# A transformer based semantic analysis of (non-English) Danish jobads

Morten Mathiasen[1], Jacob Nielsen[1] and Simon Laub[1]
[1]*EAAA, Aarhus, Denmark*
*mmat, jani, sila@eaaa.dk*

Keywords: Transformers, analyzing online job ads, skills, alignment of educational courses.

Abstract: To benefit educational adaption and guidance of the unemployed, we report on our study of automated monitorization of labor market demands by analyzing online job ads. We identify and measure two categories of competence demands, “technical competences” and “personal competences”, as well as competences described by Bloom’s taxonomy. Ads are labelled, both by humans and by natural language processing (NLP) transformers. Within all competence categories and levels of Bloom’s taxonomy we demonstrate how the automated NLP transformer process do a semantic search with the same level of precision as the humans.

## 1 INTRODUCTION

Designing educational programs and specific courses, demands substantial consideration and continual revision. Stakeholders need to address questions like what specific content should be included in the course and what competences do the students need to acquire to fit the job market after graduation?

Semantic analysis of job ads is valuable for adapting educational programs and guiding the unemployed towards the demands of the labor market. Each job ad typically asks for one or more competences, which can be identified and analyzed by reading the job ads. Natural Language Processing (NLP) is a means of automating this process (Chowdhary, 2020), and NLP transformers have generally in recent years been seen to improve semantic analysis of texts. In this study we investigate how and to what extend transformers are helpful, in the context of competences described in educational courses, for analyzing demands in Danish job ads.

### 1.1 Educational development

Traditionally, educational institutions have tried to acommodate the job marktets demands by conducting surveys and interviews with industries. However, these approaches are time-consuming, and often at risk of providing a biased insights into the general demand for competences. Additionally, changes in curricula are substantial and long processes, so there is a consistent need for rapid input into this process to prevent the initial improvements becoming redundant when finally completed. Therefore, monitoring labor market demands is essential when adapting educational programs and when guiding the unemployed towards the demands of the labor market.

### 1.2 Monitoring labor market demands

The project “European Skills, Competences, Qualifications and Occupations” (ESCO, 2023) monitor all labor markets in the Europe Union and identify all competences in demand. Statistics on job vacancies and labor market structure is monitorized by another european project ”Labour Market, including Labour Force Survey” (eurostat, 2023) which provides statistical tools for analyzing the european job market. Market research institutes such as Gartner (Gartner, 2023) provide yearly reports on emerging technologies and general market trends. Furthermore, analyses from government agencies (STAR, 2023) are valuable sources of information when identifying job market areas with shortages of qualified work force.

However, there is a lack of large-scale automated tools that monitor the changing demand of competences by the labor market. Such tools would continuously support adapting educational programs to better fit the labor market and guiding the unemployed towards the labor markets demands.

### 1.3 Job ad analytics

In Denmark, most job ads are posted online, which means that it is possible to monitor demanded competences and market changes through the ads. Automated harvesting and analysis of online job ads is therefore a means to overcome the challenges of time-consuming and small sample qualitative job market analyses. Additionally, these challenges of automation do primarily apply to analyses of job markets as the amount of job market data is much more extensive than data of courses within educational programs.

Analyses of job ads as a means to identify labor market demands is applied in multiple studies (Gromov et al., 2020; Boehler et al., 2020). Manual text mining by researchers who read, label, and analyze texts is a common approach to job ad analytics. But analyses of Danish ads supported by monolingual and multilingual tools (Strømberg-Derczynski et al., 2021)(SBERT, 2023) are also commonplace now. Automated text mining are supported by methods such as quantitative analysis based on dictionaries, and some forms of semantic analysis using machine learning are also emerging (Salloum et al., 2020; Pejic-Bach et al., 2020).

## 1.4 Transformers to monitor the labor market.

Following the introduction of NLP transformers in 2017 (Vaswani et al., 2017), they now appear to be ubiquitous in contemporary technology. BERT (Devlin et al., 2018) helps Google improve contextual understanding of unlabeled text across a broad range of tasks, and the company OpenAI (OpenAI, 2023) creates transformer based chatbot products (Adamopoulou and Moussiades, 2020) that clearly illustrate the possibilities of these models.

Transformer-based semantic textual similarity solutions have been proposed for tasks ranging from “technological trouble shooting” (Alfeo et al., 2021) to healthcare, “Clinical Sentence Pairs” (Ormerod et al., 2021) and more.

Following this development, we expect that NLP transformers can monitor demanded competences on the Danish labor market using ongoing automated analysis of online job ads. This hypothesis is the foundation of this study, which verifies the possibility of using NLP transformers for semantic analysis of demanded competences in Danish job ads.

## 1.5 Research question

Given competences described in a educational curriculum and described demands in job ads, we want to monitor the trends of necessary competences to meet the demands in the labor market.

*To what extent are NLP transformers able to identify and monitor competences in Danish job ads, matching with competences described in course materials.*

In addressing this research question, this will also indicate whether a solution based on transformer techniques can be said to be an improvement over existing techniques for job ad analytics. I.e. to what extent can transformer based semantic analyses reveal and monitorize Danish labor market demands compared to existing techniques like human manual text analysis, automated text mining by dictionaries and more traditional machine learning techniques (Salloum et al., 2020; Pejic-Bach et al., 2020).

# 2 METHOD

## 2.1 Case

Considering the possibilities of aligning job market demand and educational supply of competences, our point of departure is an investigation of Danish jobs ads that illustrate the demand for competences. Specifically, in order to examine to what extent a transformer based semantic approach can reveal and monitorize Danish job market demands, our case study revolves around Danish job ads that mentions ”Multimedia designer” from 2018-2022. These specific job ads constitute a job market area that is central to graduates from the multimedia designer education in Denmark. The job ads provide competences that are demanded by the specific job market area. We manually identified competences from a representative sample of 300 ”Multimedia designer” job ads from the period 2018-2022. The competences were categorized using Bloom’s revised Taxonomy (BRT) (Krathwohl, 2002) as our analytical framework, and to adequately identify and categorize all competences in the job ads, including competences outside the scope of the BRT framework, additional categories were created. These additional categories include personal competences, technical skills, and experience.

The categorized competences constitute the data for the following analyses.

## 2.2 Finding sentences describing competences in the course material

The focus of this study is analyses of job ads. Where the competencies we are searching for will usually come from course materials. In the testing phase we search for competencies that have already been manually analyzed. In the educational programs for danish academies and colleges in focus here, the competences taught in each course are listed in the curriculum as “learning objectives” under headlines with different taxonomic levels, ordered according to BRT (Krathwohl, 2002). A typical course will have approximately five bullet point entries under each of these taxonomic level. Easily extracted by a manual or an automated process. The match of these competences with the demand-side competences will then help us examine the educational fit.

## 2.3 Analyzing the semantic content of job ads

Sentence semantic textual similarity is the core of this project, and it revolves around sentences describing demanded competences.

Ideally, all sentences containing competences in the curriculum are matched against all similar sentences mentioned in job ads. Sentence similarity will thereby reveal a match between competences obtained through courses and competences demands in job ads. Periodization of the number of found matches reveals trends in labor market demands within the analyzed periods 2018-2022.

In order to determine sentence similarities, transformers takes center stage here. Certainly, variations over bag-of-words (BoW) techniques such as countvectorizer (NLTK, 2023b), or term frequency inverse document frequency, tf-idf (NLTK, 2023b), techniques are useful, when it comes to an analysis based on a corpus vocabulary. However, these techniques are less relevant here, when we move on to matching on semantic sentence similarity, as these techniques do not capture word positions in a text, word semantics etc.

So, we are led to the use of word embeddings that makes it possible to represent words in the form of a vector (Albrecht et al., 2020). Where the vector encodes the semantic meaning of the word, and where words with similar vectors are expected to have similar semantic meanings.

After having decided which ads to analyze, we then move on to look at sentences in the ads. Where the transformer models are used to make a measurement of similarity of semantic meanings. A similarity metric, e.g. cosine similarity, that can be used to compute similarity scores for words or texts, whether two texts have similar or more different syntax. Many pretrained transformer models based on BERT / RoBERTa networks are available. In this case, we use multilingual SBERT (SBERT, 2023) sentence transformers, which allows for calculating similarity scores for the danish curricula and course materials compared against job ad texts.

The structure of the overall method is: 1) Scrape online job ad text and preprocess the ads by extracting the parts that request competences, 2) From educational curriculum and course materials, extract sentences describing taught competences, 3) Use the transformer model to calculate the cosine similarity scores between sentences from the two sources.

The analysis takes its starting point in the categories of competences explicit stated in relevant educational curriculum and course material. And we assume that the course material is described (or reworked) in such a way that it generally matches the competency granularity in job ads. Further along in the project, additional categories of competences can be added to the analysis, if needed.

## 2.4 Comparing automated and manual labeling of competences

To test our approach, we have manually analyzed 300 multimedia job ads using a revised edition of Blooms taxonomy (Bloom et al., 1956; Krathwohl, 2002). As such, we categorized job ad demands by the categories “Knowledge”, “Comprehension”, “Application”, “Analysis”, “Synthesis”, and “Evaluation”. Additionally, for the scope of this study, two additional categories “technical competences” and “personal competences” were used to get a deeper understanding of the demands in job ads.

Two people categorized the competences in job ads and the inter-coder reliability yielded a kappa statistic of .75. Our goal is to achieve at least same precision by automated labeling.

The thorough manual analysis of the jobs ads has throughout our project been a necessary component, guiding our own understanding of achieved manual and automated labelling results. But for testing purposes, the agreed upon competencies is then only used as a starting point for the following analysis. Most of our experiments will use, a subset of, around 30 sentences, describing competencies in a certain category, matching the number of competencies we usually find described in course materials, and making both the following manual and automated test process more tractable. Next, the selected competencies are used in a new process of manual labelling of multimedia ads, which is then compared with an automated labelling. Where the hope is that the manual and automated labelling will achieve similar results.

# 3 RESULTS

## 3.1 Shortcomings of word search and Bag-of-words (BoW) techniques

Before we began using transformers for sentence similarity, we worked on various bag-of-words models, but it quickly became apparent that making small syntactically changes in ad texts, e.g. replacing one word with another with a similar meaning, would have significantly weakened our ability to identify these competences and introduce classification errors (Git, 2023). We did similar observation when searching for specific words using dictionaries as done by Pejic-Bach et al. (2020). In our test, these small changes in the words used in ads, e.g. replacing a word with another, made it impossible to match sentences representing the same semantic meaning.

## 3.2 Transformer based search for "personal competences" competences. No preprocessing of job ad texts

In all tests using transformers, we first split the job ad text into separate sentences. This was done assisted by the (Python) NLTK library (NLTK, 2023a). The library spaCy (spaCy, 2023), or functionality with regular expressions, gave similar results.

In a first test of the prototype setup, we started by searching for personal competences. Our job ad data yielded 310 sentences describing personal competences.

Different subsets of these competences where then experimentally matched against a number of subsets of job ads. Thereby giving test results with many or few hits, as expected.

In one scenario we used 32 sentences about skills, that could also be identified by the manual labelling, to various degrees in different subsets of job ads. One of these subsets of job ads, any will do, then gives us a starting point for the automated analysis. Using the multi-qa-MiniLM-L6-cos-v1 sentence transformer model (HuggingFace repository, 2022), designed for semantic search, and applicable for similarity search for Danish sentences, this approach resulted in 95 out of 100 ads are classified similarly to the manual labelling of competences (does the ad contain some of the 32 personal competences competences that we are searching for, given a 0.7 cosine similarity threshold). See figure 1.

The ROC curve (Geron, 2022) visualizes the classification with different classification thresholds.

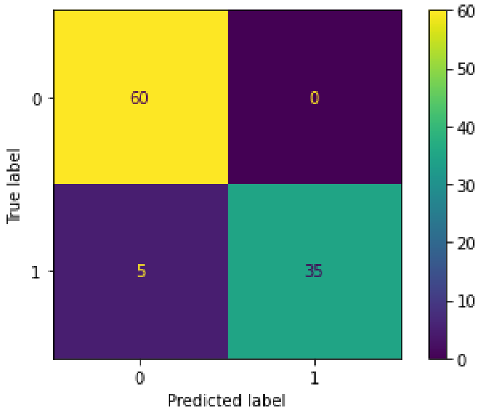

Figure 1: Confusion matrix for sentence transformer based classification of ads. Without preprocessing of the ad texts.

Where the true positive rate is given vertically, and false positive rate, horizontally. Figure 2 shows that we barely have no threshold allowing for labeling job ad sentences correctly. At most, we were able to label approximately half of all relevant sentences without reaching a high degree of false positive labeling of irrelevant sentences. Figure 2 shows that it was difficult in this setup to raise the true positive score further without also raising the false positive score.

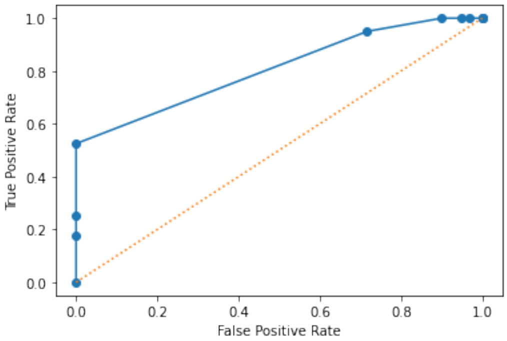

Figure 2: ROC curve without preprocessing of the ad texts, false positive rates becomes unacceptably high for high true positive rates.

## 3.3 Transformer based search for ”personal competences”. With preprocessing of job ad texts

Looking more carefully at the ads, it became apparent that many contained headers and footers that were not relevant to this analysis of competences, Therefore, they were removed. Additionally, the version of our scraping process replaces line breaks, bullet points and more html tagging with the character for space, that makes the subsequent separation of the job ad texts into sentences more difficult (for the NLTK library (NLTK, 2023a)). Correcting this, we obtained the following confusion matrix and ROC curve, using the the multi-qa-MiniLM-L6-cos-v1 transformer (HuggingFace repository, 2022)). See figure 3 and 4.

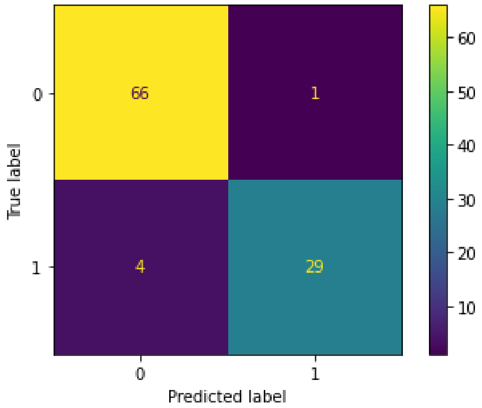

Figure 3: Confusion matrix for sentence transformer based classification of ads, using a 0.7 classification threshold upon preprocessed ad texts.

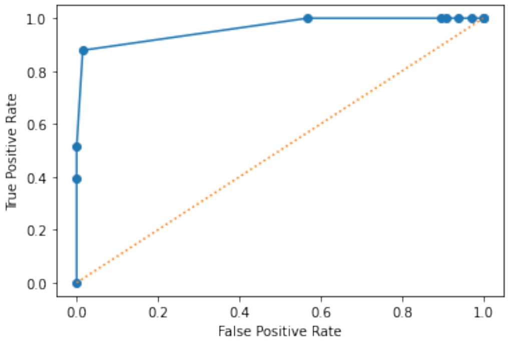

Figure 4: ROC curve for sentence transformer based classification of ads. With preprocessing of the ad texts.

The additional preprocessing demonstrates that it is possible to get to a higher true positive score before the false positive score also begins to increase.

We achieved similar results when we focused on e.g. comprehension competences, experience competences, personal competences etc. Where results from test with comprehension competences can be found in the Appendix.

# 4 DISCUSSION

## 4.1 Case: Adapting educational programs

To test the use of NLP transformers for analyzing Danish job ads, we did a case study aiming at adapting the Multimedia educational program by analyzing 300 job ad texts. We labelled competences manually and automatically by NLP transformers.

Our case study has afterwards been supporting the adaption of the Multimedia educational program, as we have identified a discrepancy in this specific supply and demand for competences. Additionally, we have outlined trends of the labor market demands which helps foreseeing future demanded competences in the upcoming years.

Other authors have investigated the area of competences extraction from job ads (Zhang et al., 2022), but our investigation of “non-English” Danish job ads, where we use transformers to categorize ads, along with the use of Blooms taxonomy (Bloom et al., 1956; Krathwohl, 2002), for the distinct purposes of alignment of courses in educational programs is novel.

And much needed in the ever changing job market, where many educational institutions experience increased pressure, from the surrounding society, to make sure that courses are aligned with the demands of the labor market. Making us confident that there is need for the analysis that our prototype can provide, even in this initial stage.

## 4.2 Job ad analytics by NLP transformers

Using a multilingual transformer model, we can then automatically run sentences (from the course material) through a batch of job ads. And generate a report that list highest similarity match for each competence, as well as number of matches above a certain threshold. Which can serve as a starting point for further analysis. Should competences in the course material be described differently, are these competences no longer in demand in the job market? Or should other competences taught in the course be highlighted?

Looking at the 300 manually labelled multimedia job ads, our prototype has been able to reproduce the labeling for categories “technical competences”, “personal competences”, “Knowledge”, “Comprehension”, “Application”, “Analysis”, “Synthesis”, and “Evaluation”, using a selection of sentences that describe these categories.

This brings us a step closer to our goal, which is to let our prototype run through jobs, and return a report on whether the ad contains sentences that describe competences equivalent to the competences we are looking for.

In handling this problem, we have been following a very general approach, which allows us to easily take in an array of sentences, and search for this set of strings (describing competences) in job ads. Where we assume that the responsibility for listing accurate sentences, describing competences in the course material, lies solely with those responsible for the courses. If the search turns up empty handed, then it is easy to add more sentences, or reformulate inaccurate sentences, that better describe the competences taught and search again. But it is of course also possible that the competences are no longer mentioned in job ads. Which is exactly the purpose of the prototype, to be able to come up with an indication of whether competences are in demand in the local job market.

Working with the classification of 300 multimedia ads, we find that the new transformer based solution is more robust than earlier versions of our classification tool. I.e. looking at a personal qualification like "taking responsibility in a project" this could, in Danish, be stated as "Du tager ansvar for dine projekter" or "Du har stor pligtfølelse omkring projekterne". Looking at the sentences, it is immediately clear that the sentences do not have many words in common, which make similarity search classification with single words, regular expressions, or various bag-of-words (BoW) techniques difficult, or impossible.

However, with the distiluse-base-multilingual-cased-v2 transformer (Reimers and Gurevych, 2019) the two sentences have a 0.75 similarity score (using cosine similarity), and is therefore similar given a 0.7 threshold for sentence similarity.

So, for such changes (to sentences about competences in the ad texts), where the semantical meaning of the sentences are preserved, we find that the transformer based solution will be able to classify the altered sentences correctly,

Online job ad texts are formatted using HTML, bullet lists, tables etc. Our NLP transformers performed much better when the ad texts were preprocessed by removing HTML tags and adding missing punctuation. So, in general we advise that the webscraped texts are preproccessed before applying NLP transformers.

## 4.3 Further research

We have considered adding a general text classification system to our system, that would be able to classify ads based on the competences described in the ads, and thereby provide valuable input to later processing steps. To some extend, this would follow in the footsteps of many other systems for text classification. The challenge is that we have approximately 14,000 described competences just in the European labor market (ESCO, 2023). Most text classification systems usually deal with far fewer classes. For instance, a classical classification task like the Reuters newswire task contains only 46 topics[1], and many classifications tasks on job ads often just classify ads as being either X or not-X. Following this pattern of binary classification, we had some success with a deep neural net technique that could classify our ads as being either IT ads, or non-IT ads. Ads looking for people with a technical background or not, etc. However, working on this, we found that classification accuracy quickly dropped as more classes for the classification task were added. Therefore, more research is needed to scale our NLP transformer solution to handle 14.000 categories (for the 2.5 million job ads) and not just BRT.

Letting our prototype analyze job ads using transformers is a rather time-consuming process, with the hardware that has so far been available in this project. Using a Tesla T4 graphic card, it takes the prototype about 2 minutes to run through 100 job ads searching for just one category of competences (comparing with approximately 30 sentences that describe that competence). Nevertheless, it is significantly faster compared to the existing manual labeling process. But, it is problematic when looking at the 2.5 million job ads that we have scraped from the internet. Therefore, we have worked on various preprocessing steps to speed up the overall processing. Regular expression search can establish whether an exact text match for a competence description is found in the ad text. Potentially, stopping the need for further processing. Processing with removal of stop words (spaCy, 2023), and lemmatization (Khyani et al., 2021), and a count of tokens in a whole document text can give us word frequencies in ad texts, which can be useful in establishing a very general classification of the ad texts, stopping further processing on ads that doesn't contain any words usually found in say multimedia designer job ads etc.

Indeed, improvements are possible by leveraging human knowledge of the ad text domain. But, focusing on a simple solution that can correctly categorize sentences, speed of processing has for now not been a major concern though. And the AI researcher Richard Sutton probably also had a point, when he

[1] https://keras.io/api/datasets/reuters,https://www.kaggle.com/datasets/nltkdata/reuters

commented: “The only thing that matters in the long run is the leveraging of computation” (Sutton, 2019). Certainly, as this project has progressed, we have only found better hardware and new improved language models very helpful.

## 5 CONCLUSIONS

This study demonstrates that NLP transformers have the capability to do semantic analysis of Danish job ad texts. Optimization led to labeling precision in the 95% range compared to human beings labeling competences in demand in the same ads. The inter-coder reliability for two people manually categorizing the same job ads competences yielded a kappa statistic of k = .75. Therefore, the findings of this paper support the claim that NLP transformers can do semantic analysis at a precision level comparable to humans. The demonstration of semantic text analysis done by NLP transformers used on Danish job ad texts enables the possibility to automate the monitorization of demanded competences at the Danish labor market. Such monitorization will benefit adaption of educational programs and guidance of employed towards vacancies.

We are now able to fully analyze smaller batches of preselected job ads. However, further improvements to our current prototype are needed before we can realisticly approach full-scale monitorization of the Danish labor market. Where a future system needs to analyze approximately 500,000 yearly Danish job ads, each to be compared with skill sets described in educational course materials, and categorized according to the approximately 14,000 competences provided by the project ”European Skills, Competences, Qualifications and Occupations” (ESCO, 2023).

## REFERENCES

Adamopoulou, E. and Moussiades, L. (2020). An overview of chatbot technology. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*, pages 373–383. Springer.

Albrecht, J., Ramachandran, S., and Winkler, C. (2020). *Blueprints for Text Analytics Using Python*. O’Reilly Media, Inc.

Alfeo, A. L., Cimino, M. G. C. A., and Vaglini, G. (2021). Technological troubleshooting based on sentence embedding with deep transformers. https://doi.org/10.1007/s10845-021-01797-w.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). *TAXONOMY OF EDUCATIONAL OBJECTIVES, The Classification of Educational Goals*. LONGMANS.

Boehler, J. A., Larson, B., and Shehane, R. F. (2020). Evaluation of Information Systems Curricula. *Journal of Information Systems Education*, 31(3):232–243.

Chowdhary, K. R. (2020). Natural language processing. In Chowdhary, K., editor, *Fundamentals of Artificial Intelligence*, pages 603–649. Springer India. 10.1007/978-81-322-3972-7_19.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

ESCO (2023). European Skills, Competences, Qualifications and Occupations. https://esco.ec.europa.eu/da/classification/skill_main/.

eurostat (2023). Labour market. https://ec.europa.eu/eurostat/web/labour-market/.

Gartner (2023). Gartner. https://www.gartner.com/.

Geron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow 3e - Concepts, Tools, and Techniques to Build Intelligent Systems*, volume 2022. O’Reilly Media, Inc., 3rd edition.

Git (2023). Project experiments and prototype code on github. https://github.com/SimonLaub/NLP_JobTrend.

Gromov, A., Maslennikov, A., Dawson, N., Musial, K., and Kitto, K. (2020). Curriculum profile: modelling the gaps between curriculum and the job market. *Educational Data Mining 2020*.

HuggingFace repository (2022). Sentence transformer. https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1/.

Khyani, D., Siddhartha, B., Niveditha, N., and Divya, B. (2021). An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology*, 22(10):350–357.

Krathwohl, D. R. (2002). A Revision of Bloom’s Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218.

NLTK (2023a). Nltk. https://www.nltk.org/.

NLTK (2023b). scikit-learn - machine learning in python. https://scikit-learn.org/.

OpenAI (2023). Openai. https://openai.com/.

Ormerod, M., Martínez Del Rincón, J., and Devereux, B. (2021). Predicting semantic similarity between clinical sentence pairs using transformer models: Evaluation and representational analysis. https://medinform.jmir.org/2021/5/e23099/.

Pejic-Bach, M., Bertoncel, T., Mesko, M., and Krstic, Z. (2020). Text mining of industry 4.0 job advertisements. *International journal of information management*, 50:416–431. Publisher: Elsevier.

Reimers, N. and Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Association for Computational Linguistics. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

Salloum, S. A., Khan, R., and Shaalan, K. (2020). A survey of semantic analysis approaches. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, pages 61–70. Springer.

SBERT (2023). Multilingual transformer models. https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models.

spaCy (2023). Industrial-strength natural language processing. https://spacy.io/.

STAR (2023). Styrelsen for arbejdsmarked og rekruttering. https://www.star.dk/.

Strømberg-Derczynski, L., Ciosici, M., Baglini, R., Christiansen, M. H., Dalsgaard, J. A., Fusaroli, R., Henrichsen, P. J., Hvingelby, R., Kirkedal, A., Kjeldsen, A. S., Ladefoged, C., Nielsen, F. Å., Madsen, J., Petersen, M. L., Rystrøm, J. H., and Varab, D. (2021). The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden. https://aclanthology.org/2021.nodalida-main.46/.

Sutton, R. (2019). The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. https://arxiv.org/abs/1706.03762/.

Zhang, M., Jensen, K. N., Sonniks, S. D., and Plank, B. (2022). Skillspan: Hard and soft skill extraction from english job postings. https://arxiv.org/abs/2204.12811/.

# APPENDIX

As further examples of transformer based search, results from searching for ”comprehension” competences, as well as results from search with the distiluse-base-multilingual-cased-v2 transformer (Reimers and Gurevych, 2019) can be found in this appendix. Both cases show results that includes preprocessing of the job ad texts.

## 5.1 Transformer based search for ”comprehension competences”. With preprocessing of job ad texts

Searching for 35 “comprehension competences” in 100 manually labelled (preprocessed) job ads gave the following ROC curve with the multi-qa-MiniLM-L6-cos-v1 transformer (HuggingFace repository, 2022). See figure 5.
Also showing for “comprehension competences” that preprocessing makes it possible to get to a relatively high true positive score before the false positive score also begins to increase.

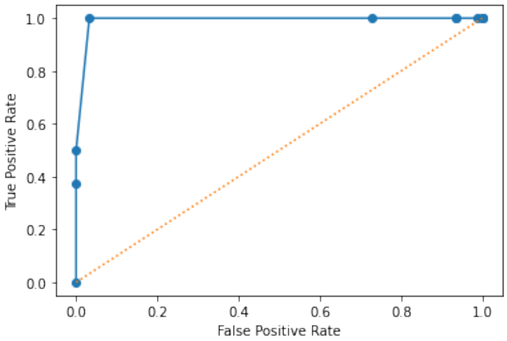

Figure 5: For the category ”comprehension”, ROC curve for sentence transformer based classification of ads. With preprocessing of the ad texts.

## 5.2 Test of other transformers for the similarity search. Here, search for “personal competences” competences.

With preprocessing of job ad texts many multilingual transformers (SBERT, 2023) seem to work. As long as they have also been trained on Danish sentences, at least to some extent. E.g. with the distiluse-base-multilingual-cased-v2 (Reimers and Gurevych, 2019) transformer, in a test where we search for 32 personal competences competences in 100 labelled job-ads, gave the following ROC curve, figure 6. Which can be compared with results shown in figure 3 and 4.

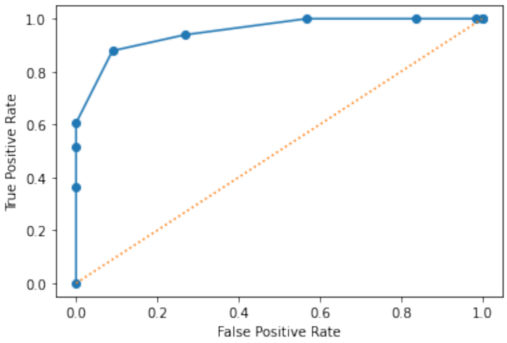

Figure 6: For the category ”personal competences”, ROC curve for sentence transformer based classification of ads. With preprocessing of the ad texts.

Figure 6 shows that it it possible to select a threshold which enable a high true positive rate, while still maintaining a relatively low false positive rate.