# A Comparative Study of NLP Models for Longitudinal Analysis of AI/ML Skill Requirements in the German Job Market

**Proposal By:** Shivang Sinha
**Date**: June 20, 2025
**Professor**: Prof. Dr Simon Werner

## 1. Motivation for the Research:

Germany's "*Industry 4.0*" strategy has created a dynamic and highly specialized job market for Artificial Intelligence (AI) and Machine Learning (ML) professionals. Understanding the specific skills required is vital for academia and industry. However, the primary source of this information which is online job advertisements that consists of unstructured text, making large-scale analysis a significant challenge for Natural Language Processing (NLP).

While prior studies (e.g., Verma et al., 2021) have analyzed job ads using traditional keyword-based methods, these techniques are rigid and often fail to capture the contextual nuances of skill requirements. The core motivation of this thesis is therefore twofold:

- From an NLP perspective: To empirically investigate and compare the effectiveness of different NLP models from classic statistical approaches to state-of-the-art transformers, for the complex task of information extraction from specialized, real-world text.
- From an application perspective: To leverage the most effective model to build a detailed, data-driven profile of the German AI/ML job market and its evolution.

This research is thus driven by a need to advance the methodology for labor market analysis while simultaneously generating valuable insights for the German economy.

## 2. Research Questions & Hypotheses

The Research questions are designed to place the NLP methodology at the center of the inquiry.

- RQ1 (Methodological Comparison): To what extent do modern NLP models—specifically a fine-tuned Transformer (BERT) and a Conditional Random Field (CRF) model—outperform a traditional keyword-based baseline for extracting skill requirements from German job advertisements?
- RQ2 (Qualitative NLP Insights): What are the qualitative differences in the types of skills identified by each model? Specifically, can transformers capture complex, context-dependent skill phrases that the other methods cannot, and what does an error analysis reveal about the architectural advantages of each approach?
- RQ3 (Application & Evolution): Using the most effective model identified in RQ1, how have the skill requirements for AI/ML roles in Germany evolved between 2019 and 2025, and how do they compare to the US market?

Our Hypotheses are:

- H1 (*Performance*): The fine-tuned Transformer (BERT) model will achieve a significantly higher F1-score on the skill extraction task compared to both the CRF model and the keyword baseline.
- H2 (*Capability*): The Transformer model will uniquely identify complex, multi-word skill phrases (e.g., "deployment of models on edge devices," "experience with GDPR-compliant data handling"), which other methods will either miss or misclassify. We expect the error

analysis to show the CRF model struggling with out-of-vocabulary terms and the keyword baseline failing on any skill not explicitly in its dictionary.

- H3 (*Market Evolution*): The analysis will reveal a statistically significant increase in demand for MLOps, LLM-related, and cloud platform skills in the 2025 dataset compared to 2019.

**3. Related Literature for Comparison**

Our work will be situated within two streams of literature:

1. Labor Market Analysis: The primary benchmark for our application-focused results will be Verma et al. (2021). We will compare our findings on the German market directly with their results from the US market.
2. NLP for Information Extraction: Our methodological comparison will be contextualized by literature on NER, comparing our model's performance to established benchmarks and studies on skill extraction from sources like the ACL Anthology.

**4. Data Collection**

A longitudinal dataset will be collected from Indeed.de to ensure comparability with related work.

1. **Contemporary Data (2025)**: using API for AI/ML job ads posted in Q1 2025. *https://docs.indeed.com/api/jobs-api/get-using-get*
2. **Historical Data (2019)**: Web scraping or Api for the same search terms from 2019.

This will result in two distinct corpora: **Germany_2019** and **Germany_2025**.

**5. Machine Learning Models & Experimental Setup**

This thesis will implement and compare three distinct models, all evaluated on the same annotated test set for a fair comparison.

1. Model 1 (Baseline): Keyword Extraction: We will replicate the method used by Verma et al. (2021). A comprehensive dictionary of skills will be curated, and we will perform regex matching on the job ad texts. This represents the non-ML baseline.
2. Model 2 (Classic ML): Conditional Random Field (CRF): A CRF is a robust statistical model for sequence labeling. We will train a CRF using features engineered from the text, such as word identity, part-of-speech tags, and word shape (e.g., capitalized, contains digits). This represents a traditional, feature-based machine learning approach.
3. Model 3 (Transformer): Fine-Tuned German BERT: We will fine-tune a pre-trained German language model (`bert-base-german-cased`) on a manually annotated dataset. As a NER model, it learns to predict a skill "tag" for each word/token in a sentence, leveraging the deep contextual understanding from its pre-training. This represents the current state-of-the-art.

**6. Measurement and Validation**

We will use a multi-faceted approach to determine if our hypotheses hold:

1. To measure H1 (*Performance*): The primary evaluation metric will be the F1-score, along with Precision and Recall, calculated for each model on a held-out, manually annotated test set. A significantly higher F1-score for the BERT model will validate this hypothesis.
2. To measure H2 (*Capability*): We will conduct a manual error analysis on a sample of 50-100 predictions from each model. We will categorize errors and identify specific linguistic phenomena (e.g., ambiguity, long-distance dependencies) that each model fails or succeeds to handle. This provides the qualitative insight into the why behind the performance differences.

3. To measure H3 (*Market Evolution*): We will use the outputs of the best-performing model (hypothesized to be BERT) to perform the final analysis. By comparing the frequency distributions of extracted skills between the 2019 and 2025 datasets using statistical significance tests (e.g., chi-squared test), we can validate our hypothesis about the evolution of skill demands.

## 7. Expected Insights for Natural Language Processing

This research is expected to yield valuable insights about NLP methods. We expect to gain a clear, quantitative understanding of the performance lift that pre-trained transformer models provide over traditional ML and non-ML methods in a specific, noisy, real-world domain. The error analysis will provide concrete evidence of how the transformer's contextual embeddings allow it to overcome the limitations of feature engineering (in CRFs) and rigidity (in keyword matching), offering a practical contribution to the field of applied NLP.

## 8. References:


- Amit Verma, Kamal Lamsal, Payal Verma (2021)- An investigation of skill requirements in artificial intelligence and machine learning job advertisements https://www.researchgate.net/publication/348937362_An_investigation_of_skill_requirements_in_artificial_intelligence_and_machine_learning_job_advertisements

- Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning (ICML 2001) http://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf