Linköping University | Department of Computer and Information Science
Master thesis, 30 ECTS | Datavetenskap
202018 | LIU-IDA/LITH-EX-A--2018/001--SE

Labelling Clinical Reports by Exploiting the Data Structure Through Topic Modelling and Active Learning

Uppmärkning av kliniska rapporter genom att utnyttja strukturen hos datan med topic modeller och active learning

Simon Lindblad

Supervisor : Marco Kuhlmann
Examiner : Arne Jönsson

External supervisor : Mikael Nilsson

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Abstract

Abstract.tex

Acknowledgments

Acknowledgments.tex

Contents

1 Introduction

The world's population is growing each year. Making healthcare more efficient and robust is of great importance in order handle the challenges that arises with a growing population. One way of increasing the efficiency as well as the quality of healthcare is to create automated systems that can aid doctors in their process. As the population is growing it's of utmost importance to ensure that the quality of diagnosis remains high, and to minimize the risk of missing some critical piece of information. Taking advantage of the available medical information is key to creating aforementioned systems.

Information pertaining to a patient's diagnosis is often in the form of written clinical reports. One example where this information could be utilized is when a doctor is writing such reports. If a system could show cases with similar features as the current one, the doctor could compare the findings and check if they have obtained an abnormal result. Being able to perform such a comparison will result in extra quality assurance in the diagnostic flow. It could also provide doctors with extra confidence in that their diagnosis is correct.

The problem systems like this would face is to identify the type of a medical report in order to make further suggestions. One approach that is commonly used for such problems is machine learning. In machine learning, you use a set of inputs and map it to some output values [2]. This is done by using data to build a, usually statistical, model.

The task of predicting a type, or class, for a given text document is called text classification. Text classification is usually solved using supervised learning [1]. In supervised text classification, you have a set of inputs, in this case text data, that already has a category assigned to it. This data is then used to fit the model so that it later make predictions for inputs that it has not yet been exposed to. A model that have been shown to be successful in text classification is Support Vector Machines (SVM) [3, 1, 4].

In order to assign fit a machine learning model to predict categories for clinical reports, we need a set of already labeled data. That is, we need to assign categories to the existing set of clinical reports. It is often the case that text data is widely available, but it is harder to come by data that is already labeled. Obtaining high quality data is important to use in machine learning systems, both in healthcare and other areas. Since the models require a sufficient amount of reports to be labeled, the task of labeling them can be cumbersome. Especially in the case of clinical data, since doctors and other clinicians time is valuable and expensive. By improving the process and the quality of data to be labeled, they can spend more time doing their job.

The field within machine learning that is focused on the task of labeling data is called active learning. It is a form of semi-supervised learning. The algorithm queries an oracle (in this case a doctor) for labels for the data points that it think will help the model improve the best. This is used when there is plenty of readily available data, but assigning labels is expensive. Since the data points to be labeled are actively selected, the models can require fewer examples than if they were selected at random. The points can be selected by considering the certainty of the models, and request to label the documents that the model is less certain about. Another approach, which has not been given as much attention, is using the underlying structure of the data to select points. The goal with this approach is that you can capture the distribution of the categories.

If you assign one of two classes to each document, you have a binary classification problem [2]. Problems where you assign one of several classes is called a multi-class classification problem. Multi-labeled classification is when you assign one or more label to each document. It type of classification that will be treated in this report is multi-labeled. Assigning several classes to a document is more time consuming than in the cases where you only need to find one option. In those cases you can stop when you have found the appropriate label. However, when a document can be assigned several classes you need to consider the entire report. This makes the use of active learning methods to enhance the labeling of documents even more useful in the multi-labeled case.

1.1 Motivation

This thesis is carried out at Sectra Medical Imaging IT Solutions AB, as a part of their research group. They are currently pursuing a research project with Region Skåne in southern Sweden. The intention behind the project is to use machine learning techniques to, among other things, be able to suggest categories to doctors while they are writing medical reports. Another case is to use the categories of documents to present doctors with medical reports that are handling similar cases from the past. With this information the doctors could get an extra quality assurance check in their diagnostic flow.

In order to build these system, you need a substantial amount of labeled clinical reports. Therefore, the purpose of this thesis is to increase the quality and efficiency of labeling these reports. This will be done by using unsupervised learning techniques such as topic models and clustering to first remove documents that aren't supposed to be labeled. That includes documents that describe patients never showing up for a scan, deceased patients and patients being moved to a different hospital, among others. For the labeling, a system is be built to use active learning in conjunction with the aforementioned unsupervised techniques to increase the quality of the labeled documents. In the work that they have done so far, the doctor that primarily worked with the labeling of reports stated that the distribution over the labeled categories are very skewed. The vast majority of labeled documents were assigned the same few categories. This in turn leads the models to require a lot of labeled samples to work well. By using active learning techniques we can reduce the number of labeled samples needed to obtain an accurate model.

1.2 Aim

The purpose with this thesis project is to evaluate different solutions to increate the quality of labeled reports, and thereby reducing the amount of them needed for a system. Resulting from this will be a complete, standalone system, for labeling reports. The reports are interactively queried so a user can label the reports that are deemed most useful by the system.

1.3 Research questions

The specific research questions that this thesis will treat is presented here. They will be the main focus of study.

1. *Is it possible filter out invalid clinical reports by using unsupervised techniques such as topic models and clustering?*
   In the dataset from Sectra, there are reports describing patients not showing up for or changing the time of their appointments, deceased patients or patients that have been ordered to another hospital. These reports does not contain any information of value from a medical point of view and should not be considered in the report labeling process.

   Unsupervised machine learning models such as topic modeling and clustering does by definition not require any labeled documents to train on. If it is possible to, without any such data, group these invalid reports together and remove them from the process before a doctor is presented with them that would be an additional hurdle removed from the process.

2. *What are the alternatives to sample documents at random in a document labeling system?*
   How we are choosing the documents to be sampled is important. If the dataset that is being sampled is skewed, i.e. some categories are a more frequent than others, our labeled set will likely follow that distribution. This will result in the system requiring a lot of labeled documents to gain a high accuracy with reports of less common categories.

   If the decision boundaries of our models can be used to pick documents that would be more informative, the number of labeled documents could be reduced and still gain the same accuracy. Another approach to selecting the documents to sample is to take advantage of the underlying structure of the data through clustering.

3. *Which of the alternatives in question 2 gives us the highest quality set of labeled reports?*
   The algorithms to be evaluated will be from the two approaches: based on the model certainty or the underlying structure of the data (or both). When choosing the algorithm to use, there are several different factors that will affect the final results and therefore needs to be taken into consideration. How well the models perform on the data is a rather obvious one – evaluating the models based on accuracy, precision, recall and f1-score. But they also need to be able to query documents in a reasonable time, if it is expensive to label reports it is likely to be expensive to wait for the reports to be queried. Choosing reports in batches and if the algorithm needs a big initial set of labeled reports are other factors that will be evaluated.

   Another indication on the quality of the labeled reports is the balance between the classes. Even though the original data may be very imbalanced, selecting samples that contains a better balance between the different samples could improve the models performance.

1.4 Delimitations

Even though the sampling strategies are evaluated objectively on the Reuters dataset, the applicability of the techniques on clinical data is only evaluated by one physician, on the one dataset provided by Sectra.

1.5 Structure of the Report

The next chapter covers the background theory that is relevant for the thesis. After the theory, the methodology used is described, which is followed by a chapter covering the results. The

method and results are then discussed in Chapter 4. Finally, Chapter 5 presents the conclusions.

2 Method

In this chapter, the method is described in a way which shows how the work was actually carried out. The description must be precise and well thought through. Consider the scientific term replicability. Replicability means that someone reading a scientific report should be able to follow the method description and then carry out the same study and check whether the results obtained are similar. Achieving replicability is not always relevant, but precision and clarity is.

Sometimes the work is separated into different parts, e.g. pre-study, implementation and evaluation. In such cases it is recommended that the method chapter is structured accordingly with suitable named sub-headings.

3 Results

This chapter presents the results. Note that the results are presented factually, striving for objectivity as far as possible. The results shall not be analyzed, discussed or evaluated. This is left for the discussion chapter.

In case the method chapter has been divided into subheadings such as pre-study, implementation and evaluation, the result chapter should have the same sub-headings. This gives a clear structure and makes the chapter easier to write.

In case results are presented from a process (e.g. an implementation process), the main decisions made during the process must be clearly presented and justified. Normally, alternative attempts, etc, have already been described in the theory chapter, making it possible to refer to it as part of the justification.

4 Discussion

This chapter contains the following sub-headings.

4.1 Results

Are there anything in the results that stand out and need be analyzed and commented on? How do the results relate to the material covered in the theory chapter? What does the theory imply about the meaning of the results? For example, what does it mean that a certain system got a certain numeric value in a usability evaluation; how good or bad is it? Is there something in the results that is unexpected based on the literature review, or is everything as one would theoretically expect?

4.2 Method

This is where the applied method is discussed and criticized. Taking a self-critical stance to the method used is an important part of the scientific approach.

A study is rarely perfect. There are almost always things one could have done differently if the study could be repeated or with extra resources. Go through the most important limitations with your method and discuss potential consequences for the results. Connect back to the method theory presented in the theory chapter. Refer explicitly to relevant sources.

The discussion shall also demonstrate an awareness of methodological concepts such as replicability, reliability, and validity. The concept of replicability has already been discussed in the Method chapter (2). Reliability is a term for whether one can expect to get the same results if a study is repeated with the same method. A study with a high degree of reliability has a large probability of leading to similar results if repeated. The concept of validity is, somewhat simplified, concerned with whether a performed measurement actually measures what one thinks is being measured. A study with a high degree of validity thus has a high level of credibility. A discussion of these concepts must be transferred to the actual context of the study.

The method discussion shall also contain a paragraph of source criticism. This is where the authors' point of view on the use and selection of sources is described.

In certain contexts it may be the case that the most relevant information for the study is not to be found in scientific literature but rather with individual software developers and open

source projects. It must then be clearly stated that efforts have been made to gain access to this information, e.g. by direct communication with developers and/or through discussion forums, etc. Efforts must also be made to indicate the lack of relevant research literature. The precise manner of such investigations must be clearly specified in a method section. The paragraph on source criticism must critically discuss these approaches.

Usually however, there are always relevant related research. If not about the actual research questions, there is certainly important information about the domain under study.

4.3 The work in a wider context

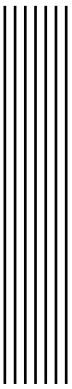
There must be a section discussing ethical and societal aspects related to the work. This is important for the authors to demonstrate a professional maturity and also for achieving the education goals. If the work, for some reason, completely lacks a connection to ethical or societal aspects this must be explicitly stated and justified in the section Delimitations in the introduction chapter.

In the discussion chapter, one must explicitly refer to sources relevant to the discussion.

5 Conclusion

This chapter contains a summarization of the purpose and the research questions. To what extent has the aim been achieved, and what are the answers to the research questions?

The consequences for the target audience (and possibly for researchers and practitioners) must also be described. There should be a section on future work where ideas for continued work are described. If the conclusion chapter contains such a section, the ideas described therein must be concrete and well thought through.



Bibliography

[1] Charu C Aggarwal and ChengXiang Zhai. "A survey of text classification algorithms". In: *Mining text data*. Springer, 2012, pp. 163–222.

[2] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Thorsten Joachims. "Text categorization with support vector machines: Learning with many relevant features". In: *European conference on machine learning*. Springer. 1998, pp. 137–142.

[4] Simon Tong and Daphne Koller. "Support vector machine active learning with applications to text classification". In: *Journal of machine learning research* 2.Nov (2001), pp. 45–66.