The Impact of Incorrect Training Sets and Rolling Collections on Technology-Assisted Review

Johannes C. Scholtes
University of Maastricht and
ZyLAB Technologies BV
Hoogoorddreef 9, 1101 BA,
Amsterdam, The Netherlands
johannes.scholtes@zylab.com

Tim van Cann
University of Maastricht

Postbus 616, 6200 MD
Maastricht, the Netherlands
timvancann@gmail.com

Mary Mack
ZyLAB North America LLC
7918 Jones Branch Drive, Suite 230
McLean, VA, 22102
United States of America
mary.mack@zylab.com

ABSTRACT

Document classification and machine learning technology in electronic discovery (eDiscovery) are gaining attention under new names such as technology-assisted review (TAR), machine-assisted review (MAR), computer-assisted review (CAR) and predictive coding [2]. Several judicial rulings have addressed typical legal concerns in relation to the quality of machine learning. In this paper, we address two of such concerns and investigate their relation with machine learning quality in more detail. The topics of interest of this paper are: (i) the impact of the quality of training documents on the overall classification results, which can be measured by investigating the impact of training supervised classifiers deliberately with wrong training samples and (ii) using machine learning in so-called rolling collections, which can be measured by investigating the quality of classification of new and unknown documents, which have not been used to extract or select machine-learning features, with existing classifiers.

A machine learning pipeline with the most-common used document feature-extraction techniques known as a bag-of-word (BoW), and Term Frequency- Inverse Document Frequency (TF-IDF) is used [9]. For feature-selection, vector logarithmic normalization and cut-off of non-used or non-relevant dimensions has been selected. The resulting data was used to train binary classifiers for each category using Support Vector Machines (SVM). Documents for the experiments came from the Reuters RCV1 corpus.

We found that in this model: (i) the impact of wrong training documents was smaller than expected: inserting up to 25% wrong training documents resulted only in 3-5% less classification quality, and (ii) that using BoW and TF-IDF based classifiers lost up to 50% in quality when used on completely new documents, such as in a rolling collection.

Keywords

Machine Learning, eDiscovery, Document Classification, Support Vector Machines (SVM), Technology-Assisted Review (TAR), Machine-Assisted Review (MAR), Computer Assisted Review (CAR), and Predictive Coding.

1. INTRODUCTION

Last year was a breakthrough year for machine-learning technology in the eDiscovery market, especially for large scale legal reviews.

Legal review is the part of an eDiscovery process where lawyers or investigators review documents and (manually) classify them in various document categories such as but not limited to privileged, confidential, and responsive. This is a very labor intensive process and always the most expensive part of a pre-trial eDiscovery. Similar (expensive) review processes of evidence material exist in internal investigations, audits, law enforcement and intelligence applications [1]. Recently, records managers or even common business users can be added to these groups, as they also need to review large data collections as part of defensible disposition- or legacy data clean-up processes [13].

As the size of electronic data collections continues to grow exponentially, it is impossible to continue reviewing documents manually [5]. Using supervised machine-learning is one of the techniques used to automate this process. Other tools are rule-based and key-word based classification, which we will not discuss in this paper.

Using machine learning has raised many legal concerns. Here, we address two of such questions: (i) What is the impact of the quality of the training documents on the overall classification results, and (ii) can we use machine learning in so-called rolling collections.

With respect to first question, parties have been requested to disclose training documents to the other side to validate the quality of the training data in a very early part of the pre-trial discovery. This may not always be in the interest of the disclosing party, as these are often the most significant documents in a case. But do we understand the exact impact of wrong training documents on the machine-learning quality? If the effect is not that large, this could change the early disclosure obligations.

With respect to the second question, it is not clear if additional collection batches require a full new training cycle or that existing classifiers can be used to classify new documents. If existing classifiers can also be used without too much loss of quality, a costly and lengthy process involving machine-learning and validation can be avoided.

To investigate the applicability of – and the effects on the machine learning results for these two legal contexts, two special machine-learning cases are investigated: (i) training supervised classifiers deliberately with wrong training samples, and (ii) the quality of classification of new- and unknown documents, which have not been used to extract- or select machine-learning features, with existing classifiers.

First, a ground truth is created. The ground truth contains the classification results of a typical machine-learning process as it is used in the legal industry. Next, the impact of (i) training the classifier with wrong documents and (ii) classifying new

documents with the classifiers will be measured by comparing the results against the ground truth.

In the next sections, the setup of the document classification pipeline will be discussed, including a detailed overview of the individual components and the experiments.

2. THE CLASSIFICTION PIPELINE

When supervised machine learning is used for automatic document-classification, a number of choices are made in relation to the translation of document content into mathematical data-representation for the machine-learning algorithm. Many variations and approaches exist. In addition, various supervised machine learning algorithms exist, all with different characteristics and quality.

In this project, the following techniques are selected to use for the machine learning process:

1. Two document feature-extraction techniques known as a (i) bag-of-word (BOW) and (ii) Term Frequency-Inverse Document Frequency (TF-IDF), and;
2. Basic document feature-selection techniques such as logarithmic normalization and selection of the relevant features by vector cut-off, and;
3. A supervised machine-learning algorithm based on Support Vector Machines (SVM) to build binary classifiers for each document category,

We will discuss these three choices in more detail in the following paragraphs.

2.1 Document Representation

To prepare a document for machine learning is not a trivial task. The document text has to be transformed into numerical data and mathematical structures. This is done by first extracting relevant features, and then to select the most relevant features.

2.1.1 Feature extraction

Feature extraction is the process of extracting relevant information from the data to create feature vectors. In eDiscovery, two feature selection schemes are used most: the Bag of Words (BoW) and the Term Frequency - Inverse Document Frequency (TF-IDF). In both cases all linguistic structure of a document are ignored and words (or terms) are seen as individual features. This may be disputable and also lead to lower quality of classification as has been addressed in [8], but in this research we restrict ourselves to these basic forms of feature extraction, as most if not all in the legal industry uses them as such.

The most basic feature extraction method is the Bag of Words (BoW). Here, text in a document is tokenized and all punctuation, numbers, and words shorter than three letters are removed. An official English stop word list is used to remove highly frequent and therefore less-discriminating words. By finding all distinct words in a collection of documents, a document vector can be produced for each document. In this document vector, dimensions representing words present in the document are converted to 1 and dimensions representing words that are not present are converted to 0. Alternatively, one can also calculate the Document Frequency (DF) of all words and replace the 1 and 0 with their respective document frequencies.

An extension of the BoW representation is the Term Frequency-Inverse Document Frequency (TF-IDF), which is calculated by multiplying the Term Frequency (TF) of a word in a document with the Inverse Document Frequency (IDF) of that word. The IDF can be calculated by dividing the size of the corpus with the number of documents the term occurs in. This approach is based on the heuristic that terms that occur in many documents are less discriminating than terms that occur only in a few documents.

2.1.2 Normalization

In order to reduce the possible large gaps between feature values normalization can be applied. Each feature is normalized between 0 and 1. Since this normalization can still create very small values a second normalization can be applied by taking the logarithm (base 10).

2.1.3 Feature selection

In order to further decrease the number of features present in the document vectors, features that seemingly do not contribute much can be removed from the document vectors. These potentially less useful features are features that are not necessary to create a difference between document categories in the machine learning phase; for example words or terms that have a high occurrence in a majority of documents.

Vector *Cut-off* is defined as:

$$Cut\text{-}off = min + (max - min) * (perc/100))$$

Where *max* is the maximum value in the document matrix *mxn* (the collection of all document vectors where *m* are the documents and *n* the features) and *min* is the minimum value in the document matrix. *perc* is a constant value chosen to be 1. If a feature has no value higher or equal to the *Cut-off*, the feaure is removed from the vector, otherwise it is kept.

This method does not remove possible outliers that hold important information but there is a risk that it may remove wanted features or keep possible unwanted features. But by keeping the *perc* value small (such as 1 in our case), this problem is avoided.

2.2 Supervised Machine Learning with SVM

Supervised machine-learning is used to construct a system that can be trained with tagged data. Each training document is tagged with the appropriate classification category. The training examples are used to create a model that can categorize new documents based on mathematical decisions. To ensure multi-category classification, for each category a separate binary model is trained which can predict with certain probability whether a document is part of the category or not. New documents are fed to all classifiers. The document then belongs to all classes for which the corresponding classifiers returns a value that is higher than a pre-set threshold.

The current leading machine learning technique in the field of text classification is Support Vector Machines (SVM). SVM builds model that separates two classes by projecting the vectors into a higher dimensional space. Next to the linear model we also used the Gaussian kernel and the sigmoid kernel. LIBSVM [4] was used for the implementation of the experiments.

3. EXPERIMENTS

3.1 Corpus

In this research, the fully annotated Reuters RCV1 [7] corpus is used. In the machine-learning research community, this corpus is one of the common standards used to evaluate the quality of automatic document classification. RCV1 has the following features:

- A little over 800,000 news articles formatted in XML format;
- About 2.5 GB uncompressed;
- Only English articles;
- Pre-annotated with 126 topic codes, 352 industry codes and 296 region codes. A document can have multiple topic, industry and/or region codes;
- The codes are hierarchical from most general to most specific category;

The corpus is partly annotated by humans; the rest has been annotated by machine learning after which the documents are checked and if necessary corrected by a selected group of professionals.

3.2 Evaluation

For the evaluation, we have used the same evaluation method used by the Legal-TREC [6] conferences, which are based on general best practice principles for measuring the quality of document classification and document retrieval in general[14] and in the application of legal review in particular [5][10]. We used the F1 score and derived 11-points precision graphs representing the quality of the classifiers [12].

3.3 Experiments Setup

The following experiments were implemented.

1. Randomly select documents from the Reuters RCV1 documents for a number of categories. 90% of the selected documents are used for training. 10% for verification of the classifiers. In addition, we also compiled an additional large test set containing randomly selected documents from the RCV1 corpus. Feature extraction was done by using BoW and TF-IDF. Feature selection was done by using vector cut-off and logarithmic normalization. For training, we used a SVM with a linear kernel. This experiment was done to obtain a ground truth with classification results.
2. The correctness of the labels given to documents can be questioned when working with humans who can make mistakes when selecting wrong documents for training. In this experiment we test how much incorrect labels are of influence on the results. We run three different experiments:
    a. We take *n* random negative labels and switch them to positive;
    b. We take *n* random positive labels and switch them to negative;
    c. We take *n* random negative and n random positive labels and inverse the labels. When injecting one-sided errors we compensate by remove *n* random documents from the injected test labels to ensure approximately 50% positive labels and 50% negative labels.

    n is chosen to vary between 0% and 25% with a step size of 5%. These are arbitrarily chosen values.
3. Additional (new and unknown) documents are randomly selected from the corpus to simulate a rolling collection. First, document vectors are constructed from documents in the known training and validation set only; as a result, the mathematical models for the document representation only contain features extracted from terms in this restricted set. This mathematical model is then used to build binary classifiers for each document category. Next, another unknown set of additional documents is used to classify with these existing classifiers. As explained before, the motivation behind this setup is that in many legal and investigative applications, not all data is available when a project starts. As a result, one cannot build a document representation that takes all features of the entire document set into consideration. New data is constantly added. This is also called a rolling collection.

4. RESULTS

4.1 Ground Truth Creation

In order to create a ground truth, first training and validation sets are randomly generated per Reuters category from RCV1 by randomly choosing 1,500 documents in the actual Reuters category (positive instances) and 1,500 documents outside the Reuters category (negative instances). From these documents 90% are used for training and 10% for validation of the classifiers. An additional test set is then created with 25,000 randomly selected documents from the entire Reuters corpus. This additional test set contains documents from within the category and outside the category.

Table 1: F1 Scores for the Ground Truth

Feature-Extraction Used	**Validation**	**Additional Test**
Bag of Words	0.9325	0.8925
TF-IDF	0.9068	0.8168

As can be seen in table 1, the results for classification are quite high. BoW is even slightly higher than the more advanced TF-IDF feature extraction. This probably has to do with our very basic choice for feature selection methods. The documents from the additional test set also score lower than the validation set. This probably has to do with how the Legal-TREC evaluation measures work: on larger document sets, it is harder for documents to get to the top of list that is recognized as properly classified as more documents compete for that ranking.

Looking at the 11-point precision in figure 1, one can obtain a more detailed insight in the quality of the classifier.

The black circle around a coordinate shows a point with a F1-score equal or higher than 0.8, the magenta circle around a coordinate shows a point with both recall and precision equal or higher than 0.8. The area where both precision and recall are over 0.8 is pretty large, which indicates a robust classifier which is not too sensitive to user-controlled thresholds.

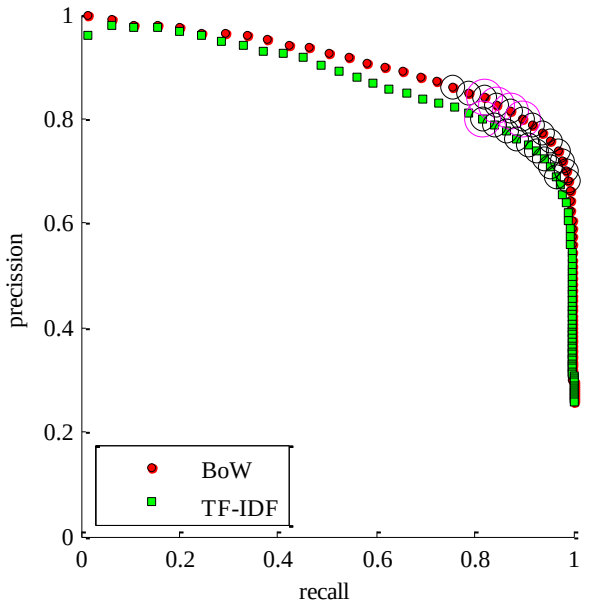

Figure 1: 11-Points Precision for the Ground Truth

4.2 The Effects of Wrong Training Data

Now that we have obtained our ground truth, we can measure the impact of incorrectly applied labels.

In this experiment, we measure the impact of 5%, 10%, up to 25% deliberately inserted wrong training documents.

Table 2 and figure 2 and figure 3 lead to a very interesting conclusion: injecting up to 25% errors in both the positive and negative training set before handing it to SVM to learn does not lead to an alarming loss in quality of evaluation. This conclusion leads to interesting application. It means that a corpus or a training set may contain incorrect labels. Incorrect labels are prone to happen when working human classifiers.

Table 2: Quality Loss F1 Scores from Injecting Training Errors

Doc. representation	Train size	0% error	25% error	Loss 0%-25%
Bag of Words	1k	0.8380	0.8112	3.20%
	3k	0.8527	0.8075	5.30%
TF-IDF	1k	0.8281	0.7989	3.45%
	2k	0.7559	0.7082	3.53%

Larger training sets suffered relatively from more loss. We expect that this has to do with the fact that the SVM model uses more incorrect data to model the classifier with larger training sets than with smaller.

Figure 2 and 3 show the detailed 11-point precision graphs of the decline in quality for the different classifiers for 5, 10, 15, 20 and 25% errors in the training sets.

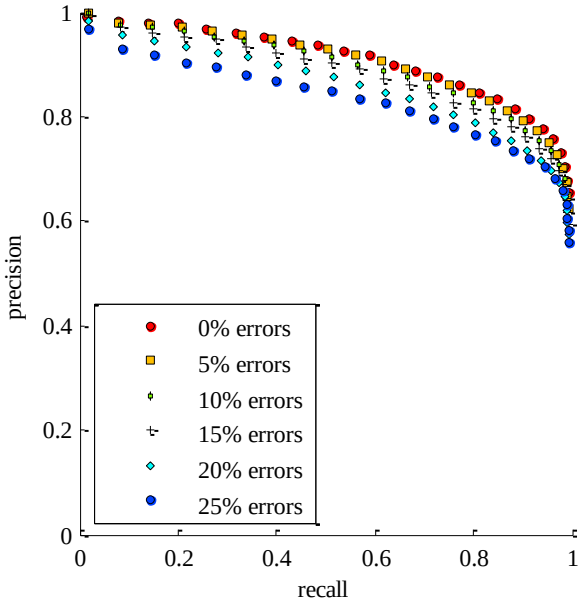

Figure 2: Quality Loss for Bag of Words

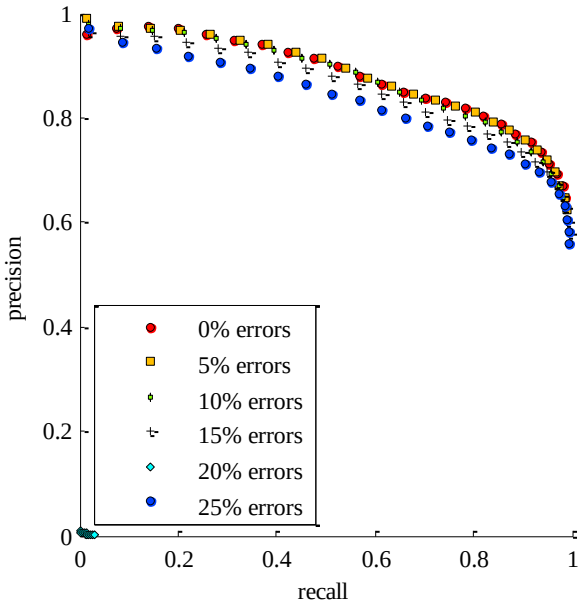

Figure 3: Quality Loss for TF-IDF

In conclusion, table 2 shows that the quality loss is very small for injecting up to 25% errors: only 3.2% to 5.3%. This means that the SVM algorithm is capable of creating a model with significant accuracy regardless of possible errors in labels in the training set. We found that boosting the errors up to 35% did lead to a complete failure of the SVM algorithm to build a proper model for the smaller training sets, but larger training sets (5,000 or

10,000) were able to build models even with 35% errors. We did not include these because we felt that such large training sets are not currently desirable in the eDiscovery context.

4.3 Rolling Collections: Using Existing Classifiers on New Data

Our next experiment leads us to the question of measuring the effects of the classification quality in the case of a rolling collection. We use a machine-learning classifier build from a very specific set of documents, on a completely different set of documents that were not part of the original construction of the document representation vectors. Looking back at the ground truth, table 3 shows the results when all documents are collected before we build our classifiers.

Table 3: F1 Scores for Static Collection

Feature-Extraction Used	Validation	Additional Test
Bag of Words	0.9325	0.8925
TF-IDF	0.9068	0.8168

Now, if we use these classifiers to classify completely new documents, we see that the performance of both the BoW and the TF-IDF document representation completely collapses.

Table 4: F1 Scores for Rolling Collection

Feature-Extraction Used	Validation	Additional Test Rolling Collection
Bag of Words	0.9872	0.3049
TF-IDF	0.9135	0.4828

These results are very relevant to the legal community, because it shows that the quality of automatic classification based machine learning, dramatically lowers in the case of rolling collections.

When using Bow or TF-IDF, to achieve validation based on precision and recall, rolling collections require a full rebuilding of the classifiers, including training and quality verification.

When TF-IDF is used to represent the content of the document, one also needs to derive the TF-IDF vectors for the entire document collection requiring significant amount of computational resources.

5. CONCLUSIONS

Injecting up to 25% errors into the BoW or TF-IDF training set before handing it to SVM to learn, does not lead to an alarming loss in quality of evaluation. As a result, one could question the necessity of disclosing the document training set for validation by the other party. Even when 25% errors are added, the classifiers still perform in the 0.8 area, which is considered to be equivalent to human reviewers.

However, in a rolling collection, when using the BoW and TF-IDF feature extraction methods, one needs to calculate the BoW and TF-IDF mathematical models over the entire document collection in order to obtain acceptable classification results. Classifying completely new documents that were not used to calculate the BoW and TF-IDF documents will lead to very low classification validation (precision and recall). In the case of rolling collections, it is recommended to recalculate-, train- and verify the entire machine learning model on the entire document collection for every addition of new documents.

6. ACKNOWLEDGMENTS

The authors wish to thank ZyLAB Technologies BV for their generous support and encouragement of this research project and for providing real-world test data as well as valuable feedback on the results.

7. REFERENCES

[1] W. Andrews and D. Logan, Early Case Assessment: E-Discovery Beyond Judges and Regulators Is About Risks, Costs and Choices, January 27 (2010).

[2] W. Andrews, D. Logan, J. Bace and S. Childs, E-Discovery SaaS and On-Premises Software Converge at Vendors as They Mature, July 29 (2010).

[3] Baron, Jason R. (2005). Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery. Sedona Conference Journal. Vol. 6, 2005.

[4] Chih-Chung Chang and Chih-Jen Lin. Libsvm. http://www.csie.ntu.edu. tw/~cjlin/libsvm/.

[5] M. Grossman and G. Cormack, Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, Richmond Journal of Law and Technology, 17(3), Spring (2011).

[6] Legal-TREC Research Program: http://trec-legal.umiacs.umd.edu/.

[7] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. RCV1: A new bench- mark collection for text categorization research. The Journal of Machine Learning Research, 5:361–397, 2004.

[8] Aisan Maghsoodi, Merlijn Sevenster, Johannes C. Scholtes and Georgi Nabaltov (2012), Automatic Sentence-based Classification of Free-text Breast Cancer Radiology Reports, 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2012).

[9] Manning, C.D., Raghavan, P. and Schütze, H. Introduction to Information Retrieval Cambridge University Press, 2008.

[10] Oard, D., Baron, J., Hedin, B., Lewis, D. and Tomlinson, S., Evaluation of Information Retrieval for E-Discovery, Artificial Intelligence and Law 18(4)347-386 (2011).

[11] Reuters RCV1 Corpus: http://trec.nist.gov/data/reuters/reuters.html

[12] Rijsbergen, C.J. van (1979). Information Retrieval. Butterworths, London.

[13] Scholtes, J.C. (2012). The Dark Side of Big Data. Solicitors Journal, 03-10-2012. United Kingdom.

[14] Voorhees, Ellen M. (Editor), Harman, Donna K. (Editor), (2005). TREC: experiment and evaluation in information retrieval. MIT Press.