Fine-tuning the SwissBERT Encoder Model for Embedding Sentences and Documents

Juri Grosjean and **Jannis Vamvas**

Department of Computational Linguistics, University of Zurich

`jurileander.grosjean@uzh.ch, vamvas@cl.uzh.ch`

Abstract

Encoder models trained for the embedding of sentences or short documents have proven useful for tasks such as semantic search and topic modeling. In this paper, we present a version of the SwissBERT encoder model that we specifically fine-tuned for this purpose. SwissBERT contains language adapters for the four national languages of Switzerland – German, French, Italian, and Romansh – and has been pre-trained on a large number of news articles in those languages. Using contrastive learning based on a subset of these articles, we trained a fine-tuned version, which we call SentenceSwissBERT. Multilingual experiments on document retrieval and text classification in a Switzerland-specific setting show that SentenceSwissBERT surpasses the accuracy of the original SwissBERT model and of a comparable baseline. The model is openly available for research use.[1]

1 Introduction

Sentence embeddings have become a valuable tool in natural language processing. Neural models are fed with sequence strings and convert them into embeddings, i.e. a numeric representation of the input text. These can be applied in a variety of contexts, e.g. information retrieval, semantic similarity, text classification and topic modeling.

SwissBERT (Vamvas et al., 2023) is a modular encoder model based on X-MOD (Pfeiffer et al., 2022), which was specifically designed for multilingual representation learning. SwissBERT has been trained via masked language modeling on more than 21 million Swiss news articles in Swiss Standard German, French, Italian, and Romansh Grischun. The model is designed for processing Switzerland-related text, e.g. for named entity recognition, part-of-speech tagging, text categorization, or word embeddings.

The aim of this work is to fine-tune the existing SwissBERT model for the embedding of sentences and short documents. Specifically, our hypothesis is that using a contrastive learning technique such as SimCSE (Gao et al., 2021) to fine-tune SwissBERT will yield a model that outperforms the base model as well as generic multilingual sentence encoders in the context of processing news articles from Switzerland.

This is evaluated on two natural language processing tasks that utilize sentence embeddings, namely document retrieval and nearest-neighbor text classification, both from a monolingual and cross-lingual perspective. Indeed, the experiments show that the fine-tuned SwissBERT, which we call SentenceSwissBERT, has a higher accuracy than baseline models. An especially strong effect was observed for the Romansh language, with an absolute improvement in accuracy of up to 55 percentage points over the original SwissBERT model, and up to 29 percentage points over the best SentenceBERT baseline.

2 Related Work

Sentence-BERT This approach introduced by Reimers and Gurevych (2019) enhances BERT and RoBERTa for generating fixed-size sentence embeddings. It investigated using the CLS-token, the mean of all output vectors (MEAN-strategy), or the max-over-time of output vectors (MAX-strategy) as sentence embeddings and found the MEAN-strategy to perform best. The method applies siamese and triplet network architectures to finetune pre-trained models, which enables them to learn high-quality sentence embeddings, e.g. for comparison via cosine similarity. The training approach entails three objective functions: classification, regression, and triplet, each with specific training structures. Data from SNLI (Bowman et al., 2015) and MultiNLI datasets (Williams et al., 2018) was used for training. Sentence-BERT has given
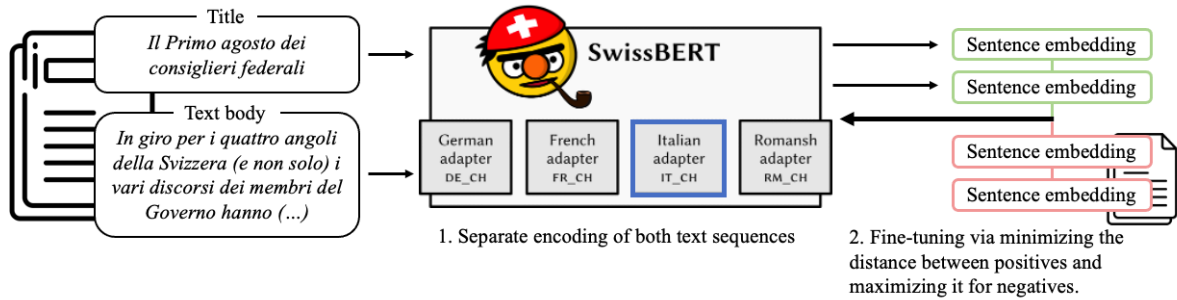
---

[1] https://huggingface.co/jgrosjean-mathesis/sentence-swissbert

Figure 1: Visualisation of the supervised SimCSE training approach.

rise to a family of popular open-source encoder models.[2]

Multilingual Sentence Embeddings   There are multiple approaches for training BERT-based encoder models for cross-lingual transfer. Reimers and Gurevych (2020) propose utilizing knowledge distillation to enhance mono-lingual models for multilingual use. Feng et al. (2022) found that harnessing pre-trained language models and fine-tuning them for cross-lingual tasks yields promising results while requiring less training data than training encoder models from scratch via multilingual language data like translations.

Contrastive Learning   This technique originally surged in training neural models to perform vision tasks, e.g. image recognition. However, it has also been shown to deliver promising results with NLP tasks. The goal is for the model to learn an embedding space in which similar data is mapped closely to each other and unalike data stays far apart. For a mini-batch of N sentences, where (h_i, h_i^+) represent a pair of semantically-related sequences, h_j a random in-batch negative, and τ the temperature hyperparameter, the training objective looks as follows:

$$- \log \frac{e^{\text{cos\_sim}(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{cos\_sim}(h_i, h_j^+)/\tau}} \qquad (1)$$

Introduced by Gao et al. (2021), the SimCSE (simple contrastive sentence embedding) framework has been found highly effective when used in conjunction with pre-trained language models. This technique can be applied using an unsupervised or a supervised training.

For the unsupervised approach, the sequences in the training data are matched with themselves to create positive matches, i.e. the cosine similarity

between both outputs (MEAN pooling or CLS) is maximized. Thanks to the dropout masks, the embeddings of identical sequences still differ slightly.

The supervised approach uses a dataset of sentence pairs with similar meanings, and an optional third entry that is contradictory in meaning to the other two (hard negative). The similarity computation is maximized for the similar sentence pairs and minimized between the positives and the negatives.

3   Fine-tuning

To fine-tune SwissBERT for sentence embeddings, we opted for a (weakly) supervised SimCSE approach without hard negatives. Analogous to the original SwissBERT, Swiss news articles serve as the training data for this. The documents are split into sequence pairs, where one sequence consists of the article's title and – if available – its lead concatenated, while the other contains the text body (see Figure 1). The title-body pairs represent (h_i, h_i^+) in the constrastive loss training objective 1.

3.1   Dataset

The fine-tuning data consists of over 1.5 million Swiss news articles obtained through the Swissdox@LiRI database[3] in German, French, Italian, and Romansh (see Table 1). All German and French articles selected from the corpus have been published between 2020 and 2023, while the Italian and Romansh media date back to 2000, because the database contains fewer articles in these languages. The news articles are pre-processed analogous to SwissBERT's original training data (Vamvas et al., 2023).

---

[2]https://www.sbert.net/

[3]https://swissdox.linguistik.uzh.ch/

Language	Documents	Tokens
German	760 350	621 107 750
French	644 416	567 688 406
Italian	63 666	35 109 282
Romansh	39 732	16 376 397
Total	**1 508 414**	**1 240 281 835**

Table 1: Composition of the dataset used for fine-tuning SwissBERT. We report the number of documents and tokens in the four languages.

3.2 Hyperparameters

The structure of the SimCSE train script provided by Gao et al. (2021)[4] was updated and adapted according to SwissBERT, i.e. adding the X-MOD model architecture configuration as well as a language switch component, so that the model would continuously adjust its adapter according to the training data language during the training process. During fine-tuning on SimCSE, we froze the language adapters and updated all the other parameters. The training data was padded / truncated to 512 tokens, so that it fits the input limit. The model was fine-tuned in one single epoch, using a learning rate of 1e-5 and the AdamW optimizer (Loshchilov and Hutter, 2019), a batch size of 512 and a temperature of 0.05, which has been recommended for SimCSE (Gao et al., 2021). We used MEAN pooling, following the findings by Reimers and Gurevych (2019).

4 Evaluation

We evaluate SentenceSwissBERT on two custom, Switzerland-related NLP tasks in German, French, Italian, and Romansh. It is measured against the original SwissBERT and a multilingual Sentence-BERT model that showed the strongest performance in the given evaluation tasks.

4.1 Dataset

For evaluation, we make use of the *20 Minuten* dataset (Kew et al., 2023), based on *20 Minuten*, one of the most widely circulated German-language newspapers in Switzerland. The articles tend to be relatively short and cover a variety of topics. Most of the documents in the dataset include a short article summary and topic tags

Given its format and features, the *20 Minuten* dataset is especially suitable for assessing SentenceSwissBERT's performance. For the evalu-

[4] https://github.com/princeton-nlp/SimCSE

Task	Language	Documents
Document retrieval	German	499
	French	499
	Italian	499
	Romansh	499
Text classification: train set	German	4 986
Text classification: test set	German	1 240
	French	1 240
	Italian	1 240
	Romansh	1 240

Table 2: Composition of the documents sourced from the *20 Minuten* dataset (Kew et al., 2023) that were employed for both evaluation tasks.

Category	Train articles	Test articles
accident	244	60
corona	1 468	367
economy	768	192
film	247	61
football	627	156
germany	250	62
social media	288	71
switzerland	300	743
ucraine war	268	66
usa	526	131
Total	**4 986**	**1 240**

Table 3: Composition of the test set of the text classification task, including the respective counts per category.

ation, all articles present in the *20 Minuten* corpus were removed from the original fine-tuning data in all languages, so that there is no overlap.

In order to expand the evaluation to French, Italian, and Romansh, the relevant parts of the articles were machine-translated via Google Cloud API (FR, IT) and Textshuttle API (RM). Using machine translation allows for a controlled comparison across languages when evaluating, since all documents share the same structure and content. Moreover, manual annotations can be automatically projected to the other languages without a need for additional annotation. A potential downside of machine translation is that the distribution of the test data does not reflect the diversity of human-written text. Tables 2 and 3 report statistics of the data we use for evaluation.

4.2 Tasks

Document retrieval For this task, the embedding of each article's summary is compared to all the articles' content embeddings and then matched by choosing the pair with the highest cosine similarity score. The performance is reported via the accuracy score, which is based on how many summaries were matched with the correct content in relation to the total number of articles processed. There is no train-test split performed for this task. It is performed monolingually (where the summary is written in the same language as the article) and cross-lingually.

Text Classification Ten categories are manually mapped from certain topic tags in the dataset. All documents without these (or overlapping) chosen topic tags are disregarded. Then, a random train-test split with a 80/20 ratio is performed once on the remaining data for every category respectively. The exact number of files per category are displayed in Table 3. Next, the text classification is carried out utilizing a nearest neighbors approach: The text body of each test article is compared to every embedding from the training data via cosine similarity. Subsequently, the topic tag of its one nearest neighbor from the training set (highest similarity) is assigned to it.

To assess cross-lingual transfer, the training data is kept in German for the assessment of each of the four languages, while the test data is machine-translated to French, Italian and Romansh. As the categories vary in frequency, the weighted average of all categories' F1-scores is reported.

4.3 Baseline Models

SwissBERT While not specifically trained for this, sentence embeddings can already be extracted from the last hidden layer of the original Swiss-BERT encoder model via MEAN pooling. The input language is specified, just like in its newly fine-tuned version. This comparison demonstrates whether there is value in fine-tuning the model specifically for sentence embeddings.

Sentence-BERT Reimers and Gurevych (2019) propose several multilingual sentence embedding models.[5] In this work, the *distiluse-base-multilingual-cased-v1* model is opted for as a baseline, as it shows the strongest performance for the

given evaluation tasks (see Appendix B). It has originally been trained following the multilingual knowledge distillation approach introduced in Section 2, using mUSE (Chidambaram et al., 2019) as teacher model and a version of the multilingual Universal Sentence Encoder (Yang et al., 2020) as the student model. This version of Sentence-BERT supports various languages, among them French, German, and Italian, but not Romansh. Unlike with SwissBERT, the input language does not need to be specified. This model has a similar number of parameters as SwissBERT (see Table 4). However, it maps to a 512-dimensional embedding space and, hence, is computationally more efficient than SwissBERT.

The other multilingual Sentence-Transformer *(paraphrase-multilingual-mpnet-base-v2)* tested is much larger (278 043 648 parameters). Although this model maps to a 768-dimensional space, analogous to SwissBERT, it performed worse in the evaluation tasks than *distiluse-base-multilingual-cased-v1* (see Appendix B). Thus, it was disregarded.

Model	Vocabulary	Parameters
Sentence-BERT	119 547	135 127 808
SwissBERT	50 262	160 101 888

Table 4: Vocabulary sizes and parameter counts of the two baseline models. The fine-tuned SentenceSwissBERT has the same size as the original model.

5 Results

Document Retrieval Results for this evaluation task are reported in Table 5. SentenceSwissBERT outperforms its base model SwissBERT, demonstrating a clear improvement compared to the original model. The largest difference is noticeable in the processing of Romansh text.

SentenceSwissBERT also obtains better results than the Sentence-BERT baseline *distiluse-base-multilingual-cased*, except for two cases. Both models achieve high accuracy in both the monolingual and cross-lingual tasks. The clearest difference can be seen for German and especially Romansh, which Sentence-BERT was not trained on.

Text classification Table 6 presents the results of this evaluation task. Again, SentenceSwissBERT tends to improve over the baselines, with the excep-

Encoder Model	Summary Language	Article Language			
		DE	FR	IT	RM
SwissBERT	DE	87.20	78.36	72.95	40.68
(Vamvas et al., 2023)	FR	86.52	84.97	78.96	40.84
	IT	83.17	80.17	84.17	33.41
	RM	46.08	39.10	43.39	83.17
Sentence-BERT	DE	91.80	90.98	**90.38**	62.53
(Reimers and Gurevych, 2019)	FR	90.78	93.19	90.78	63.36
	IT	88.12	**91.29**	91.58	65.71
	RM	70.59	73.48	73.55	73.35
SentenceSwissBERT	DE	**93.40**	**92.79**	90.18	**91.58**
	FR	**94.33**	**93.99**	90.98	**90.07**
	IT	**92.08**	90.85	**92.18**	**88.50**
	RM	**92.16**	**89.44**	**88.43**	**91.58**

Table 5: Results for the document retrieval task using the *20 Minuten* dataset (Kew et al., 2023). The accuracy score is reported. The best results per language pair are marked in bold print.

Encoder Model	Training Language	Test Language			
		DE	FR	IT	RM
SwissBERT (Vamvas et al., 2023)	DE	77.93	69.62	67.09	43.79
Sentence-BERT (Reimers and Gurevych, 2019)	DE	77.23	76.83	**76.90**	65.35
SentenceSwissBERT	DE	**78.49**	**77.18**	76.65	**77.20**

Table 6: Results for the nearest-neighbor classification task using the *20 Minuten* dataset (Kew et al., 2023). A weighted F1-score is reported and the best results are marked in bold print.

tion of Italian, where the Sentence-BERT model is slightly more accurate.

6 Discussion and Conclusion

The results confirm that contrastive learning with title–body pairs is an effective fine-tuning approach for a masked language model. Using just a subset of 1.5 million articles from the original pre-training dataset, a clear improvement on the two sentence-level tasks has been achieved.

On the one hand, we observed an effect in monolingual tasks, e.g., by matching French summaries with French articles, or by performing nearest-neighbor topic classification of German articles using German examples. On the other hand, we also evaluated cross-lingual variations of those tasks, and found a clear benefit in the cross-lingual setting as well, even though we did not use cross-lingual examples in our fine-tuning. This suggests that modular deep learning with language adapters can be combined effectively with contrastive learning.

We expect that SentenceSwissBERT will be a useful model variant for other Switzerland-related tasks that require sentence or document embeddings. For example, SentenceSwissBERT might be used for semantic search, or topic modeling based on document embeddings (e.g. BERTopic; Grootendorst, 2022). Future work could also explore whether including training data from other domains than news articles could further improve the generality of the model.

Limitations

The SentenceSwissBERT model has been trained on news articles only. Hence, it might not perform as well on other text domains. Additionally, the model input during training was limited to a maximum of 512 tokens. Thus, it may not be useful for processing longer texts. Finally, we note that we used machine-translated test data for evaluation in languages other than German.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maarten R. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Tannon Kew, Marek Kostrzewa, and Sarah Ebling. 2023. 20 minuten: A multi-task news summarisation dataset for German. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 1–13, Neuchatel, Switzerland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The multilingual language model for Switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69, Neuchatel, Switzerland. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

A   Pre-training dataset media composition

Medium	Articles	Language
lematin.ch	99 939	FR
24heures.ch	73 385	FR
tdg.ch	69 498	FR
Le Temps	63 130	FR
24 heures	62 004	FR
Tribune de Genève	57 604	FR
blick.ch	51 556	DE
rsi.ch	51 526	IT
letemps.ch	48 353	FR
rts.ch	47 397	FR
cash.ch	46 750	DE
blick.ch	43 178	FR
rtr.ch	39 732	RM
srf.ch	29 536	DE
nzz.ch	28 091	DE
tagblatt.ch	27 279	DE
luzernerzeitung.ch	23 855	DE
Aargauer Zeitung / MLZ	21 868	DE
Neue Zürcher Zeitung	18 408	DE
Le Matin Dimanche	18 352	FR
Thurgauer Zeitung	17 335	DE
Blick	14 636	DE
landbote.ch	13 089	DE
Tages-Anzeiger	13 040	DE
bazonline.ch	12 709	DE
aargauerzeitung.ch	12 309	DE
bernerzeitung.ch	12 207	DE
Zofinger Tagblatt / MLZ	11 888	DE
tagesanzeiger.ch	11 612	DE
berneroberlaender.ch	11 603	DE
thunertagblatt.ch	11 581	DE
zsz.ch	11 517	DE
L'Illustré	11 231	FR
langenthalertagblatt.ch	11 184	DE
zuonline.ch	11 120	DE
Basler Zeitung	10 895	DE
derbund.ch	10 748	DE
schweizer-illustrierte.ch	10 620	DE
Zuger Zeitung	10 557	DE
bz - Zeitung für die Region Basel	10 528	DE
handelszeitung.ch	9 790	DE
pme.ch	9 491	FR
Der Bund	9 396	DE
Werdenberger & Obertoggenburger	9 214	DE
Der Landbote	9 122	DE
Zürichsee-Zeitung	9 019	DE
fuw.ch	8 791	DE
Luzerner Zeitung	8 651	DE
Badener Tagblatt	8 435	DE
Urner Zeitung	8 284	DE
St. Galler Tagblatt	8 117	DE
Wiler Zeitung	8 003	DE

Medium	Articles	Language
Berner Zeitung	7 777	DE
Appenzeller Zeitung	7 548	DE
Zürcher Unterländer	7 425	DE
Oltner Tagblatt / MLZ	7 420	DE
badenertagblatt.ch	7 140	DE
Berner Oberländer	7 138	DE
Femina	7 106	FR
Toggenburger Tagblatt	7 032	DE
Thuner Tagblatt	6 982	DE
solothurnerzeitung.ch	6 120	DE
bzbasel.ch	5 921	DE
RTS.ch	5 914	FR
Obwaldner Zeitung	5 854	DE
Nidwaldner Zeitung	5 844	DE
TV 8	5 677	FR
Sonntagsblick	5 606	DE
Grenchner Tagblatt	5 530	DE
Solothurner Zeitung / MLZ	5 450	DE
BZ - Langenthaler Tagblatt	5 277	DE
SonntagsZeitung	5 228	DE
Limmattaler Zeitung / MLZ	5 042	DE
NZZ am Sonntag	4 991	DE
Finanz und Wirtschaft	4 962	DE
SWI swissinfo.ch	4 855	IT
Glückspost	4 621	DE
Limmattaler Zeitung	4 513	DE
limmattalerzeitung.ch	4 488	DE
rts Vidéo	4 092	FR
Die Weltwoche	4 011	DE
Bilan	3 979	FR
oltnertagblatt.ch	3 958	DE
grenchnertagblatt.ch	3 857	DE
swissinfo.ch	3 575	IT
www.swissinfo.ch	3 541	IT
swissinfo.ch	3 525	FR
PME Magazine	3 244	FR
illustre.ch	3 077	FR
Schweizer Illustrierte	3 068	DE
Handelszeitung	2 917	DE
srf Video	2 558	DE
Die Wochenzeitung	1 953	DE
bellevue.nzz.ch	1 919	DE
Thalwiler Anzeiger/Sihltaler	1 826	DE
Zuger Presse	1 781	DE
HZ Insurance	1 617	DE
Schweizer Familie	1 570	DE
weltwoche.ch	1 466	DE
Beobachter	1 446	DE
Zugerbieter	1 409	DE
Guide TV Cinéma	1 384	FR
weltwoche.de	1 267	DE
Tele	1 176	DE
Bilanz	1 085	DE
swissinfo.ch	1 004	DE
encore!	986	FR

Medium	Articles	Language
Beobachter.ch	984	DE
Das Magazin	982	DE
züritipp (Tages-Anzeiger)	882	DE
NZZ am Sonntag Magazin	823	DE
TV Star	764	DE
weltwoche-daily.ch	719	DE
bilanz.ch	596	DE
SWI swissinfo.ch	587	FR
Streaming	535	DE
HZ Insurance	529	FR
NZZ PRO Global	446	DE
Schweizer LandLiebe	441	DE
glueckspost.ch	399	DE
encore! (dt)	274	DE
Newsnet / 24 heures	227	FR
TV Land & Lüt	215	DE
NZZ Geschichte	151	DE
SI Sport	143	DE
Newsnet / Berner Zeitung	143	DE
Bolero	142	DE
boleromagazin.ch	118	DE
NZZ Folio	109	DE
beobachter.ch	107	DE
Aargauer Zeitung / MLZ	91	FR
HZ Insurance	77	IT
SI Gruen	70	DE
L'Illustré Sport	70	FR
Newsnet / Basler Zeitung	69	DE
Newsnet / Der Bund	58	DE
Bolero F	56	FR
Schweiz am Wochenende	47	FR
Badener Tagblatt	34	FR
Schweizer Versicherung	31	FR
Newsnet / Le Matin	28	FR
Newsnet / Tribune de Genève	25	FR
Schweizer Illustrierte Style	23	IT
Grenchner Tagblatt	22	FR
Oltner Tagblatt / MLZ	21	FR
Werdenberger & Obertoggenburger	21	IT
Solothurner Zeitung / MLZ	20	FR
Limmattaler Zeitung / MLZ	20	FR
Finanz und Wirtschaft	18	FR
NZZ Online	16	DE
Schweizer Versicherung	16	IT
TV4	12	DE
Limmattaler Zeitung	9	FR
rts Video	7	FR
SWI swissinfo.ch	6	DE
Newsnet / Tages-Anzeiger	6	DE
Handelszeitung	6	IT
Berner Oberländer	5	FR
Thuner Tagblatt	5	FR
berneroberlaender.ch	4	FR
Beobachter.ch	4	IT
thunertagblatt.ch	3	FR

Medium	Articles	Language
Neue Zürcher Zeitung	3	IT
cash.ch	2	FR
Blick	2	IT
Berner Zeitung	2	IT
srf.ch	2	IT
weltwoche.de	2	IT
Blick	1	FR
bernerzeitung.ch	1	FR
fuw.ch	1	FR
Sonntagsblick	1	FR
Basler Zeitung	1	FR
weltwoche.ch	1	FR
weltwoche.de	1	FR
srf.ch	1	FR
bazonline.ch	1	FR
rtr.ch	1	IT
derbund.ch	1	IT
St. Galler Tagblatt	1	IT
Die Weltwoche	1	IT
Das Magazin	1	IT
nzz.ch	1	IT
Basler Zeitung	1	IT
Schweiz am Sonntag / MLZ	1	IT
blick.ch	1	IT
Cash	1	IT
bazonline.ch	1	IT

Table 8: Composition of the dataset used to fine-tune the SwissBERT model according to medium and language.

B Evaluation results of Sentence-BERT baselines

Encoder Model	Summary Language	Article Language			
		DE	FR	IT	RM
paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2019)	DE	75.01	81.76	79.56	18.44
	FR	75.18	83.57	81.56	19.87
	IT	72.28	78.87	79.56	19.25
	RM	53.64	53.91	57.11	19.44
distiluse-base-multilingual-cased-v1 (Reimers and Gurevych, 2019)	DE	**91.80**	**90.98**	**90.38**	**62.53**
	FR	**90.78**	**93.19**	**90.78**	**63.36**
	IT	**88.12**	**91.29**	**91.58**	**65.71**
	RM	**70.59**	**73.48**	**73.55**	**73.35**

Table 9: Results for the document retrieval task using two multilingual Sentence-BERT models. The accuracy score is reported. The best results per language pair are marked in bold print.

Encoder Model	Training Language	Test Language			
		DE	FR	IT	RM
paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2019)	DE	75.42	75.64	73.88	39.38
distiluse-base-multilingual-cased-v1 (Reimers and Gurevych, 2019)	DE	**77.23**	**76.83**	**76.90**	**65.35**

Table 10: Results for the nearest-neighbor classification task using the two multilingual Sentence-BERT models. A weighted F1-score is reported and the best results are marked in bold print.