A geoscientific application of DistilBERT: automatic sorting of O&G reports

Claire Birnie, Hind Alamro, Manal Alshehri, Matteo Ravasi [KAUST]
and Xiangliang Zhang [University of Notre Dame, formerly: KAUST]

Introduction

As the O&G transition towards a more sustainable use of underground resources is ongoing, the geoscientific community is currently faced with a conundrum: "*how can we further improve our understanding and utilization of the subsurface whilst at the same time being more efficient and cost-effective?*". The key to answering this question lies in the extensive collection of public and proprietary documents (i.e., reports and presentations) that have been produced over the last 50+ years in this industry. However, the wealth of information contained in such data collections is by definition unstructured, implying the need for automated tools that could help in the mundane tasks of sorting and knowledge extraction.

In an effort to unlock the potential of our text-based data and educate future geoscientists on the use of AI-based technological developments in Natural Language Processing (NLP), in the summer of 2021 KAUST hosted a 3-days virtual summer school focused on "*Utilising unstructured data in geoscience*". Combining the expertise of individuals from the Computer Science and Earth Science schools, in the third day we sought to prove the value of pre-trained, Transformer-based NLP models on geoscientific data. In this abstract we discuss our findings and more specifically show that the pre-trained DistilBERT model can be utilised out-of-the-box in combination with clustering approaches for the task of sorting of O&G reports.

Data

In order to mimic the variety of documents that are usually produced during the lifecycle of a O&G field, from prospect evaluation all the way to plug-and-abandonment, in this work we use openly available stratigraphic reports, drilling reports, and geophysical studies. In total, 540 documents are analysed. Their document type classification is pre-determined based on the data source however these labels are not used during the training process and are only considered for evaluation purposes. We note that in practice, reports and presentations from disparate studies are sometimes placed uncategorized in the same physical location, therefore the need for an automated sorting tool.

Methodology

In this study we use an extension of the well-known BERT model (Devlin et al., 2018), DistilBERT (Sanh et al., 2019). This is a small, cheap, and fast Transformer model trained by distilling BERT base, which is shown to provide similar performance to the BERT model with 40% less parameters resulting in an overall speed-up of 60%. The pre-trained model is easily accessed via the popular HuggingFace Python library (Wolf et al., 2019).

To begin with, text sections are extracted from the available sources and cleaned to remove non alpha-numeric characters, prior to tokenisation and padding. The pre-processed text is then fed through the DistilBERT model and the resulting embedding representations are used as input to the K-Means++ algorithm of Scikit-learn (Pedregosa et al., 2011) with a preset number of clusters. After clustering the majority of document types that make up a cluster is used to categorise the cluster. For example, if 75% of the embeddings within a cluster belong to the drilling report group the cluster will be categorised as drilling.

Results

Figure 1(a) illustrates the BERT embeddings, color-coded based on the associated document category. Note that t-Distributed Stochastic Neighbor Embedding (t-SNE) has been used to reduce the dimensionality of the embeddings to two for visualisation purposes. Distinct clusters can be observed within the embedding

space. More specifically, geophysical and drilling reports both have their own relatively contained cluster, however, the litho-stratigraphic reports seem to be split over two clusters. Similarly, Figure 1(b) displays the results of K-Means clustering when selecting the number of clusters to be three, based on our prior knowledge of the number of distinct document types. Due to the large distance between the two clusters of litho-stratigraphic documents, two clusters are categorised as 'Litho-stratigraphic' while the third cluster is categorised as 'Geophysics' despite having all but four of the drilling reports within it. This is due to there being 200 geophysical studies in comparison to the 164 drilling reports analysed. As four clusters are clearly observed, the K-Means analysis is now rerun allowing for four clusters. As shown in Figure 1(c), this provides a satisfactory separation between the three different document types.

Conclusion

We have presented a use-case of Transformer-based NLP in the geoscientific domain. More specifically, our preliminary results show that DistilBERT can be easily incorporated into a workflow for automatic document sorting. Despite being trained on generic corpi, the DistilBERT embedding representations are significantly different between the three analysed report categories allowing for a relatively accurate differentiation between documents belonging to the three data sources considered in this work.

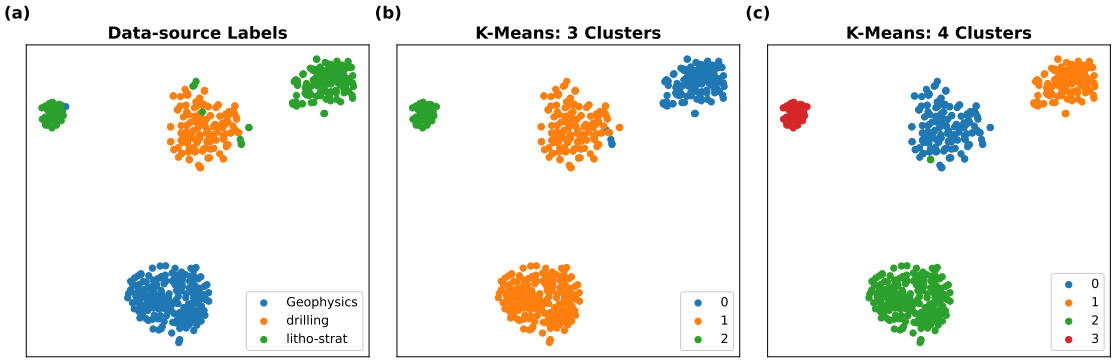

Figure 1: 2-D representation of the DistilBERT embeddings of O&G reports analysed, coloured by the different categorisation approaches. (a) represents the document categorisation based on the data-source location, (b) categorisation by K-Means with three clusters, and (c) categorisation by K-Means with four clusters.

Acknowledgements

The authors would like to thank the Norwegian Petroleum Directorate for their efforts in releasing data, as well as KAUST and Iraya Energies for supporting the summer school. All codes are available at https://github.com/cebirnie92/KAUST-Iraya_SummerSchool2021.

References

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, 2018, Bert: Pre-training of deep bidirectional transformers for language understanding: arXiv preprint arXiv:1810.04805.

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

Sanh, V., L. Debut, J. Chaumond, and T. Wolf, 2019, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter: arXiv preprint arXiv:1910.01108.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., 2019, Huggingface's transformers: State-of-the-art natural language processing: arXiv preprint arXiv:1910.03771.