The Journal of Open Source Software

AuDoLab: Automatic document labelling and classification for extremely unbalanced data

Arne Tillmann[1], **Anton Thielmann**[1], **Gillian Kant**[1], **Christoph Weisser**[1, 2], **Benjamin Säfken**[1, 2], **Alexander Silbersdorff**[1, 2], **and Thomas Kneib**[1, 2]

1 Georg-August-Universität Göttingen, Göttingen, Germany **2** Campus-Institut Data Science (CIDAS), Göttingen, Germany

DOI: 10.21105/joss.03719

Software

- Review
- Repository
- Archive

Editor: Arfon Smith

Reviewers:

- @linuxscout
- @pps121

Submitted: 27 August 2021
Published: 19 October 2021

Summary

AuDoLab provides a novel approach to one-class document classification for heavily imbalanced datasets, even if labelled training data is not available. Our package enables the user to create specific out-of-domain training data to classify a heavily underrepresented target class in a document dataset using a recently developed integration of Web Scraping, Latent Dirichlet Allocation Topic Modelling and One-class Support Vector Machines (Thielmann, Weisser, Krenz, & Säfken, 2021). AuDoLab can achieve high quality results even on higly specific classification problems without the need to invest in the time and cost intensive labelling of training documents by humans. Hence, AuDoLab has a broad range of scientific research or business real world applications. In the following, a few potential use cases will be briefly discussed that should illustrate the broad range of applications in various domains. For example AuDoLab could be used to identify emails with very specific topics such as fraud or money laundering that might have an extremely low prevalence. Similarly, AuDoLab could be used in the medical field to classify medical documents that are concerned with very specific topics such as heart attacks or dental problems. Furthermore, AuDoLab may be used to identify legal documents with very specific topics such as machine learning. Note that, the only limiting factor to the broad range of use cases, is the availability of out-of-domain training data, that can be generated via Web Scraping from IEEEXplore (IEEE Xplore, 2020), arxiv or pubmed. Given that a broad range of training documents can be obtained from these websites AuDoLab has a correspondingly broad range of applications. The following section provides an overview of AuDoLab. AuDoLab can be installed conveniently via pip. A detailed description of the package and installation and can be found in the packages repository or on the documentation website.[1]

Statement of need

Unsupervised document classification is mainly performed to gain insight into the underlying topics of large text corpora. In this process, documents covering highly underrepresented topics have only a minor impact on the algorithm's topic definitions. As a result, underrepresented topics can sometimes be "overlooked" and documents are assigned topic prevalences that do not reflect the underlying content. Thus, labeling underrepresented topics in large text corpora is often done manually and can therefore be very labour-intensive and time-consuming. AuDoLab enables the user to tackle this problem and perform unsupervised one-class document classification for heavily underrepresented document classes. This leverages the results of

[1] https://AuDoLab.readthedocs.io

one-class document classification using One-class Support Vector Machines (SVM) (Manevitz & Yousef, 2001; Schölkopf et al., 2001) and extends them to the use case of severely imbalanced datasets. This adaptation and extension is achieved by implementing a multi-level classification rule as visualised in the graph below.

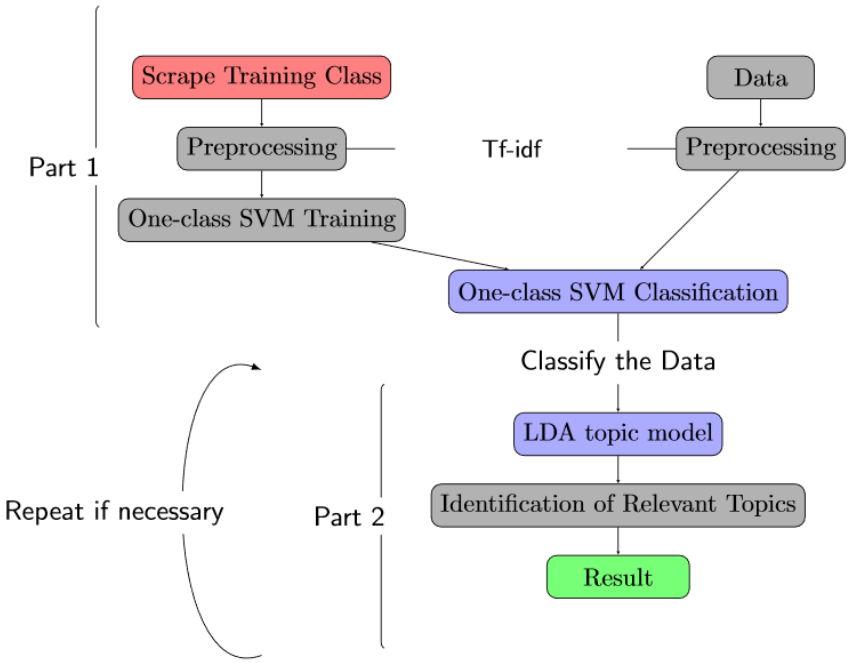

Figure 1: Classification Procedure.

The first part of the package allows the user to scrape training documents (scientific papers), ideally covering the target topic in which the user is interested, from IEEEXplore (IEEE Xplore, 2020), arxiv or pubmed. The user can search for multiple search terms and specify an individual search query and, in the case of IEEEXplore, scrape additional information as e.g. the author names or the number of citations. Thus, an individually labelled (e.g. via author-keywords) training data set is created. Through the integration of pre-labelled out-of-domain training data, the problem of the heavily underrepresented target class can be circumvented, as large enough training corpora can be automatically generated. Subsequently, the text data is preprocessed for the classification part. The text preprocessing includes common Natural Language Processing (NLP) text preprocessing techniques such as stopword removal and lemmatization. As document representations the term frequency-inverse document frequency (tf-idf) representations are chosen. The tf-idf scores are computed on a joint corpus from the web-scraped out-of-domain training data and the target text data.

The second and main part of the classification rule lies in the training of the one-class SVM (Schölkopf et al., 2001). As a training corpus, only the out-of-domain training data is used. By adjusting hyperparameters, the user can create a strict or relaxed classification rule, that reflects the users belief about the prevalence of the target class inside the target data set and the quality of the scraped out-of-domain training data. The last part of the classification rule enables the user to control the classifiers results with the help of Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) (and e.g. wordclouds). Additionally, the user can generate interactive plots depicicting the identified topics during the LDA topic modelling (Sievert & Shirley, 2014).

The second step can be reiteraded, depending on the users perceived quality of the classifica-

tion results.

Comparison with existing tools

At the moment no Python Package with a comparable functionality of AuDoLab is available, since AuDoLab is based on a novel and recently published classification prodcedure (Thielmann, Weisser, Krenz, & Säfken, 2021). Thereby, AuDoLab uses and integrates in particular a combination of Web Scraping, Topic Modelling and One-class Classifcation for which various individual packages are available. Details on the statistical methodology can be found in (Thielmann, Weisser, Krenz, & Säfken, 2021). An application of the methodology on a data set of patent data can found in (Thielmann, Weisser, & Krenz, 2021). For Topic Modelling available packages are the LDA algorithm as implemented in the package Gensim (Řehůřek & Sojka, 2010) or the package TTLocVis (Kant et al., 2020) for short and sparse text. Visual representations of the topics can be implemented with LDAvis (Sievert & Shirley, 2014). The One-class SVM classification package is availabe in Scikit-learn (Pedregosa et al., 2011). Alternative Further research could explore Deep Learning Algorithms as well (Säfken et al., 2020, 2021).

Acknowledgements

We thank the Campus-Institut Data Science (CIDAS), Göttingen, Germany for funding this project.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

IEEE Xplore. (2020). *IEEE Xplore Digital Library*. https://ieeexplore.ieee.org/Xplore/home.jsp

Kant, G., Weisser, C., & Säfken, B. (2020). TTLocVis: A twitter topic location visualization package. *Journal of Open Source Software*, *5*(54), 1–6. https://doi.org/10.21105/joss.02507

Manevitz, L. M., & Yousef, M. (2001). One-class SVMs for document classification. *Journal of Machine Learning Research*, *2*(Dec), 139–154. https://doi.org/10.1162/15324430260185574

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.5555/1953048.2078195

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. https://doi.org/10.13140/2.1.2393.1847

Säfken, B., Silbersdorff, A., & Weisser, C. (Eds.). (2020). *Learning deep: Perspectives on deep learning algorithms and artificial intelligence*. Universitätsverlag Göttingen. https://doi.org/10.17875/gup2020-1338

Säfken, B., Silbersdorff, A., & Weisser, C. (Eds.). (2021). *Learning Deep Textwork: Perspectives on Natural Language Processing and Artificial Intelligence*. Universitätsverlag Göttingen. https://doi.org/10.17875/gup2021-1608

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443–1471. https://doi.org/10.1162/089976601750264965

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.3115/v1/W14-3110

Thielmann, A., Weisser, C., & Krenz, A. (2021). One-class support vector machine and LDA topic model integration—evidence for AI patents. In N. H. Phuong & V. Kreinovich (Eds.), *Soft computing: Biomedical and related applications* (pp. 263–272). Springer International Publishing. https://doi.org/10.1007/978-3-030-76620-7_23

Thielmann, A., Weisser, C., Krenz, A., & Säfken, B. (2021). Unsupervised document classification integrating web scraping, one-class SVM and LDA topic modelling. *Journal of Applied Statistics*, *0*(0), 1–18. https://doi.org/10.1080/02664763.2021.1919063