Newspaper Signaling for Crisis Prediction

Prajvi Saxena
German Research Center for Artificial Intelligence, Saarbrücken, Germany
prajvi.saxena@dfki.de

Sabine Janzen
German Research Center for Artificial Intelligence, Saarbrücken, Germany
sabine.janzen@dfki.de

Wolfgang Maaß
German Research Center for Artificial Intelligence; Saarland University, Saarbrücken, Germany
wolfgang.maass@dfki.de

Abstract

To establish sophisticated monitoring of newspaper articles for detecting crisis-related signals, natural language processing has to cope with unstructured data, media, and cultural bias as well as multiple languages. So far, research on detecting signals in newspaper articles is focusing on structured data, restricted language settings, and isolated application domains. When considering complex crisis-related signals, a high number of diverse newspaper articles in terms of language and culture reduces potential biases. We demonstrate MENDEL – a model for multi-lingual and open-domain newspaper signaling for detecting crisis-related indicators in newspaper articles. The model works with unstructured news data and combines multiple transformer-based models for pre-processing (STANZA) and content filtering (RoBERTa, GPT-3.5). Embedded in a Question-Answering (QA) setting, MENDEL supports multiple languages (>66) and can detect early newspaper signals for open crisis domains in real-time.

1 Introduction

Monitoring newspaper sources has a significant impact on companies, health organizations, and civil defense in preparing for and responding to emerging trends, conflicts, and crises situations effectively (Elliott and Timmermann, 2016; Dim et al., 2021). Nonetheless, up until now, newspaper signaling is not applied in crisis and risk management in practice as it is challenging due to the amount of unstructured data, media, cultural bias (Hanitzsch et al., 2020), and multiple languages (Asr and Taboada, 2019). So far, research on detecting potential signals for crisis-related events based on newspaper articles is focusing on structured data (Hassanzadeh et al., 2022; Huang et al., 2022, 2020; Sakaki et al., 2010; Rasouli et al., 2020; Asif et al., 2021), restricted language settings (Luca Barbaglia and Manzan, 2023), and isolated application domains, e.g., mobility, finances (Dim et al., 2021; Agrawal et al., 2022). The existing systems that monitors newspapers (GAIA-X, 2022; Eurostat, 2023) only aggregates and visualize current news, they don't posses advanced language processing capabilities. However, MENDEL distinguishes itself by processing unstructured newspaper data in real-time, and its use of large language models for multi-lingual content filtering and context-based crisis forecasting, offering a broader and more dynamic approach to early crisis signaling by encouraging organizational preparedness and supporting a rapid and effective response (Bundy et al., 2017).

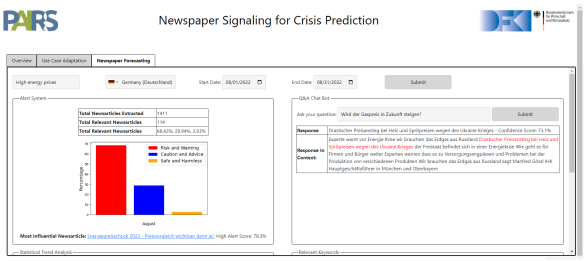

Figure 1: Demonstration of the user interface used for real-time newspaper signaling for detecting energy-related crisis signals.

The major challenge in using newspaper articles for crisis prediction is the management of unstructured data. Most approaches tackle this challenge by specifying a domain ontology for converting unstructured data into structured data, e.g., (Agrawal et al., 2022); facing all the well-known disadvantages in performance, flexibility, and openness with regard to domains. Thus, multi-lingual issues as well as media and cultural biases in newspaper articles written by diverse journalists for diverse newspapers in diverse countries on the same events cannot be captured (Hanitzsch et al., 2020). Furthermore, the media landscape in each country has an influence on how events are portrayed (Kalogeropoulos et al., 2019). Therefore,

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations), pages 166–173
June 16-21, 2024

considering a high number of diverse newspaper resources in terms of language and culture reduces potential biases when seeking to identify complex crisis-related signals in newspaper articles. In this paper, we demonstrate MENDEL – a model for multi-lingual and open-domain newspaper signaling for detecting crisis-related indicators in newspaper articles (cf. Figure 1). Our model works with unstructured data of newspaper articles and combines multiple state-of-the-art transformer-based models (Vaswani et al., 2017) for pre-processing (STANZA (Qi et al., 2020)) and content filtering (XLM-RoBERTa (Conneau et al., 2020), GPT-3.5 (Brown et al., 2020)). MENDEL supports multiple of the most spoken languages in the world (>66) (e.g., Mandarin Chinese, Spanish, English, Hindi, Arabic). The model is able to detect newspaper signals for open domains, e.g., energy, finances, and supply chains, that can be directly adjusted by the user in terms of keywords. One appeal of the model is the usage of purely unstructured data on newspaper articles for processing signals in real-time in contrast to other approaches mixing those data with already existing structured data suffering from lower quality and timeliness, e.g., Wikidata (Hassanzadeh et al., 2022; Li et al., 2022; Shane E. Halse and Caragea, 2018; Mai and Quan, 2020). Furthermore, MENDEL covers a 2-step filtration pipeline based on RoBERTa (Conneau et al., 2020) and Cosine similarity for determining domain relevance and crisis reference enabling an extensive filtration of articles with high accuracy and less redundancy in distinction from other approaches, that use no filtration, restricted approaches and only sentiment analyses[1] (Agrawal et al., 2022). The model provides the risk and warnings, statistical trends of the crisis and is exemplified within a QA system as a natural language assistant in risk and crisis management (cf. Figure 1)[2]. We were able to evaluate the proposed approach by means of a set of newspaper articles (total: 18,673 news articles) in terms of performance in identifying potential signals for economic recession and energy-related crisis situations (i.e., availability and costs of energy like gas, oil, coal, solar, wind, supply chain disruption, mobility, etc) in Germany.

[1]https://eventregistry.org/products/intelligence/

[2]Link to demo video: https://youtu.be/q2UTeQsBnDc

2 Multi-lingual and Open-domain Newspaper Signaling for Crisis Prediction

We present MENDEL, a model for multi-lingual and open-domain newspaper signaling for crisis prediction powered with a real-time alert system, statistical trend visualization, and a QA chat-bot. In this paper, we explore crises defined as periods of substantial instability that disrupt the normal functioning of systems, leading to notable consequences; specifically, we concentrate on events that are predictable but challenging to influence, e.g. supply chain disruptions, mobility, rise in energy prices, economic crisis etc as outlined in various crisis taxonomies, including (Gundel, 2005). The framework consists of four main modules (cf. Figure 2): Data acquisitions, Data-processing pipeline, Two-stage data filtration, and Context-based reasoning and forecasting. MENDEL operates on domain-specific keywords given by a user with additional parameters such as language, country, and time frame. This serves as input to the data acquisition module which consists of several components, charged with generating domain-specific keywords, extracting newspaper articles, and a data parser. Outputs of these components are fed to the data-processing module which handles data cleaning by removing stopwords, punctuation followed by tokenizer and lemmatizer. This module processes the data that can be directly fed to our two-stage data filtration, which is responsible to filters the articles based on users specified crisis domain and by finding future and present tenses in articles. Outputs are further passed to the Context-based Reasoning and Forecasting module. It generates, risks and warnings for the filtered articles and also provide statistical trend visualization for six months period. It also provides relevant keywords and QA chat-bot. Overall, MENDEL pipeline uses completely unstructured data, i.e., newspaper articles, supports open crisis domain, multiple languages and provide early signaling and warnings.

To introduce the proposed approach in the demo, we give a short example course of detecting crisis-related indicators in newspaper articles, starting with the domain name provided by the user e.g. 'high energy prices' and ending with identified domain-specific newspaper signals. For the following, imagine a user, such as a company, looking for signals of energy-related crises due to reduced availability or rising costs of energy such as gas,

oil, coal, supply chain disruption and mobility, etc. MENDEL would provide them with alert trends in the form of risk and warnings for rising energy prices. Additionally, it provides visualizations of the statistical crisis trend over 6 months period.

2.1 Data Acquisition

The data acquisition pipeline receives the specified domain name directly from the interface, i.e. 'high energy prices', which is given as a raw string input by the user. The keyword expansion model then generates a list of relevant keywords for the specified domain, e.g., energy shortage, energy cost surge, energy demand, and electricity blackouts. To do this, we are utilizing the openAI's[3] generative pre-trained transformer 3.5 (GPT-3.5) (Brown et al., 2020) model, which broaden the search scope and improve the precision of extracted articles over the manually collected list of keywords (Kyröläinen and Laippala, 2023). We then extract news articles for the curated list of relevant keywords using event registry news API[4]. It allows to obtain access to real-time as well as archive news articles. Additional filters such as time, date, country, and language can be provided from the user interface. Extracted articles are then passed through a data parser which extracts specific data from the entire news article such as title, URL, published date, and first four paragraphs; as typically an abstract of the article is given at the beginning. This helps to reduce the text size of the individual article and improves processing time.

2.2 Data Processing

Our data processing module mainly performs the pre-processing of the extracted news articles as shown in (cf. Figure 2). The data cleaning module handles the removal of special characters, converting them to lowercase, and removing duplicates and missing values. Followed by this, all the stop words and punctuation from the data is removed. STANZA (Qi et al., 2020) is used for tokenization and lemmatization of the data. It tokenizes and splits sentences, each of these sentences contains a list of tokens which is then converted to it's lemmatized form. In the example, the parsed news articles from section 2.1 are processed by removing punctuation, stop words, etc., and converting them into tokens and lemma form. The decomposed lemma represents the output as a set of single-word tokens, e.g., ['Household', 'will', 'face', 'energy-expensive', 'winter', ..., 'economic', 'stress'].

2.3 Two-Stage Data Filtration

News articles extracted from news API[5] generally contain irrelevant and noisy data. Hence, it is important to filter the articles based on the user's specified crisis domain. In this research, we propose a two-stage data filtering. First, to filter articles based on the use case, and second, to focus on articles that are related to future warnings. The domain-based articles matching module utilizes a state-of-the-art multilingual RoBERTa model (Conneau et al., 2020) to perform filtering based on the desired domain. For this, we derive embedding vectors of the articles by using its learned embedding representations from the RoBERTa model. Further we generate the embedding vectors of all the domain-relevant keywords and use cosine similarity to check if the embedding vectors of articles and domain keywords are close to each other. In this filtering stage, only articles with a cosine similarity greater than 70% arbitrary threshold are retained while the rest are discarded. Subsequently, we need to focus on articles related to future warnings, hence we propose a filtration method to only get articles that are in the future tense and present tense, and reject the articles which are in the past tense. To achieve this, we use the pre-trained zero-shot multilingual XLM-RoBERTa (Conneau et al., 2020) model to classify articles according to their tenses. We set up an arbitrary threshold of 70% combining both future and present tense confidence and discard the remaining articles.

For our example, we first input the sentence: [household will face energy expensive winter....economic stress] to the RoBERTa encoder along with the domain keywords (high energy prices, energy shortage, energy cost surge, energy demand, etc) and then we compute their likeness using cosine similarity. Here, we get a similarity of 93.7% which preceded the threshold of 70%, hence we pass this to the second filtration step where we give it to our tense-based classifier and got the confidence score of 94.1% for future and present tense. As the confidence is higher than the threshold of 70%, we include this article in our data set.

[3]https://openai.com/api/

[4]https://www.newsapi.ai

[5]https://www.newsapi.ai

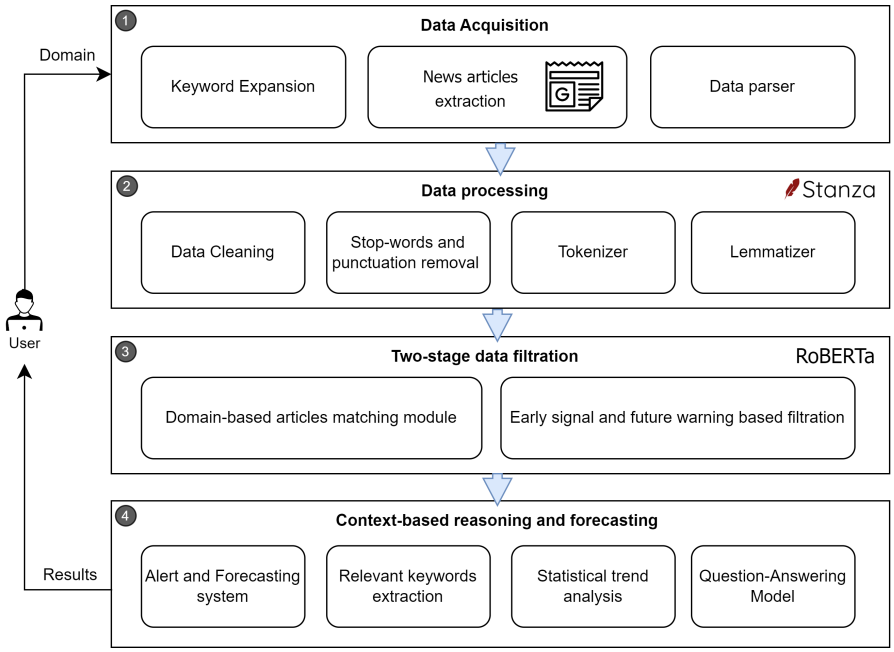

Figure 2: Architecture diagram of MENDEL: A model for multi-lingual and open-domain newspaper signaling for detecting early crisis-related indicators in the newspaper article.

2.4 Context-based Reasoning and Forecasting

Finally, the context-based reasoning and forecasting module receives the final set of embedding from two-stage data filtration2.3. Our alert system is based on the powerful GPT-3.5 (Brown et al., 2020) which classifies the article into three categories 'risk and warning', 'caution and advice', and, 'safe and harmless'. The classified articles are then statistically analyzed by computing the percentage and average confidence score of 'risk and warnings' articles to calculate the alerts. The predictions are delivered to users through the user interface along with visualization of the statistical crisis trend over 6 months period. For the 'high energy prices' example, users will be shown a percentage of 'risk and warnings' signals found with their mean confidence scores, and a graph showing the alert trends for 6 months. If the alerts percentage is high and the graph shows an exponential growth pattern, it is most likely to have an upcoming crisis related to 'high energy prices'. We further use KeyBERT [6] to extract the relevant keywords from the high-risk and warning articles. The resulting relevance for the selected example is ("gas price spikes: 92%", "energy-expensive: 72%", "supply chain disruptions: 78%", mobility: 43%]). Moreover, we also provide an extractive QA chatbot (Conneau et al., 2020; Rajpurkar et al., 2016) for users to interact and ask any questions related to filtered articles. It meets interactive needs of the users and provide more dynamic and responsive way to access crisis-related information.

3 Implementation and Evaluation

Based on the proposed model MENDEL (cf. Figure 2), we implemented a newspaper signaling service for crisis and risk management [7]. The service architecture has been deployed using Python and Django while the client side interface has been designed as a web interface using HTML, CSS, and JavaScript. The system accepts keywords by the user in the form of plain text along with specific countries, languages, and time frames. In response, the system presents domain-specific crisis signals and highlights the most influential articles. Moreover, it delivers alerts with confidence and severity levels, along with the relevant keywords[8].

3.1 Settings

We evaluated the performance of the signaling service in identifying potential signals for economic recession and energy-related crisis situations (i.e., availability and costs of energy like gas, oil, coal, solar, and wind). Here, we selected a subset of

[6]https://github.com/MaartenGr/KeyBERT

[7]Link to GitHub repo: https://github.com/InformationServiceSystems/pairs-project/tree/main/Modules/NewspaperSignaling

[8]Demo video: https://youtu.be/q2UTeQsBnDc

past and ongoing crises in Germany and represented with event (E) i.e., economic recession in mid-2023[9] (e_1), huge increase of energy prices in 2022[10] (e_2), and immense raise of gasoline prices in 2021[11] (e_3). The objective of the experiment was the identification of early signals regarding these crises in newspaper articles with high percentage and within notable time. We defined three-time horizons for test runs: four (t_{-1}), eight (t_{-2}), and twelve (t_{-3}) weeks before the crisis event occurred. Furthermore, the final set of articles went through the 'alert and forecasting' component and calculated the potential alerts by performing text classification. To assess the performance of our alert and forecasting model, we recruited three human annotators for creating an annotated text corpora due to the lack of crisis newspaper benchmarks. We focused on our crisis events (E) and generated ground truth from annotators. We provided clear instructions and examples for classification to ensure consistent labeling. They classified the articles into three categories: 'risk and warning', 'caution and advice', and 'safe and harmless'. These categories were motivated by existing crisis-related labels used in datasets like CrisisBench (Alam et al., 2021). The ground truth was then determined by selecting the majority label from the three annotator's inputs.

3.2 Data

To conduct the experiment, we collected the news articles for the three past events (E) using the event registry news API. The data acquisition module extracted a total of 18,673 articles in real-time from 01.07.2021 to 31.05.2023 (Table 1 displays article counts across different modules). For event (e_1) we input the keyword 'economic recession' and retrieved 12,265 articles. Following, for the event (e_2) we used the keyword 'high energy prices' and received 5,839 articles. Lastly, for the event (e_3) we got 569 articles for the keyword 'high gas prices'. The majority of articles consisted of German and English languages, but we also found multiple other languages such as Russian, Bulgarian, Spanish, Slovenian, Czech, Indonesian, and Chinese. The processed and parsed articles were fed to the two-stage data filtration, which narrowed down the relevant articles to a total of 4,002. On average, 67.37% of collected articles were irrelevant to the selected domains and future signals, highlighting the importance of the two-stage data filtration.

We also prepared a dataset for testing the performance of our 'alert and forecasting' system. We took a subset of the output of 2-stage filtered data for crisis events (E) and created datasets of total of 319 articles (e_1: 115, e_2: 100, e_3: 104) for annotations.

3.3 Results

Table 2 shows the performance of the signaling service in identifying newspaper signals for time intervals of 4 (t_{-1}), 8 (t_{-2}), and 12 (t_{-3}) weeks in advance of crisis events (E). For each point in time $t \in T$, we examined the monthly growth trend of risk and warning signals in newspaper articles, i.e., the Risk and Warning percentage ($RW\%$). It is defined as the percentage of articles classified as 'risk and warning' by our alert and forecasting module. Results show a generally growing trend in the frequency of risk and warning signals as indicators for recession and energy-related crises.

MENDEL was able to detect early signals at all points in time interval T, i.e., 4, 8, and 12 weeks before the crisis event. When points in time $t \in T$ are marked bold in Table 2, the signaling service detected newspaper signals for the respective crisis events $e \in E$. However, for the event e_3 at (t_{-3}) the $RW\%$ was weak maybe due to the fact that raising gasoline prices are quite popular and volatile compare to e_1, e_2. As people are familiar with volatile gasoline prices, the need for communicating about this issue is lower than with respect to immensely raising prices for electricity. e_1, e_2. Overall, the results of the run-time study indicate a positive evaluation of the newspaper signaling service implementing MENDEL.

To verify the quality of the results reported and to evaluate our alert and forecasting model performance we used the generated ground truth data. By experimenting with multiple state-of-the-art classification models to identify the most effective model for crisis signaling. Due to high-class imbalance in the results of events (e) with 'risk and warning' being the dominant class, we adopted the micro F1 score for the evaluation metric (Takahashi et al., 2022). Table 3 illustrates the model comparison, where the GPT-3.5 model outperformed other models, while Bart (Lewis et al., 2020) and De-

[9] https://www.dw.com/en/recession-in-germany-what-does-that-mean/a-63444401

[10] https://tradingeconomics.com/germany/electricity-price

[11] https://take-profit.org/en/statistics/gasoline-prices/germany/

Event (E)	#articles	#language	#German	#English	#data_processing	#2-stage_filtration
e_1	12265	02	9864	2399	12263	3371
e_2	5839	06	5486	298	5836	482
e_3	569	06	450	64	564	149

Table 1: Distribution of extracted and processed articles across different stages of MENDEL for all events (E).

Event (E)	Date of event (t_0)	Description	t_{-1}	$RW\%$	t_{-2}	$RW\%$	t_{-3}	$RW\%$
e_1	05/23	Economic Recession.	**04/23**	80.73	**03/23**	60	**02/23**	74.44
e_2	09/22	Peak in electricity price.	**08/22**	75.5	**07/22**	68.42	**06/22**	66.6
e_3	10/21	Peak in gasoline price.	**09/21**	73.7	**08/21**	59.09	07/21	30

Table 2: Results of run-time study for evaluating the performance of MENDEL in identifying newspaper signals for time intervals of 4 (t_{-1}), 8 (t_{-2}), and 12 (t_{-3}) weeks in advance of economic recession and past energy-related crisis events E between July 2021 and May 2023 in Germany. Domain-specific keyword: *['Economic recession', 'High energy prices', 'High gas Prices']*. (Legend: $RW\%$ = risk and warning percentage)

BERTaV3 (Laurer et al., 2024) also demonstrated promising results. Therefore, we used GPT-3.5 (Brown et al., 2020) for our alert and forecasting system.

Table 3: Performance of 'alert and forecasting' model for classifying the articles in 'risk and warning', 'caution and advice', and 'safe and harmless' categories. For the Events (E), numbers reported are micro-averaged f1 scores on different text classification models based on human-generated ground truth labels.

Model	e_1	e_2	e_3
$\text{XLM-RoBERTa}_{\text{Large}}$	0.46	0.37	0.27
DeBERTaV3	0.65	0.58	0.61
BART	0.6	0.68	0.63
$\text{GPT}_{3.5}$	**0.75**	**0.8**	**0.79**

4 Conclusion

We considered real-time newspaper signaling for detecting crisis-related indicators based on purely unstructured data. So far, research on detecting signals in newspaper articles is focusing on structured data, restricted language settings, and isolated application domains, giving little attention to the thereby induced potential biases. We introduced MENDEL – a model for multi-lingual and open-domain newspaper signaling for detecting crisis-related indicators in newspaper articles. The model works with unstructured data from newspaper articles and combines multiple transformer-based models for pre-processing (STANZA) and content filtering (XLM-RoBERTa, GPT-3.5). Embedded in a question-answering setting, MENDEL supports multiple spoken languages in the world (>66) and is able to detect newspaper signals for open domains in real time. We were able to evaluate the proposed approach by identifying potential signals for events (E), economic recession, and energy-related crisis situations. In terms of performance, we evaluated our alert and forecasting model by creating human-annotated data and achieved up to 80% average micro-F1 score. We also were able to identify the potential signals for recession and energy-related crisis, from four (t_{-1}), eight (t_{-2}), and twelve (t_{-3}) weeks before the crisis event occurred.

Ethics statement and limitations

MENDEL aims to make it easier to comprehend news articles about growing and rapidly updating crises, as it can be challenging for humans to keep up with emerging issues from extensive unstructured news data. It is not intended to make predictions, but rather to offer early warning signs of impending crises that would take humans too much time to detect. Verification of crisis warnings is a task that our system does not undertake and that we consider for future work. Our approach does not prove that all crises can be predicted with the same level of performance. It is highly influential on the quality and quantity of news articles due to its capability to deal only with unstructured data.

Acknowledgement

This work was partially funded by the German Federal Ministry of Economics and Climate Protection (BMWK) within the research project PAIRS (grant number: 01MK21008B) and by Saarland Ministry for Economics, Innovation, Digital and Energy (MWIDE) and European Regional Development Fund (ERDF) within the research project INTE:GRATE.

References

Garima Agrawal, Yuli Deng, Jongchan Park, Huan Liu, and Ying-Chih Chen. 2022. Building knowledge graphs from unstructured texts: Applications and impact analyses in cybersecurity education. *Information*, 13(11).

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 923–932. AAAI Press.

Amna Asif, Shaheen Khatoon, Md Maruf Hasan, Majed Alshamari, Sherif Abdou, Khaled Elsayed, and Mohsen Rashwan. 2021. Automatic analysis of social media images to identify disaster type and infer appropriate emergency response. *Journal of Big Data*, 8.

Fatemeh Torabi Asr and Maite Taboada. 2019. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jonathan Bundy, Michael D. Pfarrer, Cole E. Short, and W. Timothy Coombs. 2017. Crises and crisis management: Integration, interpretation, and research development. *Journal of Management*, 43(6):1661–1692.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Chukwuma Dim, Kevin Koerner, Marcin Wolski, and Sanne Zwart. 2021. Hot off the press: News-implied sovereign default risk. Available at SSRN: https://ssrn.com/abstract=3955052 or http://dx.doi.org/10.2139/ssrn.3955052.

Graham Elliott and Allan Timmermann. 2016. Forecasting in economics and finance. *Annual Review of Economics*, 8:81–110.

Eurostat. 2023. Businesses in the manufacturing sector. Available at `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Manufacturing_statistics_-_NACE_Rev._2&oldid=502915`.

GAIA-X. 2022. Gaia-x architecture document, release 22.04. Available at `https://gaia-x.eu/wp-content/uploads/2022/06/Gaia-x-Architecture-Document-22.04-Release.pdf`.

Stephan Gundel. 2005. Towards a new typology of crises. *Journal of contingencies and crisis management*, 13(3):106–115.

Thomas Hanitzsch, Josef Seethaler, Elizabeth A Skewes, Maria Anikina, Rosa Berganza, Incilay Cangöz, Mihai Coman, Basyouni Hamada, Folker Hanusch, Christopher D Karadjov, et al. 2020. Worlds of journalism: Journalistic cultures, professional autonomy, and perceived influences across 18 nations. In *The global journalist in the 21st century*, pages 473–494. Routledge.

Oktie Hassanzadeh, Parul Awasthy, Ken Barker, Onkar Bhardwaj, Debarun Bhattacharjya, Mark Feblowitz, Lee Martie, Jian Ni, Kavitha Srinivas, and Lucy Yip. 2022. Knowledge-based news event analysis and forecasting toolkit. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5904–5907. International Joint Conferences on Artificial Intelligence Organization. Demo Track.

Lida Huang, Gang Liu, Tao Chen, Hongyong Yuan, Panpan Shi, and Yujia Miao. 2020. Similarity-based emergency event detection in social media. *Journal of Safety Science and Resilience*, 2.

Lida Huang, Panpan Shi, Haichao Zhu, and Tao Chen. 2022. Early detection of emergency events from social media: a new text clustering approach. *Natural Hazards*, 111:1–25.

Antonis Kalogeropoulos, Jane Suiter, Linards Udris, and Mark Eisenegger. 2019. News media trust and news consumption: factors related to trust in news in 35

countries. *International Journal of Communication*, 13:22.

Aki-Juhani Kyröläinen and Veronika Laippala. 2023. Predictive keywords: Using machine learning to explain document characteristics. *Frontiers in Artificial Intelligence*, 5.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-shyuan Chiang, Tuan Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. COVID-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–144, Dublin, Ireland. Association for Computational Linguistics.

Sergio Consoli Luca Barbaglia and Sebastiano Manzan. 2023. Forecasting with economic news. *Journal of Business & Economic Statistics*, 41(3):708–719.

Trung Mai and Tho Quan. 2020. Ontology-based sentiment analysis for brand crisis detection on online social media. *Science Technology Development Journal - Engineering and Technology*, 3:First.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Elham Rasouli, Sajjad Zarifzadeh, and Amir Jahangard Rafsanjani. 2020. Webkey: a graph-based method for event detection in web news. *J. Intell. Inf. Syst.*, 54(3):585–604.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 851–860, New York, NY, USA. Association for Computing Machinery.

Anna Squicciarini Shane E. Halse, Andria Tapia and Cornelia Caragea. 2018. An emotional step toward automated trust detection in crisis social media. *Information, Communication & Society*, 21(2):288–305.

Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. 2022. Confidence interval for micro-averaged f1 and macro-averaged f1 scores. *Applied Intelligence*, 52.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30. Curran Associates, Inc.