”Listening In”: Social Signal Detection for Crisis Prediction

Sabine Janzen
Deutsches Forschungszentrum
für Künstliche Intelligenz (DFKI)
sabine.janzen@dfki.de

Prajvi Saxena
Deutsches Forschungszentrum
für Künstliche Intelligenz (DFKI)
prajvi.saxena@dfki.de

Sebastian Baer
Deutsches Forschungszentrum
für Künstliche Intelligenz (DFKI)
sebastian.baer@dfki.de

Wolfgang Maaß
Deutsches Forschungszentrum
für Künstliche Intelligenz (DFKI)
wolfgang.maass@dfki.de

Abstract

Crises send out early warning signals; mostly weak and difficult to detect amidst the noise of everyday life. Signal detection based on social media enables early identification of such signals supporting pro-active organizational responses before a crisis occurs. Nonetheless, social signal detection based on Twitter data is not applied in crisis management in practice as it is challenging due to the high volume of noise. With OSOS, we introduce a method for open-domain social signal detection of crisis-related indicators in tweets. OSOS works with multi-lingual Twitter data and combines multiple state-of-the-art models for data pre-processing (SoMaJo) and data filtration (GPT-3). It excels in crisis domains by leveraging fine-tuned GPT-3^{FT} (Curie) model and achieving benchmark results in the CrisisBench dataset. The method was exemplified within a signaling service for crisis management. We were able to evaluate the proposed approach by means of a data set obtained from Twitter (X) in terms of performance in identifying potential social signals for energy-related crisis events.

Keywords: Social signal detection, Crisis prediction, Social media, Open-domain

1. Introduction

Crises send out early warning signals; mostly weak and difficult to detect amidst the noise of everyday life (Diks et al., 2019; Fu and Zhu, 2020). For instance, before the financial crisis of 2008, there were several signals indicating that the global financial system was becoming unstable, e.g., a significant increase in the use of complex financial instruments, the amount of debt being taken, or the increase in housing prices. In crisis management, signal detection mechanisms aim to enable an early identification of such signals as well as pro-active organizational responses before a crisis occurs (Elsubbaugh et al., 2004; Hensgen et al., 2003; Paraskevas and Altinay, 2013; Parnell and Crandall, 2021; Wolbers et al., 2021). In this context, we define signals as indicators or pieces of information that may suggest the occurrence or likelihood of a crisis event, e.g., data points, patterns, trends, or anomalies (Imran et al., 2015). By monitoring information sources, such as social media, companies, government, and health organizations as well as civil defense are able to detect early signals and emerging trends for being prepared with respect to potential conflicts and crisis events (Reuter et al., 2018; Saroj and Pal, 2020). Here, the analysis of tweets can be a useful approach due to the large amount of data generated on Twitter by private users, companies, and organizations[1]. Nonetheless, up til now, social signal detection based on tweets is not applied in crisis management in practice as it is challenging due to the high volume of noise and irrelevant tweets (Barbosa and Feng, 2010; Daniel et al., 2017), language and cultural bias (Kruspe et al., 2021), as well as the limited context of tweets induced by their restricted length (Bonaretti, 2018). So far, research on detecting signals for crisis events based on tweets is focusing on restricted language settings (Alharbi and Lee, 2022) and isolated application domains, e.g., early detection of COVID-19 outbreaks (Cheng et al., 2021; Gharavi et al., 2020), earthquakes

[1]For instance, in 2020, the number of active Twitter users was 348 million posting 500 million tweets per day (source: https://www.statista.com/statistics/303681/twitter-users-worldwide/).

(Bügel and Zielinski, 2013; Poblete et al., 2018), or situational awareness for emergency agencies (Cameron et al., 2012).

A major challenge in using tweets for crisis prediction is the handling of noise and irrelevant tweets. Most approaches tackle this challenge by applying natural language processing techniques, such as sentiment analyses (Aslan et al., 2023; Mir, 2023), and content analyses (Mir, 2023; Terpstra, 2012); showing disadvantages due to the limited scope of training data and thus, low performance in terms of accuracy.

In this work, we present OSOS – a method for open-domain social signal detection of crisis-related indicators in tweets. Our method works with multi-lingual Twitter data and combines multiple state-of-the-art models for data pre-processing (SoMaJo) (Proisl and Uhrig, 2016) and data filtration (GPT-3^{FT} (Curie)) (Brown et al., 2020). It supports most of the spoken languages in the world (e.g., Spanish, English). The method is able to detect social signals in tweets for open domains, e.g., energy, finances, and supply chains, that can be directly adjusted by the user in terms of keywords and crisis data obtained by Twitter.

One appeal of the method is the data filtration approach, a combination of fine-tuned GPT-3 model for classification and sentiment analyses to filter out non-relevant tweets in real time. The fine-tuned GPT-3^{FT} (Curie) achieves benchmarking performance in classification on the CrisisBench dataset (Alam et al., 2021) with an accuracy of 88.2% compared to other large language models such as XLM-RoBERTa, DistilBERT, etc. Due to the ability of OSOS to capture a multi-language setting, a high number of tweets can be integrated. This enables the integration of cross-cultural perspectives reducing biases when applying tweet analysis for social signal detection (Barbieri et al., 2015). Additionally, OSOS captures trends by observing the exponential growth in the number of filtered tweets for finding early signals of crises.

The method was exemplified within a signaling service for risk and crisis management. We were able to evaluate the proposed approach by means of a data set of 46.963 unprocessed tweets by Twitter posted in Germany, the UK, and Spain from 01.01.2020 to 31.12.2022 in terms of performance in identifying potential social signals for energy-related crisis events.

2. Signal Detection in Crisis Management

Crisis refers to a time of great instability, where the normal functioning or equilibrium of a system is disrupted, often leading to significant consequences. As crises can take many forms, there are several crisis taxonomies given in the literature. According to (Gundel, 2005), in this paper, we focus on crisis events that are characterized as predictable but hardly influenceable, e.g., a rise in energy prices, and disruptions in supply chains. These events may not directly lead to civil unrest or an increase in mortality, but they do represent significant disruptions in economic stability and can have cascading effects on various sectors, potentially leading to larger crises if not managed appropriately. Understanding those crises is crucial for preemptive crisis management that covers four successive phases: mitigation, preparedness, response, and recovery (Lauras and Comes, 2015). Detecting early signals is the first step for mitigating potential crisis events and being more resilient to crises by encouraging organizational preparedness and supporting a rapid and effective response (Bundy et al., 2017; Mitroff, 1988). Here, signals are characterized as data indicating a deviation from normality, e.g., in the form of exceeding thresholds or abnormal user behavior patterns (Imran et al., 2015).

Social media platforms have seen a huge increase in the number of users in recent years and nowadays play an active role in the daily business of private users, companies, and government agencies (Auxier and Anderson, 2021; Reuter et al., 2018). In the context of crisis events, social media platforms form extended social systems that enable the dissemination of crisis-relevant information within and between official and public channels. That makes the information-seeking and self-organizing behavior of users visible, and thus noticeable and traceable (Palen et al., 2009). Taking the perspective of information retrieval, in social media a signal is represented as a change in the number of posts discussing a topic at a given time (Dou et al., 2012; Imran et al., 2015). Several studies on natural disaster detection (Earle et al., 2011; Li and Rao, 2010), emergency response (Bügel and Zielinski, 2013; Cameron et al., 2012; Imran et al., 2015), prediction of COVID-19 or flu outbreaks (Gharavi et al., 2020) have shown that collecting, processing, and analyzing social media data is a useful approach for detecting early signals of crises (Achrekar et al., 2011; Ashktorab et al., 2014).

3. Open-Domain Social Signal Detection for Crisis Prediction

This serves as input for the **Data Pre-processing** which handles data cleaning including the removal of stopwords, punctuations, and duplicates as well as performs tokenization and sentence parsing (cf. Figure 1). Resulting data are fed into the **Data Filtration** which performs extensive filtering of tweets using a state-of-the-art GPT-3 model (Brown et al., 2020) for

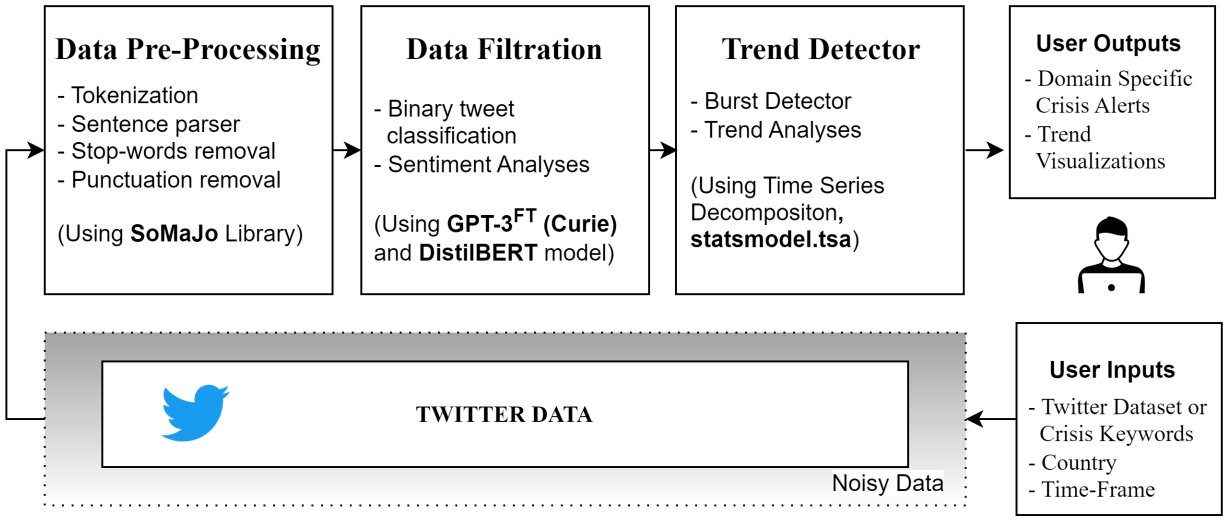


Figure 1: Method for open-domain social signal detection of crisis-related indicators in tweets (OSOS)

text classification and DistilBERT (Sanh et al., 2019) for sentiment analyses. Outputs of this step are transferred to the **Signal Detector** which is charged with trend analyses using burst detection (Imran et al., 2015; Zhang, 2006). That means the frequency of relevant tweets is monitored and time series with exponential growth in frequency are identified (cf. Figure 1).

To introduce the proposed approach, we will give an example course of social signal detection of crisis-related indicators in Twitter data, starting with the domain-specific keywords provided by the user, e.g., *"Blackout OR (rising energy costs) OR (high energy costs) OR (energy shortage) OR (supply security AND energy) OR (energy crisis) OR (energy supply), etc"*. The example course ends with alerts for signals indicating a potential energy-related crisis situation in the future. We apply OSOS on tweets extracted using the Twitter API[2] that provides actual and historical data in real-time. We extracted tweets from Germany (#21,251), the UK (#7,016), and Spain (#18,696) starting from 2020 to 2022. For the following, imagine an end user, e.g., a government organization or company looking for signals for impending energy-related crisis situations due to reduced availability or rising costs of energy such as gas, oil, coal, etc.

3.1. Data Pre-processing

Pre-processing of extracted tweets is done by removing special characters, converting them to lowercase, and deleting duplicates and empty strings in tweets (cf. Figure 1). Following this, all punctuation is removed from the data. By applying the tokenizer library SoMaJo (Proisl and Uhrig, 2016), tweets are split into sentences and tokens, e.g., *['new', 'the', 'energy', 'crisis', 'isnt', 'just', 'european', 'problem', 'it', 'threatens', 'to', 'raise', 'prices', 'for', 'millions', 'around', 'the', 'globe']*.

3.2. Data Filtration

As the extracted tweets often contain irrelevant and noisy data, it's crucial to filter the tweets based on the user-specified crisis domain. Thus, we need a classification model that can accurately classify tweets within the crisis domain context. To achieve that, we leverage the power of the pre-trained GPT-3 model (GPT-3^{PT}) (Brown et al., 2020), well known for its capabilities in handling large-language tasks. However, for more effective tweet classification, we need a model that is tailored to work with the crisis context. Therefore, we fine-tune the GPT-3 (Curie) model (GPT-3^{FT}) on the CrisisBench dataset (Alam et al., 2021) enhancing the capability of GPT-3^{PT} [3] classifier to effectively identify crisis-related tweets. The dataset for fine-tuning consists of 156,899 tweets, split into the train (#109,796), dev (#16,008), and test (#31,095) sets.

The main objective of this fine-tuned GPT-3 model (GPT-3^{FT}) is to binary classify the crisis tweets into two categories: 'informative' and 'not_informative'. All tweets classified into the informative categories are the ones that concern the crisis. Furthermore, we also perform sentiment analysis using DistilBERT (Sanh et al., 2019) and calculate the sentiments of each tweet

[2]https://developer.twitter.com/en/docs/twitter-api

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

('positive', 'negative', 'neutral'). DistilBERT offers an advantage by being a transformer architecture (Vaswani et al., 2017) and is 60% faster, and 40% smaller than the traditional BERT model while retaining 97% of its language understanding capabilities. Also, negative and neutral sentiments are commonly associated with crisis-related events or issues people face (Lambret and Barki, 2017). Hence, we specifically focus on identifying tweets with 'negative' or 'neutral' sentiments.

After performing these filtrations, the output of this module is a data set of tweets classified as 'informative' w.r.t. crisis events and having "negative" or "neutral" sentiments. All other tweets are discarded as irrelevant tweets. For our example, we first input the tweets: [new the energy crisis isnt just a european problem it threatens to raise prices for millions around the globe] to our fine-tuned GPT-3^{FT} (Curie) model. It classifies it as 'informative' with a high confidence score of 95.14% and is categorized as 'negative' sentiment. Since it meets both criteria ('informative' and 'negative/neutral'), it can now proceed to the next step of signal detection.

3.3. Signal Detector

The Signal Detector applies a burst detection method for identifying early crisis signals (Imran et al., 2015; Zhang, 2006). It takes the output from the Data Filtration (cf. Figure 1), maps frequencies of tweets to time series on a daily basis, and extracts the trend component of the time series. Afterward, trends are analyzed with respect to periods of exponential growth. To determine these exponential periods, log scaling is applied to trend component values, thus periods of exponential growth can be detected as periods of linear growth on the scaled values. Linear growth is given when scaled values differ by 10%, i.e., they grow by an almost constant value.

Definition 1. *A signal is defined by a period of exponential growth that holds for a minimum of seven days with an arbitrary threshold of 10% in (tweet) frequency change.*

Such a period triggers an alert for a detected crisis-related social signal.

4. Implementation and Evaluation

Based on the proposed method OSOS (cf. Figure 1), we implemented a signaling service for open-domain social signal detection in risk and crisis management[4]. The system accepts keywords by users describing the domain of interest in the form of plain text and can be configured with respect to preferred countries, languages, and time intervals. In order to be able to process tweets from Twitter for detecting social signals, a pipeline has been deployed by using Python 3.9, PyTorch[5], and the X respectively Twitter API[6]. Having confirmed the keywords, the resulting pipeline returns potential signals of an upcoming crisis event if any including visualizations of detected signals as shown in Figure 2.

4.1. Setting

To evaluate our approach, we conducted a run-time study with the implemented signaling service. The objective of this study was to evaluate the performance of the service in identifying potential social signals for energy-related crisis events (i.e., decrease of availability and increase of costs of energy like gas, oil, coal, solar, and wind). As keywords for the domain of interest, we determined *['Blackout', 'rising energy costs', 'high energy costs', 'energy shortage', 'supply security AND energy', 'energy crisis', 'energy supply', etc]* according to (Vrana et al., 2023) and (Commission, 2022). The set of keywords is characterized by different languages. We applied a data set of 46,963 unprocessed tweets by Twitter (X) from 01.01.2020 to 31.12.2022 posted in Germany, the UK, and Spain consisting of multiple languages. Table 1 shows details of the dataset with respect to the number of tweets after pre-processing and sentiment analysis as well as the distribution among various languages. For evaluation, we identified three energy-related crisis events E in history that were predictable but hardly influenceable according to (Gundel, 2005). The crisis events E took place at a time t_0 between June 2022 and September 2022 in Germany, the UK, and Spain:

- e_1: Peak in electricity price (571.2 EUR/MWh): In Germany, electricity price reached an all-time high on 25th August 2022[7]. A major reason for this is the Russia-Ukraine war, that led to drastic fluctuations in prices and overall market instability.
- e_2: Peak in electricity price (481.38 GBP/MWh): In the UK, electricity prices reached an all-time high starting in August-September 2022[8].
- e_3: Peak in gasoline price (2.29 EUR/L): In Spain, there was a significant rise in gasoline prices in June 2022. The initial rise in prices was driven by the post-pandemic recovery in gas demands and

[4]Link to the code: https://github.com/InformationServiceSystems/pairs-project/tree/main/Modules/OSOS

[5]https://pytorch.org/

[6]https://developer.twitter.com/en/docs/twitter-api

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

[8]https://tradingeconomics.com/united-kingdom/electricity-price

Event (E)	#tweets	#data_processing	#languages	#English	#German	#Spanish	#data_filtration
e_1	21,251	20,243	39	15,003	4,240	247	3,036
e_2	7,016	7,011	24	6,950	5	6	937
e_3	18,696	18,572	39	15,023	1,338	673	3,069

Table 1: Distribution of tweets across different stages of OSOS for all events (E). Total extracted tweets (#tweets), post-processing (#data_processing), total languages (#languages), and count of 'informative' tweets with 'negative' or 'neutral' sentiment (#data_filtration)

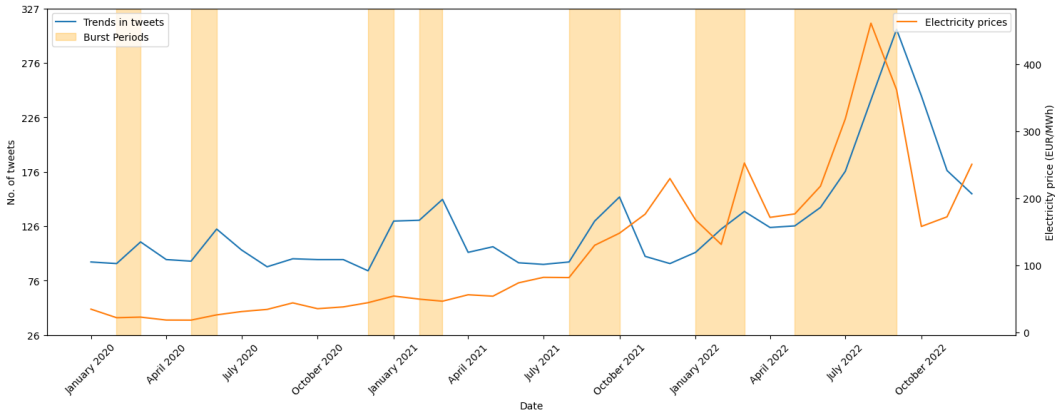
aggravated in early 2022 due to the Russia-Ukraine war[9].

The performance of the system in detecting signals for those events E was evaluated for time intervals of 4 (t_{-1}), 8 (t_{-2}), and 12 (t_{-3}) weeks in advance of the crisis event $e \in E$ (cf. Table 2).

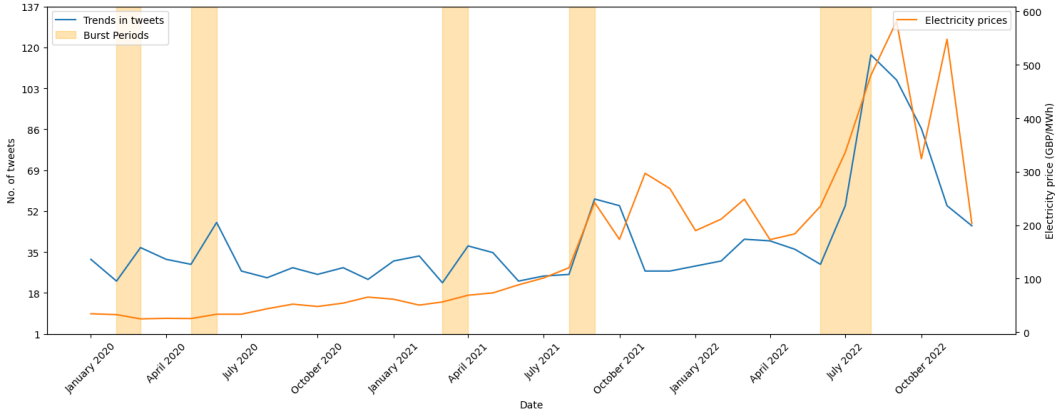
4.2. Results

In summary, after data pre-processing and filtration (cf. Figure 1), the final data set for the run-time study covered 7042 tweets (e_1: 3036; e_2: 937; e_3: 3069). Figure 2 shows the performance of the signaling service in identifying social signals for time intervals of 4 (t_{-1}), 8 (t_{-2}), and 12 (t_{-3}) weeks in advance of energy-related crisis events E between January 2020 to December 2022 in Germany, UK and Spain. For each point in time $t \in T$, we examined the exponential growth in the frequency of tweets for the events (E), i.e., the frequency change ($FC\%$). When points in time $t \in T$ are marked bold in Table 2, the signaling service detected social signals for the respective crisis event $e \in E$ based on the frequency change (FC%) in relevant tweets that was $\geq 10\%$ and continued for a minimum of seven days. Results show a general growing trend in the frequency of tweets relevant as indicators for energy-related crises with respect to Germany, the UK, and Spain (cf. Table 2). The peaks for the events (E) were reached in late Summer 2022 (cf. Figure 2). (e_1: 08.2022; e_2: 09.2022; e_3: 06.2022). For events e_1, the signaling service was able to detect early social signals at all points in time interval T, i.e., 4, 8, and 12 weeks before the crisis event with the significant increase in FC(%) (cf. Table 2). Also, burst periods and signal captured by OSOS is visualized in Figure 2a. For event e_2, the signaling service detected early social signals 12 weeks before the crisis event with a 35.0% frequency change (FC%) (cf. Table 2), 8 weeks prior with a notable 90.0% increase but couldn't capture the trend 4 weeks ahead. However, interestingly, in September 2022 (cf. Figure 2b), the electricity price had a significant downfall after reaching its peak, which was also captured by the OSOS 4 weeks ahead with -15.0% FC%. Comparing e_1 and e_2 with e_3, we see smaller changes in frequencies of tweets in e_3 (cf. Table 2). This is raising interesting research questions as the reasons could be issues in sampling for signal detection (7 days trend score) but also underlying effects of communication in perception of different crisis types. According to (Malecki et al., 2021) and (Coombs, 2004), social and cultural factors, immediacy, uncertainty, familiarity with similar crisis situations, personal control, trust in institutions and media, etc can shape the response to people in risk and crisis communication. Furthermore, aspects of attributions of cause and responsibility to organizations in crisis situations can have a positive or negative influence on the need for communicating about those events in social media (Schwarz, 2012). Applying these approaches to crisis communication, we intend to investigate the effect in e_3. In this case, i.e., event e_3, the signaling service captured the overall dynamic of increasing (tweet) frequency, but identified changes laid below the threshold until t_{-1}, so no signals until 4 weeks before the event occurred were detected. Furthermore, we evaluated the performance of the Tweet classification model for the data filtration module (cf. Subsection 3.2) using the test set of CrisisBench dataset (Alam et al., 2021), consisting of 31,095 crisis-related tweets. We experimented with multiple state-of-the-art text classification models to determine the best-performing model for the data filtration module. These models were fine-tuned and tested on the CrisisBench dataset, and evaluated on metrics such as accuracy, precision, recall, and f1 scores. As shown in Table 3, GPT-3^{FT} model outperformed other models for all the metrics, reporting accuracy (0.882) and f1 score (0.905). Overall, the results of the run-time study and model performance evaluation indicate a positive evaluation of the signaling service implementing OSOS regarding its ability to identify potential social signals before a crisis event occurs. This early identification of signals based on social media data enables companies, governments, and health organizations for pro-active organizational responses as well as data-driven decision-making in expectation of conflicts and crisis events.

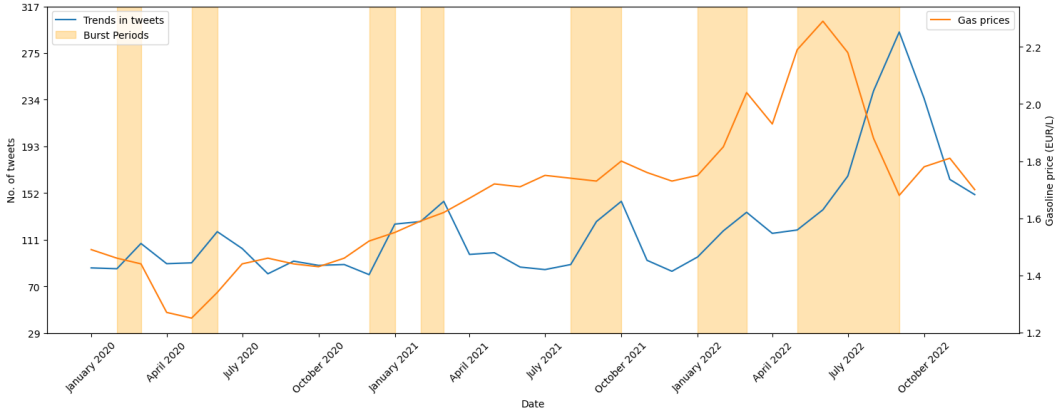
[9]https://tradingeconomics.com/spain/gasoline-prices



(a) Crisis event e_1:'Peak in electricity price', Germany, Jan 2020 - Dec 2022.



(b) Crisis event e_2:'Peak in electricity price', the UK, Jan 2020 - Dec 2022.



(c) Crisis event e_3:'Peak in gasoline price', Spain, Jan 2020 - Dec 2022.

Figure 2: Visualization of detected social signals for potential crisis events (E) with respect to the data set of tweets posted in Germany, England (UK) and Spain between 2020 and 2022.

Table 2: Results of run-time study for evaluating the performance of OSOS in identifying social signals for time intervals of 4 (t_{-1}), 8 (t_{-2}), and 12 (t_{-3}) weeks in advance of energy-related crisis events E between January 2020 and December 2022 in Germany, UK and Spain based on a data set of tweets (N = 46.963). (Legend: FC = Frequency change) Reporting averaged FC(%) score for the time intervals in (T) .

Event (E)	**Date of event (t_0)**	**Description**	t_{-1}	**FC(%)**	t_{-2}	**FC(%)**	t_{-3}	**FC(%)**
e_1 (in Germany)	08/22	Peak in electricity price	**07/22**	89.0	**06/22**	45.0	**05/22**	23.0
e_2 (in UK)	09/22	Peak in electricity price	08/22	-15.0	**07/22**	90.0	**06/22**	35.0
e_3 (in Spain)	06/22	Peak in gasoline price	**05/22**	24.0	04/22	4.0	03/22	-25.0

Table 3: Classification results of different models on predicting informativeness in CrisisBench dataset (Acc = accuracy, P = precision, R = recall, F1 = F1 score, GPT-3^{PT} = GPT-3 pre-trained, GPT-3^{FT} = GPT-3 fine-tuned)

Model	**Acc**	**P**	**R**	**F1**
Monolingual model				
CNN	0.828	0.827	0.828	0.828
fastText	0.821	0.820	0.821	0.820
BERT	0.873	0.872	0.873	0.872
DistilBERT	0.872	0.871	0.872	0.871
RoBERTa	0.880	0.879	0.880	0.879
Multilingual model				
GPT-3^{PT}	0.5992	0.8167	0.4464	0.5773
BERT-m	0.879	0.879	0.879	0.879
DistilBERT-m	0.873	0.872	0.873	0.872
xlm-RoBERTa	0.879	0.879	0.879	0.879
GPT-3^{FT}(Curie)	**0.882**	**0.897**	**0.912**	**0.905**

4.3. Limitations

OSOS is able to detect early crisis-related signals for diverse domains in multi-lingual tweets by using multiple state-of-the-art models for data pre-processing (SoMaJo) and data filtration (GPT-3^{FT} (Curie)) in combination with burst detection. The approach does not prove that all crisis events can be predicted with the same level of performance. As mentioned before, we focused on crisis events that are characterized as predictable but hardly influenceable, e.g., a rise in energy prices (Gundel, 2005). That means only crisis events can be predicted that are sending out processable signals in advance. The model is fine-tuned on the most spoken languages (English, French, Spanish, German, Italian, Portuguese, Greek, Bulgarian, Russian, Turkish, Arabic, Japanese, Corsican, Tagalog, Vietnamese, Indonesian, Chinese, Hindi, Urdu, etc) and based solely on Twitter data. Therefore, besides cultural biases as mentioned before, OSOS might face the issue of poor quality and increasing prevalence of misinformation and "fake tweets" which can influence the results. As the main data source for our model is Twitter posts, results could be manipulated by targeted mass postings of false or malicious tweets. This can be mitigated by combining OSOS with models that process further information sources for signal detection, e.g., newspaper articles, stock market data, business reports, etc. Future work will further involve verification techniques to minimize these risks by investigating advanced techniques such as transformer-based neural networks, which leverage context-rich embeddings, graph-based methods that analyze information dissemination patterns, and multimodal approaches that cross-verify textual content with associated media. These methods will be integrated into the OSOS approach, providing a comprehensive countermeasure against misinformation. Furthermore, empirical user studies are planned with the extended OSOS method for evaluating its performance in social signal detection of crisis-related indicators with domain experts from industry and civil defense.

5. Ethics Statement

Concerning the EU guidelines on ethical AI (High-Level Expert Group on Artificial Intelligence, 2019), we consider potential risks and ethical issues associated with the proposed approach, particularly in relation to the principles of (1) respect for human autonomy, (2) prevention of harm, (3) fairness, and (4) explicability.

(1) Respect for human autonomy: Our approach aims to support users, particularly companies, government, and health organizations in proactively identifying to crisis events. It respects human autonomy by providing

organizations with decision-making support rather than making decisions on their behalf. In this case, users keep full and effective self-determination over themselves having meaningful opportunities for human choice. The intention of OSOS is to "support humans by providing emerging trends and signals of upcoming crisis situations, and aim for helping the society". (High-Level Expert Group on Artificial Intelligence, 2019).

(2) Prevention of harm: AI systems should protect human dignity as well as mental and physical integrity. The proposed approach as well as the environment it operates in is safe and secure as it utilizes individual tweet data and does not generate any text but rather just filters tweets into pre-defined categories. It prevents unauthorized access and malicious use, and it understands and addresses potential biases. It takes precautions while integrating the large corpus of Twitter data to not create unintended consequences or asymmetries of power or information.

(3) Fairness: The EU guidelines on ethical AI High-Level Expert Group on Artificial Intelligence, 2019 describe a substantive and a procedural dimension of fairness. OSOS fully commits to both dimensions. Regarding the procedural dimension, the proposed approach entails the ability to contest against decisions made by OSOS (High-Level Expert Group on Artificial Intelligence, 2019), i.e., to challenge tweets that were labeled as informative.

(4) Explicability: We prioritize explicability for the decision-making process, although the OSOS model uses the GPT-3 model which is considered a black box like many deep learning models. The model may lack an explanation as to why it classified the tweet as 'informative' or 'not-informative', However, due to our fine-tuning we have explicitly extended the model to work for the crisis domain restricting its output to generate classification labels, ensuring no harmful or irrelevant text is generated. Also, we tried to bridge the gap for its black box nature and provide traceability, audibility, and transparency in the communication of the system's capabilities (High-Level Expert Group on Artificial Intelligence, 2019).

6. Conclusion

In crisis management, signal detection based on social media enables early identification of crisis-related signals supporting pro-active organizational responses before a crisis occurs. Nonetheless, up until now, social signal detection based on Twitter data has not been applied in crisis management in practice as it is challenging due to the high volume of noise. Focusing on predictable but hardly influenceable crisis events, we introduced OSOS, a method for open-domain social signal detection of crisis-related indicators in tweets. OSOS works with multi-lingual Twitter data and combines multiple state-of-the-art models for data pre-processing (SoMaJo) and data filtration (GPT-3^{FT} (Curie)), achieving benchmark results in CrisisBench dataset. It supports multiple of the most spoken languages in the world (e.g., Spanish, English) and is able to detect social signals using burst detection in tweets for open domains, e.g., energy, finances, and supply chains. OSOS was exemplified within a signaling service for risk and crisis management. We were able to evaluate the proposed approach by means of a data set of 46.963 unprocessed tweets by Twitter posted in Germany, the UK, and Spain from 01.01.2020 to 31.12.2022 in terms of performance in identifying social signals for energy-related crisis events (i.e., decreasing availability and rapidly increasing costs of energy like gas, oil, coal, solar, wind).

Acknowledgement

This work was partially funded by the German Federal Ministry of Economics and Climate Protection (BMWK) under the contract 01MK21008D.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2011). Predicting flu trends using twitter data. *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, 702–707.

Alam, F., Sajjad, H., Imran, M., & Ofli, F. (2021). Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 923–932. https://ojs.aaai.org/index.php/ICWSM/article/view/18115

Alharbi, A., & Lee, M. (2022). Classifying Arabic crisis tweets using data selection and pre-trained language models. *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, 71–78. https://aclanthology.org/2022.osact-1.8

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. *ISCRAM*, 269–272.

Aslan, S., Kiziloluk, S., & Sert, E. (2023). Tsa-cnn-aoa: Twitter sentiment analysis using cnn optimized via arithmetic optimization algorithm. *Neural Computing & Applications*, 1–18.

Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center*, *1*, 1–4.

Barbieri, F., Ronzano, F., & Saggion, H. (2015). Do we criticise (and laugh) in the same way? automatic detection of multi-lingual satirical news in twitter. *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. *Coling 2010: Posters*, 36–44.

Bonaretti, D. (2018). Effective use of twitter data in crisis management: The challenge of harnessing geospatial data.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners.

Bügel, U., & Zielinski, A. (2013). Multilingual analysis of twitter news in support of mass emergency events. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, *5*(1), 77–85.

Bundy, J., Pfarrer, M. D., Short, C. E., & Coombs, W. T. (2017). Crises and crisis management: Integration, interpretation, and research development. *Journal of Management*, *43*(6), 1661–1692. https : / / doi . org / 10 . 1177 / 0149206316680030

Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency situation awareness from twitter for crisis management. *Proceedings of the 21st international conference on world wide web*, 695–698.

Cheng, I. K., Heyl, J., Lad, N., Facini, G., & Grout, Z. (2021). Evaluation of twitter data for an emerging crisis: An application to the first wave of covid-19 in the uk. *Scientific Reports*, *11*(1), 19009.

Commission, E. (2022). Eu's response to the energy challenges. https://europa.eu/eurobarometer/surveys/detail/2912

Coombs, W. T. (2004). Impact of past crises on current crisis communication: Insights from situational crisis communication theory. *The Journal of Business Communication (1973)*, *41*(3), 265–289.

Daniel, M., Neves, R. F., & Horta, N. (2017). Company event popularity for financial markets using twitter and sentiment analysis. *Expert Systems with Applications*, *71*, 111–124.

Diks, C., Hommes, C., & Wang, J. (2019). Critical slowing down as an early warning signal for financial crises? *Empirical Economics*, *57*, 1201–1228.

Dou, W., Wang, X., Skau, D., Ribarsky, W., & Zhou, M. X. (2012). Leadline: Interactive visual analysis of text data through event identification and exploration. *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 93–102.

Earle, P. S., Bowden, D., & Guy, M. (2011). Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of geophysics*, *54*(6), 708–715.

Elsubbaugh, S., Fildes, R., & Rose, M. B. (2004). Preparation for crisis management: A proposed model and empirical evidence. *Journal of contingencies and crisis management*, *12*(3), 112–127.

Fu, K.-w., & Zhu, Y. (2020). Did the world overlook the media's early warning of covid-19? *Journal of Risk Research*, *23*(7-8), 1047–1051.

Gharavi, E., Nazemi, N., & Dadgostari, F. (2020). Early outbreak detection for proactive crisis management using twitter data: Covid-19 a case study in the us. *arXiv preprint arXiv:2005.00475*.

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

Hensgen, T., Desouza, K. C., & Kraft, G. D. (2003). Games, signal detection, and processing in the context of crisis management. *Journal of Contingencies and Crisis Management*, *11*(2), 67–77.

High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy ai [Accessed: 2023-05-08].

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 1–38.

Kruspe, A., Kersten, J., & Klan, F. (2021). Review article: Detection of actionable tweets in crisis events. *Natural Hazards and Earth System Sciences*, *21*(6), 1825–1845. https://doi.org/10.5194/nhess-21-1825-2021

Lambret, C., & Barki, E. (2017). Social media crisis management: Aligning corporate response strategies with stakeholders' emotions online. *Journal of Contingencies and Crisis Management*, *26*. https://doi.org/10.1111/1468-5973.12198

Lauras, M., & Comes, T. (2015). Special issue on innovative artificial intelligence solutions for crisis management. *Engineering Applications of Artificial Intelligence*, *46*(Part B, SI), p–287.

Li, J., & Rao, H. R. (2010). Twitter as a rapid response news service: An exploration in the context of the 2008 china earthquake. *The Electronic Journal of Information Systems in Developing Countries*, *42*(1), 1–22.

Malecki, K. M., Keating, J. A., & Safdar, N. (2021). Crisis communication and public perception of covid-19 risk in the era of social media. *Clinical infectious diseases*, *72*(4), 697–702.

Mir, A. (2023). Exploring the perceived opinion of social media users about the ukraine–russia conflict through the naturalistic observation of tweets. *Social Network Analysis and Mining*, *13*. https://doi.org/10.1007/s13278-023-01047-2

Mitroff, I. I. (1988). Crisis management: Cutting through the confusion. *MIT Sloan Management Review*, *29*(2), 15.

Palen, L., Vieweg, S., Liu, S. B., & Hughes, A. L. (2009). Crisis in a networked world: Features of computer-mediated communication in the april 16, 2007, virginia tech event. *Social Science Computer Review*, *27*(4), 467–480.

Paraskevas, A., & Altinay, L. (2013). Signal detection as the first line of defence in tourism crisis management. *Tourism Management*, *34*, 158–171. https://doi.org/https://doi.org/10.1016/j.tourman.2012.04.007

Parnell, J. A., & Crandall, W. R. (2021). What drives crisis readiness? an assessment of managers in the united states: The effects of market turbulence, perceived likelihood of a crisis, small-to medium-sized enterprises and innovative capacity. *Journal of Contingencies and Crisis Management*, *29*(4), 416–428.

Poblete, B., Guzmán, J., Maldonado, J., & Tobar, F. (2018). Robust detection of extreme events using twitter: Worldwide earthquake monitoring. *IEEE Transactions on Multimedia*, *20*(10), 2551–2561.

Proisl, T., & Uhrig, P. (2016). Somajo: State-of-the-art tokenization for german web and social media texts. *Proceedings of the 10th Web as Corpus Workshop*, 57–62.

Reuter, C., Hughes, A. L., & Kaufhold, M.-A. (2018). Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of HumanÄìComputer Interaction*, *34*(4), 280–294. https://doi.org/10.1080/10447318.2018.1427832

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. https://doi.org/10.48550/ARXIV.1910.01108

Saroj, A., & Pal, S. (2020). Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction*, *48*, 101584. https://doi.org/https://doi.org/10.1016/j.ijdrr.2020.101584

Schwarz, A. (2012). How publics use social media to respond to blame games in crisis communication: The love parade tragedy in duisburg 2010. *Public Relations Review*, *38*(3), 430–437.

Terpstra, T. (2012). Towards a realtime twitter analysis during crises for operational crisis management.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/ARXIV.1706.03762

Vrana, V., Kydros, D., Kotzaivazoglou, I., & Pechlivanaki, I. (2023). Eu citizens' twitter discussions of the 2022-23 energy crisis: A content and sentiment analysis on the verge of a daunting winter. *Sustainability*, *15*(2), 1322.

Wolbers, J., Kuipers, S., & Boin, A. (2021). A systematic review of 20 years of crisis and disaster research: Trends and progress. *Risk, Hazards & Crisis in Public Policy*, *12*(4), 374–392.

Zhang, X. (2006). *Fast algorithms for burst detection*. New York University.