

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

Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår. Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art. Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart. För ytterligare information om Linköping University Electronic Press se förlagets hemsida <http://www.ep.liu.se/>.

Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances. The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility. According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement. For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: <http://www.ep.liu.se/>.

Abstract

Abstract.tex

Acknowledgments

Acknowledgments.tex

Contents

Abstract	iii
Acknowledgments	iv
Contents	v
List of Figures	vi
List of Tables	vii
1 Introduction	1
1.1 Motivation	2
1.2 Aim	2
1.3 Research questions	3
1.4 Delimitations	3
1.5 Structure of the Report	3
2 Theory	4
2.1 Text Processing using Unsupervised Techniques	4
2.2 Text Classification	7
2.3 Active Learning	9
2.4 Evaluation Metrics	14
2.5 Related Work	14
3 Method	16
3.1 Frameworks and Tools	16
3.2 Datasets	16
4 Results	18
5 Discussion	19
5.1 Results	19
5.2 Method	19
5.3 The work in a wider context	20
6 Conclusion	21
Bibliography	22

List of Figures

2.1	Diagram of the LDA model.	5
2.2	(a) to (e) shows iterations of K-means until convergence. In (e) it can be seen that the new centroids capture the same documents as the previous iteration, and we have converged.	8
2.3	An overview of an active learning system	9

List of Tables

2.1	Confusion matrix for explaining true positives, false positives, true negatives and false negatives	14
-----	---	----



1 Introduction

The world's population is growing each year. Making healthcare more efficient and robust is of great importance in order to handle the challenges that arise 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 [6]. 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) [17, 1, 32].

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 [6]. 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 5. Finally, Chapter 6 presents the conclusions.



2 Theory

In this section the theory behind the techniques used during the thesis work will be presented. The first part will go through techniques used to process the data and perform an exploratory analysis. After that text classifications, primarily with SVM, will be covered. The last section contains an overview of the field of active learning, as well as a comparison of some different active learning techniques for multi-labeled data.

2.1 Text Processing using Unsupervised Techniques

Techniques in machine learning that does not require you have a categorized or labeled set of data is called unsupervised. They use the structure of the data to obtain the information to use when processing it. When it comes to text data, there are a few common methods and techniques that are unsupervised, and can be used for different purposes. Examples of such techniques are *topic modeling* and *clustering*. Another interesting technique is word2vec, that is used to produce word embeddings.

When working with text, it needs to be represented in a way that allows the models to work with it effectively. *Bag-of-words* (BoW) is one of the more common representations when performing text analysis. Using BoW the text is represented as a multi-set. That is, a document is represented by the number of occurrences of the different words. The representation of a document therefore becomes very high-dimensional, there is one dimension for each word in the vocabulary. Like the name implies, the positions of the words are not taken into account, they are viewed as if they were taken from a bag. Another drawback from this approach is that a word in a written language can be used to express several different thoughts, and one thought can be expressed using several different words. However, it is easy to work with, and is used when performing topic modeling among other things.

One way to incorporate positional information into the representation is the use of n-grams. Instead of storing information pertaining to one term, information is stored with regards to n consecutive terms. Considering the text “Pattern Recognition and Machine Learning” using a bigram (n-gram with $n=2$), would result in the tokens: “Pattern Recognition”, “Recognition and”, “and Machine”, and “Machine Learning”.

Topic Modeling

A topic model is a statistical model for finding topics within text [11]. The topics build upon the probability that a certain word would occur in a text about a given topic, on the basis of terms occurring together. For example, if the topic represents United States politics, words such as “government”, “Trump”, “Reagan”, “Senate”, or “Medicaid” are more likely to appear than “sailboat” or “sweater”. Any given document can then contain a topic with some probability. This can be viewed as fuzzy clustering, and that the document has a degree of membership in a topic or cluster [11]. The most common topic model in use is Latent Dirichlet Allocation (LDA) [11]. Another topic model that preceded LDA is Probabilistic latent semantic analysis (PLSA) [15]. However, PLSA has been shown to be more prone to overfitting than LDA [11].

In the rest of the report, the following notation will be used:

- D denotes a corpus of M documents: $D = \{w_1, w_2, \dots, w_M\}$.
- The number of topics is K . Each topic is indexed by i .
- N_d is the number of terms in document d .
- N_i is the number of terms in topic i .
- V denotes the number of words in the vocabulary.

Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a statistical model, where abstract topics in the model are defined as distributions over words [7]. LDA is based on a generative process, a model of which can be seen in Figure 2.1. The circles in this figure represent random variables. Dependencies between these random variables are shown with arrows, and if a variable is observed it is shaded in the figure. In this model, the only observed variable is the words in the document. Parts of the model are surrounded by a rectangle to show that the part is repeated several times.

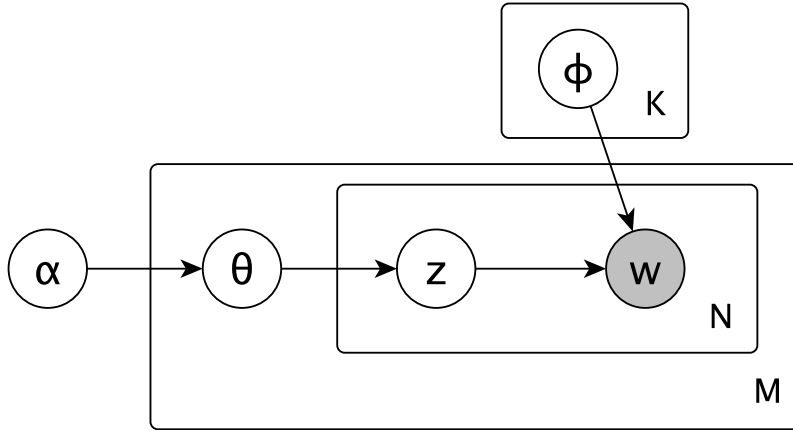


Figure 2.1: Diagram of the LDA model.

The generation of a corpus is done with the following steps [11, 7]:

- **Draw a distribution over the words for each topic.** A sample ϕ_i is drawn from a symmetric Dirichlet distribution with parameter β . This sample represents the distribution of terms for the topic i .

$$\Phi_i \sim Dir(\beta) \quad (2.1)$$

$$p(\Phi_i|\beta) = \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \prod_{v=1}^V \phi_{iv}^{\beta-1} \quad (2.2)$$

Here, Γ is the gamma function.

- **Draw a distribution over the topics for each document.** A sample θ_d is drawn from a Dirichlet distribution with parameters α . This sample represents the distribution of topics for document d .

$$\Theta_d \sim Dir(\alpha) \quad (2.3)$$

$$p(\Theta_d|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{di}^{\alpha_i-1} \quad (2.4)$$

- For each token with index n :
 - **Draw a topic assignment z_{dn} for the token index n .** z_{dn} is drawn from the distribution over topics for each document. That is, z_{dn} is drawn from a multinomial distribution using θ_d as a parameter.

$$z_{dn} \sim Multinomial(\Theta_d) \quad (2.5)$$

$$p(z_{dn} = i|\Theta_d) = \theta_{di} \quad (2.6)$$

- **Draw a token w_{dn} .** The token w_{dn} is drawn from the topic distribution assigned to the index n . That is, w_{dn} is drawn from a multinomial with parameter $\phi_{z_{dn}}$.

$$w_{dn} \sim Multinomial(\Phi_{z_{dn}}) \quad (2.7)$$

$$p(w_{dn} = v|z_{dn} = i, \Phi_i) = \phi_{iv} \quad (2.8)$$

The LDA model identifies topics from different terms that occur together. Consider the case where an LDA model has been used to learn a number of topics. Two terms that frequently occur together are then likely to be in the same topic. So, if the same word has been used to express different thoughts, and the word has the same probability in two topics, the words that it co-occurs with can be used to differentiate between the different thoughts.

The task of learning the LDA model is a Bayesian Inference problem. We have several variables that we cannot observe: the word distribution for the topics (ϕ_i), the topic assignments for the tokens (z), and the topic distribution for the documents (θ_d). The only observed variables are the words in the document. We have to approximate the posterior distribution using some sampling method, since it cannot be inferred automatically [7].

There exist a few algorithms that can be used to learn topics for the LDA model. Two of these that has shown to be able to extract useful topics from text are *collapsed Gibbs sampling* [14] and *variational Bayes* [7]. In collapsed Gibbs sampling, θ and ϕ are marginalized out. It works by repeatedly sampling the topic assignment z_{dn} for each token, conditioned on the assignments for the other tokens. Variational Bayes works by using simpler single-variable models to approximate the LDA. As a consequence, it disregards any dependencies between the variables. This is the approach used in the original LDA paper [7].

Collapsed Gibbs

To be written

Text Clustering

Cluster analysis is commonly defined as finding groups in a given dataset. The members of these groups are determined to be similar by a similarity measure [18, 2]. Since text data is sparse, but yet have a very high dimensionality. With one dimension per term in the dictionary, it is not uncommon with dimensions in the order of 10^5 . For this reason, some of the more naive clustering algorithms does not work as well for text data [2].

In distance-based clustering, a similarity function is used to measure the closeness between two text documents. For the purpose of measuring the similarity between text objects, the cosine similarity function is most commonly used [2]. Two different approaches to distance-based clustering are distance-based partitioning, and agglomerative hierarchical clustering.

K-means Clustering

When using the k-means clustering algorithm, the clusters are based upon an initial set of k representatives. A simple approach to k-means clustering can be seen as:

1. Select K seeds from the original dataset
2. Assign the rest of the documents to one of these seeds, based how how similar they are by the similarity function
3. Before each new iteration, select a new centroid for each cluster. This should be the point that is the best central point for the cluster.
4. Repeat step 2 and 3 until convergence.

A visualization of this can be seen in Figure 2.2. One advantage that K-means has over K-medoid is that it requires a small number of iterations, especially compared to K-medoid [2, 29]. However, K-means is rather sensitive to the selection of initial seeds. One approach is to just select them randomly, or selecting them based on the result of another lightweight clustering method. A frequently used method is k-means++, that has been shown to improve both the speed and accuracy of k-means clustering [4].

Word Embeddings with Word2Vec

To be written

2.2 Text Classification

Text classification is a widely studied field within Computer Science. It is an important problem in supervised machine learning, and it is the task of assigning one or more classes to a given text document [1]. The problem is mainly approached with supervised machine learning. That is, with a dataset that consists of a collection of text documents, where each document has one or more classes assigned to it. With the help of these labels, a classification model is fitted to the data. The goal of this is for the model to be able to correctly assign a class to a a previously unseen text document. Some of these classification models can also produce a probability of a document being of a certain class. Other models are based on the concept of a margin that separates the classes, and the distance between a data point and a margin can be used to indicate how certain the model is of the assigned label [32]. Example of use cases for text classification is categorization of news articles, document retrieval and

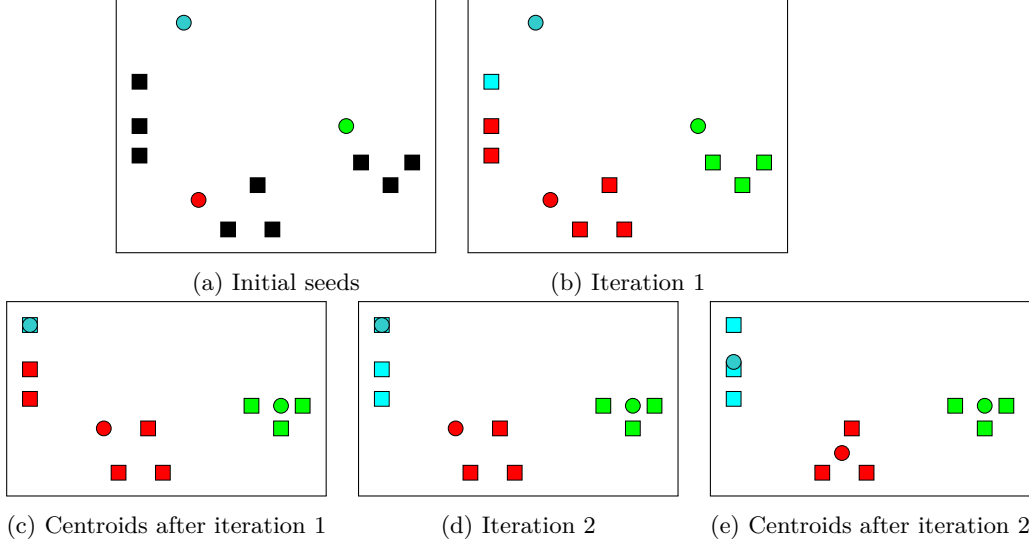


Figure 2.2: (a) to (e) shows iterations of K-means until convergence. In (e) it can be seen that the new centroids capture the same documents as the previous iteration, and we have converged.

email filtering. There exists several different models for classifying text. Decision trees, neural networks and Support Vector Machines (SVM) are some have been previously applied to the text domain with successful results [2]. In this thesis, SVM are the main focus, since they have been studied extensively in the context of active learning.

Support Vector Machines

SVMs work by implicitly map the training data to a feature space [6]. The goal is that the data should be linearly separable in the feature space, even if it is not in the input space. In the case of binary classification, a point is classified by the linear model:

$$y(x) = w^T \phi(x) + b \quad (2.9)$$

The sign of $y(x)$ determines the label assigned to x .

SVMs work by trying to find the hyperplane that maximizes the margin. That is, the distance between any point and the decision boundary should be as large as possible. The hyperplane that gives us the maximum margin can be found by:

$$\arg \min_{w,b} \frac{1}{2} \|w\|^2 \quad (2.10)$$

In order to allow for better generalization, and for data that isn't completely linearly separable, SVMs make use of slack variables ξ_n to penalize points that are close to the decision boundary [6]. A parameter $C > 0$ controls how much effect the slack variables will have. The equation with the slack variables becomes:

$$\arg \min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_n \xi_n. \quad (2.11)$$

A smaller C -value allows more points to be misclassified, in order to achieve better generalization.

Multi-Label Classification

Multi-label classification is the type of text classification where one instance can be associated with multiple labels. It is a generalized version of the multi-class classification problem, where you have more than 2 labels, but each document is only assigned one [33].

A common way of solving multi-label classification problems is the Binary Relevance method [25, 8, 22]. It is a way of transforming the multi-label classification problem into several different binary ones. With Binary Relevance you fit one classification model per label in your data. Each of these classifiers are then predicting whether or not the document is associated with the corresponding label or not.

2.3 Active Learning

Conventional machine learning systems use a set of available data to find a hypothesis that can explain the patterns. The purpose of active learning is to allow a system to *select* the data that it wants labeled, and therefore the data it wants to be trained on [30]. An active learning system samples a document to be labeled from a pool of unlabeled data, and then queries an oracle (often a human annotator) to get the label for that document. By being able to decide what data to label and use, the goal is that the system can achieve better results, and that the data will be of higher quality. A model of the active learning system can be seen in Figure 2.3.

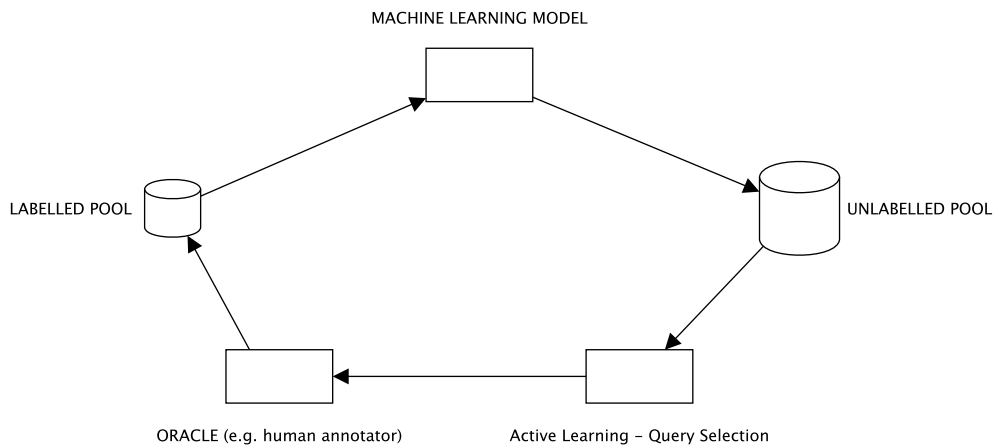


Figure 2.3: An overview of an active learning system

In several different domains, data is readily available and easy to come by. But even if the data is abundant, labels for the data is often harder and more expensive to come by [30], especially when it comes to multi-label problems.

The next section will describe different ways to access the documents in active learning systems, followed by some theory of how the samples relate to the hypothesis space. After that some concrete methods for selecting the samples to queried are described and compared.

Pool-Based Sampling

The main focus is how to select the samples to be labeled. There are different sampling methods in use, and which one is more appropriate depends on how the data can be accessed. Pool-based sampling is motivated by the assumption that there exists a large ready pool of data, where only a small portion is labeled [19, 30]. The samples to be labeled is then selecting by evaluating the entire pool of unlabeled data, and select the most appropriate one based

on a defined utility measure. If the entire pool is large, a subset could be used instead. For applied active learning, pool-based sampling seems to be the most popular choice, but there are some alternatives that have been used in theoretical settings such as stream-based selective sampling. The difference between stream-based selective sampling and pool-based sampling is the individuality of the decisions in stream-based selective sampling, where you draw one sample at a time from an input source and make the decision whether or not to query a label for it. For text applications, where a set of data is often readily available, pool-based sampling is often the more appropriate option since you can consider the entire dataset. Pool-based is therefor the sampling technique that will be considered in this paper.

Searching the Hypothesis Space

In machine learning hypothesis is a specific configuration of a model, the purpose of which is to predict outputs on new instances of data by generalizing the training data. One hypothesis can, for example, be an SVM model with specific values for the parameters. The set of all possible hypothesis that we are working with the *hypothesis space*. Following the SVM example, the hypothesis space would be the set of SVM's with the different values that are possible for the parameters. The hypothesis space is defined as:

$$\mathcal{H} = \left\{ f \mid f(x) = \frac{w * \phi(x)}{|w|} \right\} \quad (2.12)$$

where $w \in \mathcal{W}$ and \mathcal{W} is our parameter space

The version space the subset of the hypothesis space The subset of the hypothesis space that in the feature space separates the data is called the version space, which is defined as:

$$\mathcal{V} = \left\{ f \in \mathcal{H} \mid y_i f(x_i) > 0 \forall i \in \{1 \dots n\} \right\} \quad (2.13)$$

So the version space therefore represents the different hypothesis that make correct predictions on the training data. Under the assumption that one of the hypothesis can fully separate the data, the version space shrinks when more labeled data is acquired. So for new labeled instances the hypotheses in the version space will give better predictions for the training data. Based on this, an active learning algorithm should aim to reduce the size of the version space with each new sample, optimally make it half the size in each iteration.

There exists a useful relationship between the feature space \mathcal{F} and the hypothesis space \mathcal{H} called the *version space duality* [32, 34]. It states that hyperplanes in the hypothesis corresponds to points in the feature space, and the other way around. So by selecting points to be labeled, constraints can be enforced on the hypothesis that form the version space.

One approach to this is called that has shown to be successful is *Uncertainty Sampling* [30]. The idea behind SVM is to find a hyperplane that separates two classes in a binary classification with the maximum margin. Out of the different hyperplanes in the hypothesis space, the version space contains those that can successfully separate the data. We want to select the points in the feature space that will reduce the amount of the valid hypotheses the most. Since SVMs tries to find the support vectors that maximizes the decision boundary in the feature space, separating the two classes. Considering this in \mathcal{H} , it will be analogous to the hypothesis in the center of the hypothesis space encompassed constraints set by the labeled points. What Uncertainty Sampling is predicting the values for the unlabeled points, and then choose the one that it is most uncertain about, the one closest to the decision boundary, to be labeled. Based on the version space duality, it is a good approximation for dividing the version space in two.

Binary Version Space Minimization

Binary Minimization is a generalization of uncertainty sampling, to make it work with multi-label data. The approach taken is to decompose the multi-label problem to several binary one-vs-rest tasks, like discussed in 2.2. The unlabeled point that is chosen for labeling is then the one with the smallest SVM margin across all the binary classification tasks. By doing this, it does not incorporate the multiple labels into the decision process, but treats all classes individually and equally.

Maximum Loss Reduction with Maximum Confidence

Maximum Loss Reduction with Maximum Confidence (MMC) was developed by Yang et al [35]. The goal of the technique is to find the samples that will reduce the expected model loss the most, and select this sample for labeling. These are the basic notations that will be used when explaining the MMC approach:

- The labeled dataset: D_L .
- The unlabeled dataset: D_U .
- Possible query set: D_S .
- Optimal query set: D_S^* .
- The classification function that is trained on dataset D_L : f_{D_L} .
- A data point: x , and its label: y .
- The loss of on data point x : $L(f_{D_L}(x))$.
- The expected loss of the model: $\widehat{\sigma_{D_L}}$.

The expected model loss that MMC is trying to reduce can be defined as follows [35]:

$$\widehat{\sigma_{D_L}} = \int_x \left(\sum_{y \in Y} L(f_{D_L})P(y|x) \right) P(x) dx \quad (2.14)$$

It is hard to estimate $P(x)$, so it is instead measured over all the samples in D_U . This results in another estimate:

$$\widehat{\sigma_{D_L}} = \frac{1}{|D_U|} \sum_{x \in D_U} \sum_{y \in Y} L(f_{D_L})P(y|x) \quad (2.15)$$

After a set of data points D_S has been labeled, the new dataset $D'_L = D_L + D_S$ is obtained. Under the assumption that any $x \in D_U - D_S$ has an equal effect on a model trained on the datasets D_L and D'_L , we get the following equation for the reduction of the expected loss [35]:

$$D_S^* = \arg \max_{D_S} (\widehat{\sigma_{D_L}} - \widehat{\sigma_{D'_L}}) = \arg \max_{D_S} \left(\sum_{x \in D_S} \sum_{y \in Y} (L(f_{D_L}) - L(f_{D'_L}))P(y|x) \right) \quad (2.16)$$

In their paper, Yang et al. considers the process of finding the greatest reduction in two steps: finding a good estimate for the conditional probability $p(y|x)$, and finding a way to assess the loss reduction of a multi-label classifier.

It is unfeasible for a query strategy to provide a an estimation for all possible label combinations. If there are n different labels, there will be 2^n different label combinations. In order to estimate the conditional probability $p(y|x)$, MMC uses an approach that first estimates the number of labels for a given data point, and then uses that estimate to select the most probable labels. Consider the case where a data point has m labels. Since we can obtain

the probability from our classification model for each label, we can sort them in descending order. The first m labels are then likely to have a high probability, while the rest a rather low probability.

Yang et al. [35] describes the process of estimating the number of labels as follows:

1. Use the classification model to obtain the probabilities for each label for all the data samples.
2. For each data sample, sort and normalize the probabilities for all the labels.
3. Using the labeled dataset, train a logistic regression classifier with the sorted and normalized probabilities as features, and the number of labels as target.
4. With the trained logistic regression model, predict the number of labels for the samples in the unlabeled pool.

After obtaining the predicted number of labels m for a sample x , the estimate for $p(y|x)$, denoted \hat{y} , is then obtained by selecting m the most probable labels based on the original classification models output.

The only task that's left is now to estimate the loss for the multi-label classifier. By using the model where there are k different binary classifiers for a problem with k labels, the model loss can be calculated by adding the loss for the different binary classifiers like:

$$L(f) = \sum_{i=1}^k L(f^i) \quad (2.17)$$

where the loss of a single binary classifier is denoted as $L(f^i)$. With this definition, it only remains to define the measure of loss on a single binary classifier. The measurement that is used by MMC is to estimate the model loss by the size of the version space of the SVM [32, 35]. The version space's size can be computed with Equation 2.13. However, computing this for each possible label set is expensive. By using the heuristic from [31], an approximation of the version space with the added label can be obtained from the current SVM classifiers' margin. The reduction rate after adding a the data point (x, y^i) , where $y^i \in -1, 1$, can be expressed as follows [35, 31]:

$$\frac{L(f_{D_L}^i(x, y^i))}{L(f_{D_L}^i)} \approx \frac{V_{D_L}^i(x, y^i)}{V_{D_L}^i} \approx \frac{1 + y^i f_{D_L}^i(x)}{2} \quad (2.18)$$

$L(f_{D_L}^i)$ does not involve the sample selected for labeling, so by writing the loss reduction as:

$$\begin{aligned} L(f_{D_L}) - L(f_{D'_L}) &= \sum_{i=1}^k (L(f_{D_L}^i) - L(f_{D'_L}^i)) \\ &= \sum_{i=1}^k (L(f_{D_L}^i) (1 - \frac{L(f_{D'_L}^i)}{L(f_{D_L}^i)})) \end{aligned} \quad (2.19)$$

it can be seen that focusing on the reduction rate is sufficient. By incorporating the result from Equation 2.18, the following approximation for the reduction rate is obtained:

$$\sum_{i=1}^k \left(\frac{1 - y^i f_{D_L}^i(x)}{2} \right) \quad (2.20)$$

The only thing that remains is now to combine the estimation of $p(x|y)$ with the estimate for loss reduction. The resulting equation, called maximum loss reduction with maximal confidence, is [35]:

$$D_S^* = \arg \max_{D_S} \left(\sum_{x \in D_S} \sum_{i=1}^k \left(\frac{1 - \hat{y}^i f_{D_L}^i(x)}{2} \right) \right) \quad (2.21)$$

Adaptive Active Learning

In their paper, Li et al. [20] presents two approaches to active learning:

- Max-Margin Uncertainty Sampling
- Label Cardinality Inconsistency

These two techniques are then combined in a weighted fashion into what they call Adaptive Active Learning.

Max-Margin Uncertainty Sampling

The idea behind *Max-Margin Uncertainty Sampling* comes from the observation that multi-label classification prediction is mainly about separating the positive labels from the negative labels [20]. That is, separating the labels gets assigned to an instance and the ones that does not. In order to model the uncertainty of the prediction of the prediction on a datapoint, Li et al. suggest the usage of a global separation margin to separate the negative labels from the positive.

The positive labels for a data point x is defined as those where $\text{sign}(f'_{D_L}(x))$ is positive. The separation margin is defined as:

$$\begin{aligned} \text{sep_margin}(x) &= \min_{i \in \hat{y}^+} f'_{D_L}(x) - \min_{i \in \hat{y}^-} f'_{D_L}(x) \\ &= \min_{i \in \hat{y}^+} |f'_{D_L}(x)| + \min_{i \in \hat{y}^-} |f'_{D_L}(x)| \end{aligned} \quad (2.22)$$

where \hat{y}^+ denotes the set of predicted labels that are positive on the instance, and the \hat{y}^- denotes the negative ones.

The data point that the model is the most uncertain about is then the one with the smallest margin. Li et al. define their global measure, max-margin prediction uncertainty, as:

$$u(X) = \frac{1}{\text{sep_margin}(X)} \quad (2.23)$$

Label Cardinality Inconsistency

Label Cardinality Inconsistency is based on that the underlying distribution is the same for the labeled and unlabeled data. In a multi-label dataset, the *label cardinality* is defined as the average number of labels assigned to each class [33]. The selection strategy that Li et al. based on this measures the Euclidean distance between the number of assigned predicted labels on x , and the label cardinality of the labeled data:

$$c(x) = \left\| \sum_{i \in \hat{y}^+} 1 - \frac{1}{N_L} \sum_{y \in Y_L} \sum_{i \in y^+} 1 \right\|_2 \quad (2.24)$$

where N_L is the number of labeled samples, Y_L are the labels for those samples, and y^+ are the positive labels in y .

Integration - Adaptive Active Learning

Since *max-margin uncertainty sampling* and *label cardinality inconsistency* complement each other, an integration method is used:

$$q(x, \beta) = u(x)^\beta \cdot c(x)^{1-\beta} \quad (2.25)$$

where β is a parameter controlling the weight put on the two measures. This parameter is chosen by in each iteration evaluating a discrete set of values, for example $\{0, 0.1, 0.2, \dots, 1\}$.

Then selecting β based on the most informative sample amongst the . Equation 2.25 shows the *approximate generalization error*, which is used to select the sample.

$$\epsilon(x) = \sum_{x \in D_U} \max_{i \in f(x)^+} (1 - f^i(x)) + \max_{i \in f(x)^-} (1 + f^i(x)) \quad (2.26)$$

where $f(x)^+$ and $f(x)^-$ are the predicted positive labels, and negative labels, respectively. So, the sample is then selected by:

$$x^* = \arg \min_{x \in D_U} \epsilon(x) \quad (2.27)$$

2.4 Evaluation Metrics

For classification of information retrieval systems the typical evaluation metrics to use are *precision*, *recall* and *recall* [16]. We define the following metrics in terms of *true positives*, *false positives*, *true negatives* and *false negatives*. How they are defined can be seen in table 2.1. Data points that are correctly classified are then either *true positives* or *true negatives*.

	Correct P	Correct N
Predicted P	True Positive	False Positive
Predicted N	False Negative	True Negative

Table 2.1: Confusion matrix for explaining true positives, false positives, true negatives and false negatives

Precision is the percentage of the results found by the system that are correct [27]. Recall is the percentage of correct results in the dataset that are found by the system. Precision and recall are defined as follows by Van Rijsbergen [26]:

$$Precision = \frac{tp}{tp + fp} \quad (2.28)$$

$$Recall = \frac{tp}{tp + fn} \quad (2.29)$$

F-score is the harmonic mean between recall and precision, and is defined as [27]:

$$F = \frac{2 * precision * recall}{precision + recall} \quad (2.30)$$

Another metric that is used for evaluation is *accuracy*, which is defined as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (2.31)$$

These metrics mainly work

2.5 Related Work

Using topic modeling for various clinical applications has been done before. Topic modeling have been a popular approach for this purpose since clinical data often is in the form of free text. The resulting topics can be interpreted by humans, which allow doctors to get more insight into the system. Sarioglu et al. [28] to represent clinical reports with topic vectors in order to classify them. Chan et al. [10] used topic models to analyze patient records and clinical reports from cancer patients. In their paper, they found relationships between the content of the notes on the patients, with the data that was available on the patients' genetic

mutations. The interpretability of topics generated from an LDA model was studied on clinical reports by Arnold et al [3]. They evaluated how interpretable topics were based on how many topics the model used.

Active learning has been researched in text classification with different approaches. They can be seen as two categories: searching through the hypothesis space by using the uncertainty of a model, or by exploiting the structure of the data through clustering [12].

One of the common baselines for active learning is uncertainty sampling [19]. That simply queries the label for the data point the model is most uncertain about. In [12] hierarchical clustering is used in an active learning system. The labels are queried from clusters where there is a lot of uncertainty when it comes to the majority label. By pruning the tree of clusters while querying for labels the goal is to obtain a pruning where each node mostly contains one label.

In [23] they also take advantage of a clustering to select the samples to be labeled in a two-class environment. They take advantage of that the data points closest to the centroids are the most important ones, and that most data points in one cluster have the same label. What this approach has in common with a lot of the current research is that it is treating single-label or binary classification problems, which cannot be directly applied to a multi-label scenario.

Research in [9] is dealing with the multi-label problem. That is the paper that developed the *binary version space minimization* strategy that is described in section 2.3. The MMC strategy that is described in section 2.3, and the adaptive active learning strategy in section 2.3 are also techniques for managing the multi-label problem. For image classification, there has been some methods develop, for example [21, 24]. In [21] the goal is to, after making predictions, selecting the sample with the biggest mean loss. However experiments have shown that this is not as suitable for text classification [35]. In [24], the approach is to use pairs of labels and samples to present to the annotator, and the aim is to minimize the Bayesian classification error. Due to the fact that labeling for text classification is more time consuming than image classification, since you have to read an entire text, this approach is not suitable for text classification [35].

Active learning has been used to deal with the problem of imbalanced datasets before. In a binary classification setting, Ertekin et al. [13] used uncertainty sampling with SVMs to get a more balanced dataset. In [5], Attenberg et al. uses density based active learning to improve the class balance. However, it does not attempt to apply this in a multi-label setting.



3 Method

The task of making a better system for labeling clinical reports was approached with several text mining techniques, support vector machines and a few active learning querying strategies. At first, the framework and tools used in the system are described, followed by a description of the provided dataset. Finally, the experiments used to answer the research questions are presented.

3.1 Frameworks and Tools

The entire system was written in Python. The motivation behind this choice was mainly that, when it comes to machine learning and text mining, most of the existing infrastructure at Sectra is using Python. This, in combination with the fact that there exists several tools for these purposes in Python, such as *numpy*, *nlTK*, *scikit-learn* and *gensim*.

However, when it comes to the active learning, there does not seem to be a proven mainstream library that contains a set of readily available algorithms. In order to achieve better integration between the active learning system and the existing infrastructure at Sectra, as well as making adaptations such as the number of items queried at each iteration, an active learning framework was created. The ground for this framework were the algorithms presented in Section 2.3.

3.2 Datasets

In this thesis, two different datasets were used. The dataset provided by Sectra, as well as Reuters-21578. The latter was used to be able to simulate a multi-label labeling process, to evaluate how well the different strategies work before being integrated into Sectra's system. Since the vast majority of the dataset from Sectra was unlabeled, this could not effectively be done using only that.

The set of reports provided by Sectra contained [X] different entries, where [Y] were initially labeled. However, those labels were subject to change, so they were mainly used to see if there were any correlation between the labels and clusters during the exploration phase. A sample report can be seen in [FIGURE]. It contains the fields:

FIELD

The work used the ReportText field for all the text analysis. The labels that were initially assigned to these reports were:

LABEL

The distribution of labels among these initially labeled reports can be seen in [FIGURE]. Note that this is only a count of the individual labels, and the multi-label nature of the labeling is not taken into account in the histogram.

The Reuters-21578 newswire dataset is widely used when it comes to text classification research, and provides a good multi-label benchmark that can be used to compare how well certain techniques perform to other papers. All experiments used the *ModApte* split of the dataset, which is commonly used and readily available. It splits the dataset into a defined set of training and test documents, containing 7.769 and 3.019 entries respectively. This split contains a subset of the categories, specifically 90 different ones. Since the clinical dataset from Sectra only contained [X] different categories, this would not mirror that very well, so instead the 15 most common categories of those were taken out. The distribution of them can be seen in [FIGURE].

A decorative element consisting of several thin, vertical black lines of varying heights, creating a stylized 'L' shape or a series of vertical bars.

4 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.



5 Discussion

This chapter contains the following sub-headings.

5.1 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?

5.2 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 (3). 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.

5.3 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.



6

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] Charu C Aggarwal and ChengXiang Zhai. “A survey of text clustering algorithms”. In: *Mining text data*. Springer, 2012, pp. 77–128.
- [3] Corey W Arnold, Andrea Oh, Shawn Chen, and William Speier. “Evaluating topic model interpretability from a primary care physician perspective”. In: *Computer methods and programs in biomedicine* 124 (2016), pp. 67–75.
- [4] David Arthur and Sergei Vassilvitskii. “k-means++: The advantages of careful seeding”. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.
- [5] Josh Attenberg and Seyda Ertekin. “Class imbalance and active learning”. In: *inde Imbalanced Learning: Foundations, Algorithms, and Applications* (2013), p. 101149.
- [6] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [7] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.
- [8] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. “Learning multi-label scene classification”. In: *Pattern recognition* 37.9 (2004), pp. 1757–1771.
- [9] Klaus Brinker. “On active learning in multi-label classification”. In: *From Data and Information Analysis to Knowledge Engineering*. Springer, 2006, pp. 206–213.
- [10] Katherine Redfield Chan, Xinghua Lou, Theofanis Karaletsos, Christopher Crosbie, Stuart Gardos, David Artz, and Gunnar Ratsch. “An empirical analysis of topic modeling for mining cancer clinical notes”. In: *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE. 2013, pp. 56–63.
- [11] Steven P Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. “Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond”. In: *Mining text data*. Springer, 2012, pp. 129–161.
- [12] Sanjoy Dasgupta and Daniel Hsu. “Hierarchical sampling for active learning”. In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 208–215.

- [13] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. “Learning on the border: active learning in imbalanced data classification”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, pp. 127–136.
- [14] Thomas L Griffiths and Mark Steyvers. “Finding scientific topics”. In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.
- [15] Thomas Hofmann. “Probabilistic latent semantic analysis”. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, pp. 289–296.
- [16] Jing Jiang. “Information extraction from text”. In: *Mining text data*. Springer, 2012, pp. 11–41.
- [17] Thorsten Joachims. “Text categorization with support vector machines: Learning with many relevant features”. In: *European conference on machine learning*. Springer. 1998, pp. 137–142.
- [18] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.
- [19] David D Lewis and William A Gale. “A sequential algorithm for training text classifiers”. In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc. 1994, pp. 3–12.
- [20] Xin Li and Yuhong Guo. “Active Learning with Multi-Label SVM Classification”. In: *IJCAI*. 2013, pp. 1479–1485.
- [21] Xuchun Li, Lei Wang, and Eric Sung. “Multilabel SVM active learning for image classification”. In: *Image Processing, 2004. ICIP’04. 2004 International Conference on*. Vol. 4. IEEE. 2004, pp. 2207–2210.
- [22] Oscar Luaces, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. “Binary relevance efficacy for multilabel classification”. In: *Progress in Artificial Intelligence* 1.4 (2012), pp. 303–313.
- [23] Hieu T Nguyen and Arnold Smeulders. “Active learning using pre-clustering”. In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 79.
- [24] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, and Hong-Jiang Zhang. “Two-dimensional active learning for image classification”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, pp. 1–8.
- [25] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. “Classifier chains for multi-label classification”. In: *Machine learning* 85.3 (2011), p. 333.
- [26] C.J Rijsbergen. *Information Retrieval, 2nd Edition*. 1979.
- [27] Erik F Tjong Kim Sang and Fien De Meulder. “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition”. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (2003).
- [28] Efsun Sarioglu, Kabir Yadav, and Hyeon-Ah Choi. “Topic modeling based classification of clinical reports”. In: *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*. 2013, pp. 67–73.
- [29] Hinrich Schütze and Craig Silverstein. “Projections for efficient document clustering”. In: *ACM SIGIR Forum*. Vol. 31. SI. ACM. 1997, pp. 74–81.
- [30] Burr Settles. “Active learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1 (2012), pp. 1–114.
- [31] Simon Tong. *Active learning: theory and applications*. Stanford University, 2001.

- [32] 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.
- [33] Grigorios Tsoumakas and Ioannis Katakis. “Multi-label classification: An overview”. In: *International Journal of Data Warehousing and Mining* 3.3 (2006).
- [34] Vladimir Vapnik. *Statistical learning theory*. 1998. Wiley, New York, 1998.
- [35] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. “Effective multi-label active learning for text classification”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2009, pp. 917–926.