Submitted on: 19.07.2018

Machine learning for production of Dewey Decimal

Svein Arne Brygfjeld
National Library of Norway, Oslo, Norway.
E-mail address: svein.arne.brygfjeld@nb.no

Freddy Wetjen
National Library of Norway, Oslo, Norway.
E-mail address: Freddy.wetjen@nb.no

André Walsøe
National Library of Norway, Oslo, Norway.

Abstract:

Based on Open Source software and existing metadata and content, the National Library of Norway has carried out a series of experiments to study automatic classification of articles based on the Dewey Decimal Classification system. Various platforms and models for machine learning has been used. The results indicate machine learning is a suitable environment for semi-automated or fully automated production of DDC. Furthermore, they show that training of machine learning platforms may be enforced by using artificial documents.

Keywords: Machine Learning, Dewey Decimal Classification, Automatic classification.

1 INTRODUCTION

The National Library of Norway (NLN) has a national role as a memory institution for all types of published, both in printed as well as born digital, information. Among these are published scientific articles. Articles are received, catalogued and preserved at NLN. This activity has been going on for years, resulting in a relatively large collection of both content and metadata.

The need for effective cataloging is well recognised within the organisation, and various approaches are tested to reduce the amount of work spent on the cataloging. One of these approaches are the use of machine learning to automate parts of the cataloging work.

Over the last years, one has seen a significant general development and improvement of software for machine learning, exemplified in software like IBM Watson, TensorFlow from Google, several open source python libraries and others. Applications on many areas have been demonstrated.

A modern, digital library represent an ideal environment for ML. A digital library has a well-organized digital content as well as well-formed metadata in high quality. Both these assets are valuable contributions to ML.

2 ABOUT NORART

Norart holds metadata for more than 600.000 Norwegian and Nordic published articles [NLN 2007]. It is the main tool at the NLN for description of articles, both scientific and others. More than 36.000 of these are in digital format, with links from the metadata to the article on the Internet.

3 THE METADATA

The metadata in Norart is based on a Norwegian locale of MARC, NORMARC [NLN 2018]. Articles are given a classification based on DDC [OCLC 2018].

4 THE CONTENT

The content used in the experiments is based on articles published in Norwegian/Nordic scientific journals. They are born digital, and typically available as PDF or HTML.

5 TERMINOLOGY AND DEFINITIONS

ML platform: The technical platform used for machine learning. In this context of this article, the definition is limited to the software made for machine learning.
Model: The specific approach to using the ML platform.
Performance: The ability of the ML platform to deliver correct answers to relevant questions.
Training: The process of ingesting information into the ML platform and improving the performance of the platform
Test set: The information set used to verify the performance of the ML platform
Training set: The set of information used for training the ML platform
Testing: Verifying the performance of the trained ML platform. Testing is based on measuring the performance of the ML platform for a predefined test set.
Document: The content of the information set

6 ASSUMPTIONS

Based on the state of software, and the potential learning platform, one may assume that:

- Machines should be able to learn to do automatic classification based on DDC
- The precision should improve as the number of digits in the DDC grows[1]

For many use cases, one may expect that the number of objects in the learning set is smaller than needed. To compensate for the lack of content, one may make artificial documents by exchanging words in existing documents with other words. The assumption is that:

- Artificial documents will improve machine learning for text classification

It is assumed that the choice of model for learning will have influence on the precision, as well as the size of the learning set and the precision of the classification.

6.1 Artificial documents

Size and quality of data influences machine learning models. In our experiments, we detected this problem early. To achieve a larger dataset, we generated additional artificial documents based on the original documents based on methods described by Xiang Zhang et al. [Zhang, X 2016]. Different techniques were applied:

- Interchanging text within document and between documents of the same category
- Using words with similar meaning from pregenerated wordlists from various sources

Such changes are done to up to 20% of the original text. The result of this is much larger amount of documents. This is, of course, only done with the training documents.

It is well-known that the size and quality of the training set influences the performance of the ML platform. In some cases, the size of the available training set may be smaller than needed. In our experiments, the lack of documents has been compensated by making what we call artificial documents. These are new documents based on existing (original) documents, and they are produced by exchanging some words with new words, typical synonyms. Such changes are done to up to 20% of the original text. The result of this is much larger amount of documents.

7 METHOD

7.1 Performance factors

Four factors with an assumed significant influence of the performance are tested in the experiment. These are:

- The model
- The size of the training set
- The number of DDC digits, and
- The relative number of artificial documents as compared to real documents.

[1] As the number of digits grows, the implication is that the classification will be more precise.

7.2 Models

Various models is known to support various use cases. In the experiments carried out, four models have been used. These are:

- Multi-Layer Perceptrons (MLP)
- Convolutional Neural Networks (CNN)
- Fasttext, and
- Logistic Regression (LR)

7.3 Training set vs test set

For each of the training cycles, a training set and a test set is defined. The test set is in general 20% of the size of the training set. Training has been done based on a training set with 50, 100, 200 and 400 documents.

7.4 Training cycle

For each of the defined scenarios, a training cycle is carried out. The definition of a given training cycle includes the following steps:

1. Choice of model and ML platform
2. Definition of training set, including metadata and a set of documents
3. Definition of test set, including metadata and a set of documents
4. Definition of performance measures
5. Training
6. Testing
7. Evaluation

The process steps of the training circle is shown in Figure 1.

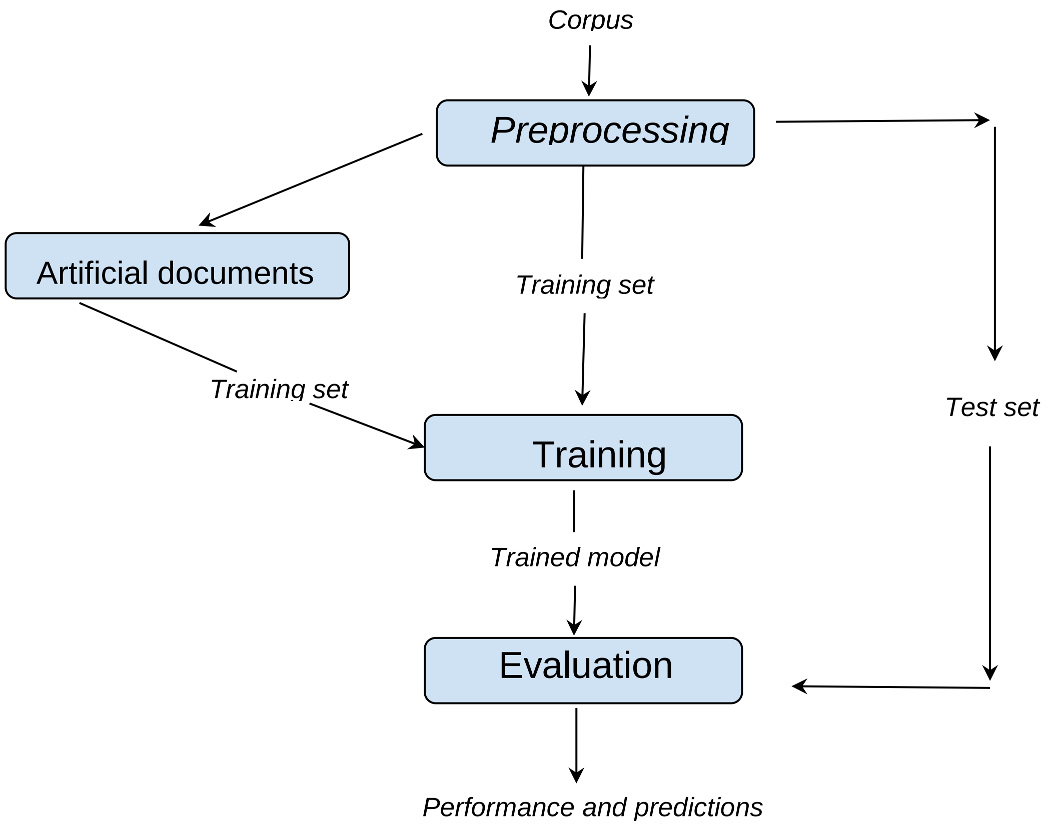

Figure 1

Each training cycle is carried out on all the models and for all the chosen number of documents for training.

A principle setup for the training and test setup is shown in the Figure 2, and in the potential use is also included.

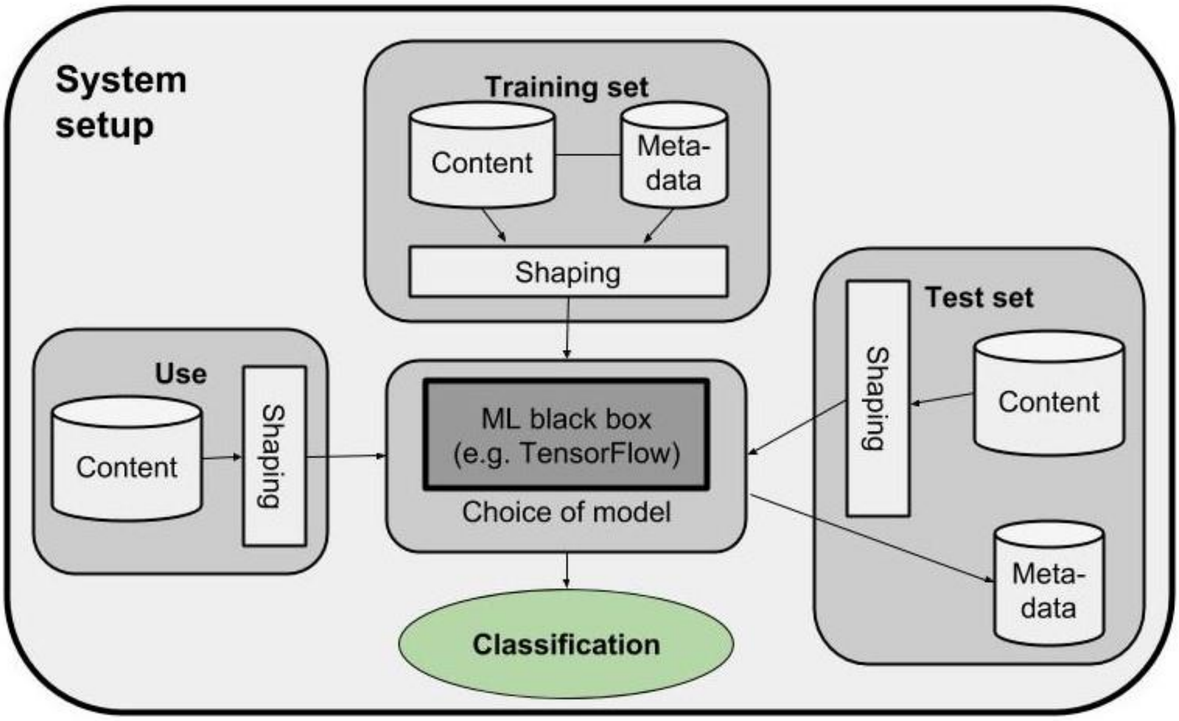

Figure 2

8 DISCUSSION

8.1 Size of training set

The size of the training set has a significant influence on the performance of a model. When the number of available documents in the training set is small, we can compensate by producing artificial documents. Figure 3 shows the performance depending on the number of documents in the training set, and the difference with and without artificial documents. The latter is indicated with a +sign.

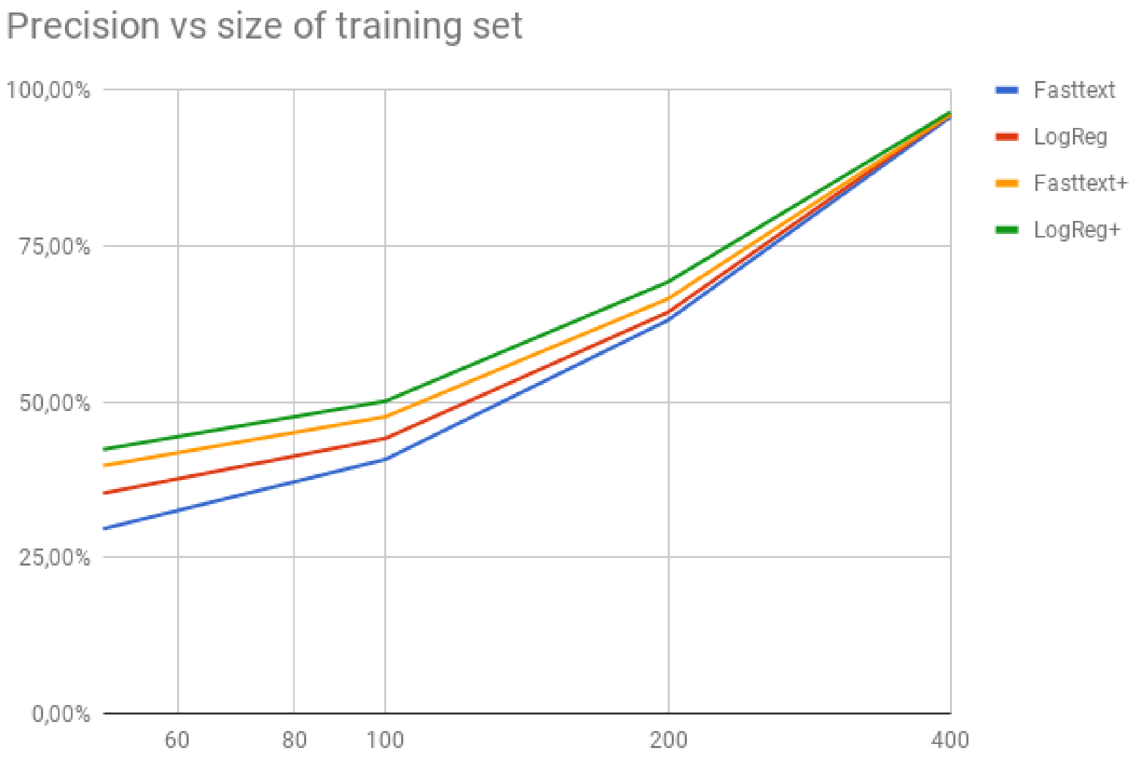

Figure 3

The performance of all models improves significantly as the size of the training set grows, and the difference between the performance of the different models become lesser as size grows.

When the available training set is too small, performance may be significantly improved by adding artificial documents. As an example, for LogReg with a training set of 50 documents, introducing 10% artificial documents improves lifts the precision from 35% to 42%.

In a real environment it would be logical to improve/enforce training by using feedback from users, e.g. based on user behaviour in a digital library user experience.

8.2 Precision of prediction

As the number of digits in the DDC classification grows, one implication is that the subject represented by the DDC number is more precise. Based on a very limited set of experiments, it is shown that as the classification becomes more precise (more DDC digits), the performance of the classification becomes better. This is shown in Figure 4, where the variation of the performance for 3, 4 and 6 DDC digits is shown for all four models, and with a constant of 200 documents in the training set.

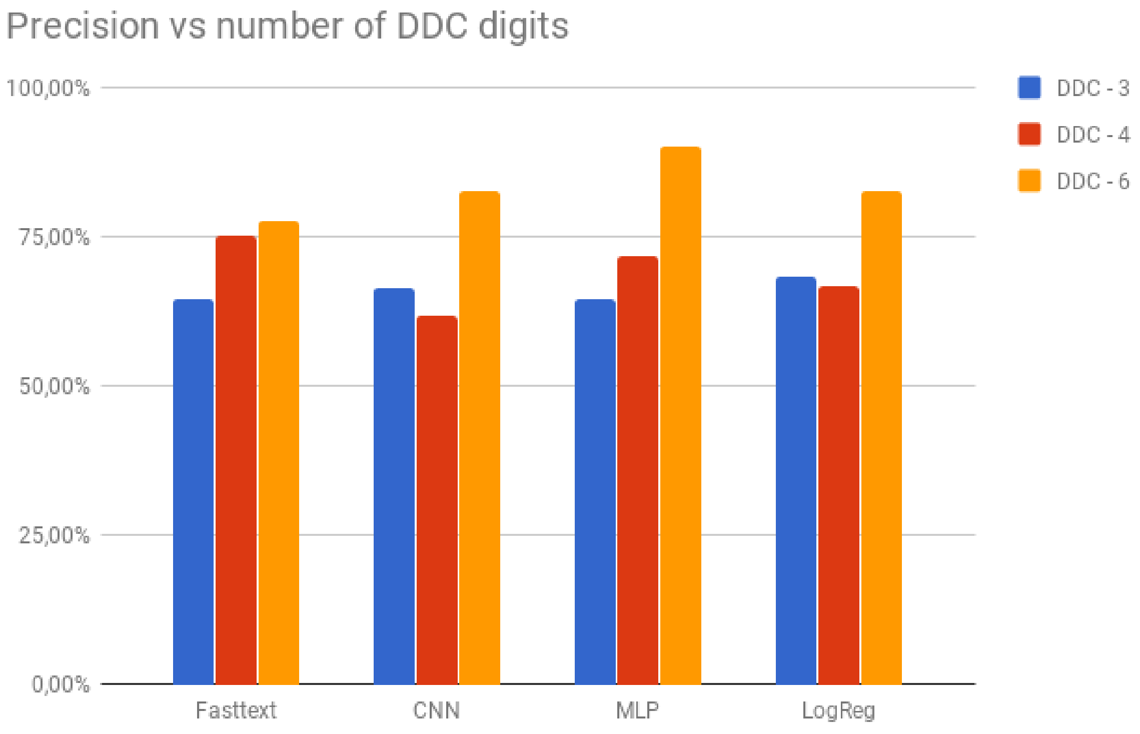

Figure 4

The experiments have been carried out for a limited set of classifications, and the results may change with a more complex environment.

9 CONCLUSIONS

The experiments verify that the size of the training set is a significant factor for improving the performance of a model. With a relatively limited size in the training set, a precision level better than 95% was achieved for all the tested models.

As the content in general becomes digital, it is realistic to establish training sets of relevant size and character. Combined with mature and available ML technology, this points in the direction of semi-automatic or fully automatic classification.

For situations where the available set of training data is relatively small, it is possible to improve the performance by the use of artificial documents. As the training set becomes larger, such improvements seems to have less significance.

10 ACKNOWLEDGMENTS

We would like to thank the group of skilled librarians at the National Library of Norway. They have contributed with their knowledge and expertise to the understanding and use of Dewey Decimal Classification, and to the relationship between

11 REFERENCES

NLN. 2007. NORART. [ONLINE] Available at: http://nabo.nb.no/trip?_b=baser&_q=100&_s=E&navn=norart&title=&fag=&CCL=&_BOOL=AND . [Accessed 19 July 2018].

NLN. 2018. NORMARC. [ONLINE] Available at: https://bibliotekutvikling.no/ressurser/kunnskapsorganisering/verktoykasse-forkunnskapsorganisering/marc-formater/normarc/. [Accessed 19 July 2018].

OCLC. 2018. Dewey Decimal Classification summaries. [ONLINE] Available at: https://www.oclc.org/en/dewey/features/summaries.html. [Accessed 19 July 2018].

Zhang, X, 2016. Character-level Convolutional Networks for Text Classification. Character-level Convolutional Networks for Text Classification, [Online]. arXiv:1509.01626v3 [cs.LG], 1-9. Available at: https://arxiv.org/abs/1509.01626v3 [Accessed 19 July 2018].