Using GPT-4 for Text Analysis: Insights from English and German Language News Classification Tasks

Viktor Suter, Miriam Meckel

Institute for Media and Communication Management, University of St. Gallen, St. Gallen, Switzerland
viktor.suter@unisg.ch, miriam.meckel@unisg.ch

Abstract

Large language models are rapidly becoming an essential tool for social scientists. In particular, they have the potential to completely change the way researchers approach text analysis. In this study, we use the GPT-4 model to classify the content of newspaper articles and assess their sentiment. To do this, we collect headlines and leads from U.S. and German newspapers (n = 1,629) on how generative AI is represented in major news media outlets in both countries and inductively develop coding instructions based on this data. We then feed the data and instructions to GPT-4 and a human coder to compare their outputs and assess validity. We find that the coding procedure is highly reliable, with substantial to near perfect agreement between the human coder and GPT-4. We also find that reliability decreases for more complex constructs and is modestly lower for classification tasks performed in German than in English. Based on this analysis, we argue that LLMs offer powerful new approaches to text analysis that cross methodological divides between qualitative and quantitative approaches to empirical text analysis.

Introduction

Content analysis is a widely adopted research method in communication studies and the social sciences in general. In qualitative content analysis, people read texts and interpret their meaning, while quantitative approaches often use computer technology to extract meaning from texts (Brady 2019; Lazer and Radford 2017). Computational tools include a heterogeneous variety of approaches. They range from dictionary-based techniques that match words in texts to predefined lists to assign texts to categories (e.g., Lind et al. 2019) to statistically complex machine learning tools such as BERT models that can be used, among other things, to detect hate speech in online conversations (e.g., Mozafari, Farahbakhsh, and Crespi 2020). Computational methods are evolving rapidly and often reflect the state-of-the-art in social science research. They too have been invaluable in analyzing the vast amounts of textual data available in digital form that are not amenable to qualitative methods due to their sheer volume.

Although computational methods have proven extremely useful, they have drawbacks. Researchers who opt for machine-dependent methods sacrifice a degree of transparency and control. Unless researchers have a high level of technical expertise, they may struggle to understand how algorithms operationalize the properties they purport to measure (Kantner and Overbeck 2020). In addition, computationally measured properties often map only vaguely onto existing social science theories or concepts (Baden et al. 2022; Nguyen et al. 2020). Validating machine-generated outputs is therefore crucial (Grimmer and Stewart 2013). Yet, validation frequently involves ex post procedures, which pose their own challenges in terms of data interpretation and biases (Baden et al. 2022). Another limitation is that advanced computational methods require manually labeled training data. To create training datasets, researchers need to hire, train, and coordinate annotators, often consisting of student coders (Gilardi, Alizadeh, and Kubli 2023). Acquiring training data can therefore be costly for a researcher in terms of time and money.

In this paper, we argue that the introduction of large language models (LLMs), such as ChatGPT, offers powerful new approaches to text analysis. Fresh research has already pointed to the considerable potential of LLMs for the social sciences (Bail 2023). LLMs have been shown to accurately estimate the ideological leanings of politicians (Wu et al., 2023), effectively emulate public opinion in silicon samples (Argyle et al. 2023), aid in text analysis tasks, such as annotating tweets (Gilardi, Alizadeh, and Kubli 2023; Heseltine and Clemm Von Hohenberg 2023), and to meaningfully augment social science research pipelines (Ziems et al. 2024). Our analysis contributes to this field of research by illustrating the benefits and usefulness of LLMs for content analysis. We build on a new sample of 1,629 headlines and leads published on U.S. and German newspaper websites over a one-year period. We inductively develop a codebook that contains seven topic and one sentiment classification tasks and submit these to GPT-4 to classify the data. We manually code a portion of the data and compare this subset with the output from the GPT-4 model to validate the results. We observe substantial agreement between the manually and computationally coded outputs across tasks and languages. By way of this study, we show that LLMs address some of the aforementioned shortcomings of existing text analysis approaches. Additionally, we aim to demonstrate that LLMs are not only potent but also user-friendly tools, accessible

to researchers with limited knowledge of computational approaches. This accessibility makes the boundaries between quantitative and qualitative methods more permeable, offering ample opportunity for cross-fertilization.

The remainder of the paper is organized as follows: The next section details the data collection and analysis procedures, specifically explaining the development of the prompts in the codebook and how we passed them to GPT-4 via the OpenAI API. In the results section, we first present the validation metrics and then move on to briefly examine the thematic findings of the content analysis. Finally, in the discussion, we consider how LLMs are changing textual analysis, particularly at the intersection of qualitative and quantitative methods.

Data and Methods

We conducted the analysis on a set of articles published by U.S. news outlets (specifically, The New York Times, The Wall Street Journal, The Los Angeles Times, USA Today, and The Washington Post) and German news outlets (specifically, Frankfurter Allgemeine, Die Zeit, Süddeutsche Zeitung, Die Welt, and Bild) from September 2022 to August 2023. All these outlets have large readerships in their respective countries. We created the dataset with an eye to how generative AI is represented in these major news media outlets in both countries. Although we are primarily interested in the methodological implications of our research design, generative AI as topic adds a timely and relevant backdrop to our analysis. The dataset was compiled by querying the search engines on the outlet-specific websites for the terms 'generative AI,' 'large language model,' 'chatbot,' 'GPT,' 'automated text generation,' and 'automated image generation.' These search terms were translated into German to query German news sources. For each article in the search results, we scraped the publicly available publication date, the title, the lead, and its URL employing the BeautifulSoup python package. After the initial data retrieval, we filtered the data to the set that falls within the required time period and then conducted a manual review of each article's headline and lead to ensure the relevance and alignment with the thematic focus on generative AI. In total, this led to a dataset that includes 874 articles from the US and 775 articles from Germany. The raw data is available at www.github.com/viktor-suter/-News-Classification-GPT4.git.

Figures 1 and 2 present some descriptive statistics that summarize the main features of the dataset. Figure 1 offers an overview of the total number of articles by outlet and country. Furthermore, Figure 2 depicts the weekly variation in the number of articles published, segmented by country, and includes a trend line. This trend line shows a steep increase in media coverage following the public release of the ChatGPT chatbot in late November 2022, a trend that is evident in both German and US media. The data also indicates that as of August 2023, media coverage continues to mount in Germany, while in the US it begins to taper off after a period of heightened attention.

Germany		USA	
Website	Count	Website	Count
Frankfurter Allgemeine	238	New York Times	287
Die Zeit	174	Wall Street Journal	278
Süddeutsche Zeitung	146	Los Angeles Times	109
Die Welt	133	USA Today	103
Bild	64	Washington Post	97
Total	755	Total	874

Figure 1: Article Counts by German and US News Outlets

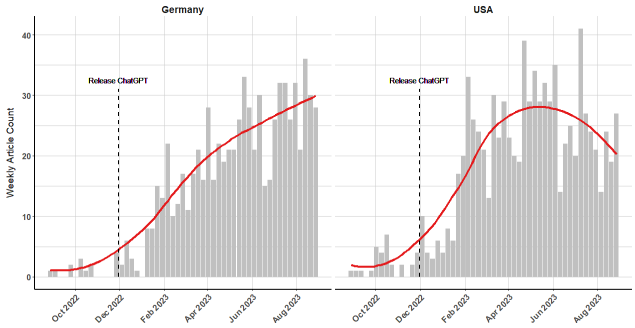

Figure 2: Trends in Weekly Article Counts in Germany and the USA

Developing Classification Tasks

Our analysis aimed to determine whether a headline and its corresponding lead were relevant to a set of seven topics and to identify the sentiment expressed. To achieve this, we developed a codebook with detailed definitions and instructions for classifying the data. The topics covered include existential risk, misinformation, regulation, work, creative industries, education, and the stock market. The sentiment could be classified as positive, negative, or neutral. Though the choice of these classifications was inspired by prior research on the public perception of AI (e.g., Cave, Coughlan, and Dihal 2019; Fast and Horvitz 2017), the codes were developed inductively, that is, by manually reading and capturing key constructs contained in the text corpus. This process of the formulation of instructions, called prompt engineering in the context of LLMs, bears a close resemblance to the development of a codebook in qualitative research. Both aim to standardize the understanding of coded constructs and their conceptual boundaries among different coders, as well as to define rules for interpretation and provide contextual cues.

In developing the coding instructions, we followed Törnberg's (Törnberg 2023) guidelines. Separate instructions were defined for each topic and sentiment classification. These instructions defined the information to be extracted from the text, relevant contextual details, and the expected response pattern. The topics were implemented as binary constructs, i.e., if a text was relevant to a topic, it was coded as '1', otherwise as '0'. The sentiment construct, on the other hand, could assume three values: '0' for negative, '1' for neutral, and '2' for positive. This approach resulted in a general template for the implementation of the classifi-

cation tasks. For example, the instruction for the work topic was realized as follows:

> "Classify the following text as either relevant ('1') or not relevant ('0') to the topic of 'Work', focusing on the impact of Artificial Intelligence (AI) on labor relations and employment. A text is relevant if it addresses any of the following issues: AI making human jobs easier, AI freeing humans from repetitive and boring tasks, AI and automation relegating humans to menial jobs, AI leading to the displacement of human jobs, the potential of AI to cause large-scale unemployment. Respond with '1' for relevant texts or '0' for texts that do not relate to these criteria. Do not provide any explanation for your choice." (see Appendix A.1.2)

The instructions were first developed in English based on a manual reading of both the American and German datasets. After validation (discussed below), the English instructions were translated into German. Full versions of both the English and German classification instructions, outlining the specific coding guidelines, can be found in the Appendix (Sections A.1 and A.2, respectively). We then used the OpenAI API to interact with the gpt-4 model, prompting it with instructions from the codebook, feeding it our data, and col- lecting its responses. For readers unfamiliar with these steps, Törnberg (Törnberg 2023) provides a comprehensive and easy-to-follow guide to setting up access to the API, load- ing the data, and running the prompts.

Validation

To ensure that we capture valid constructs, we tested several variations of the instructions on a small sample of 30 observations each from the US and German datasets. We performed a separate run for each topic, sentiment, and language, providing the same set of instructions to both a human coder (the first author of this paper) and GPT-4. Between each coding run with GPT-4, we initiated a new session to make sure that it did not retain any memory of past interactions, ensuring that its classifications were indeed zero-shot. By comparing the outputs of the human coder and GPT-4, we were able to examine cases of disagreement, refine our understanding of the constructs we were seeking to measure, and tweak the instructions accordingly. After several iterations, we settled on the set of instructions detailed in the Appendix. We then applied these instructions to the entire dataset. At the same time, we extracted a 10% non-replacement random sample from both the German (n=76) and American (n=87) datasets, which were manually classified by the human coder. We used this subset of data to compute the validation metrics, including percentage agreement - the percentage of cases where GPT-4 and the human coder assigned the same category - and Cohen's Kappa. Cohen's Kappa is a more robust measure of intercoder agreement than percentage agreement because it accounts for chance agreement between coders. The outcome of this validation process is presented in the Results section.

Results

The results focus on two main aspects: first, the validation metrics, and second, the thematic insights. Our main focus is on the validation metrics, which establish the accuracy and reliability of the classification tasks. The thematic findings are presented primarily as an additional plausibility check of the coding procedure, but they also offer some interesting insights into the media discourse on generative AI.

Validation Metrics

Tables 1 and 2 below show the percentage agreement and Cohen's Kappa statistics for the German and U.S. samples, respectively, that we randomly selected from our data. The tables provide a comparative view of the intercoder reliability for the constructs included in our classification tasks. As mentioned earlier, work, education, regulation, misinformation, stock market, creative industries, and existential risk are binary categories, whereas sentiment is rated on a ternary scale. Beginning with the German news articles, percentage agreement is high for all constructs, ranging from 81.5% for sentiment to 98.6% for education. Cohen's Kappa values range from 0.661 for existential risk, suggesting substantial agreement, to 0.916 for stock market, indicating almost perfect agreement. In the US sample, percentage agreement is high across constructs as well, ranging from 88.5% to 98.8%. Cohen's Kappa values are also relatively high ranging from 0.788 for existential risk to 0.946 for the stock market. Generally, the overall high percentage agreement and good Cohen's Kappa values suggest that the coding procedure is reliable and that the GPT-4 model produces results consistent with the human coder.

Although the coding procedure is generally reliable, three aspects stand out: the existential risk construct, the sentiment construct, and language differences. First, the existential risk construct shows high percentage agreement across samples, but Cohen's Kappa is comparatively low. This discrepancy likely results from the rare occurrence of existential risk in the datasets (3.7% in the U.S. and 8.1% in Germany, as detailed in Thematic Insights), leading to an imbalance that reduces Kappa. The low Kappa value may therefore not indicate poor intercoder reliability, but rather the statistical challenge of a low-frequency category. Second, the sentiment construct has both comparatively low percentage agreement and Cohen's Kappa. This may be due to the ternary nature of the construct, which increases coding complexity as coders have one more option to consider compared to binary constructs. In addition, the coding of sentiment involves a strong element of subjective judgment. These aspects are likely to contribute to lower scores. Nevertheless, the Kappa values (0.707 for Germany and 0.812 for the US) still indicate substantial to almost perfect agreement. Third and finally, consider the differences between languages. While percentage agreement is consistently high in both samples, Cohen's Kappa is slightly lower for German constructs, and this difference becomes larger for existential risk and sentiment. This suggests that issues of complexity and infrequency of constructs may be amplified by a slightly poorer performance of GPT-4 in German than in English.

Table 1: Validation Metrics for the German Sample

Constructs	Pct. Agr.	Cohen's Kappa
Work	97.3%	0.860
Education	98.6%	0.850
Regulation	96.0%	0.834
Misinformation	97.3%	0.874
Stock Market	98.6%	0.916
Creative Industries	97.3%	0.843
Existential Risk	98.6%	0.661
Sentiment	81.5%	0.707

Table 2: Validation Metrics for the US Sample

Constructs	Pct. Agr.	Cohen's Kappa
Work	96.5%	0.869
Education	98.8%	0.903
Regulation	94.2%	0.813
Misinformation	98.8%	0.903
Stock Market	98.8%	0.946
Creative Industries	94.2%	0.822
Existential Risk	97.7%	0.788
Sentiment	88.5%	0.812

Thematic Insights

After testing the validity of our coding instructions on the subsample, we extended them to the full dataset. Figure 3 shows bar charts detailing the frequency of different constructs and their sentiment distributions. The data show parallel trends in the coverage of generative AI in both countries. Regulation, work, and creative industries emerge as the most discussed topics, though their order differ between countries. While topics at the lower end of the prevalence spectrum are closely clustered in both datasets, we note a marked difference in the representation of education, which accounts for 10.2% in the German dataset but only 5.9% in the US dataset. Existential risk is rarely discussed in both datasets, but less so in the US at 3.7% compared to 8.1% in Germany. The sentiment analysis reveals a general tendency towards neutrality in both countries, though this tendency is more pronounced in the US, while at the same time the proportion of negative sentiment is higher in the German context. Positive sentiment is at very similar levels between the two countries. In addition, a closer look at the sentiment within specific topics shows that existential risk, misinformation, and regulation are overwhelmingly seen in a negative light in either country. Education, work, and creative industries show a more balanced sentiment profile for both countries. Notably, the German data reflect a much more positive view of the stock market than the U.S. data, where sentiment is more balanced, although it remains the most positively viewed issue in the U.S.

These thematic findings are broadly consistent with general expectations based on the nature of the constructs. For instance, misinformation, associated with the spread of false or fabricated information, tends to foster distrust and confusion. Discussions of existential risks highlight concerns about the long-term negative impact of AI on human well-

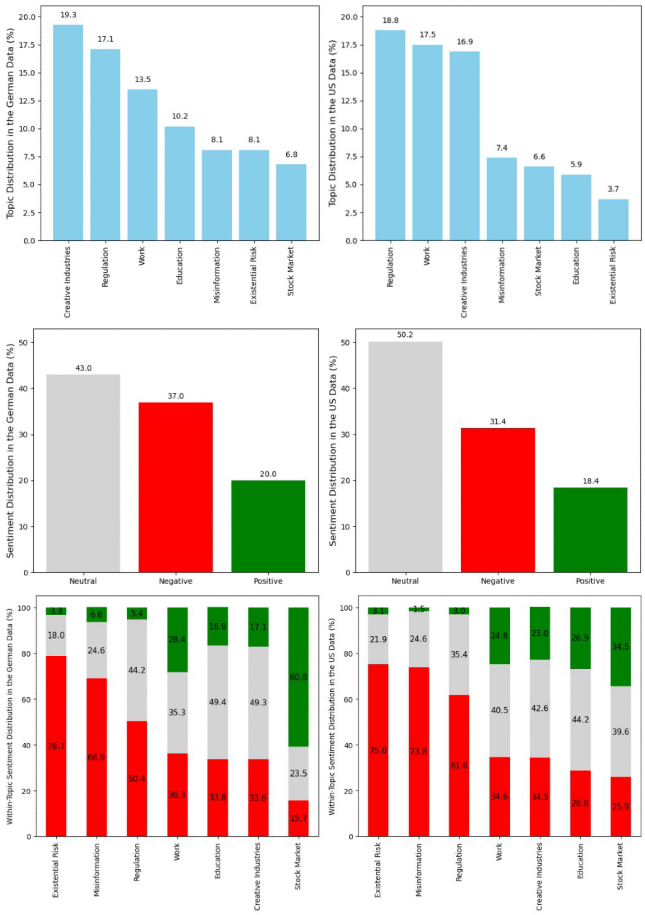

Figure 3: Bar Charts Showing Topic Prevalence and Sentiment Distribution in Newspaper Headlines and Leads about Generative AI from Germany (left) and the United States (right) for September 2022 - August 2023

being, while regulation often addresses social problems created by AI, such as privacy and ethical issues. These areas are inherently associated with uncertainty, risk, and challenge, and are therefore likely to generate negative sentiment. Similarly, the data collection period of this study coincided with a bull market in technology shares, which explains the predominantly positive stock market coverage. Though no substitute for systematic validation, linking findings to market and societal trends underscores the plausibility of the insights derived from the coding approach presented.

Discussion and Concluding Remarks

In this study, we demonstrate the potential of LLMs for text analysis by showing that GPT-4 produces accurate text classifications that are consistent with those of human coders. This has potentially crucial implications for empirical social science text analysis. Traditionally, the analysis of large text corpora has required considerable coding and computational skills. While these skills remain important, LLMs like GPT-4 significantly lower the barrier, allowing researchers to op-

erationalize constructs using natural language rather than relying solely on complex computer coding or opaque off-the-shelf algorithms. This approach may be more intuitive and accessible to many researchers, especially those who prefer qualitative research designs. Furthermore, while qualitative methods excel at in-depth text analyses, they are often constrained to small-n samples and criticized for limited reliability. By incorporating LLMs into research designs, it is possible to develop a set of codes based on theory or inductively, as demonstrated in this study, and then apply these codes to a wider range of texts. This not only enhances scale but can be used to complement a research process of qualitative reading and analysis.

The results of this study are encouraging, but there are certain caveats to the method presented here. First, we observed that while GPT-4 effectively handles precisely specified categorizations, its reliability seems to decrease as constructs become more complex. Second, we noted poorer validation metrics when GPT-4 coded German as opposed to English language constructs. Although these performance differences were small, it may be the case that other languages result in greater performance disparities. Testing the performance of GPT-4 or other LLMs in non-Germanic language contexts provides an opportunity for further exploration. Third, it is critical to note the role that human effort played in achieving high agreement rates. The iterative process of prompt development and strategy adjustment was essential to bring GPT-4's performance up to human standards. While the LLM can substantially augment the coding process, it is not a direct replacement for human coders. Future studies should investigate the circumstances under which the LLM route is beneficial and when manual coding might be more efficient. While GPT-4 demonstrates impressive capabilities, its application to content analysis should be viewed as a collaborative effort between humans and AI, where human expertise remains central to achieving high-quality results.

Moreover, it is important to acknowledge that LLMs are known to exhibit various types of bias, such as stereotypical associations or negative sentiments towards some social groups (Bender et al. 2021). While computers can introduce this type of machine bias, the results of analyses performed by humans should not be taken as unbiased truth either. For example, as Kahneman et al. (Kahneman, Sibony, and Sunstein 2021) point out, humans tend to introduce "noise," i.e., inconsistency and variability, into their judgments. Both sources of error, AI bias and human noise, are liable to negatively affect outcomes. Instead of treating either machine or human-driven analysis as the one best method, a more thorough integration of human and computational approaches provides researchers with a reciprocal and iterative human-machine learning process. Such approaches have the potential to mitigate the inherent limitations of each method and improve the quality and reproducibility of analyses of phenomena relevant to social scientists across research domains.

References

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Baden, C.; Pipal, C.; Schoonvelde, M.; and Van Der Velden, M. A. C. G. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16(1): 1–18.

Bail, C. A. 2023. Can Generative AI Improve Social Science? *SocArXiv*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Brady, H. E. 2019. The Challenge of Big Data and Data Science. *Annual Review of Political Science*, 22(1): 297–323.

Cave, S.; Coughlan, K.; and Dihal, K. 2019. "Scary Robots": Examining Public Responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 331–337.

Fast, E.; and Horvitz, E. 2017. Long-Term Trends in the Public Perception of Artificial Intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

Grimmer, J.; and Stewart, B. M. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): 267–297.

Heseltine, M.; and Clemm Von Hohenberg, B. 2023. Large Language Models as a Substitute for Human Experts in Annotating Political Text.

Kahneman, D.; Sibony, O.; and Sunstein, C. R. 2021. *Noise: A flaw in human judgment*. Little, Brown Spark, first edition.

Kantner, C.; and Overbeck, M. 2020. Exploring Soft Concepts with Hard Corpus-Analytic Methods. In Reiter, N.; Pichler, A.; and Kuhn, J., eds., *Reflektierte algorithmische Textanalyse*, 169–190. De Gruyter.

Lazer, D.; and Radford, J. 2017. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, 43(1): 19–39.

Lind, F.; Eberl, J.-M.; Heidenreich, T.; and Boomgaarden, H. G. 2019. When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction. *International Journal of Communication*, 13.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8).

Nguyen, D.; Liakata, M.; DeDeo, S.; Eisenstein, J.; Mimno, D.; Tromble, R.; and Winters, J. 2020. How We Do Things With Words: Analyzing Text as Social and Cultural Data. *Frontiers in Artificial Intelligence*, 3: 62.

Törnberg, P. 2023. How to use LLMs for Text Analysis (Version 1). *arXiv*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.

APPENDICES

A.1 Codebook (English)

A.1.1. Sentiment

Negative / Neutral / Positive: Your task is to classify the sentiment of newspaper texts about Artificial Intelligence (AI) based on the emotions and language used. Label a text as negative if it expresses emotions such as dissatisfaction, sadness, anger, or frustration, or if it contains criticism, unfavorable comparisons, or highlights problems or drawbacks associated with AI. Label a text as positive if it expresses emotions such as happiness, satisfaction, excitement, or gratitude, or if it contains praise, positive comparisons, or highlights solutions or benefits associated with AI. Label a text as neutral if it lacks overtly positive or negative emotional language and presents objective statements or information that neither praises nor criticizes AI. Also rate as neutral texts with an equal mix of positive and negative sentiments. Respond with the number that represents the sentiment of the text: 0 for negative, 1 for neutral, 2 for positive. Do not provide any explanation for your choice.

A.1.2. Topics

Work: Classify the following text as either relevant ('1') or not relevant ('0') to the topic of 'Work,' focusing on the impact of Artificial Intelligence (AI) on labor relations and employment. A text is relevant if it addresses any of the following issues: AI making human jobs easier, AI freeing humans from repetitive and boring tasks, AI and automation relegating humans to menial jobs, AI leading to the displacement of human jobs, the potential of AI to cause large-scale unemployment. Respond with '1' for relevant texts or '0' for texts that do not relate to these criteria. Do not provide any explanation for your choice.

Education: Classify the following text as either relevant ('1') or not relevant ('0') to the topic of 'Education', focusing on the impact of Artificial Intelligence (AI) on educational institutions. A text is relevant if it addresses any of the following issues: changes in learning and teaching at schools or universities due to technological change, highlighting the need for new approaches to assessment and grading, automatic tutoring or grading, personalized learning. Respond with '1' for relevant texts or '0' for texts that do not relate to these criteria. Do not provide any explanation for your choice.

Regulation: Classify the following text as either relevant ('1') or not relevant ('0') to the topic of 'Regulation', focusing on the impact of artificial intelligence (AI) on government policy and supervision. A text is relevant if it addresses any of the following: governments planning and implementing policies related to AI and the handling of personal information, the transparency of AI decision-making processes, how AI-generated content is treated under intellectual property rights, or calls for bans targeting AI applications. Respond with '1' for relevant texts or '0' for texts that do not address these criteria. Do not provide any explanation for your choice.

Misinformation: Classify the following text as either relevant ('1') or not relevant ('0') to the topic of 'Misinformation', focusing on the impact of artificial intelligence (AI) on the creation and dissemination of misleading information. A text is relevant only if it directly addresses one of the following: the use of AI to create deepfakes or other synthetic media intended to deceive; the role of AI in undermining the credibility and trustworthiness of news sources and disseminated information. Texts that simply mention AI in a general digital media context without explicitly discussing aspects of misinformation are not relevant. Respond with '1' for texts that address these specified aspects of AI-related misinformation, or '0' for those that do not. Do not provide any explanation for your choice.

Existential Risk: Classify the following text as either relevant ('1') or not relevant ('0') to the topic 'Existential Risk', focusing on the risks of human extinction and complete loss of human control due to advances in Artificial Intelligence (AI). A text is relevant only if it explicitly addresses scenarios such as: the potential for advanced AI leading to human extinction, or AI evolving to the point where humans completely lose control over it. Texts that address lesser risks, such as economic disruption, privacy issues, or non-catastrophic consequences, are not relevant. Respond with '1' for texts that address these specified existential risks or '0' for those that do not. Do not provide any explanation for your choice.

Stock Market: Classify the following text as either relevant ('1') or not relevant ('0') to the topic 'Stock Markets', focusing on the impact of Artificial Intelligence (AI) and technology companies on stock prices and trading. A text is relevant if it discusses any of the following: stock prices, sales figures, profits and losses of technology companies, and the role of AI in stock analysis and investment advice. Texts that contain news about technology companies that are not about financial or stock market performance are not relevant. Respond with '1' for relevant texts or '0' for texts that are not relevant. Do not provide any explanation for your choice.

Creative Industries: Classify the following text as either relevant ('1') or not relevant ('0') to the topic 'Creative Industries', focusing on the impact of Artificial Intelligence (AI). A text is relevant if it discusses any of the following: AI's influence on musicians, writers and authors, painters, visual artists, and designers, as well as actors and performing artists, and the art market. It is also relevant if it addresses AI's role in creating interactive media experiences for video games, social media, or streaming services. Respond with '1' for relevant texts or '0' for texts that do not address these criteria. Do not provide any explanation for your choice.

A.2 Codebook (German)

A.2.1. Sentiment

Negativ / Neutral / Positiv: Deine Aufgabe ist es, die Stimmung von Zeitungsartikeln u¨ber Ku¨nstliche Intelligenz (KI) anhand der verwendeten Emotionen und Sprache zu klassifizieren. Kennzeichne einen Text als negativ, wenn er Gefu¨hle wie Unzufriedenheit, Traurigkeit, Wut oder Frustration ausdru¨ckt oder wenn er Kritik, negative Vergleiche entha¨lt oder Probleme oder Nachteile im Zusammenhang mit KI hervorhebt. Kennzeichne einen Text als positiv, wenn er Gefu¨hle wie Glu¨ck, Zufriedenheit, Freude oder Dankbarkeit ausdru¨ckt oder wenn er Lob, positive Vergleiche entha¨lt oder Lo¨sungen oder Vorteile im Zusammenhang mit KI hervorhebt. Bewerte einen Text als neutral, wenn er keine offensichtlichen positiven oder negativen Emotionen ausdru¨ckt und objektive Aussagen oder Informationen entha¨lt, die KI weder loben noch kritisieren. Bewerte auch Texte als neutral, die eine Mischung aus positiven und negativen Emotionen enthalten. Gib die Zahl an, die der Stimmung des Textes entspricht: 0 fu¨r negativ, 1 fu¨r neutral, 2 fu¨r positiv. Begru¨nde deine Bewertung nicht.

A.2.2. Themen

Arbeit: Klassifiziere den folgenden Text als relevant (’1’) oder nicht relevant (’0’) fu¨r das Thema ”Arbeit” mit Schwerpunkt auf den Auswirkungen von Ku¨nstlicher Intelligenz (KI) auf Arbeitsverha¨ltnisse und Bescha¨ftigung. Ein Text ist relevant, wenn er eines der folgenden Themen behandelt: KI erleichtert menschliche Arbeit; KI befreit Menschen von repetitiven und langweiligen Aufgaben; KI und Automatisierung degradieren Menschen zu niederen Ta¨tigkeiten; KI fu¨hrt zur Verdra¨ngung menschlicher Jobs; das Potenzial von KI, Massenarbeitslosigkeit zu verursachen. Texte, die KI in einem allgemeinen Kontext behandeln, sich aber nicht explizit auf eines der oben genannten Themen beziehen, sind nicht relevant. Antworte mit ’1’ fu¨r relevante Texte oder mit ’0’ fu¨r nicht relevante Texte. Begru¨nde deine Antwort nicht.

Bildung: Klassifiziere den folgenden Text entweder als relevant (’1’) oder nicht relevant (’0’) fu¨r das Thema ’Bildung’, wobei der Schwerpunkt auf den Auswirkungen von Ku¨nstlicher Intelligenz (KI) auf Bildungseinrichtungen liegt. Ein Text ist relevant, wenn er eines der folgenden Themen anspricht: Vera¨nderungen des Lernens und Lehrens an Schulen oder Universita¨ten aufgrund des technologischen Wandels, Hervorhebung der Notwendigkeit neuer Ansa¨tze bei der Bewertung und Benotung, automatisierte Betreuung oder Benotung, personalisiertes Lernen. Antworte mit ’1’ fu¨r relevante Texte oder mit ’0’ fu¨r Texte, die sich nicht auf diese Themen beziehen. Begru¨nde deine Wahl nicht.

Regulierung: Klassifiziere den folgenden Text entweder als relevant (’1’) oder nicht relevant (’0’) fu¨r das Thema ”Regulierung”, wobei der Schwerpunkt auf den Auswirkungen von Ku¨nstlicher Intelligenz (KI) auf Politik und staatliche Regulierung liegt. Ein Text ist relevant, wenn er eines der folgenden Themen anspricht: staatliche Maßnahmen in Bezug auf KI, insbesondere Politik im Zusammenhang mit Datenschutz oder Transparenz von KI-gestu¨tzten Entscheidungsprozessen; die Behandlung von KI-generierten Inhalten im Kontext von Rechten des geistigen Eigentums; oder Diskussionen u¨ber Verbote bestimmter KI-Anwendungen. Texte, die KI in einem allgemeinen Kontext behandeln, oder Sicherheitslu¨cken oder Aspekte von Falschinformationen beziehen sich aber nicht explizit auf eines der oben genannten Themen beziehen, sind nicht relevant. Antworte mit ’1’ fu¨r relevante Texte oder mit ’0’ fu¨r nicht relevante Texte. Begru¨nde deine Antwort nicht.

Falschinformationen: Klassifiziere den folgenden Text entweder als relevant (’1’) oder als nicht relevant (’0’) fu¨r das Thema ”Falschinformationen” mit Schwerpunkt auf den Auswirkungen von Ku¨nstlicher Intelligenz (KI) auf die Erstellung und Verbreitung irrefu¨hrender Informationen. Ein Text ist nur dann relevant, wenn er direkt auf einen der folgenden Punkte eingeht: den Einsatz von KI zur Erstellung von Deepfakes oder anderen synthetischen Medien, die der Ta¨uschung dienen; die Rolle von KI bei der Untergrabung der Glaubwu¨rdigkeit und Vertrauenswu¨rdigkeit von Nachrichtenquellen und vero¨ffentlichten Informationen. Texte, die KI lediglich in einem allgemeinen Kontext digitaler Medien erwa¨hnen, ohne explizit auf Aspekte der Desinformation einzugehen, sind nicht relevant. Antworte mit ’1’ fu¨r Texte, die diese spezifischen Aspekte von KI-bezogener Fehlinformation ansprechen, oder mit ”0” fu¨r Texte, die dies nicht tun. Begru¨nde deine Wahl nicht.

Existenzielle Risiken: Klassifiziere den folgenden Text entweder als relevant (’1’) oder als nicht relevant (’0’) fu¨r das Thema ’Existenzielle Risiken’, wobei der Schwerpunkt auf den Risiken des mo¨glichen Untergangs der Menschheit oder des vollsta¨ndigen Verlusts menschlicher Kontrolle durch Fortschritte in der Ku¨nstlichen Intelligenz (KI) liegt. Ein Text ist nur dann relevant, wenn er sich explizit mit existenziellen Szenarien auseinandersetzt, wie etwa dem Potenzial fortgeschrittener KI, zum Untergang der Menschheit zu fu¨hren, oder der Entwicklung einer KI, die dazu fu¨hrt, dass der Mensch die Kontrolle u¨ber Maschinen und Technologie verliert. Texte, die sich mit weniger gravierenden Risiken wie wirtschaftlichen Verwerfungen, Fragen der Privatspha¨re oder anderen nicht katastrophalen Folgen befassen, sind nicht relevant. Antworte mit ’1’ fu¨r Texte, die existenzielle Risiken behandeln, oder mit ’0’ fu¨r Texte, die dies nicht tun. Begru¨nde deine Wahl nicht.

Aktienma¨rkte: Klassifiziere den folgenden Text entweder als relevant (’1’) oder nicht relevant (’0’) fu¨r das Thema ’Aktienma¨rkte’, wobei der Schwerpunkt auf den Auswirkungen von Ku¨nstlicher Intelligenz (KI) und Technologieunternehmen auf Aktienkurse und -handel liegt. Ein Text ist relevant, wenn er einen der folgenden Punkte behandelt: Aktienkurse, Umsatzzahlen, Gewinne und Verluste von Technologieunternehmen oder die Rolle von KI in der Aktienanalyse und Anlageberatung. Texte, die Meldungen u¨ber Technologieunternehmen enthalten, in welchen es nicht um Finanzen oder Aktienkurse geht, sind nicht relevant. Antworte mit ’1’ fu¨r relevante Texte oder mit ’0’ fu¨r nicht relevante Texte. Begru¨nde deine Wahl nicht.

Kreativbranchen: Klassifiziere den folgenden Text entweder als relevant (’1’) oder nicht relevant (’0’) fu¨r das Thema ’Kreativbranche’ mit Schwerpunkt auf den Auswirkungen von Ku¨nstlicher Intelligenz (KI) auf Kunst- und Kreativschaffende. Ein Text ist relevant, wenn er eines der folgenden Themen behandelt: Der Einfluss von KI auf Musiker, Schriftsteller, Maler oder Schauspieler. Texte sind auch relevant, wenn sie die Rolle von KI bei der Schaffung interaktiver Medienerlebnisse fu¨r Videospiele, soziale Medien oder Streaming-Dienste behandeln. Texte, die KI in einem allgemeinen Kontext behandeln, aber nicht explizit auf einen der oben genannten Berufe oder Branchen Bezug nehmen, sind nicht relevant. Antworte mit ’1’ fu¨r relevante Texte oder mit ’0’ fu¨r nicht relevante Texte. Begru¨nde deine Wahl nicht.