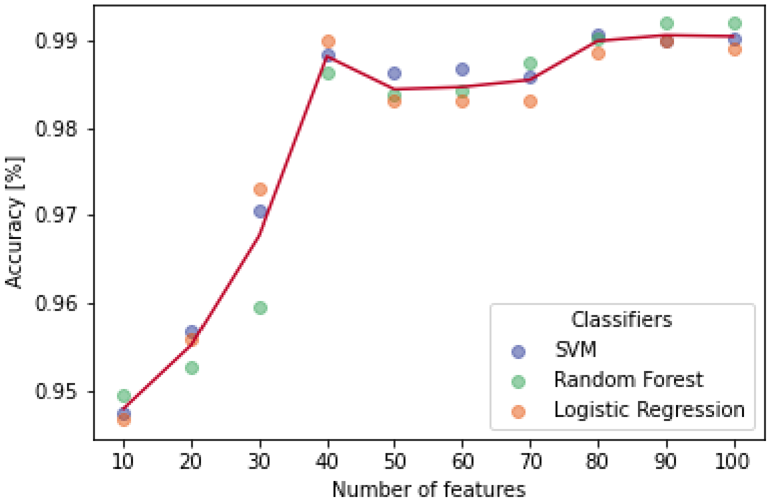


Figure 3. Mean accuracy performances *of the three classifiers displayed for each value of total number of features. Each data point is the mean value of twenty values obtained by shuffling the data.*

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

3.2. Genus Extension

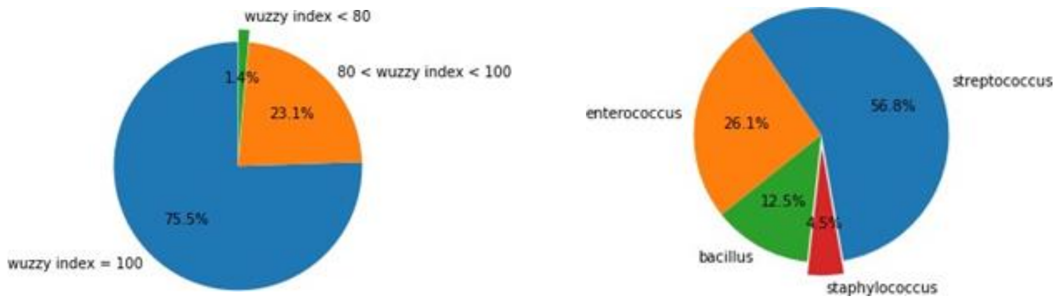
The sample of clinical notes contained 107 abbreviated genera followed by species. After that, the system elaborated the notes, all 107 genera were extended and completely matched with the indications of the expert.

3.3. Microorganism Detection

The sample of available clinical notes was composed of 499 texts, 276 *(55,3%)* of them actually presented the name of a microorganism. We performed two tests:

1. First, we introduced all 499 notes into the module for microorganism extraction.
2. Then we introduced only the notes that actually contained the microorganisms.

The system correctly identified all the microorganisms names in both cases. In detail, it found 416 genera of microorganisms, the majority of them (321) with a Wuzzy index of 100. This is also a consequence of the process of genus extension.



(a) Similarity indexes percentage distribution of genera.

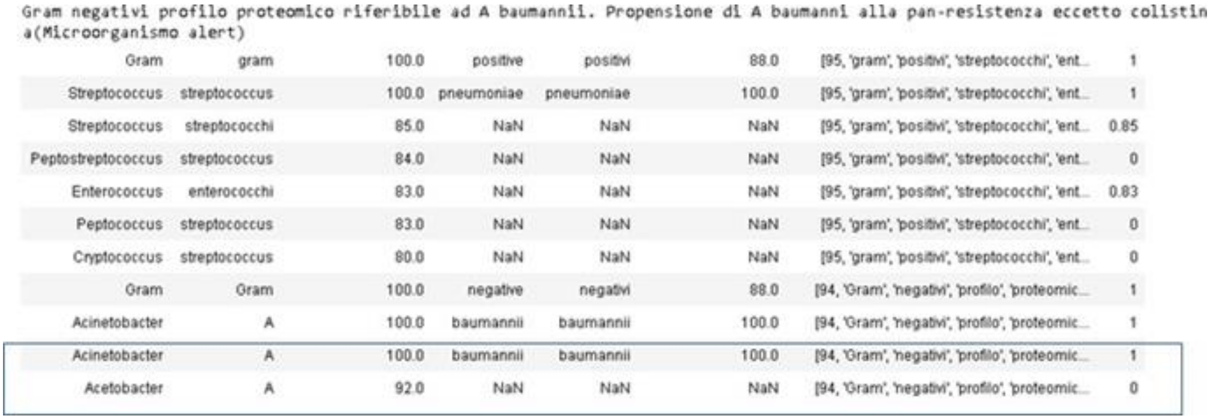
(b) Percentage distribution of genera which found with a low indexes

Figure 4. *System performance on genera extraction.*

Figure 4b shows that '*Staphylococcus*' is the microorganism genus with the lowest score, and in particular, it tends to have a very low similarity index between 76 and 80, if a species is not specified. Indeed, frequently we did not find just the strictly scientific term, but also the Italian term in the notes, because *Staphylococcus* are among the most widespread bacteria, so it is common to mention it in the context of common discourse.

This behavior impacts the similarity index, in particular *Staphylococcus* and 'stafilococco/stafilococchi' (that are the Italian terms referable to the *Staphylococcus* genus) have 14 and 12 letters respectively; only 9 letters coincide, so they have a Levenshtein distance of 5 (5 changes are needed to transform the first word into the other one). On the other hand, species never showed a Wuzzy index lower than 88.

Finally, we introduced a weight parameter, i.e., a decimal value between 0 and 1, associated to each couple of genus and species, or only to the genus if present alone. This process highlighted the maximum Wuzzy indexes, because the same word (genus and/or species) could be associated with more than one genus/species. For example, the genera *Acetobacter* and *Acinetobacter* have a similarity index of 92, which is quite high. In order to identify the correct genus, we compared the following token with all the species of that specific genus present in the vocabulary. If a matching was found (with Wuzzy index over 98), then that genus got a weight equal to 1, and the others a weight of 0.



```
Gram negativi profilo proteomico riferibile ad A baumannii. Propensione di A baumanni alla pan-resistenza eccetto colistin
a(Microorganismo alert)
               Gram           gram    100.0    positive     positivi    88.0  [95, 'gram', 'positivi', 'streptococchi', 'ent...     1
      Streptococcus  streptococcus    100.0  pneumoniae   pneumoniae   100.0  [95, 'gram', 'positivi', 'streptococchi', 'ent...     1
      Streptococcus   streptococchi    85.0         NaN          NaN     NaN  [95, 'gram', 'positivi', 'streptococchi', 'ent...  0.85
 Peptostreptococcus  streptococcus     84.0         NaN          NaN     NaN  [95, 'gram', 'positivi', 'streptococchi', 'ent...     0
       Enterococcus   enterococchi     83.0         NaN          NaN     NaN  [95, 'gram', 'positivi', 'streptococchi', 'ent...  0.83
        Peptococcus  streptococcus     83.0         NaN          NaN     NaN  [95, 'gram', 'positivi', 'streptococchi', 'ent...     0
       Cryptococcus  streptococcus     80.0         NaN          NaN     NaN  [95, 'gram', 'positivi', 'streptococchi', 'ent...     0
               Gram           Gram    100.0    negative     negativi    88.0  [94, 'Gram', 'negativi', 'profilo', 'proteomic...     1
      Acinetobacter              A    100.0   baumannii    baumannii   100.0  [94, 'Gram', 'negativi', 'profilo', 'proteomic...     1
      Acinetobacter              A    100.0   baumannii    baumannii   100.0  [94, 'Gram', 'negativi', 'profilo', 'proteomic...     1
        Acetobacter              A     92.0         NaN          NaN     NaN  [94, 'Gram', 'negativi', 'profilo', 'proteomic...     0
```

Figure 5. Example of the system output *provided for the input the sentence that can be seen at the beginning of the figure. The displayed columns correspond to: genus from vocabulary, specific word in text which the genus matches to, genus Wuzzy index, species from vocabulary, word in text which the species matches to, species Wuzzy index, clinical note divided into tokens, weight.*

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. due to a spelling error, then the algorithm will return both genera but the weight will be equal to 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 one achieved a mean accuracy value of 99.06% considering all the three classifiers, and the second one extracted all the names of microorganisms reported by the experts from the hospital. We should consider however that during this second phase some ambiguities can be found. Indeed, there are several microorganisms with identical species and genera, which begin with the same letter. If one of such cases 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 *Streptococcus* genus and *Staphylococcus* genus. *Staphylococcus intermedius* is quite frequent in animals, but in very few cases, it is reported as a human pathogen, and most of them are associated with an exposure to animals, especially dogs. Contrary, *Streptococcus intermedius* is one of the major causes of brain abscesses, but just very few cases of brain abscesses are annually documented in Italy. In fact, the incidence 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 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 described the microorganism, we can consider the 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 at any viewpoint of a 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 systems from the perspectives of the involved domains, its architectural composition/decomposition and the related development process of implementable information and communication 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 have been checked individually and manually in order to avoid the illicit dissemination of personal data. In the near future, the correct structuring of the electronic health record (which enables in its 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 following section.

5. Future Work – Challenges and Solutions

Collecting and storing as well as retrieving representational objects of facts, systems and processes is always a matter of the language and related grammar used to perform those actions. As simpler and more constrained a language and the ruling grammar is, which is equivalent to the expressivity of languages/ontologies, as easier can the outcome be processed. 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 at 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, but also the related constraints. Figure 6 illustrates the related business system according to ISO 23903. The domains involved are clinical domains, managing patients' care, supporting facilities such as laboratories and microbiological departments to provide diagnostic services, but also regulatory domains (policy domains) such as legal affairs, administration, security and privacy management, etc. 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.

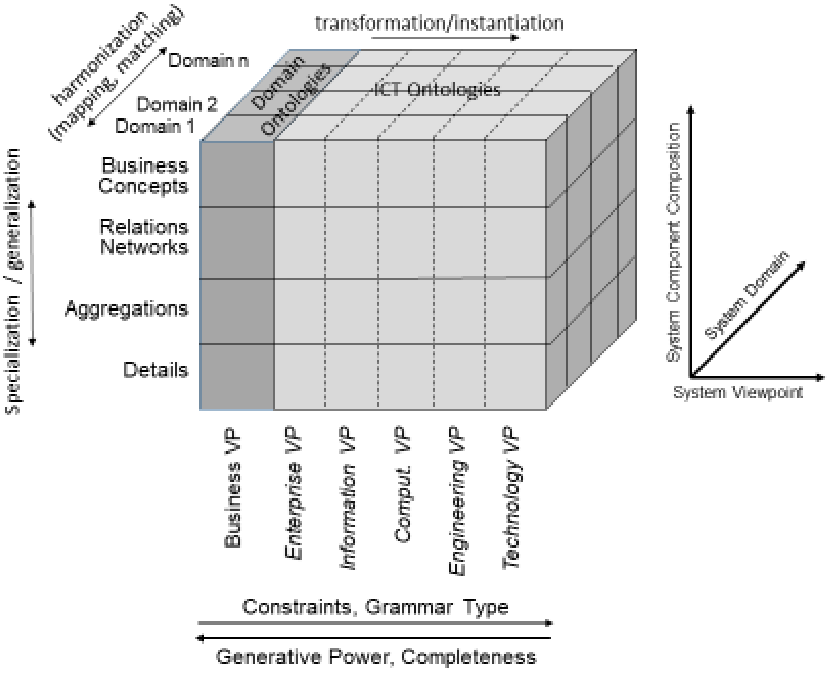


Figure 6. Generic business system representation according to ISO 23903, including the language/grammar characterization according to Chomsky [44]

Technology View and Engineering View are represented by data, Computational View and Information View 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 the domain knowledge using domains 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].

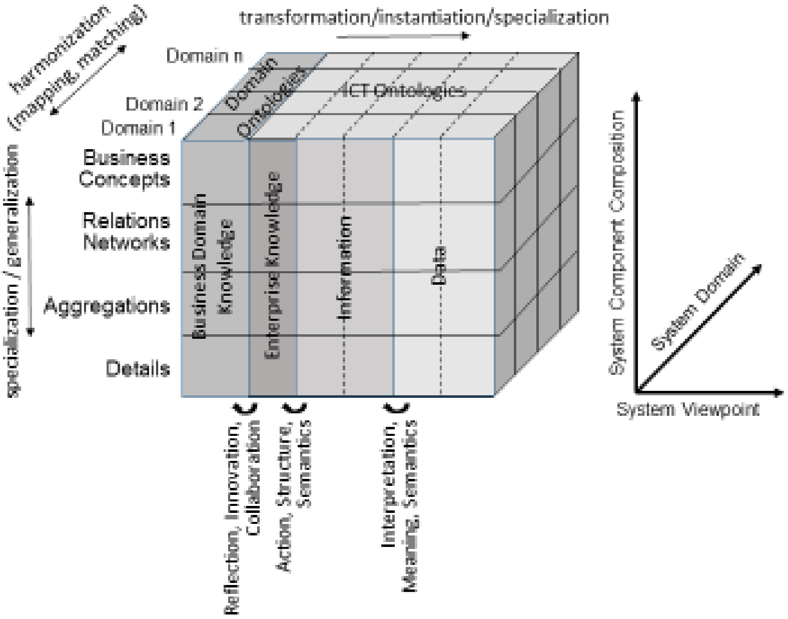

harmonization (mapping, matching)
transformation/instantiation/specialization
Domain n
Domain 2
Domain 1
Domain Ontologies
ICT Ontologies
Business Concepts
Relations Networks
Aggregations
Details
specialization / generalization
Business Domain Knowledge
Enterprise Knowledge
Information
Data
System Component Composition
System Domain
System Viewpoint
Reflection, Innovation, Collaboration
Action, Structure, Semantics
Interpretation, Meaning, Semantics

Figure 6. 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 the knowledge space in question. Not just the naming of entities, but the underlying concepts and comprehensive relations. More details on the challenges and solutions can be found, e.g., in [48] or [49], but also in the introductory paper to this volume [50].

Conclusions

The main aim of this work was developing a NLP pipeline to support the automatic extraction of the microorganisms' names and important information contained in microbiological notes of culture reports. We decided to pre-process the notes before the extraction process by removing non 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 with the bag of words technique. All the microorganisms present were extracted correctly, so the main goal was achieved. Finally, considering that our vocabulary is based on international nomenclature standards, the presented pipeline can be applied to similar laboratory notes from other hospitals all over the national territory.
A next step of this project will be to automatically extract from the same clinical notes also the antibiotic prescription. In particular, a key information that should be considered is the sensitivity of the specific microorganism to each single antibiotic tested. In the context of healthcare transformation towards pHealth or even 5P medicine *(personalized, preventive, predictive, participative, precision medicine),* we have to advance the system further.

References

1. https://www.epicentro.iss.it/
2. Revelas, Angela. "Healthcare–associated infections: A public health problem." *Nigerian medical journal: journal of the Nigeria Medical Association* 53.2 (2012): 59.
3. Huys, Geert, et al. "Intra-and interlaboratory performance of antibiotic disk-diffusion-susceptibility testing of bacterial control strains of relevance for monitoring aquaculture environments." Diseases of aquatic organisms 66.3 (2005): 197-204.
4. Adamu, J. Y., et al. "Antimicrobial susceptibility testing of Staphylococcus aureus isolated from apparently healthy humans and animals in Maiduguri, Nigeria." International Journal of Biomedical and Health Sciences 6.4 (2021).
5. Magiorakos, A-P., et al. "Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance." Clinical microbiology and infection 18.3 (2012): 268-281.
6. Basak, Silpi, Priyanka Singh, and Monali Rajurkar. "Multidrug resistant and extensively drug resistant bacteria: a study." Journal of pathogens 2016 (2016).
7. https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/0909_TER_The_Bacterial_Challenge_Time_to_React.pdf

8. Centers for Disease Control and Prevention (U.S.);National Center for Emerging Zoonotic and Infectious Diseases (U.S.);National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (U.S.);National Center for Immunization and Respiratory Diseases (U.S.); Antibiotic resistance threats in the United States, 2013;2013; https://stacks.cdc.gov/view/cdc/20705
9. Ang, Hui, and Xuan Sun. "Risk factors for multidrug-resistant Gram-negative bacteria infection in intensive care units: A meta-analysis." International journal of nursing practice 24.4 (2018): e12644.
10. Siwakoti, Shraddha, et al. "Incidence and outcomes of multidrug-resistant gram-negative bacteria infections in intensive care unit from Nepal-a prospective cohort study." Antimicrobial Resistance & Infection Control 7.1 (2018): 1-8.
11. Tosi, Martina, et al. "Multidrug resistant bacteria in critically ill patients: a step further antibiotic therapy." Focused Issue: Infections in ICU Guest Editors: George Dimopoulos, MD, PhD Yuetian Yu, MD Zhongheng Zhang, MD, MM (2019).
12. Da Costa, Paulo Martins, Luís Loureiro, and Augusto JF Matos. "Transfer of multidrug-resistant bacteria between intermingled ecological niches: the interface between humans, animals and the environment." International journal of environmental research and public health 10.1 (2013): 278-294.
13. Saud, Bhuvan, et al. "Multidrug-resistant bacteria from raw meat of buffalo and chicken, Nepal." Veterinary medicine international 2019 (2019).
14. Rahman, Md, et al. "Isolation and molecular characterization of multidrug-resistant Escherichia coli from chicken meat." Scientific Reports 10.1 (2020): 1-11.
15. Jeżak, Karolina, and Anna Kozajda. "Occurrence and spread of antibiotic-resistant bacteria on animal farms and in their vicinity in Poland and Ukraine." Environmental Science and Pollution Research (2021): 1-27.
16. G. Huys, et al. “Biodiversity of chloramphenicol-resistant mesophilic heterotrophs from Southeast Asian aquaculture environments” Research in Microbiology, 2007, vol. 158, pp. 228-235
17. Pham, Thi Thu Hang, et al. "Analysis of antibiotic multi-resistant bacteria and resistance genes in the effluent of an intensive shrimp farm (Long An, Vietnam)." Journal of environmental management 214 (2018): 149-156.
18. Higuera-Llantén, Sebastián, et al. "Extended antibiotic treatment in salmon farms select multiresistant gut bacteria with a high prevalence of antibiotic resistance genes." PLoS One 13.9 (2018): e0203641.
19. Grimson, William, et al. "Specifying an open clinical laboratory information system." Computer methods and programs in biomedicine 50.2 (1996): 95-109.
20. Aller, Raymond D. "Software standards and the laboratory information system." American journal of clinical pathology 105.4 Suppl 1 (1996): S48-53.
21. Gazzarata, Roberta, et al. A terminology service compliant to CTS2 to manage semantics within the regional HIE., European Journal of Biomedical Informatics 13.1 (2017).
22. https://www.omg.org/
23. https://www.hl7.org/
24. Matheny, Michael E., et al. Detection of blood culture bacterial contamination using natural language processing., AMIA Annual Symposium Proceedings. Vol. 2009. American Medical Informatics Association, 2009.
25. Maganti, Nenita, et al. Natural language processing to quantify microbial keratitis measurements., Ophthalmology 126.12 (2019): 1722-1724.
26. Fu, Sunyang, et al. Automated detection of periprosthetic joint infections and data elements using natural language processing., The Journal of Arthroplasty 36.2 (2021): 688-692.
27. Gazzarata, Roberta, et al. A SOA based solution for MDRO surveillance and improved antibiotic prescription in the Abruzzo region., pHealth 2019. IOS Press, 2019. 49-54.
28. Mora, Sara, et al. "A NLP Pipeline for the Automatic Extraction of Microorganisms Names from Microbiological Notes." pHealth. 2021.
29. https://pandas.pydata.org/
30. Bird, Steven. NLTK: the natural language toolkit., Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. 2006.

31. Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
32. Van Rossum, G. (2020). The Python Library Reference, release 3.8.2. Python Software Foundation.
33. Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
34. Waskom, Michael L. "Seaborn: statistical data visualization." Journal of Open Source Software 6.60 (2021): 3021.
35. Hunter, John D. "Matplotlib: A 2D graphics environment." Computing in science & engineering 9.03 (2007): 90-95.
36. https://github.com/seatgeek/thefuzz
37. Ghosh, Sourish, Anasuya Dasgupta, and Aleena Swetapadma. "A study on support vector machine based linear and non-linear pattern classification." 2019 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2019.
38. Chapelle, Olivier, Patrick Haffner, and Vladimir N. Vapnik. "Support vector machines for histogram-based image classification." IEEE transactions on Neural Networks 10.5 (1999): 1055-1064.
39. Kleinbaum, David G., et al. Logistic regression. New York: Springer-Verlag, 2002.
40. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
41. Linne´, Carl von. Systema naturae. Ed 10 (1758): 551.
42. V. Mondain, et al. “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” European Journal of Clinical Microbiology & Infectious Diseases, 2018, pp. 1-6, https://doi.org/10.1007/s10096-018-3202-1
43. International Organisation for Standardisation. ISO 23903:2021 Interoperability and integration reference architecture – Model and framework. ISO: Geneva; 2021.
44. Chomsky Hierarchy in Theory of Computation. https://www.geeksforgeeks.org/chomsky-hierarchy-in-theory-of-computation/
45. Krogstie J. Business Information Systems Utilizing the Future Internet. LNBIP 2011;90:1-18.
46. Aamodt, Agnar, and Mads Nygård. "Different roles and mutual dependencies of data, information, and knowledge—an AI perspective on their integration." Data & Knowledge Engineering 16.3 (1995): 191-222. DOI: https://doi.org/10.1016/0169-023X(95)00017-M
47. International Organisation for Standardisation. ISO/IEC 21838:2021 Information technology – Top-level ontologies (TLO). ISO: Geneva; 2021.
48. Blobel, Bernd, Pekka Ruotsalainen, and Frank Oemig. "Why interoperability at data level is not sufficient for enabling pHealth?." pHealth. 2020.
49. Blobel, Bernd, et al. "Transformation of Health and Social Care Systems—An Interdisciplinary Approach Toward a Foundational Architecture." Frontiers in Medicine 9 (2022): 802487. doi: 10.3389/fmed.2022.802487
50. Blobel B, et al. Designing and Managing Advanced, Intelligent and Ethical Health Ecosystems. In this volume.