Development of a multiclass document classifier

About Sosias Data
We are a Data-Mining startup founded by three students from TUM. We mine publicly available data of companies and sell it aggregated to our customers through our Web-App.

About the project
We are mining daily gigabytes of data that need to be analyzed continuously. Currently, we are only scratching the surface of our data’s potential – we want to change this with you! Your task will be to develop a model to classify documents based on a standardized industry-classification.
We don’t want you to reinvent the wheel but want you to use existing techniques to achieve the best results.

Scope of the project
1. Develop a working prototype
2. Give a short description of the used techniques and methodologies

Expected outcome
1. Working Prototype (via e.g. GitHub Repo, Colab Notebook – whatever you feel comfortable with, but please not Dropbox ;-)
2. Short informal report with a performance evaluation.

What you can expect from us
You will be assigned to one of our founders who will take care of you during the semester via regular 1 to 1 meetings (online or offline) and support you in the project. We are very flexible in terms of working time and will of course stay within the boundaries of the IDP’s ECTS limit.

What we expect from you
If you can work independently and want to do something hands-on, then we are your partner! Besides the regular meetings you will have complete freedom in terms of time management or the choice of used techniques. Wheter you are using a rule-based approach or training a BERT model is completely up to you! It’s your project!

Interested? Write us!
Send us a short mail answering us why you want to work on this project specifically and what your knowledge in the field of NLP is (lectures, Praktika, etc..). CV and grades are optional – we value your motivation and drive most!
Send it all to Benedikt Reinhard (Please call me Ben), breinhard@sosias-data.de.