Using Titles vs. Full-text as Source for Automated Semantic Document Annotation

Lukas Galke
ZBW—Leibniz Information Centre for Economics, Kiel
l.galke@zbw.eu

Florian Mai
Kiel University
stu96542@mail.uni-kiel.de

Alan Schelten
Kiel University
stu111405@informatik.uni-kiel.de

Dennis Brunsch
Kiel University
deb@informatik.uni-kiel.de

Ansgar Scherp
ZBW—Leibniz Information Centre for Economics, Kiel
a.scherp@zbw.eu

ABSTRACT

We conduct the first systematic comparison of automated semantic annotation based on either the full-text or only on the title metadata of documents. Apart from the prominent text classification baselines kNN and SVM, we also compare recent techniques of Learning to Rank and neural networks and revisit the traditional methods logistic regression, Rocchio, and Naive Bayes. Across three of our four datasets, the performance of the classifications using only titles reaches over 90% of the quality compared to the performance when using the full-text.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → **Document analysis**;

KEYWORDS

Multi-label classification; document analysis; semantic annotation

ACM Reference Format:
Lukas Galke, Florian Mai, Alan Schelten, Dennis Brunsch, and Ansgar Scherp. 2017. Using Titles vs. Full-text as Source for Automated Semantic Document Annotation. In *Proceedings of K-CAP 2017: Knowledge Capture Conference, Austin, TX, USA, December 4–6, 2017 (K-CAP 2017),* 4 pages.
https://doi.org/10.1145/3148011.3148039

1 INTRODUCTION

In contrast to the full-text, the documents' metadata is directly available on the Linked Open Data cloud, accessible in RDF format, and can be processed with no legal barriers for semantic annotation. Conducting semantic annotations by using only the title is challenging, since the title is short and thus carries only little information compared to the full-text. The process of semantic annotation is a multi-label classification task, not only one but a set of concepts is needed to appropriately describe the semantics of the document. We run an extensive series of experiments to compare established methods and recent methods from machine learning for multi-label document classification. The goal is to decide whether it is possible to reach a comparable classification performance when using only the title of the documents. All the compared approaches operate on the underlying machine learning level which makes a comparison with prevalent end-to-end ontology tagging systems such as SOLR ontology tagger[1] and MAUI[2] difficult. We instead show that despite not using the hierarchical properties of the thesaurus, the presented methods outperform the best-performing methods that do make use of the hierarchy such as the ones of our own prior work [3]. Apart from the well-known multi-label classification baseline k-nearest neighbors (kNN) [11] and support vector machines (SVM), we revisit traditional text classification methods such as Naive Bayes, Rocchio, and logistic regression (LR). We also include the prominent Learning to Rank (L2R) approach [5], as well as a modern variant of neural networks motivated by the success of the Deep Learning field. In the past, algorithms of the lazy learner family such as kNN used to dominate multi-label classification tasks on such datasets with a large number of target classes [3, 8, 10]. However, we show that eager learners such as logistic regression and feed-forward neural networks outperform lazy learners. Lazy learners as well as Learning to Rank need to store and traverse $O(N_{\text{training examples}} \cdot N_{\text{features}})$ space to predict the labels for a single new document at test time. In contrast, most eager learners have the benefit of $O(N_{\text{parameters}})$ time complexity to predict a label set for an unseen document, which is important when applying an automated semantic annotation process for on-the-fly enrichment of metadata on the Linked Open Data cloud. Summarized, the contributions of this work are: (1) To the best of our knowledge, the first large-scale systematic comparison of multi-label classifiers applied to either the full-text or only the titles of documents. (2) Results that show that eager learners such as neural networks and linear models outperform lazy learners even when a high amount of possible labels is considered. (3) We offer evidence that using only the title for high-dimensional multi-label classification is a reasonable choice for semantic annotation of resources where only metadata is available. Our technical report [1] reveals more details on the compared approaches including the hyper-parameters as

K-CAP 2017, December 4–6, 2017, Austin, TX, USA

ACM ISBN 978-1-4503-5553-7/17/12…$15.00
https://doi.org/10.1145/3148011.3148039

[1] https://www.opensemanticsearch.org/solr-ontology-tagger
[2] https://github.com/zelandiya/maui-standalone

well as an overview of the field. We have published the full source code of our text processing pipeline on GitHub[3].

2 EXPERIMENTAL SETUP

Datasets. We have conducted our experiments on four datasets of English documents: two datasets are obtained from scientific digital libraries in the domains of economics and political sciences along with two news datasets from Reuters and New York Times. For each document in the datasets, there are manually created gold-standard annotations provided by domain experts, who work as professional subject indexers in the corresponding organizations. In addition, each dataset provides a domain-specific thesaurus that serves as controlled vocabulary of the gold-standard. Its concepts are used as target labels in our multi-label document classification task. The thesaurus also offers sets of concept-specific phrases (i. e., `skos:prefLabel` and `skos:altLabel` in case of SKOS format) that are used for concept extraction from the documents' full-text and titles [2]. The *economics* dataset consists of 62, 924 documents and is provided by ZBW — Leibniz Information Centre for Economics. The annotations are taken from the Standard Thesaurus Wirtschaft (STW) version 9[4], which is a controlled domain-specific thesaurus for economics and business studies maintained by ZBW. The thesaurus contains 6, 217 concepts with 12, 707 concept-specific phrases. From these concepts, 4, 682 are used in the corpus and thus considered in the multi-label classification task. Each document is annotated by domain experts with on average 5.26 labels (SD: 1.84). The *political sciences* dataset has 28, 324 documents. Similar to the economics dataset, we made a legal agreement for the political sciences dataset with the German Information Network for International Relations and Area Studies[5] that is providing the documents. The labels are taken from the thesaurus for International Relations and Area Studies[6], which contains 9, 255 concepts (and an equivalent number of concept-specific phrases, i. e., there are no alternative phrases). From these concepts, 7, 234 are used in the corpus. Each document in the dataset has on average 8.07 labels (SD: 3.03). The *Reuters RCV1-v2* dataset contains 805, 414 articles. We chose articles where both the titles and the full-text of the documents are available. From this set of documents, we randomly selected 100, 000 articles to match the scale of the scientific corpora. In our experiments, we employ the thesaurus re-engineered from the Reuters dataset by Lewis et al. [6]. The thesaurus contains 117 concepts and a total of 173 concept-specific phrases. From these concepts, 101 are used in the corpus. Each document was annotated with on average 3.21 (SD: 1.41) labels. The *New York Times Annotated Corpus Dataset* (NYT) contains 1, 846, 656 articles. Each article has two sets of annotations, created by a professional indexing service and annotations which were added by the authors using a semi-automatic system. We used the annotations provided by the indexing service because it is reasonable to expect that they are more consistent and of higher quality (cf. [4]). As for the Reuters dataset, we chose a random subset of 100, 000 documents containing both full-text and titles. The number of concepts in the NYT dataset is 25, 226. From these concepts, 6, 809 are used in our random sample.

[3]https://github.com/Quadflor/quadflor
[4]http://zbw.eu/stw/versions/9.0/about.en.html
[5]http://www.fiv-iblk.de/eindex.htm
[6]http://www.fiv-iblk.de/information/information_thesaurus.htm

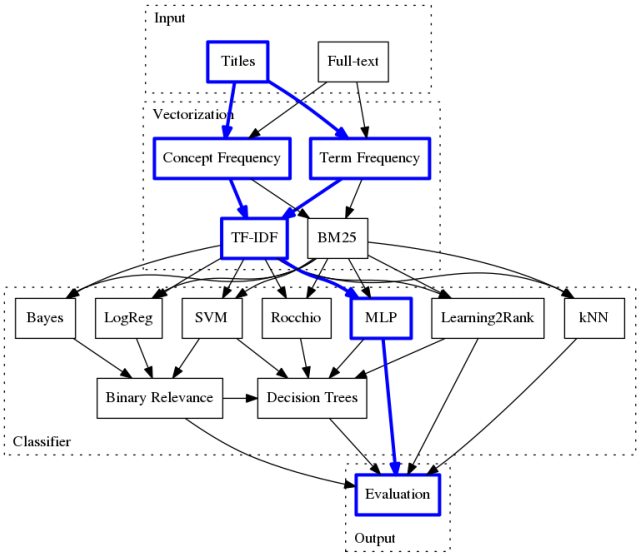

Figure 1: Configurable text-processing pipeline used for our experiments. The best-performing strategy is emphasized.

Each document is annotated with on average 2.53 (SD 1.78) labels. Like the political sciences dataset, each concept consists of only a single concept-specific phrase.

Vectorization methods. We compare the different vectorization methods of the input text as shown in Figure 1 (see our TR [1] for details). One vectorization is based on term frequencies (TF-IDF) and the other is based on concept frequencies (CF-IDF) (cf. [2]). We experiment with the re-weighting method BM25 using term frequencies and BM25C using concept frequencies. The concatenation of both terms and concepts is denoted by CTF-IDF and BM25CT, respectively. As classifier, we employ kNN with cosine distance. The performance of kNN relies on the assumption that documents are well represented by the features and that similar documents have similar labels. Therefore, its classification performance is a good indicator for the quality of the features.

Classification methods. After determining the best-performing vectorization method, we compare classifiers from the lazy learner family as well as eager learners. We leverage the generalized linear models (optimized by SGD [12]) SVMs and logistic regression to perform multi-label classification with binary relevance, i. e., training a separate classifier for each class. To adapt the Learning to Rank approach and the multi-layer perceptron to multi-labeling, we consider using thresholds as well as stacking with decision trees (cf. [4]). We also experiment with stacking the decision trees on top of binary-relevance logistic regression. Please note, we keep all hyperparameters fixed across all experiments and datasets.

Preprocessing. Prior to counting terms and extracting concepts, both the input text and the concept-specific phrases of the thesauri are subject to preprocessing steps. This includes discarding all characters except for sequences of alphabetic characters with a length of at least two. Words connected with a hyphen are joined

(i. e., the hyphen is removed). Detected words are lower-cased and lemmatized based on the morphological processing of WordNet [9].

Evaluation. For evaluation, we separate each dataset into 90% training documents and 10% test documents and perform a 10-fold cross-validation, such that each document occurs exactly once in the test set. Hence for each test document, we compare the predicted labels with the label set of the gold standard and evaluate the F_1 measure. The F_1 measure is the harmonic mean between precision, i. e. true positives w.r.t false positives, and recall, i. e. true positives w.r.t false negatives. When no label is predicted, the precision is set to zero. The F-scores are averaged over the test documents. We chose this sample-based F_1 measure over class-averaged or global variants because it is closest to an assumed application, where each individual document needs to be annotated as good as possible. Finally, we report the mean sample-based F-score over the ten folds of the cross-validation. Please note, there is a possibility that all documents annotated with a specific label fall into one test set. Despite no training data is available for these labels, we do *not* exclude those from our evaluation metric.

3 RESULTS

Results for Vectorization Methods. Table 1 shows the results for the text vectorization experiment. The term-based vectorization method TF-IDF performs consistently better than the purely concept-based vectorization CF-IDF methods on both the titles and the full-text. The differences in the F-scores ranges from 0.003 on Economics to 0.307 F-score on Reuters. When combining the term vector with the concept vector, the performance is at least as good as the other text vectorization methods and in many cases yields better results. This is more noticeable on titles than on full-texts. BM25 re-weighting does not improve the results compared to TF-IDF neither in case of the titles nor the full-text. Rather, we observe a decrease in performance by up to 0.13. These experiment using a nearest neighbor classifier indicates that CTF-IDF is the best vectorization method. Henceforth, we use CTF-IDF for comparing the performance of the classifiers.

Results for Classifiers. The results of comparing the different classifiers are documented in Table 2. The generalized linear models SVM and logistic regression are close to each other. The difference is no more than 0.04 for any dataset. Considering Learning to Rank, we observe that the technique yields consistently lower scores than the multi-layer perceptron. Overall, the eager learners SVM, LR, L2R and MLP outperform both Naive Bayes and the lazy learners Rocchio and kNN. Among all classifiers, MLP dominates on all datasets apart from NYT on titles, where LRDT achieves a .021 higher score. While the stacked decision tree module increases the F-scores of logistic regression on all datasets with fewer than 100 documents per label (all but Reuters), the impact of the stacking method is inconsistent for the Learning to Rank and MLP approaches. It is noteworthy that there are cases where a classifier performs better on the title data than the same classifier applied on the full-text data. These are Bernoulli Bayes on the Reuters dataset and RocchioDT on the economics dataset. As a general rule, however, full-texts generate higher scores than the titles. Comparing different classifiers across titles and full-text, we can make the observation that some

Table 1: Sample-averaged F-scores of the text vectorization methods with using kNN as common classifier

Input	Vectoriz.	Econ.	Polit.	RCV1	NYT
Full-text	TF-IDF	0.406	0.269	0.758	0.394
Full-text	BM25	0.370	0.230	0.740	0.370
Full-text	CF-IDF	0.402	0.266	0.451	0.367
Full-text	BM25C	0.296	0.161	0.423	0.236
Full-text	CTF-IDF	**0.411**	**0.272**	**0.761**	**0.406**
Full-text	BM25CT	0.377	0.231	0.742	0.379
Titles	TF-IDF	0.351	0.201	0.709	0.238
Titles	BM25	0.349	0.196	0.687	0.230
Titles	CF-IDF	0.303	0.183	0.275	0.105
Titles	BM25C	0.304	0.172	0.193	0.073
Titles	CTF-IDF	**0.368**	**0.212**	**0.717**	**0.242**
Titles	BM25CT	0.364	0.208	0.693	0.239

classifiers trained on titles outperform others that were trained on the full-text. Apart from the NYT corpus, the eager learners LR, LRDT, and MLP on titles are superior to kNN on full-texts. Finally, we compare the F-scores of the best-performing multi-layer perceptron on titles with its scores obtained on full-text. On the NYT dataset, 58% of the F-score is retained when using only titles. On the political sciences and economics datasets, the retained F-score is 83% and 91%, respectively. On the Reuters dataset, the MLP using solely titles retains 95% of the F-score that is obtained with full-text.

4 DISCUSSION AND CONCLUSION

The results show that multi-label classification of text documents can be reasonably conducted using only the titles of the documents. Over all datasets, the multi-layer perceptron on titles retains 82% of the F-score obtained on full-text. This gives an empirical justification for the value of automated semantic document annotation using metadata. From the first experiment, we find that combining words with extracted concepts as features is preferable over one of them alone. Concepts hold valuable domain-specific semantic information. The term frequency on the other hand, holds implicit information which is as well important for correct classification. Eager learners are, by design, capable of learning which terms or concepts need to be associated to the respective class. The results show that also lazy learners benefit from this joint representation. The second experiment shows that eager learners such as logistic regression and MLP consistently outperform lazy learners for multi-label classification. This result extends recent advancements in multi-labeling [7] towards document classification scenarios with many possible output labels and only few examples per class.

Inspecting the results for titles and full-text, the best-performing classifiers still perform better on the full-text. This is not surprising since the full-text holds considerably more information (including the title). However, for all datasets apart from the NYT dataset,

Table 2: Sample-averaged F-scores for classification methods with using the best vectorization method CTF-IDF

Input	Classifier	Econ.	Polit.	RCV1	NYT
Full-text	kNN (*baseline*)	0.411	0.272	0.761	0.406
Full-text	Bayes (Bernoulli)	0.318	0.191	0.657	0.281
Full-text	Bayes (Multinom.)	0.235	0.207	0.703	0.349
Full-text	SVM	0.481	0.319	0.852	0.554
Full-text	LR	0.485	0.322	0.851	0.556
Full-text	L2R	0.431	0.328	0.727	0.435
Full-text	MLP	**0.519**	**0.373**	**0.857**	0.569
Full-text	RocchioDT	0.291	0.225	0.645	0.393
Full-text	LRDT	0.498	0.339	0.843	0.562
Full-text	L2RDT	0.415	0.280	0.751	0.421
Full-text	MLPDT	0.492	0.340	**0.857**	**0.578**
Titles	kNN	0.368	0.212	0.717	0.242
Titles	Bayes (Bernoulli)	0.301	0.179	0.708	0.233
Titles	Bayes (Multinom.)	0.254	0.178	0.699	0.214
Titles	SVM	0.426	0.272	0.804	0.325
Titles	LR	0.429	0.274	0.803	0.326
Titles	L2R	0.419	0.296	0.699	0.296
Titles	MLP	**0.472**	**0.309**	**0.812**	0.332
Titles	RocchioDT	0.335	0.219	0.584	0.252
Titles	LRDT	0.451	0.279	0.796	**0.353**
Titles	L2RDT	0.428	0.261	0.730	0.25
Titles	MLPDT	0.457	0.277	0.808	0.340

the difference in F-score of the best-performing MLP is small. The difficulties in classifying the documents in the NYT dataset can be explained by a characteristic that the titles consist only of 4 words on average. There may be a lower bound on the title length to perform the classification task, since a short title limits the amount of available information and thus prohibits discrimination. From the other datasets, we can state that an average of 7 words per title leads to at least 80% retained F-score. Thus, it would require further investigation to understand the specific influence of the title length on the classification performance. The complexity of a multi-labeling problem depends on the number of available documents per label, independent of whether the full-text or the titles are used. Especially binary-relevance classifiers suffer from conservative label assignments (high precision, low recall), when many negative examples and only few positive examples are presented during training. While the results of the stacked decision tree module are inconsistent for MLP and L2R, it does alleviate the conservative assignments problem of binary-relevance, when only few documents per label are available.

In our experiments over four large-scale real-world corpora covering a broad range of domains (economics, political sciences and news), we did not limit the complexity by excluding rare labels and kept all independent variables as well as hyperparameters fixed. In our prior work [3], we have used the thesaurus hierarchy to model label dependencies which improves the classifications obtained by kNN. Despite not making use of the hierarchy anymore, we are able to achieve even higher absolute F-scores using eager learning techniques and supplying term features in addition to extracted concepts. We can therefore drop the constraint of a hierarchical organization among the labels. Due to this minimal amount of requirements and invariant configurations of the text processing pipeline, we can expect our findings to generalize to a wide range of other corpora.

To validate the practical impact of the experimental results, we have conducted a qualitative assessment of the experimental results in an expert workshop with three subject indexing specialists at ZBW, the national library for economics in Germany. The experts state that titles can be sufficient for classification of scientific documents. They further noted that titles contain less information than what an intellectual indexer has available when manually conducting the classification tasks for the documents. They also pointed out that researchers carefully chose their titles for findability. The experts argued that reasonably good automatic indexing based on titles is valuable since it does not raise legal problems compared to processing full-text as discussed in the introduction. We conclude that using the documents' title for automated semantic annotation is not only technically possible with a high quality but also valuable from a practical point of view.

Acknowledgements. This research was co-financed by the EU H2020 project MOVING (contract no 693092). We thank T. Rebholz, G. Schädle, and A. O. Kempf from ZBW for valuable discussions.

REFERENCES

[1] L. Galke, F. Mai, A. Schelten, D. Brunsch, and A. Scherp. 2017. Using Titles vs. Full-text as Source for Automated Semantic Document Annotation. *ArXiv e-prints* (May 2017). arXiv:cs.DL/1705.05311
[2] Frank Goossen, Wouter IJntema, Flavius Frasincar, Frederik Hogenboom, and Uzay Kaymak. 2011. News personalization using the CF-IDF semantic recommender. In *Web Intelligence, Mining and Semantics*. ACM.
[3] Gregor Große-Bölting, Chifumi Nishioka, and Ansgar Scherp. 2015. A comparison of different strategies for automated semantic document annotation. In *Knowledge Capture*. ACM.
[4] Andreas Heß, Philipp Dopichaj, and Christian Maaß. 2008. Multi-value classification of very short texts. In *Advances in Artificial Intelligence*. Springer.
[5] Minlie Huang, Aurélie Névéol, and Zhiyong Lu. 2011. Recommending MeSH terms for annotating biomedical articles. *Am. Medical Informatics Association* 18, 5 (2011).
[6] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Machine Learning Research* 5 (2004).
[7] Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale Multi-label Text Classification. Revisiting Neural Networks. In *Machine Learning and Knowledge Discovery in Databases*. Springer.
[8] Eleftherios Spyromitros, Grigorios Tsoumakas, and Ioannis Vlahavas. 2008. An empirical study of lazy multilabel classification algorithms. In *Artificial Intelligence*. Springer.
[9] Princeton University. 2010. About WordNet. wordnet.princeton.edu. (2010).
[10] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* 26, 8 (2014), 1819–1837.
[11] Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition* 40, 7 (2007).
[12] Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*. ACM.