MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190242

Improving the Prescription Process Information Support with Structured Medical Prospectuses Using Neural Networks

Oana Sorina Chirila[a], Ciprian Bogdan Chirila[b], Lăcrămioara Stoicu-Tivadar[a]

[a] *Department of Automation and Applied Informatics, University Politehnica Timişoara, Timişoara, Timiş, Romania,*
[b] *Department of Computers and Information Technology, University Politehnica Timişoara, Timişoara, Timiş, Romania*

Abstract

To provide the best treatment, a physician needs information about both the patient and the medicines matching the patient status and improving it. In this article, we present three methods for structuring the sections of medical prospectuses using neural networks. To structure the information from a medical prospectus we use 3 web sources with structured data from sections (with names sections from prospectuses and with uniformized names of sections) to train as input for neural networks. The tests were conducted on Romanian prospectuses. After running the three algorithms, the prospectuses were compared in terms of accuracy and execution time for each source. It was concluded that the accuracy is higher in convolutional networks and in the case of uniform name sections. The output data is used in applications with decision support for the treatment, matching best treatment with the patient's status.

Keywords:

Prescriptions, neural networks (computer), drug information services

Introduction

Prescribing medicines for certain illnesses in as correct manner as possible is a challenge for all doctors and healthcare providers worldwide. The number of illnesses and medicines is higher and higher, and new treatments for new diseases are rare. The lack of sufficient information on treatments and the lack of uniformity of existing data on medication, as well as the lack of tools to compare and verify interactions between patient medication, creates a vulnerable situation. Prior to providing a specific treatment, a physician needs data such as: medical history, diagnosis, and complete information about the appropriate medication for the diagnosis.

The drug prospectus contain very important information for the doctor. For example, in a Nurofen 200 mg prospect, there are several sections, such as "What is Nurofen 200 mg and What is used for," "Warnings and precautions," "What should you avoid when you are taking this medicine?", "Nurofen 200 mg with other medicines", "Pregnancy, breast-feeding and fertility", "How to take Nurofen 200 mg", "Possible side effects", "Containing 200 mg nurofen". These data are useful for the physician to not prescribe wrong medication that may interact with another drug / disease / allergy / condition of the patient. For example, in the prospectus of Nurofen 200 mg, it is emphasized that a patient with gastro-duodenal ulcer cannot take this medicine. The same data is found in any medical prospect, but each manufacturer names the sections differently or locates them in different order. Any doctor has access to online prospectuses, but they need extended knowledge about the medication and for this knowledge they have to read the full prospectus for the right information each time they consult the information about the drug, especially for new ones. This means a lot of invested time for this process.

Information and medical data are both structured and unstructured. Most structured information is available in English. Various databases with structured information have been created and are easily included in medical applications or used by physicians to provide effective treatment for patients. Many researchers have begun to develop algorithms or to use new technologies to create the largest possible structured information. Structured information is easier to read and requires less time to be found. The researchers also seek to extract important data from various documents to use them in specific areas or to build for physicians structured databases with the extracted information. .

In this age of technology, we are confronted with a great deal of medical information coming from different sources. For physicians to have access to this ocean of information, structuring and compaction of data is needed. Important data can come from our daily activities, from the internet or even from clinical staff. Figure 1 shows the diversity of sources from which important clinical information can be obtained to be used later to treat and improve the medical condition of a patient. The goal is to integrate and structure information from these sources as to make it more easily available to medical units [1].

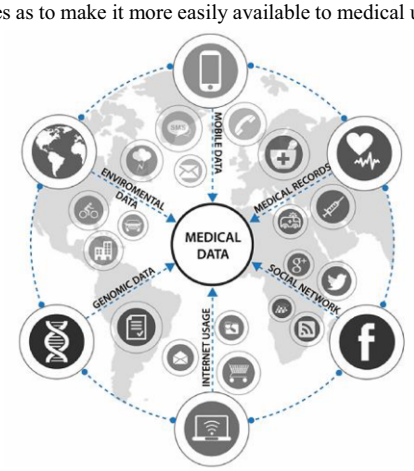

Figure 1– Sources of Medical Data [1]

A lot of work has been done on natural language processing (NLP) in English. This also applies to the processing of medical texts where the most robust data structuring and processing is for English texts. Deléger, Grouin, and Zweigenbaum [2] describe the implementation of a system for extracting

medicines and adjacent information from French texts based on an originally algorithm used for English medical text. The system is based on special dictionaries of medical terms and extraction rules. The texts resulted from 17,412 French electronic health records (EHRs) from the Cardiology Unit of the French University Hospital registered between 2004 and 2006. This dataset was divided into two data bodies: a corpus of development of 17,362 documents used to implement the system and a test body of 50 documents. The test body contains 253 drugs plus associated information items. The extraction rules were designed using English rules and examples of the development corpus. The dosing, form, frequency, duration and reason for which the drug was administered were extracted from the documents. The evaluation of the algorithm on the 50 documents obtained an F-measure score of 86.7%.

In Xu and Wang [3], a simple and accurate learning algorithm was developed to extract drug-disease pairs from 20 million biomedical summaries available on MEDLINE. The authors' process for extracting these data consists of the following steps:

1. Obtaining and analyzing the MEDLINE corpus
2. Creating lexicons for diseases and medicines
3. Correlation of MEDLINE sentences with disease and drug entities
4. Find patterns of treatments
5. Extract extra pairs from MEDLINE using selected patterns
6. Perform a semantic analysis of extracted drug-disease pairs.

The authors have used around 100 million sentences extracted from MEDLINE abstracts published between 1965 and 2010. A total of 34,305 pairs of single-disease treatments were extracted, most of which were not included in the existing structured databases. The algorithm of the authors reached a precision of 0.904 in extracting all pairs.

Extracting information about medicinal products from clinical texts is very important for EHR research. Jiang *et al.* [4] present the implementation of Java MedEX, an existing Unstructured Information Management Architecture (UIMA) data extraction system. In addition to showing earlier developments, the authors included new coding modules in the MedEx-UIMA system that map the data of drugs extracted with the RxNorm concepts. The MedEx-UIMA system consists of two main components: the clinical text extraction module and a standardization module encoding RxNorm (RxNorm concept unique identifiers) and normalizes TIMEX3 [5] frequency information. The information extraction module is a Java implementation of an older MedEx version then implemented in Python. The authors processed 826 documents with both systems (MedEx-UIMA and MedEx-Python) reaching similar results with both systems. Using the 300 annotated drug entries manually, the MedEx-UIMA system obtained the F-measure score for accuracy of 98.5%.

Casillas *et al.* [6] present a model for extracting allergic drug reactions from medical electronic records. The purpose of the paper is to extract this data from medical texts into Spanish. The authors developed two methods: a rule-based and a machine-based method. Both methods include semantic knowledge derived from FreeLing-Med, a software explicitly developed for medical text analysis. The corpus of text used in this research is composed of EHR documents written by doctors from Galdakao-Usansolo Hospital. First, a simple approach has been attempted to co-ordinate the terms of the EHR with SNOMED CT terms for allergy detection. This approach is effective but has not proven appropriate for medical texts written by physicians in which the type of allergy writing is different from the standard. Follow a rule-based algorithm has better results (70% accuracy) than the first simple approach (30% accuracy). The best approach with great results was the machine learning-Inferred Classifier method that gave a precision score of 88%.

Despotou *et al.* [7] uses NLP techniques to interpret UK NICE BNF drug recommendations that are provided as free text. The NLP component, MetaMap, identifies and interprets the semantic meaning of concepts in medical texts. NICE BNF provides structured definitions for drug-related issues but uses raw text for instructions on indications for drugs such as "500 mg three times a day" or "300-900 mg every 4-6 hours; maximum 4g per day." The authors analyze the dosing instructions and using NLP techniques identify types of semantic expressions and investigate how they can be generalized and used in specific rules to be applied in medical IT systems that contribute to e-prescribing.

Over the last few years, as presented in this section, a lot of applications were developed to support structuring and embedding information in medical applications. Databases have been created with a wealth of drug-related information, but most research has been made in and for English-speaking countries.

In this study, we use neural networks and the associated algorithms to structure Romanian medical prospectuses in order to extract the information from each section for further use in medical applications. We use the algorithms to train the neural networks since there are no other alternatives that we can use for Romanian text to extract certain sections from the prospectuses. This kind of processing works in cases where translate coding or databases is missing. In the following we present the methods, the results, and conclusions of the study.

Methods

Deep learning methods or neural networks have recently influenced many areas, including the processing of natural language. These methods are constantly improved with algorithms and increased performance compared to what exists in each field. A number of tools have been developed to enable the implementation of deep learning such as: Caffe, DeepLeaning4J, Eblearn, Keras, Neon, Scikit-learn, TensorFlow, Theano, etc. These tools attempt to optimize different aspects in learning or developing deep learning algorithms. These deep learning software tools have begun to receive a great deal of attention from the research community and are being increasingly developed to allow the formation of deep networks with thousands of parameters. Developers are trying to continually improve these tools to attract as many users as possible and to promote research [8].

In the current study we use medical prospectuses in Romanian language extracted from three online sources and structured on sections by extraction algorithms that we propose for this action [9].

In this study we use a method of training the neuronal networks with the names of selected already structured sections through various neural network algorithms and test them by providing other texts to which the trained network predicts the correct section name. The purpose of categorizing the text is to categorize the texts into categories that are easier to access. Each text can be categorized into one or more categories. We use neural networks to learn the classifiers in the examples and automatically categorize other documents in the same categories. In order to obtain this structure, we use three neural network models: Vector Support Machine Classifier from Scikit-learn library, Naïve Bayes Classifier from Scikit-learn

library, and 1D Sequence-Model Convolution Networks with sequential model from the Keras library. Figure 2 presents the workflow obtaining this structure.

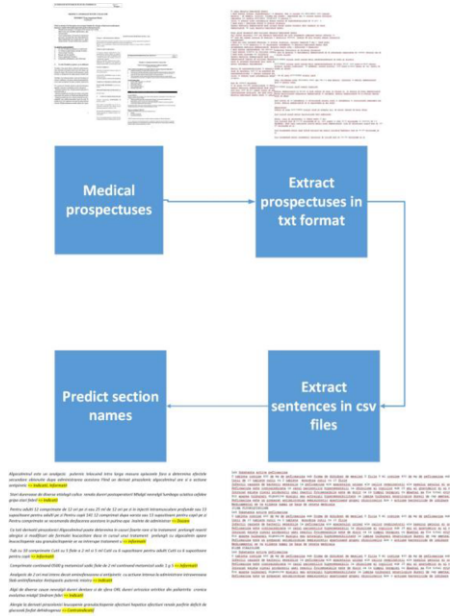

Figure 2 – Workflow for Structuring Prospectuses

Support Vector Machine (SVM) is a supervised classification algorithm that is extensively and successfully used for text classification task. An SVM represents the examples as points mapped in space so that the examples of the separated categories are divided by a clear gap that is as broad as possible [10].

Naive Bayesian is a simple and efficient classifier to implement NLP because it supposes that all the words of the documents are independent one of each other. The Naïve Bayes Classifier is the simplest probabilistic classifier used to classify text documents. The Naïve Bayes method is a module classifier based on probability and the probability of known conditionality. The basic idea is to use common words and category probabilities to estimate the class of a particular document [11].

Convolutional Neural Network (CNN) is a class of artificial neural networks that use convolutive layers to filter inputs for useful information. Convolution involves combining input data (a feature map) with a convolution kernel (filter) to form a map of the transformed features. Convolutional filters are modified based on the learned parameters to extract the most useful information. Text convolution is used for convolutional 1D networks [12].

To learn these three types of neural network models we use the structured data from the three medical prospecting websites: HelpNet, Pharmacists' Page and CSID. Table 1 presents the figures found in each data file.

After creating the files for each source as well as for mixed sources, we ran the three previously specified algorithms one by one to train the network section names for each text and then predict the section names for other text files. For training and testing, we use combinations between the above-mentioned sources. In order to run the algorithms, we use a computer with Intel Core i5-6400 processor of 2.70 GHz and 8 GB RAM and the processing time was quite good. Each run calculates the accuracy of the results and the running time. The outcomes of the process are presented in the Results section.

Table 1 – Number of Data Used in Neural Network Structuring

Name of the source	Number of prospectuses	Number of sections
CSID	3814	44834
Help Net	2820	18336
Pharmacists' Page	1513	22851

For all the algorithms, the first step consists of creating the learning and testing files containing the information of interest, specifically the label and the text associated with that tag. The learning files contain 70% of this information, and the test files 30%. For each solution we use specific items, as follows. The SVM (Support Vector Machine) algorithm uses the LinearSVC (Linear Support Vector Classification) classifier to train the network, and the prediction function is called on the test file. CNN uses the Sequential model in the Keras library to train the network, using the 'Dense layer' and calling the 'relu' activation function. We use 3 iterations (epochs) per learning file. Since we use textual classification, we only need a layer of convolution, so we use 1D Convolution. In the Naive Bayes algorithm, the text is transformed firstly as an array of elements, followed by a normalized expression of frequent terms and calling the NB multinomial classifier for the data previously obtained.

Each classifier is evaluated by calculating the accuracy and F1 score.

One of the main issues we had to deal with was the unevenness of the section names for the same type of section, different names for each drug may appear. For example, in the warning section, the following names may be found: caution, special warnings, warnings, warnings and precautions, precautions for use, warnings and special precautions, precautions, warning. To overcome this problem, we began aligning these sections creating files with similar section names and reference names. In [13] we propose a method of refining prospectuses sections to make them uniform.

To refine section names, we identified the most important section names: 'Indications', 'Contraindications', 'Dosage', 'Pregnancy and Breastfeeding', 'Active Substance', 'Composition', and any other section in the 'Information' section. We restored prospectus sources using just these section names and we ran the three algorithms again for all sources and their combinations. The outcomes of the process will be presented in the Results section.

Following the SVM algorithm running on data from the three sources, accuracy scores range from 1.99% in different training and testing sources and up to 56.78% on the same data source both for training and for testing. Execution time depends on the size of the text and the number of section names in the training source so that in the sources where the data from the CSID site is used, the running time is longer, the section names being varied and the prospectuses being more.

When running the Naive Bayes algorithm on the same data, the running time has dropped a lot, and accuracy is higher for all tests. For the algorithm, when referring to data with many different section names (CSID), we train the network only for fewer sections, choosing the most common ones. In this case, the results were more accurate, between 45.41% and 68.44%.

When running the convolutional network, the accuracy increased even more, but running time increased significantly as compared with the previous methods. The best accuracy was

recorded for training and test data for the CSID source (77.47%). With this method we had to lower more the number of training sections for an optimal algorithm operation. When we use CSID source to train the neural network we select only the most common 350 names from sections since the array that is created in the algorithm is very large and the algorithm cannot be run on a usual computer.

The percentage of accuracy is lower when providing new sources of information different from those with which the network was trained. Section names are very different from one source to another. Thus, in the CSIDs there are approximately 11,300 unique names for different sections, in the Pharmacists page there are around 450 unique names of sections, and in Help Net there are around 350 unique names of sections.

After the sections were reduced and uniformized, the accuracy increased greatly after running the SVM algorithm. The algorithm run times have also decreased, obtaining an accuracy of 38.26% for two different sources and an accuracy of up to 93.08% for training and testing data from the same source.

After running the Naive Bayes algorithm for data sources with uniform sections the accuracy increased significantly, and the running time of the algorithm decreased. The accuracy in this case ranges between 71.86% and 85.05%, both of which are this time for sections of the same data source.

Following the run of the convolution algorithm, the accuracy is highest, but the times stay high. Accuracy ranges between 77.62% and 91.31%. As can be seen, the accuracy of all source combinations is over 77%, which means that the correct prediction rate of the sections is the best.

As can be seen in the results presented with data uniformity, the accuracy result increased. Training the algorithms with the most complex and complete databases, the accuracy will increase. For what we need in medical applications, namely as accurate as possible, the best algorithm tested in this paper for accuracy in both non-uniform and uniform sections is the convoluted neural network. Due to the fact that it is working with multiple layers to drive information, the accuracy is the highest.

Discussion

The first algorithm that was run was the Support Vector Machine on the three sources and their combinations. The first run was on data with non-homogeneous sections, and the second run was on data with homogeneous sections. Figure 3 shows a comparative diagram of the two results in terms of accuracy.

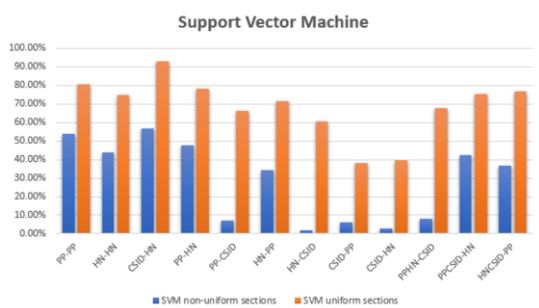

Figure 3 – SVM Accuracy

The first comparison is between the type of source used for the SVM algorithm, non-uniform sources in the blue version on the graphs, and sources with uniformized sections in the orange version on the graphs. As can be seen, the accuracy in the second case has increased significantly, especially if different data sources were used for training and testing and the running time decreased due to the number of low sections. For this algorithm, a relatively higher accuracy was obtained only for uniform sections.

The average accuracy for SVM on non-uniform sections is 28.52%, and for SVM on uniform sections the average increased to 68.58%, with an increase of 40.06%.

The median of accuracy for SVM on non-uniform sections is 35.65%, and for SVM on median uniform sections is 73.20%.

The second comparison was made between data used, with a different number of sections for the Naive Bayes algorithm. Figure 4 shows a comparative diagram of the two results in terms of accuracy.

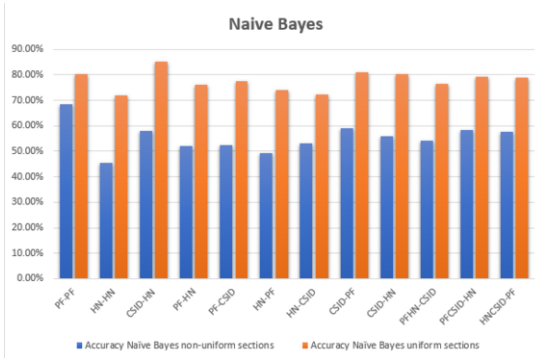

Figure 4 – Naïve Bayes Accuracy

In this case, the accuracy between the two types of sources used differed less. And in this case, the accuracy was bigger and more constant, as can be seen in the graph for the uniform sections. The average accuracy for Naive Bayes on non-uniform sections is 55.28%, and for Naive Bayes average uniform sections increased to 77.76%, with an increase of 22.48%.

The median of accuracy for Naive Bayes on non-uniform sections is 54.93%, and for Naive Bayes on uniform sections is 78.17%.

The last comparison between the accuracy of the different three algorithms on data sources was done for the convolutional network algorithm. Figure 5 presents a comparative diagram of the two results in terms of accuracy.

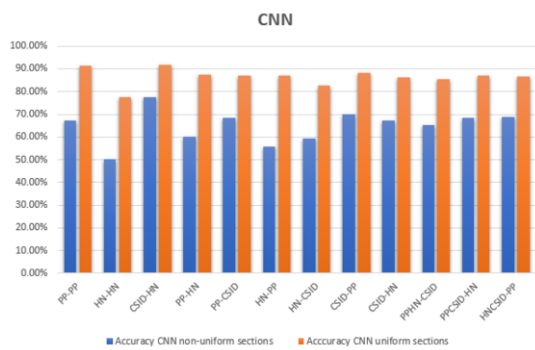

Figure 5 – CNN Accuracy

In this case, the accuracy was also good for non-uniform sections, but increased over 85% in most cases to uniform sections. The average of CNN accuracy for non-uniform

sections is 64.87%, and for CNN on uniform sections the average grew to 86.55%, with an increase of 21.68%.

The median of accuracy for CNN on non-uniform sections is 67.10%, and for CNN on uniform sections is 86.97%.

As can be seen, the best accuracy was in the case of convolutional neural networks and on uniformized sections. In Figure 6, we can see comparatively the accuracy of the algorithms on the non-uniform databases, and in Figure 7 we can see comparatively the accuracy of the algorithms on the uniform databases.

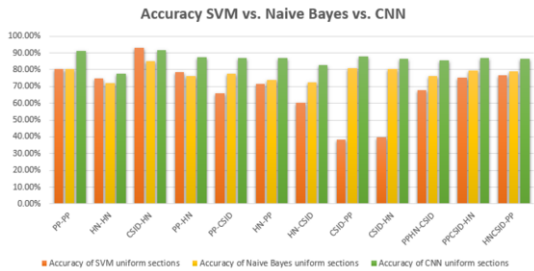

Figure 6 – Comparative Accuracy on Non-Uniform Sections

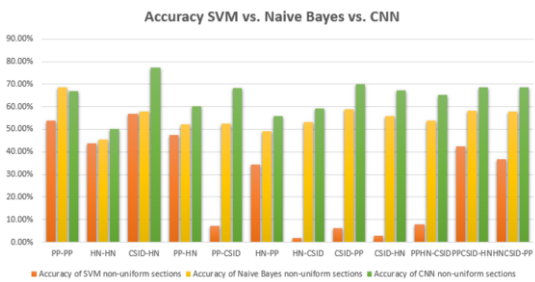

Figure 7 – Comparative Accuracy on Uniform Sections

Conclusions

Structuring medical information is very important to help doctors find the best treatments for patients and to create systems with alerts when drugs can interact with certain diseases / allergies or other medicines that the patient already has. In this study, we have chosen to structure the medical prospectuses in Romanian because in this language there are no databases with full structured information about the drugs. The extracted information can then be used to create databases that are added to the prescription modules. In the future, we will test algorithms on medical prospectuses from other languages in order to create a general model for structuring medical prospectuses. Future work also focuses on more solutions to be multilingual and adapt easy to situations in which there are no structured databases or translated codes with drug information.

The results are encouraging and the solution is easy to implement as a module in healthcare information systems.

References

[1] M. Ghassemi, L. A. Celi and D. J. Stone, State of the art review: the data revolution in critical care, *Critical Care (2015) 19:118*

[2] L. Deléger, C. Grouin, P. Zweigenbaum, Extracting Medication Information from French Clinical Texts, *Studies in Health Technology and Informatic*s, MEDINFO 2010, 160:949-953

[3] R. Xu and Q. Wang, Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing, *BMC Bioinformatics*, 2013:14:181

[4] M. Jiang, Y. Wu, A. Shah, P. Priyanka, J. C. Denny, H. Xu, Extracting and standardizing medication information in clinical text – the MedEx-UIMA system, *AMIA Jt Summits Transl Sci Proc. 2014* :37-42

[5] J. Pustejovsky, J. Castaño, R. Ingria, et al. TimeML: Robust Specification of Event and Temporal Expressions in Text, *Fifth International Workshop on Computational Semantics*. 2003

[6] A. Casillas, K. Gojenola, A. Perez and M. Oronoz, Clinical text mining for efficient extraction of drug-allergy reactions, *IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016*

[7] G. Despotou, I. Korkontzelos, N. Matragkas, E. Bilicia and T. N. Arvanitis, Structuring Clinical Decision Support Rules for Drug Safety Using Natural Language Processing, *Data, Informatics and Technology: An Inspiration for Improved Healthcare, IOS Press*, 2018

[8] S. Bahrampour, N. Ramakrishnan, L. Schott, M. Shah, Comparative Study of Deep Learning Software Frameworks, *Cornell University Library*, Computer Science, 2016

[9] O. S. Lupşe and L. Stoicu-Tivadar, Extracting and Structuring Drug Information to Improve e-Prescription and Streamline Medical Treatment, *Applied Medical Informatics, Vol. 40, No. 1-2* /2018: 7-14

[10] Scikit-learn, http://scikit-learn.org, accessed in July 2018

[11] R. M. Chezian, C. Kanakalakshmi, Performance Evaluation of Machine Learning Techniques for Text Classification, *Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications*, 27th March 2015

[12] N. Ackermann, Introduction to 1D Convolutional Neural Networks in Keras for Time Sequences, Good Audience, https://blog.goodaudience.com, accessed in August 2018

[13] O. S. Lupşe and L. Stoicu-Tivadar, Supporting Prescriptions with Synonym Matching of Section Names in Prospectuses, *Stud Health Technol Inform*. 2018;251:153-156

Address for correspondence

Chirila Oana Sorina, email: oana.chirila@aut.upt.ro