A Comparison of “X” Sentiment Analysis Investigating the Impact of COVID-19 on “Essential Jobs”

Ali Vahdatnia[1,†], Danoosh Peachkah[1,*,†] and Michael Tiemann[1,2]

[1]*University of Koblenz, Department of Computer Science, Germany*
[2]*Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany*

Abstract

This paper investigates transformations of “Essential Jobs” during the COVID-19 pandemic through sentiment analysis of social media data, specifically focusing on “X” posts. The study employs a comprehensive methodology consisting of traditional and modern sentiment analysis tools as well as advanced deep learning approaches to examine job-related sentiments across English and German languages. The research demonstrates that the “Twitter-XLM-RoBERTa” model outperforms other sentiment analysis tools in both base and enhanced implementations, challenging the assumption that deep learning enhancements necessarily improve sentiment analysis performance. The findings indicate significant variations between “Essential Job” designations. However, the high proportion of “No-Data” classifications and linguistic variability between English and German datasets suggest methodological limitations.

Keywords

Essential Jobs, “X” Social Media Analysis, COVID-19 Pandemic, Sentiment Analysis, Deep Learning, Cross-lingual Analysis, Job Classification

1. Introduction

Social media has emerged as a potent instrument for capturing and shaping public opinion during significant events. Platforms such as “X” (formerly “Twitter”) can be seen as real-time indicators of societal sentiment. These platforms facilitate the aggregation of dynamic insights into public sentiment and reactions, thereby offering researchers unprecedented opportunities to study human behavior and opinion at scale.

In order to comprehend the substantial and unstructured data generated on social media platforms such as “X”, researchers employ techniques such as sentiment analysis, which categorizes posts as positive, negative, or neutral. The efficacy of this process is augmented through the integration of machine learning and Natural Language Processing (NLP), a combination fostering the extraction of emotional meaning from textual data. NLP facilitates the nuanced interpretation of user-generated content, while machine learning models can learn from labeled datasets to classify new posts with a high degree of accuracy. Deep learning, a subset of machine learning, has further advanced this capability by more effectively handling the complexity of human language, especially with the introduction of Large Language Models (LLMs). These models have been demonstrated to possess the capacity to comprehend implicit sentiments and contextual meaning, rendering them particularly well-suited for the analysis of extensive social media datasets with enhanced precision and scalability.

These methods help using social media data to study public opinion on societal changes, such as the evolving definition of Essential Jobs during the course of the pandemic. Conventional data collection methodologies, such as surveys, have encountered significant limitations, as they are not only time- and ressource-intensive. Social media provides a viable alternative, offering the capacity to capture real-time, spontaneous public reactions. This paper addresses the research gap on why more and

AISEER 2025 International Workshop on AI in Society, Education, and Educational Research, European Conference on Artificial Intelligence, ECAI, October 25-30, 2025, Bologna, IT

[*] Corresponding author.

[†] These authors contributed equally.

tiemann@bibb.de (M. Tiemann)
0000-0001-7136-2744 (M. Tiemann)

more jobs were regarded as essential during the pandemic by conducting a comprehensive analysis of sentiments expressed on “X” with respect to Essential Jobs, building upon extant research on the German workforce. The objective of this study is to find hints on whether shifts in what constitutes an Essential Job were institutionally driven or arose “organically” from public discourse. However, we focus on two particular research questions:

- Which sentiment analysis method is the most effective for assessing changes in sentiment towards Essential Jobs? (RQ1)
- How might contemporary deep learning algorithms improve the performance of the sentiment analysis process? (RQ2)

2. Related Work

Tiemann et al. [1] look into the societal and economic assessment of Essential Jobs during the COVID-19 pandemic. The analysis involves comparing two lists of Essential Jobs, namely the Berlin List — Essential Jobs before and in the initial phase of the pandemic — and the Extended List — jobs that were added to the Essential Jobs list as the pandemic continued. They find differences in wages, prestige, workload, and degrees of qualification, analysing data from the 2018 BIBB/BAuA Employment Survey [**?**] to understand these differences across occupations throughout the pandemic. The results emphasize discrepancies in remuneration and occupational prestige between Essential and Non-Essential Jobs. One gap in research is thus to investigate whether jobs had been included in the Essential Jobs list “organically” over the course of the pandemic which could have shown in changing sentiments towards them or in the discussions surrounding them. By using sentiment analysis techniques, as well as incorporating deep learning methods on “X” data, we want to deepen our understanding on sentiments towards occupations as the initial analyses did only rely on standard methods for this. Other literature studying Essential Jobs focus on the effects of the COVID-19 pandemic on workers classified as Essential and Non-Essential, as an example see van Zoonen et al. [2].

Miah et al. [3] explored a novel system for cross-lingual sentiment analysis utilizing transformers and Large Language Models in an ensemble architecture. The study focuses on the efficacy of pre-trained sentiment analysis models such as Twitter-RoBERTa-Base-Sentiment-Latest, BERT-base-multilingual-uncased-sentiment, and GPT-3 from OpenAI. A hybrid paradigm for examining comments on YouTube social media was presented by Jelodar et al. [4]. The framework utilizes semantic and sentiment analysis approaches to identify meaningful latent topics and levels of sentiment in user comments. It employs Latent Dirichlet Allocation (LDA) for topic modeling and VADER for sentiment analysis. The study of Albahli et al. [5] aims to analyze public sentiment about COVID-19 vaccinations by using “X” data and implementing a deep learning method. The research utilizes historical and real-time data obtained via web scraping from “X”, as well as the VADER sentiment analysis tool. The authors introduce a model as a solution to the shortcomings of previous studies. Badi et al. [8] conducted an extensive investigation of the sentiment analysis of “X” posts on COVID-19 vaccinations, with a special focus on AstraZeneca and Pfizer.

In order to explore the predictability of trends using recognized patterns over “X” social media sentiment analysis, Di Tollo et al. [7] employ a combination of stochastic neural networks, the BERT model as an NLP technique, and an external evolutionary algorithm to optimize parameters for robustly accurate predictions. According to Hameleers [9], while the digital media ecosystem has evolved into a crucial component of society and one of the most demanding communication tools, it also creates an environment subject to the strategic exploitation of platform architectures and algorithmic systems to magnify misleading content. These platforms and tech companies serve as crucial intermediaries that can either restrict or amplify disinformation through their algorithmic design and content moderation policies.

3. Data and Methods

3.1. Data

This paper is centered on the analysis of public opinion shared on X. Due to modifications in X policies in 2023, it has become impracticable to retrieve the most recent X posts. It is fortunate that research involving German occupations in the same year had already been conducted, as this thesis will build upon this previous study [10]. The X data from that study will be re-analysed, as it was the most recent data available at the time.

The data collection process was executed in accordance with the protocol established in the preceding study (see [10]). The process was initiated with the access of X's API. The dataset under consideration contains approximately 3.5 million tweets related to occupations, accompanied by supplementary metadata.

3.2. Methods

The initial step in this research will be the conceptualization and implementation of a variety of sentiment analysis tools, as well as their subsequent evaluation to determine the most effective tool. Subsequently, deep learning algorithms will be integrated with the previously mentioned sentiment analysis tools to assess the impact of deep learning models on the aforementioned tools. This analysis will identify the most effective tool for determining whether a job is essential or not.

Effectiveness of Sentiment Analysis Tools (RQ1) To address the first research question a process of conceptualizing and implementing the tools is needed. This process involves a multi-step approach, incorporating various sentiment analysis methods and addressing the challenges posed by multilingual data. Ultimately, to address finding the most effective tool, a comprehensive performance evaluation will be developed and conducted.

Impact of Deep Learning on Sentiment Analysis Tools (RQ2) This section examines the integration of deep learning models with previously introduced sentiment analysis tools in order to respond to the second research question. Studies show that deep learning models are indeed practical for both sentiment analysis and emotion classification tasks, and they can outperform conventional machine learning in general for these tasks. Particularly, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) variants have become increasingly popular for sentiment analysis tasks due to their state-of-the-art performance, as many studies have successfully applied these deep learning models for sentiment analysis (Kastrati et al., 2024). The endeavor focuses on harnessing the capabilities of both deep learning and machine learning techniques to enhance the accuracy and robustness of sentiment analysis.

4. Analysis and Evaluation

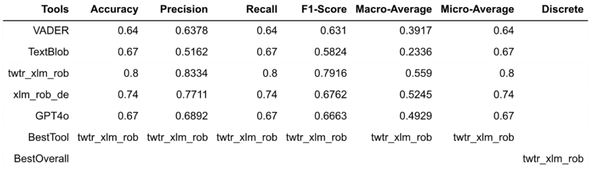
4.1. Evaluation Results of the Sentiment Analysis Tools (RQ1)

Measuring the performance of the sentiment analysis tools is the first mandatory step in analyzing and evaluating their outcomes. This is accomplished through the utilization of continuous measurements, including MSE, MAE, R-squared, and correlation coefficients, as well as discrete metrics, such as accuracy, precision, recall, and F1-score. For brevity, this paper only reports detailed results on the discrete metrics.[1]

A cumulative score is introduced to give a fair perspective on the best sentiment analysis tool overall. However, to experiment with the impact level and magnitude of the metrics, they are fed to the score individually and then at once, meaning that first, only the first three error metrics, R-squared and

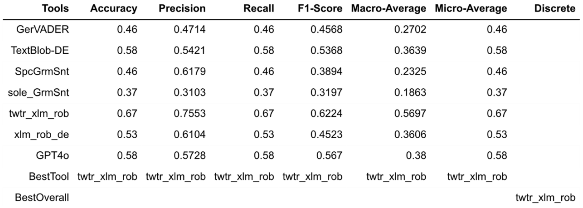
[1] Detailed results for the continuous metrics can be obtained via the authors.

Table 1
Discrete Evaluations of English Tools (RQ1)



Tools	Accuracy	Precision	Recall	F1-Score	Macro-Average	Micro-Average	Discrete
VADER	0.64	0.6378	0.64	0.631	0.3917	0.64	
TextBlob	0.67	0.5162	0.67	0.5824	0.2336	0.67	
twtr_xlm_rob	0.8	0.8334	0.8	0.7916	0.559	0.8	
xlm_rob_de	0.74	0.7711	0.74	0.6762	0.5245	0.74	
GPT4o	0.67	0.6892	0.67	0.6663	0.4929	0.67	
BestTool	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	
BestOverall							twtr_xlm_rob

Table 2
Discrete Evaluations of German Tools (RQ1)



Tools	Accuracy	Precision	Recall	F1-Score	Macro-Average	Micro-Average	Discrete
GerVADER	0.46	0.4714	0.46	0.4568	0.2702	0.46	
TextBlob-DE	0.58	0.5421	0.58	0.5368	0.3639	0.58	
SpcGrmSnt	0.46	0.6179	0.46	0.3894	0.2325	0.46	
sole_GrmSnt	0.37	0.3103	0.37	0.3197	0.1863	0.37	
twtr_xlm_rob	0.67	0.7553	0.67	0.6224	0.5697	0.67	
xlm_rob_de	0.53	0.6104	0.53	0.4523	0.3606	0.53	
GPT4o	0.58	0.5728	0.58	0.567	0.38	0.58	
BestTool	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	twtr_xlm_rob	
BestOverall							twtr_xlm_rob

correlation metrics were taken into account, and then they were included stepwise to reach a cumulative score. Interestingly, the best tool for both languages didn't change even after the inclusions.

For the English dataset, the initial evaluations showed that the Twitter-XLM-RoBERTa ('twtr_xlm_rob' model) consistently outperformed other tools across all continuous metrics where its superior performance can be attributed to its strong correlation with labels and lower error rates. The high correlation coefficients indicate a strong positive relationship between predicted and actual sentiment scores. The low error metrics suggest that Twitter-XLM- RoBERTa predictions are closer to the true labels compared to other tools. GPT4o showed the second-best performance in the continuous evaluation, though its performance was not as strong as Twitter-XLM-RoBERTa, it still demonstrated a good correlation with labels and relatively low error rates, outperforming the remaining tools.

For the German part, GPT4o emerged as the top performer in the continuous evaluation, achieving the highest scores in most metrics, which indicates its superior ability to predict sentiment scores that closely align with the true labels, showing both high correlation and low error rates. Despite trailing GPT4o by a small margin, the Twitter- XLM-RoBERTa displayed strong predictive capabilities, especially in minimizing absolute errors, and placed second in the continuous evaluation.

Similar to the continuous perspective, a cumulative score is calculated considering all key classification metrics, including accuracy, precision, recall, F1-score, macro-average, and micro-average, to have a thorough assessment of classification performance.

As with the continuous evaluation, the Twitter-XLM-RoBERTa again emerged as the top performer across all discrete metrics, indicating the tool's strength in classifying "X" posts into the correct sentiment categories, while its performance was consistent across different sentiment classes (Table 1). Due to the outstanding classification performance of XLM-RoBERTa-German, it outperformed the other tools in the discrete evaluation, yet not as strong as Twitter-XLM-RoBERTa, and thus earned the second-best ranking.

For German "X" posts (Table 2), the Twitter-XLM-RoBERTa again surpassed all other tools, demonstrating its ability to properly classify sentiments across multiple groups while maintaining a balanced performance in both precision and recall. Even while GPT4o was powerful and ranked second in the discrete examination, it was not as dominant as it was in the continuous evaluation. This revealed that while it excels in predicting sentiment scores on a continuous scale, its performance in discrete classification is good but not superior to Twitter-XLM-RoBERTa.

Focusing on the ROC curves, for the tools employed for English "X" posts, the Twitter- XLM-RoBERTa consistently outperforms the others, with the highest AUC values across different classes, ranging from 0.69 to 0.82 (a good to excellent performance). The XLM-RoBERTa-German and GPT4o came next, as

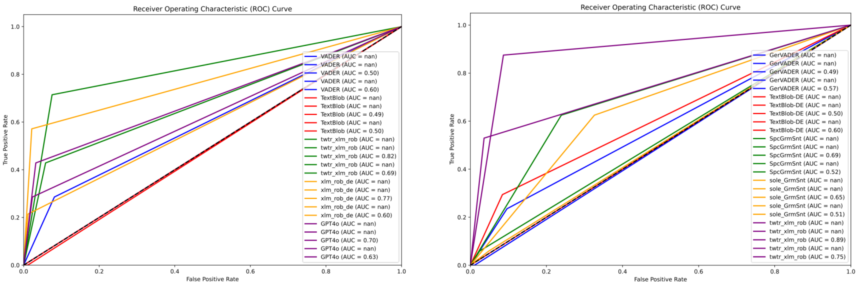
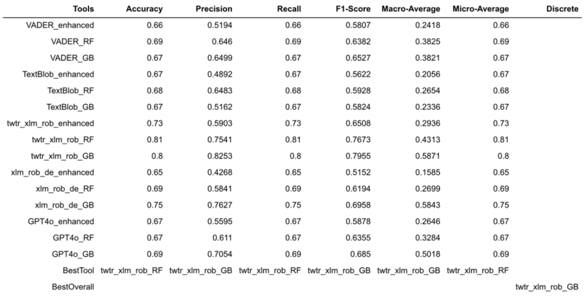


Figure 1: ROC Curves of English Tools (RQ1, left); ROC Curves of German Tools (RQ1, right)

Table 3
Discrete Evaluations of Enhanced English Tools (RQ2)



Tools	Accuracy	Precision	Recall	F1-Score	Macro-Average	Micro-Average	Discrete
VADER_enhanced	0.66	0.5194	0.66	0.5807	0.2418	0.66	
VADER_RF	0.69	0.646	0.69	0.6382	0.3825	0.69	
VADER_GB	0.67	0.6499	0.67	0.6527	0.3821	0.67	
TextBlob_enhanced	0.67	0.4892	0.67	0.5622	0.2056	0.67	
TextBlob_RF	0.68	0.6483	0.68	0.5928	0.2654	0.68	
TextBlob_GB	0.67	0.5162	0.67	0.5824	0.2336	0.67	
twtr_xlm_rob_enhanced	0.73	0.5903	0.73	0.6508	0.2936	0.73	
twtr_xlm_rob_RF	0.81	0.7541	0.81	0.7673	0.4313	0.81	
twtr_xlm_rob_GB	0.8	0.8253	0.8	0.7955	0.5871	0.8	
xlm_rob_de_enhanced	0.65	0.4268	0.65	0.5152	0.1585	0.65	
xlm_rob_de_RF	0.69	0.5841	0.69	0.6194	0.2699	0.69	
xlm_rob_de_GB	0.75	0.7627	0.75	0.6958	0.5843	0.75	
GPT4o_enhanced	0.67	0.5595	0.67	0.5878	0.2646	0.67	
GPT4o_RF	0.67	0.611	0.67	0.6355	0.3284	0.67	
GPT4o_GB	0.69	0.7054	0.69	0.685	0.5018	0.69	
BestTool	twtr_xlm_rob_RF	twtr_xlm_rob_GB	twtr_xlm_rob_RF	twtr_xlm_rob_GB	twtr_xlm_rob_GB	twtr_xlm_rob_RF	
BestOverall							twtr_xlm_rob_GB

both tools demonstrated good to moderate performances, though their performances were not even comparable to Twitter-XLM-RoBERTa. Regarding German “X” posts Twitter- XLM-RoBERTa showed the best performance with an AUC of 0.89 and the ROC curve consistently above all others, indicating superior discrimination ability across all thresholds. The second-best performer with an AUC of 0.75 was GPT4o, also exhibiting good discriminative power.

4.2. Evaluation Results of the Enhanced Tools (RQ2)

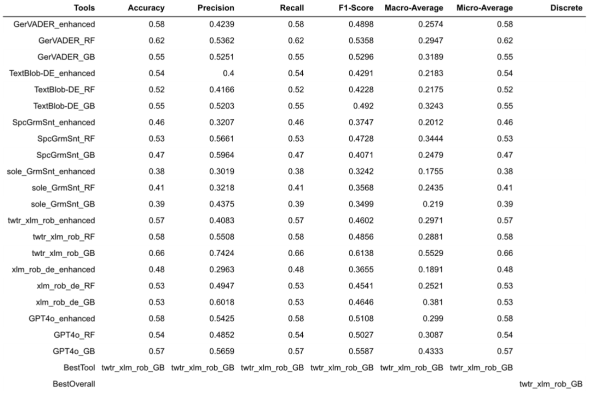
In the next step the results of the most efficient model are evaluated to ensure they are accurate, besides realizing the opportunity to compare it to the results of RQ1. The relative performance of the enhanced sentiment analysis tools is therefore evaluated using the same metrics as in section 4.1, with the same continuous and discrete metrics.

The continuous evaluation results show varying performance across enhanced sentiment analysis tools and their variants. The gradient-boosting (twtr_xlm_rob_GB) enhancement of the Twitter-XLM-RoBERTa and GPT4o models demonstrate superior performance in capturing the nuances of sentiments on a continuous scale across various metrics and evaluation approaches, indicating that both models are effective in predicting sentiment scores accurately, and consistently outperforming other tools across various metrics. Moreover, it demonstrates that the ensemble variant of each tool tends to perform better than the base-enhanced counterparts using the CNN-LSTM model.

Regarding the discrete viewpoint (Table 3), the gradient-boosting form of the enhanced Twitter-XLM-RoBERTa along with the random forest (twtr_xlm_rob_RF) variant consistently outperforms other tools in discrete classification on almost all performance metrics, followed by the gradient-boosting version of the XLM-RoBERTa-German model which shows competitive performance in terms of accuracy and other metrics.

The results of the ROC curves and AUC values, as evident in Figure 2, show that the enhanced Twitter-XLM-RoBERTa with both random forest and gradient-boosting ensemble emerges as the overall best tool for sentiment analysis, demonstrating superior performance across all sentiment classes. Their ability to distinguish between different sentiment classes is significantly better than other models, as

Table 4
Discrete Evaluations of Enhanced German Tools (RQ2)



Tools	Accuracy	Precision	Recall	F1-Score	Macro-Average	Micro-Average	Discrete
GerVADER_enhanced	0.58	0.4239	0.58	0.4898	0.2574	0.58	
GerVADER_RF	0.62	0.5362	0.62	0.5358	0.2947	0.62	
GerVADER_GB	0.55	0.5251	0.55	0.5296	0.3189	0.55	
TextBlob-DE_enhanced	0.54	0.4	0.54	0.4291	0.2183	0.54	
TextBlob-DE_RF	0.52	0.4166	0.52	0.4228	0.2175	0.52	
TextBlob-DE_GB	0.55	0.5203	0.55	0.492	0.3243	0.55	
SpcGrmSnt_enhanced	0.46	0.3207	0.46	0.3747	0.2012	0.46	
SpcGrmSnt_RF	0.53	0.5661	0.53	0.4728	0.3444	0.53	
SpcGrmSnt_GB	0.47	0.5964	0.47	0.4071	0.2479	0.47	
sole_GrmSnt_enhanced	0.38	0.3019	0.38	0.3242	0.1755	0.38	
sole_GrmSnt_RF	0.41	0.3218	0.41	0.3568	0.2435	0.41	
sole_GrmSnt_GB	0.39	0.4375	0.39	0.3499	0.219	0.39	
twtr_xlm_rob_enhanced	0.57	0.4083	0.57	0.4602	0.2971	0.57	
twtr_xlm_rob_RF	0.58	0.5508	0.58	0.4856	0.2881	0.58	
twtr_xlm_rob_GB	0.66	0.7424	0.66	0.6138	0.5529	0.66	
xlm_rob_de_enhanced	0.48	0.2963	0.48	0.3655	0.1891	0.48	
xlm_rob_de_RF	0.53	0.4947	0.53	0.4541	0.2521	0.53	
xlm_rob_de_GB	0.53	0.6018	0.53	0.4646	0.381	0.53	
GPT4o_enhanced	0.58	0.5425	0.58	0.5108	0.299	0.58	
GPT4o_RF	0.54	0.4852	0.54	0.5027	0.3087	0.54	
GPT4o_GB	0.57	0.5659	0.57	0.5587	0.4333	0.57	
BestTool	twtr_xlm_rob_GB	twtr_xlm_rob_GB	twtr_xlm_rob_GB	twtr_xlm_rob_GB	twtr_xlm_rob_GB	twtr_xlm_rob_GB	
BestOverall							twtr_xlm_rob_GB

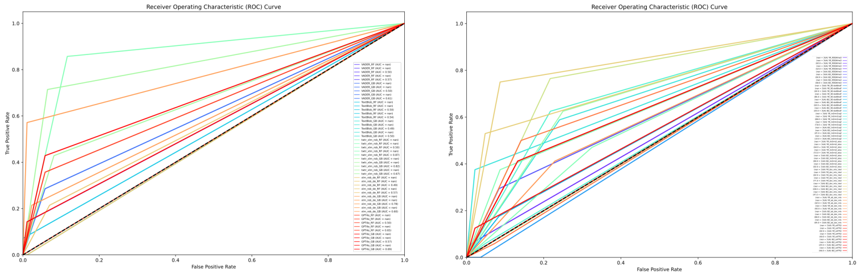


Figure 2: ROC Curves of English Tools (RQ2, left); ROC Curves of German Tools (RQ2, right)

evidenced by the consistently showing the highest AUC values across different classes, often above 0.80, and curves furthest from the diagonal line. The GPT4o families exhibit competitive performance, particularly for certain sentiment classes.

The GPT4o model with both gradient-boosting and random forest enhancements performs best across all metrics in the continuous evaluation. Technically, it shows the highest correlation coefficients and the best error metrics – lowest MSE, MAE, and RMSE. While the Twitter-XLM-RoBERTa model with gradient boosting was not as strong as the GPT4o models, it still shows good performance, particularly in terms of Pearson correlation and error metrics.

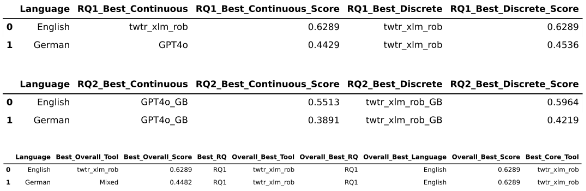
In the discrete evaluation (Table 4), the Twitter-XLM-RoBERTa model with gradient boosting performs best overall, demonstrating particularly strong precision and a good balance between precision and recall, as reflected in its high F1-Score. The same model with base enhancement comes next, with slightly lower but still impressive scores. The GPT4o model with gradient boosting is the third-best performer in this category, indicating strong performance, particularly in terms of F1-Score and Macro-Average.

Based on the ROC curves shown in Figure 2 and related AUC values, the best tool appears to be Twitter-XLM-RoBERTa with gradient boosting with ROC curves consistently above others for most classes and AUC values close to 0.8, indicating strong discriminative power across all sentiment classes. The ROC curves for the gradient- boosting type of GPT4o are also among the top performers, with considerable AUC values, showcasing decent performance for some classes. The Twitter XLM-RoBERTa model with base enhancement also shows competitive performance, with ROC curves often close to the top two models.

4.3. Summary

In order to discover the best model overall, the results of the first research question regarding the evaluations of the sentiment analysis tools without enhancement from external deep learning and machine learning models will be called ERQ1. They are compared to the evaluation results of the

Table 5
Top-Performer Between Base and Enhanced Tools



	Language	RQ1_Best_Continuous	RQ1_Best_Continuous_Score	RQ1_Best_Discrete	RQ1_Best_Discrete_Score
0	English	twtr_xlm_rob	0.6289	twtr_xlm_rob	0.6289
1	German	GPT4o	0.4429	twtr_xlm_rob	0.4536

	Language	RQ2_Best_Continuous	RQ2_Best_Continuous_Score	RQ2_Best_Discrete	RQ2_Best_Discrete_Score
0	English	GPT4o_GB	0.5513	twtr_xlm_rob_GB	0.5964
1	German	GPT4o_GB	0.3891	twtr_xlm_rob_GB	0.4219

	Language	Best_Overall_Tool	Best_Overall_Score	Best_RQ	Overall_Best_Tool	Overall_Best_RQ	Overall_Best_Language	Overall_Best_Score	Best_Core_Tool
0	English	twtr_xlm_rob	0.6289	RQ1	twtr_xlm_rob	RQ1	English	0.6289	twtr_xlm_rob
1	German	Mixed	0.4482	RQ1	twtr_xlm_rob	RQ1	English	0.6289	twtr_xlm_rob

enhanced sentiment analysis tools (here ERQ2) to reach a thorough examination of the second research question. This will be done by investigating the results of both Continuous and Discrete aspects of both English and German tools, which would strengthen the quality of the examination.

For English, the best overall tool comes from ERQ1, namely the Twitter-XLM-RoBERTa model. This tool achieved the highest evaluation score, outperforming its enhanced variants in ERQ2. Notably, it is the best performer for both continuous and discrete metrics in ERQ1, demonstrating its robustness across different evaluation criteria. The German language analysis revealed a more complex picture, as the best overall performance came from ERQ1, but with a mixed result. In fact, GPT4o performed best for continuous metrics, while Twitter-XLM-RoBERTa excelled in discrete metrics.

The analysis reveals that the best overall tool across both languages is Twitter-XLM- RoBERTa, which demonstrates no significant improvements over the baseline sentiment analysis tools (Table 5). The analysis of core tools further proves that it is indeed the best overall core tool due to appearing with the maximum number of 5 times among the top-performing models in both base and enhanced CNN-LSTM with machine learning ensemble models across both languages.

5. Conclusions and Outlook

The Twitter-XLM-RoBERTa base model ultimately emerged as the best sentiment analysis tool due to demonstrating superior performance overall in both English and German datasets, highlighting the effectiveness of a robust pre-trained language model and transformer- based architecture and enabling effective cross-lingual sentiment detection. In a next step, a comprehensive assessment framework was implemented to determine whether or not jobs can be predicted as being Essential Jobs. The approach entailed utilizing the elected Twitter- XLM-RoBERTa model to conduct sentiment analysis on the "X" posts for all available datasets and aggregate the outcomes for each unique job title to use as a predictor of job essentiality. Two different methods were introduced to assess the essentiality of job titles: The *Berlin Method* solely focuses on the comparison of the test subject to the corresponding ground truth of the initial list of essential jobs. The *Temporal Method* compares the changes to the list of essential jobs over the pandemic's duration to the initial list on the same dataset. And *Both Methods* which is the combination of the two methods, acting as a meta-classifier of the essentiality assessment mechanism, to make the final decisions based on the results of the other two methods.

The findings obtained from the final assessment reveal a significant degree of complexity and variability, underscoring the challenges inherent in sentiment-based assessments, specifically in less-popular and intricate contexts such as job essentiality. The following content offers an in-depth interpretation of the findings:

Berlin Method: The Berlin Method exhibited a consistent trend across both languages, with the majority of unique job titles of both English and German versions of the Extended List being categorized as No-Data, suggesting a potential bias or overgeneralization in the classification process. The next highest proportion of jobs were allocated to the Essential category, with about half of the No-Data amount, highlighting the permissible performance of the method in discovering Essential Jobs. The overall outcome of the Berlin Method may reflect the inherent limitations of the ground truth dataset or the aggregation mechanism, which could have failed to capture delicate variations in sentiment for certain job titles.

Temporal Method The Temporal Method results revealed greater variability between the two languages. While the Extended List predominantly classified jobs as No- Data in both languages, the rest of the German dataset jobs revealed somewhat similar trends between the rest of the classes, while the English dataset classified nearly no job as essential. This disparity could be attributed to restricted classification criteria, variations in the temporal distribution of the datasets, differing patterns of sentiment evolution across languages, or the impact of language-specific nuances on sentiment scoring. The relatively balanced distribution across the remaining classes in both languages otherwise may suggest that the Temporal Method failed to capture a broader range of sentiment dynamics compared to the Berlin Method.

Both Methods The "Both Methods" approach demonstrated significant diversity in categorization between Essential and Non-Essential classes. The No-Data category appeared to be the dominant class as in most of the previous cases, for the English Added dataset, almost one-third of the remaining items were classified as Essential and a limited amount of the jobs to Non-Essential, while the German dataset exhibited equal job classification in the Essential and Non-Essential categories. This suggests that the integration of Berlin and Temporal Methods, while robust in combining complementary perspectives and resulting in depicting a better classifications, may still be influenced by the inherent biases or limitations of each approach. Notably, the excess of No-Data besides variation of Non-Essential classifications between languages underscores the difficulty of definitively identifying job essentiality through sentiment analysis alone.

This research has made significant contributions to understanding the dynamics of Essential Job classifications during the COVID-19 pandemic through advanced sentiment analysis of social media data and highlights both methodological advances and significant analytical challenges. While Twitter-XLM-RoBERTa has emerged as the superior sentiment analysis tool within all competitors on both the English and German datasets, the results indicate that adding deep learning enhancements unexpectedly failed to surpass baseline performance. The new dual-stream comparative evaluation approach, which merges Berlin-based reference analysis with temporal sentiment evolution, sheds light on the intricate nature of Essential Job classifications. Significant methodological constraints, including concerns with data quality, language variation, and threshold setting, are highlighted through the considerable proportion of classifications becoming No-Data and substantial distinctions between the outcomes of English and German datasets. While the established framework seems promising in analyzing job essentiality through sentiment analysis, its practical application faces considerable constraints, particularly in predicting prospective Essential Jobs. Future developments ought to focus on enhanced job title grouping strategies, refined multilingual model adaptation, and more sophisticated comparative analysis design to minimize classification ambiguity. Despite these limitations, this research establishes a foundational framework for investigating job essentiality through sentiment analysis during global crises, providing valuable insights into the intersection of public sentiment and social policy determination.

In this study it was decided not to concentrate on job title pre-processing, separation, grouping, but instead execute just the minimal amount of data cleaning and pre-processing. This decision ended up with the study's conclusion being based on raw job titles, leading to significant job title variation. Future research could investigate how prediction of job essentiality could be improved by more rigorous data pre-processing and grouping of job titles. Filtering jobs and focusing on those with a minimum amount of tweets related to them could also prove interesting, as well as job-specific sentiment analysis instead of aggregated sentiments for groups of jobs. Lastly, fine-tuning for multilingual contexts could also improve results.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Tiemann, S. Udelhofen, L. Fournier, What social media can tell us about essential occupations, in: INFORMATIK 2023 - Designing Futures: Zukünfte gestalten, Gesellschaft für Informatik e.V., Bonn, 2023, pp. 1983–1992. doi:10.18420/inf2023_198.

[2] W. van Zoonen, R. E. Rice, C. L. Ter Hoeven, Sensemaking by employees in essential versus non-essential professions during the covid-19 crisis: A comparison of effects of change communication and disruption cues on mental health, through interpretations of identity threats and work meaningfulness, Management Communication Quarterly 36 (2022) 318–349.

[3] M. S. U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, M. Mridha, A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm, Scientific Reports 14 (2024) 9603.

[4] H. Jelodar, Y. Wang, M. Rabbani, S. B. B. Ahmadi, L. Boukela, R. Zhao, R. S. A. Larik, A nlp framework based on meaningful latent-topic detection and sentiment analysis via fuzzy lattice reasoning on youtube comments, Multimedia Tools and Applications 80 (2021) 4155–4181.

[5] S. Albahli, M. Nawaz, Tsm-cv: Twitter sentiment analysis for covid-19 vaccines using deep learning, Electronics 12 (2023) 3372.

[6] M. Kastrati, Z. Kastrati, A. Shariq Imran, M. Biba, Leveraging distant supervision and deep learning for twitter sentiment and emotion classification, Journal of Intelligent Information Systems 62 (2024) 1045–1070.

[7] G. di Tollo, J. Andria, G. Filograsso, The predictive power of social media sentiment: Evidence from cryptocurrencies and stock markets using nlp and stochastic anns, Mathematics 11 (2023) 3441.

[8] H. Badi, I. Badi, K. El Moutaouakil, A. Khamjane, A. Bahri, Sentiment analysis and prediction of polarity vaccines based on twitter data using deep nlp techniques, Radioelectronic and Computer Systems (2022) 19–29.

[9] M. Hameleers, Disinformation as a context-bound phenomenon: Toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination, Communication Theory 33 (2023) 1–10.

[10] A. Vahdatnia, D. Peachkah, Monitoring of digitization and sustainability on twitter, in: INFORMATIK 2023 - Designing Futures: Zukünfte gestalten, Gesellschaft für Informatik e.V., Bonn, 2023, pp. 1973–1982. doi:10.18420/inf2023_197.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.