Few Shot Classification for Labeling of Medieval and Early Modern Charter Texts

Kovács, Tamás
tamas.kovacs@uni-graz.at
Universität Graz, Austria

Aoun, Sandy
sandy.aoun@uni-graz.at
Universität Graz, Austria

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Austria

Nicolaou, Anguelos
anguelos.nicolaou@gmail.com
Universität Graz, Austria

Luger, Daniel
daniel.luger@uni-graz.at
Universität Graz, Austria

Atzenhofer-Baumgartner, Florian
florian.atzenhofer-baumgartner@uni-graz.at
Universität Graz, Austria

Lamminger, Florian
florian.lamminger@uni-graz.at
Universität Graz, Austria

Decker, Franziska
franziska.decker@uni-graz.at
Universität Graz, Austria

The Monasterium.net [1] comprises more than 600,000 charters from the Middle Ages and Early Modern Era, originating from diverse European regions and languages. These records provide comprehensive insights into various facets of life during historical periods that pertain to particular occurrences of legal proceedings. The approach we are taking to facilitate the filtration of descriptive texts and transcriptions in Monasterium.net involves the allocation of semantic categories to the legal acts documented therein, including but not limited to confiscation, donation, and property sale. The aim of the model is to exhibit efficient and effective generalization to novel, unobserved instances within identical categories. The unsupervised learning algorithm is utilized for the task of classifying data without any labeled training samples and an indeterminate number of categories. The main objective of the model is to dynamically allocate instances to an indeterminate set of potential categories. This model must leverage the intrinsic information of the labels.

Text classification is a widely used technique with numerous applications, such as document categorization, sentiment analysis, and others (Joachims 1998; Sebastiani 2002; Yang / Pedersen 1997) . Numerous methodologies are available for addressing this fundamental issue; however, a significant proportion of them necessitate substantial quantities of annotated data to be effective (Kowsari et al. 2019). Gathering annotations for a particular use case is frequently one of the most costly components of any machine learning undertaking. The need for methods that optimize limited data sets is becoming more prevalent.

Over the past decade, the field of Computer Vision has adopted a popular technique for zero-shot learning: utilizing featurizers to integrate images and all possible class labels into their associated latent representations (Lampert / Nickisch / Harmeling 2009; Socher et al. 2013). This approach allows for the development of a linear projection that aligns image and label embeddings using a training set based on a sample of given labels. Consequently, any label (seen or unseen) and any image can be embedded within the same latent space, enabling the measurement of the distance between them. Although this method has proven effective in Computer Vision, it is not directly applicable to Natural Language Processing.

Natural Language Processing (NLP) relies on advanced embedding techniques that encode both data and class names within a single space using a unified model, thereby eliminating the need for data-intensive alignment phases. Pooled word vectors, for example, have been extensively utilized for years (Veeranna et al. 2016) . Recent advancements in sentence embedding models, such as various Sentence BERT models, have facilitated the generation of sequence and label embeddings by fine-tuning pooled multilingual sequence representations for greater conceptual complexity (Reimers / Gurevych 2019; Lavi / Medentsiy / Graus 2021). However, these models were not developed to understand single-word representations like label names, and our label embeddings would not be discernible in other word-level embedding techniques such as GloVE or Word2vec (Mikolov et al. 2013; Pennington / Socher / Manning 2014). This highlights the distinctive challenges and techniques required in NLP compared to Computer Vision when dealing with embeddings and the alignment of data and class labels.

Labeled examples are necessary to identify an intersection between the sequence model and word2vec label representations. A small subset of Monasterium offers a comprehensive representation of legal acts through historical abstracts. Our methodology employs these abstracts to construct a word2vec model, using the most frequent words (excluding stop-words) from the word2vec dictionary of the abstracts as labels. Subsequently, the algorithm generates an intermediate least-squares projection matrix for given label embeddings, based on their corresponding data embeddings. Through the application of a variation of L2 regularization, the weights are shifted toward the identity matrix, rather than reducing their norm. The projection matrix ultimately provides dimensionality reduction by implementing an additional transformation to the pre-trained Sentence BERT embeddings of both sequences and labels, yielding appropriate text classification. To further augment classification performance, especially in cases of limited labeled data, we integrate the concept of prototypical few-shot classification. This approach empowers our model to effectively generalize to new classes, even when presented with a limited number of labeled examples. By amalgamating projection matrix-based dimensionality reduction and pre-trained Sentence

BERT embeddings with prototypical few-shot classification, we can establish a more robust and effective text classification model capable of handling limited labeled data and performing optimally across various legal acts represented in the Monasterium dataset.

Notes

1. https://www.monasterium.net

Bibliography

Joachims, Thorsten (1998): *Text categorization with Support Vector Machines: Learning with many relevant features*. in: Nédellec, Claire / Rouveirol, Céline (eds.): *Machine Learning: ECML-98*. Berlin, Heidelberg: Springer. 137–142.

Kowsari, Kamran / Jafari Meimandi, Kiana / Heidarysafa, Mojtaba / Mendu, Sanjana / Barnes, Laura / Brown, Donald (2019): "Text Classification Algorithms: A Survey", in: *Information* 10 (4): 150. 10.3390/info10040150.

Lampert, Christoph H. / Nickisch, Hannes / Harmeling, Stefan (2009): *Learning to detect unseen object classes by between-class attribute transfer*. in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 951–958.

Lavi, Dor / Medentsiy, Volodymyr / Graus, David (2021): *conSultantBERT: Fine-tuned Siamese Sentence-BERT for Matching Jobs and Job Seekers*. arXiv. http://arxiv.org/abs/2109.06501.

Mikolov, Tomas / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): *Efficient Estimation of Word Representations in Vector Space*. arXiv. http://arxiv.org/abs/1301.3781.

Pennington, Jeffrey / Socher, Richard / Manning, Christopher (2014): *Glove: Global Vectors for Word Representation*. in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics. 1532–1543. http://aclweb.org/anthology/D14-1162.

Reimers, Nils / Gurevych, Iryna (2019): *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv. http://arxiv.org/abs/1908.10084.

Sebastiani, Fabrizio (2002): "Machine Learning in Automated Text Categorization", in: *ACM Computing Surveys* 34 (1): 1–47. 10.1145/505282.505283.

Socher, Richard / Ganjoo, Milind / Manning, Christopher D / Ng, Andrew (2013): *Zero-shot learning through cross-modal transfer*. in: Burges, C.J. / Bottou, L. / Welling, M. / Ghahramani, Z. / Weinberger, K.Q. (eds.): *Advances in neural information processing systems*. Curran Associates, Inc.

Veeranna, Sappadla Prateek / Nam, Jinseok / Menc#a, Eneldo Loza / Furnkranz, Johannes (2016): "Using Semantic Similarity for Multi-Label Zero-Shot Classification of Text Documents", in: *Computational Intelligence*:

Yang, Yiming / Pedersen, Jan O. (1997): *A Comparative Study on Feature Selection in Text Categorization*. in: *Proceedings of the Fourteenth International Conference on Machine Learning* (= ICML '97). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 412–420.