Received: 8 September 2023 | Accepted: 15 August 2024

DOI: 10.1111/deci.12653

DECISION SCIENCES
A JOURNAL OF THE DECISION SCIENCES INSTITUTE
WILEY

SPECIAL ISSUE ARTICLE

Unsupervised news analysis for enhanced high-frequency food insecurity assessment

Cascha van Wanrooij[1,2] | **Frans Cruijssen**[1,2] | **Juan Sebastian Olier**[3]

[1]Department of Econometrics and Operations Research, School of Economics and Management

[2]Zero Hunger Lab, Tilburg University, Tilburg, The Netherlands

[3]Department of Cognitive Science and Artificial Intelligence, School of Humanities and Digital Sciences, Tilburg University, Tilburg, The Netherlands

Correspondence
Cascha van Wanrooij, Department of Econometrics and Operations Research, School of Economics and Management, Zero Hunger Lab, Tilburg University, Tilburg, The Netherlands.
Email: c.vanwanrooij@tilburguniversity.edu

Funding information
Kickstart AI

Abstract
This article introduces an artificial intelligence (AI)-based system for forecasting food insecurity in data-limited settings, employing unsupervised neural networks for topic modeling on news data. Unlike traditional methods, our system operates without relying on expert assumptions about food insecurity factors. Through a case study in Somalia, we show that the method can yield competitive performance, even in the absence of traditional food security indicators such as food prices. This system is valuable in supporting expert assessments of food insecurity, unlocking a wealth of untapped information from news outlets, and offering a path toward more frequent and automated food insecurity monitoring for timely crisis intervention.

KEYWORDS
food insecurity, news analysis, Somalia, time series forecasting, unsupervised topic modeling

1 | INTRODUCTION

Food insecurity, affecting more than 700 million people globally, is a widespread issue around the world (FAO, 2022). To effectively coordinate humanitarian interventions during food crises, it is crucial to objectively measure or classify the severity of food insecurity. The Integrated Food Security Phase Classification (IPC) system was developed for this purpose by the Food and Agriculture Organization (FAO) of the United Nations. The IPC system classifies regional acute food insecurity into five phases, ranging from minimal food insecurity (IPC 1) to famine (IPC 5). IPC uses evidence-based methods like weather indicators and food consumption surveys. Classifications are made through a collaborative process involving stakeholders such as local authorities and experts from various backgrounds. However, due to the time-consuming nature of this process, classifications are only published periodically, usually at intervals of 6 months (IPC Global Partners, 2021).

1.1 | Stakeholder perspective

This research project was conducted as part of a university lab that maintains strategic partnerships with organizations like the FAO and the World Food Programme (WFP) to tackle food insecurity through data science and artificial intelligence (AI). We first learned about the challenges with IPC through initial scoping discussions with FAO staff in South Sudan. They highlighted the 6-month lag between IPC updates as a major drawback limiting timely crisis response.

To explore possibilities for improving the IPC approach, in 2019 the FAO established an internal working group called Advanced Technology and Artificial Intelligence (ATARI). Our team has taken part in their meetings, and their recommended focus areas were the basis of our research project. In 2021, ATARI released a report that confirms that the IPC process has not used the full potential of technological developments yet (Armstrong et al., 2021). New opportunities from emerging technologies, notably AI, are analyzed

through a framework that considers impact and feasibility. The report notes that although consensus-based human analysis should remain core to the IPC, new technologies to improve data management, analysis efficiency, and global coverage and frequency must be explored together with the academic community. In particular, the ATARI working group identified natural language processing (NLP) of news sources as a promising area to provide more frequent interim updates in between IPC releases, enabling faster interventions in case of deteriorating food security. Specifically, the report advocates technologies like AI/machine learning (ML) for forecasting, anomaly detection, web scraping, and alternative data sources, given their high impact potential in the medium- and short term.

Our project therefore aims to demonstrate exactly this—how AI can be integrated into the IPC process to reduce cost and improve the frequency of IPC forecasts, without modification of the core IPC process. By showcasing large-scale scraping of public news data for food security forecasts, we hope to motivate broader technology integration as per the ATARI report's recommendations.

In addition, we have directly engaged with the FAO's Chief Scientist Dr. Ismahane Elouafi, and other FAO senior staff to validate these needs and obtain feedback on our approach. Dr. Elouafi has confirmed the potential of our results to augment IPC and invited us to present them at FAO's annual innovation conference.

1.2 | Academic background

Balashankar et al. (2023) have shown that utilizing news sources can provide valuable information for predicting and classifying food insecurity using the IPC system. It has been found that by incorporating news features, such as those relating to food prices and rainfall data, a linear model's power to predict future food crises (IPC phase 3+) significantly improves. The news features in question track the number of times keywords relating to food insecurity are mentioned in news articles. This article aims to expand on this approach in two ways.

First, we take advantage of recent developments in NLP. Rather than simply counting the frequency of specific keywords, we perform unsupervised topic modeling using BERTopic (Grootendorst, 2022) on news articles. This allows us to extract more context from news articles and to identify latent or hidden topics in the text data. In contrast, a keyword search using frame-semantic parsing only identifies explicitly mentioned topics (Das et al., 2014). Because we perform unsupervised topic modeling, we do not require expert-curated input. As such, this method allows one to analyze food insecurity from a broader perspective by considering diverse factors, including those not previously thought to relate to food insecurity.

Second, we include non-English news sources in our analysis. These news sources enrich our corpus with more localized information than would be possible when restricting ourselves to articles in English. This can provide a more nuanced and localized understanding of food insecurity in specific regions.

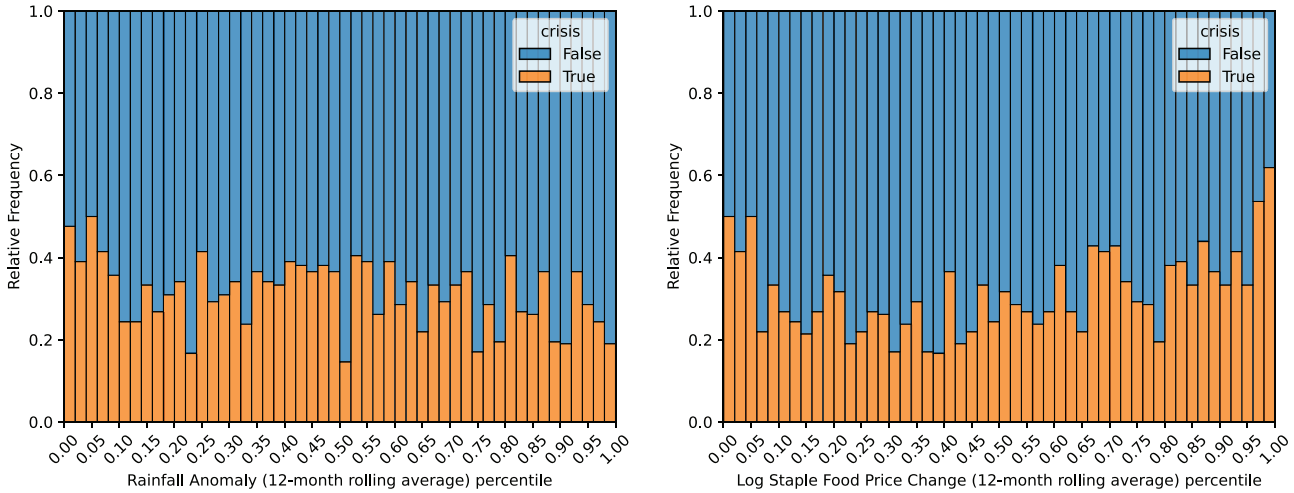
We apply the methodology developed in this article to a case study in Somalia. This country has been facing chronic food insecurity due to droughts, ongoing conflict, and instability related to civil war. The instability in the country has resulted in a lack of reliable data on local socioeconomic conditions, such as food prices, making Somalia difficult for the standard IPC process, as the country shows the highest IPC forecast error of any African country. Additionally, Somali, the native language of Somalia, has a relatively limited online presence. This is a common issue for many languages natively spoken in food-insecure nations. It is, therefore, interesting to explore how we can still take advantage of the possibilities offered by large language models (LLMs).

The article is organized as follows; In Section 2, we explore why a broader look at food insecurity beyond the usual factors is necessary. In Literature, we explore the current literature on using AI to model food insecurity. In Section 4, we provide a brief overview of the field of topic modeling. In Section 5, we explain how we have obtained a corpus of news articles, and how such an approach may be replicated. In Section 6, we describe the methodology to obtain topic features and use them to predict the IPC variable at varying lead times. In Section 7, we compare these topic features to existing data sources to check the soundness of our method. In Section 8, we present prediction results of our approach compared to a baseline and IPC predictions made by experts. Finally, we conclude by formulating both policy and research recommendations.

2 | MOTIVATION

In addition to using traditional sources of information, analysis of news articles may be an additional way to measure the impact or outcome of a shock on people's livelihoods and, by extension, their food security status. For example, traditional drought measurements may focus on comparing rainfall data against historical trends. However, the extent to which a drought leads to an actual deterioration of food insecurity depends on various other contextual factors, such as a region's technological development, reliance on agriculture, political stability, and access to food markets. These factors can, in principle, still be measured and incorporated into a model that predicts food insecurity. Other factors, however, such as the competence of local authorities or the impact of past events on a region's ability to deal with shocks, may be poorly reflected in existing data sources and are difficult to measure directly.

Figure 1 illustrates this point for Somalia by showing two district-specific risk factors of food insecurity: rainfall and food prices. For each percentile of the rainfall/food price data, we calculate the ratio of food crises (IPC > 3) versus no food crises a posteriori or not. As can be seen in both figures,



(a) Rainfall deviations (b) Food price changes

FIGURE 1 Relative histograms of district-specific rainfall deviations (from long-running seasonal means) and food price changes in Somalia. The histograms are split on whether a food crisis directly succeeds an observation. The data are derived from Andree et al. (2020).

only for extremely low rainfall and extremely high food price increases does the probability of a food crisis occurring reach or exceed 50%. In fact, for almost any value of annual rainfall, it is more likely that no food crises will follow. Even if, in a given district, the average annual food price change exceeds the 95th percentile, and average annual rainfall is lower than the 5th percentile, we only see a 52.6% probability of a food crisis occurring. Further contributing to the difficulty in predicting food crises, is that a significant portion of food crises occur without the presence of low rainfall and/or high food prices.

Similar observations can be made for other risk factors, including agronomic indicators. This signifies the "data gap" causing the difficulty in modeling food crises. As we look at the year directly preceding a food crisis, one may note that predicting, rather than nowcasting food crises is even more difficult.

Therefore, we argue for a broader perspective on food insecurity analytics. Previous studies have tried to include features such as conflict fatalities, derived from the Armed Conflict Location & Event Data Project database (ACLED) (Raleigh et al., 2010), into a model for predicting food crises, unfortunately, results are inconclusive (see Section 3).

In cases where the causes of food insecurity are not fully understood, the stories of (local) journalists may together give a more accurate picture of the impact of an event, as journalists can speak to sources (first-hand) and ask them how an event impacts their livelihood. A local journalist may also be more attuned than food security experts to the cultural situation of a region, which could yield a more accurate assessment of a regional situation than traditionally used data sources such as standardized food consumption surveys. Additionally, news sources are published at a higher frequency than other data sources, allowing us to capture food insecurity developments within a matter of days rather than months.

3 | LITERATURE

In addition to IPC classifications, food insecurity forecasts are released by the Famine Early Warning Systems Network (FEWS NET) up to 8 months ahead. According to the FEWS NET (FEWS NET, 2023), such expert-made forecasts are made by using an eight-step scenario development plan that first identifies the relevant factors contributing to food insecurity and then makes key assumptions on how these factors will influence household income and food sources, eventually settling on a projected IPC phase.

Given the significant time investment that is needed to make these projections, effort has been taken to (partially) automate this process. Thus, in this section, we explore earlier work on predicting food insecurity using ML and aim to identify gaps that our method can close. In Table 1, we provide an overview of key publications and some of their characteristics.

Wang et al. (2020), Westerveld et al. (2021), and Lentz et al. (2019) explicitly mention the difficulties of a lack of ground truth data in training their models. It is therefore difficult to compare the approaches quantitatively, as different metrics, dependent variables, and contexts are used. Some papers yield contrasting results, for example, Wang et al. (2020) and Westerveld et al. (2021) on the issue of conflict as a predictor of food insecurity. One commonality between these papers is that they all mention the presence of strong state dependence in food insecurity, except for Andree et al. (2020), which leaves out state dependence variables altogether. In addition, few papers provide comparisons with

TABLE 1 Publications using machine learning to nowcast/forecast food insecurity.

Publication	Year	Geographic scope	Food insecurity measure	Objective	Methodology
Mwebaze et al. (2010)	2010	Uganda, village-level	Caloric intake below 1800kcal	Nowcasting	Causal structure learning
Okori and Obua (2011)	2011	Uganda, village-level	Caloric intake below 1800kcal	Nowcasting	Support Vector machines, K-nearest neighbours
Lentz et al. (2019)	2019	Malawi, village-level	rCSI, HDDS, FCS	Forecasting	Linear regression
Wang et al. (2020)	2020	15 countries, national-scale	% in IPC 1, 2, 3+	Nowcasting/Forecasting	Linear regression
Andree et al. (2020)	2020	21 countries, district-level	IPC 3+ dummy	Forecasting	Random forest
Westerveld et al. (2021)	2021	Ethiopia, district-level	IPC phase transitions	Forecasting	Tree boosting/bagging
Martini et al. (2022)	2022	71 countries, district-level	rCSI, FCS	Nowcasting/Forecasting	Tree boosting
Balashankar et al. (2023)	2022	21 countries, district-level	IPC 3+ dummy	Forecasting	Random forest on news features
Ahn et al. (2023)	2023	9 countries, district-level	IPC level (linearly interpolated)	Nowcasting	Transformer neural network on news articles

other papers but rather rely on a specific benchmark to verify the usefulness of the model. In general, this research topic is still in its infancy. Finally, given the volatile context of food insecurity, it is not surprising that all papers recommend using the models in conjunction with current expert contextual understanding.

In general, the biggest limitation we observe in earlier work, with the exception of Balashankar et al. (2023) and Ahn et al. (2023), is that although different modeling methods are being applied, the underlying data remain roughly the same. Specifically, while climatological causes of food insecurity are well-captured by rainfall and indicators such as Normalized Difference Vegetation Index (NDVI), other causes of food insecurity—such as economic conditions or political instability, are only captured through indirect indicators like ACLED fatality count (which may be underreported), or through low-quality indicators, such as food prices that require interpolation due to significant missing data.

Based on the literature, and considering the earlier work by Balashankar et al. (2023), we formulate the following gaps to be addressed:

- **Model generalizability**: The studies often focus on specific regions or countries. This raises questions about the generalizability of the models, as the data used may be difficult to obtain in other countries. In addition, all studies listed that do not predict the IPC are conducted based on food surveys, which may be costly to attain for different periods. Future research should thus focus on developing more universally applicable models, foregoing the need to gather specific contextual data.
- **Comparison with expert forecasts**: Only Andree et al. (2020) and Balashankar et al. (2023) provide direct comparisons with expert forecasts. Thus, it is hard to assess the exact added value of the models above currently used solutions.
- **Unsupervised indicators**: All studies found use a set of features that is informed by prior knowledge of causes of food insecurity. As such, there is no potential of finding new drivers of food insecurity that were unconsidered.

In addition, the promising results of studies using contextual information, such as Lentz et al. (2019), motivate searching for a method that allows a model to make predictions using contextual information, while still maintaining flexibility to be applied to different countries. While we focus on Somalia as a case study, we demonstrate that a model utilizing only news features performs competitively with a model utilizing contextual data. Thus, nothing in our method is particular to Somalia. The same method can be applied in any other (food insecure) country, without the need for expensive on-the-ground data collection.

Our article is inspired by the work of Balashankar et al. (2023), which uses keyword search on news sources to expand the data set used in Andree et al. (2020). The motivation is that high-frequency indicators of developments in food-insecure nations are often not available or of low quality. By adding keyword counts to the data set, significant performance gains can already be made over methods that only incorporate the traditional tabular data sources. The potential success of this method has been implied by earlier works, which note the importance of high-frequency indicators for model performance (Lentz et al., 2019; Westerveld et al., 2021). In this article, however, we want to explore the viability of a completely unsupervised approach. The work of Balashankar et al. (2023) is able to automatically pick up on semantic causes of food insecurity. However, it is unable to utilize factors as of yet unconsidered in food insecurity. By not relying on predefined keywords, we reduce human bias in the modeling process.

We believe our approach can improve high-frequency food insecurity prediction. By incorporating local news sources, we ideally obtain a higher geographical granularity. Addi-

tionally, the use of unsupervised topic modeling simplifies the relatively elaborate method of Balashankar et al. (2023).

A recently published study proposes a method called HungerGist (Ahn et al., 2023); this method is broadly similar to ours, where aggregated sentence embeddings from news articles are used to predict the IPC 1 month ahead of time in nine African countries. The conclusion from this article is broadly similar, with slight but consistent improvements over a linear baseline in some countries. Unfortunately, no further direct comparison can be made, as Somalia was not considered, and no experiments with longer lead times were made.

4 | LITERATURE ON TOPIC MODELING

This section provides a brief review of the field of topic modeling, its goal, different topic modeling methods, and applications. Please refer to literature reviews for a more extensive discussion of the field (Churchill & Singh, 2022; Kherwa & Bansal, 2019; Vayansky & Kumar, 2020) and to the paper introducing BERTopic for a more recent rundown of neural methods in topic modeling (Grootendorst, 2022).

Topic modeling, an area in ML and NLP, aims to discover topics in a collection of documents (corpus) automatically. These topics, generally represented as a set of relevant words, are determined (statistically) based on the frequency and co-occurrence of patterns in a corpus (Kherwa & Bansal, 2019). Topic models are inherently unsupervised and are able to learn the semantic structure of a corpus without labeled data (Churchill & Singh, 2022).

The inception of topic modeling can be traced back to the late 1990s with the advent of latent semantic analysis (LSA) (Churchill & Singh, 2022). This method introduces the idea of using a "bag-of-words" approach, wherein each document is represented as a vector over a vocabulary $(w_1, w_2, \ldots, w_n) \in \mathbb{R}^n$, where w_i is the number of occurrences of the ith word in the vocabulary. In other words, the order of the words in a document does not matter.

Topic modeling gained further traction with the introduction of probabilistic methods such as latent Dirichlet allocation (LDA) by Blei et al. (2003). The strength of LDA compared to LSA lies in its ability to assign documents to found topics in a probabilistic manner. In essence, each document is given a distribution over topics. First, one selects the number of topics k. Then, the topic distribution for each document is assumed to be drawn from the Dirichlet allocation distribution. The distribution's likelihood is maximized to arrive at the most likely parameters for the given corpus, which results in a set of topics (Blei et al., 2003).

One drawback of LDA is the sparsity of word occurrences. A document may contain words that are only prevalent in that specific document, or new texts may contain many words that the model has not been trained on (Blei et al., 2003). One way to address this is by the use of LLMs, which can assign more meaningful representations to text than bag-of-words methods.

In recent years, LLMs, capable of modeling long text sequences, have resulted in significant advancements in many domains of NLP (Min et al., 2023). The advantage of LLMs is that they can model an entire text at once, taking into account the specific relations between words in the text, rather than a simple unordered collection of words as is the case in LDA. A text is first "tokenized," broken down into smaller subpieces, for example, words or subwords. The tokens are then given a unique vector embedding. One class of LLMs is called an "encoder model." This model learns to map a text (i.e., a sequence of token embeddings) to a semantically meaningful embedding space (Phuong & Hutter, 2022). Thus, two texts that use completely different words, might be embedded similarly if their content matches.

A recent method, called BERTopic (Grootendorst, 2022), greatly simplifies topic modeling by exclusively relying on document embeddings by encoder models to perform topic creation. The topics found are thus solely identified by latent semantic document structure rather than the occurrence of specific words. BERTopic consists of four distinct steps for a given corpus of n documents:

1. **Embedding**: Generate an embedding for each document using an encoder model. In the end, one will obtain n high-dimensional vectors, one for each document.
2. **Dimension reduction**: Reduce the dimensionality of the embeddings to make the next step of clustering easier. In the end, one obtains n low-dimensional vectors. For this step, we use principal component analysis (PCA).
3. **Clustering**: Apply a clustering method to the low-dimension vectors based on the Euclidean distances between the vectors. In the end, you end up with groups of clustered low-dimensional vectors. For this step, we use K-means.
4. **Identification**: Make the clusters of vectors identifiable by incorporating common information in each document group, for example, by looking at words that are relatively more common in one document group than another. The algorithm we use here is Term Frequency–Inverse Document Frequency (TF-IDF). The TF-IDF is based on the following ratio:

$$\frac{\text{Term Frequency}}{\text{Document Frequency}} = \frac{\text{Frequency of a term in the topic cluster}}{\text{Frequency of a term in the entire corpus}}.$$

For each topic cluster, we select the four keywords that have the highest TF-IDF ratio. This means these keywords have the highest amount of mentions in the articles of the cluster, relative to their mentions in the total corpus. As such, TF-IDF selects keywords highly unique to the cluster, rather than keywords that are common everywhere.

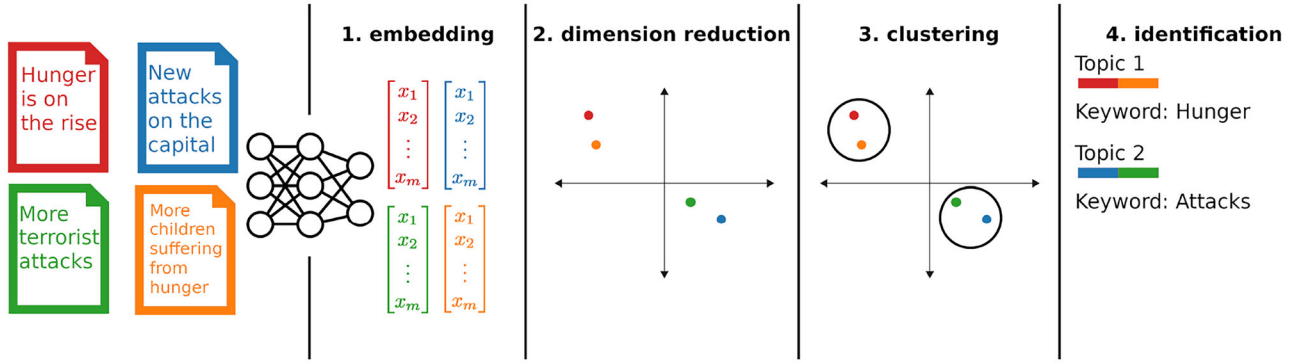


FIGURE 2 All four steps of BERTopic stylistically visualized, only four greatly simplified news articles are shown.

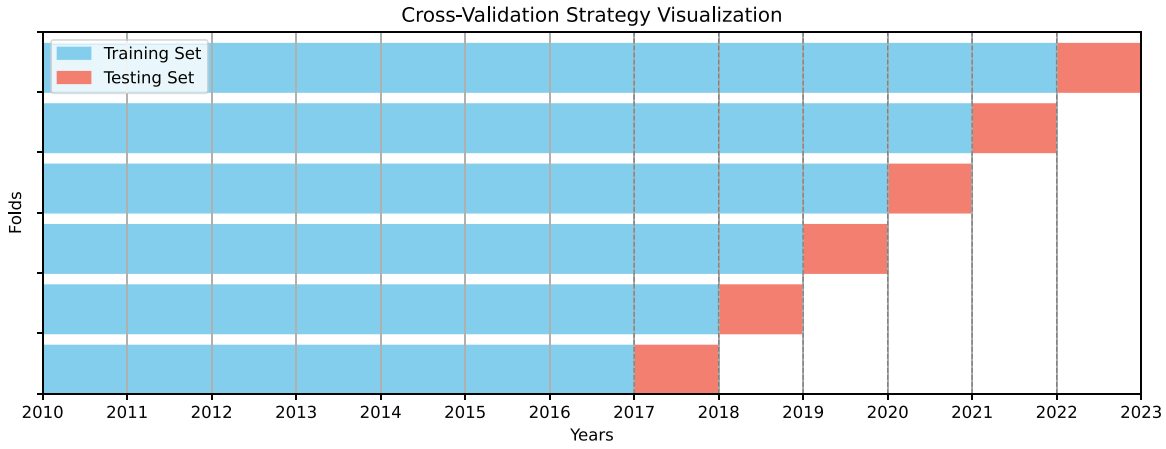


FIGURE 3 The cross-validation method used.

We want to stress that the identification step does not influence how the documents are grouped. So while it may look like BERTopic operates on keyword similarity, those keywords are merely used to make the topics easily understandable. The topics are grouped purely based on semantic similarity, as identified by the encoder model. Figure 2 provides a visual representation of the four-step process.

One great advantage of BERTopic is its flexibility; the methods at one step can be swapped for another without changing the other steps. For example, for the clustering algorithm, one can choose from methods like HDBSCAN and KMeans. On average, BERTopic performs favorably on topic coherence and diversity compared to earlier methods of topic modeling using news articles and tweets (Grootendorst, 2022), motivating our choice for BERTopic in this study. In the next section, we will elaborate on how we obtained our data set of news articles. Following that, we will explain how we use BERTopic as part of forecasting the IPC.

5 | OBTAINING NEWS ARTICLES AT SCALE

To obtain our corpus, the retrieval of a large number of news articles is automated using web-scraping techniques. We initially consider as candidates all news sources on Somalia currently online and that are listed by the Stanford library.[1] Sources that only offer access to recent articles, that is after 2019, are excluded, as are news sources that offer a limited number of articles (less than 1000). The reason for this is twofold. First, using a temporal holdout strategy (shown in Figure 3), recent sources available only in the holdout set may introduce a distributional shift between the training and holdout sets, potentially impairing model performance. Second, there is the practical concern of the difficulty in enumerating all sources. While there are only a few large news sources in Somalia, many smaller ones exist. Including these would make data collection much more time-intensive and harder to reproduce. Additionally, to keep reproduction relatively simple, we consider only news websites that have a mechanism to access historical articles easily, such as an archive search. Without these mechanisms, one would have to rely on methods such as web crawling, with no way of verifying whether all articles of a source have been obtained. Following this process, we arrived at the news sources listed in Table 2. It is worth noting that smaller or nonaccessible/online sources

[1] https://web.archive.org/web/20230328202236/ https://library.stanford.edu/africa-south-sahara/browse-topic/african-news-online-african-newspapers-countries-list/somalia

TABLE 2 Table of news sources.

News source	URL	Language	% of articles
All Africa	https://allafrica.com/	en	13.87
Allbanaadir	https://www.allbanaadir.org/	en/so	2.16
BBC Somali	https://www.bbc.com/somali/	so	7.13
Berbera News	https://www.berberanews.net/	so	3.41
Caasimada	https://www.caasimada.net/	so	14.07
Daljir	https://www.daljir.com/	so	4.51
Garowe Online	https://www.garoweonline.com/	en/so	6.76
Hiiraan Online	https://www.hiiraan.com/	en/so	25.19
Hiiraanweyn	https://www.hiiraanweyn.net/	en/so	14.16
Horn Observer	https://hornobserver.com/	en	0.43
Puntland Post	https://puntlandpost.net/	so	2.44
Somaliland Sun	https://somalilandsun.com	en/so	2.75
Wardheer News	https://wardheernews.com/	en	3.12

we skipped might significantly differ from the larger sources we considered, as highly local (e.g., region-specific) sources are often smaller and may be less prone to government censorship. Thus, while these sources are small, they might significantly alter model predictions.

Unlike other sources in Table 2, All Africa is a news aggregator and producer of African news in the English language. Most of the Somalian news articles from there are sourced from the Shabelle Media Network.[2] We collected articles from all sources by using the Python `aiohttp` library to request site contents, then using the Python `BeautifulSoup` library to extract article contents from the HTML.

In collecting these data, we made efforts to comply with ethical guidelines and best practices, including following the guidelines set out in the website's `robots.txt` file when available. Note that any data we have obtained are publicly available from the relevant websites at the time of publishing. In addition, we have made every effort to handle and protect the data responsibly and we tried to prevent overloading the servers by limiting scraping concurrency as well as by applying an exponential back-off mechanism.

It is important to acknowledge that news sources may contain significant biases, particularly in countries with high levels of state censorship, such as Somalia. We aim to create a diverse corpus of news articles from multiple independent news sources to mitigate these biases. By aggregating information from different sources, we aim to minimize the impact of biases in individual news publications by increasing the diversity of viewpoints. While it is hard to measure this bias directly, we demonstrate in Online Appendix B that there is a noticeable difference in the topics talked about between sources, lending credence to the idea that incorporating different news sources increases data set diversity. Additionally, in Online Appendix C, we demonstrate that an approach that includes all news articles is generally superior to one that leaves out a single source on prediction performance, lending further credence to the idea of minimizing biases. In Section 7, we demonstrate the correspondence between the topic features derived from our proposed news analysis and existing data sources on food insecurity, which, although also not free from bias, serve as the benchmark for currently employed methods. This correspondence helps to further validate the insights obtained from news analysis and ensure their reliability.

Instead of storing articles as a whole, the articles are broken down into paragraphs. Additionally, we store the title and the date of the article. We have attempted to localize each news article to one of Somalia's 18 regions where applicable (see Figure 7 in the Online Appendix for an annotated map). While we find evidence for the success of our localization approach, we observe deteriorated performance when including localized news article data in our method for predicting IPC. The localization approach and its effects are left out of the main text and are elaborated on in Online Appendix D.

The Somali articles have been translated to English using the Google Translate API[3] as a preprocessing step. We also considered using multilingual sentence encoders such as LABse (Feng et al., 2022) or LASER (Artetxe & Schwenk, 2019). From the source Hiiraan Online (see Table 2), we mined paraphrased articles in English and Somali. Ideally, we want articles with the same content to get assigned the same topic, regardless of language. Thus, we compared the different embedding approaches followed by clustering according to these criteria using the mutual information score and the Rand index. Here, we find it gives better results to use Google Translate preprocessing paired with an English-only sentence encoder, over directly encoding sentences using a multilingual model.

6 | METHODS

After obtaining our corpus of news articles, we want to turn the information in the news articles into features used to predict food insecurity. The process can be broken down into the following steps:

1. **Application of BERTopic**: We apply BERTopic to the entire corpus of news articles as described in Figure 2. Using K-means clustering with $K = 100$, we categorize the articles into 100 distinct topics. We present an overview of all topics and their frequencies for a model fitted up to 2022 in Table 7 in the Online Appendix.[4]
2. **Monthly topic count**: For each month in our data set, we count the number of articles associated with each of the 100 topics. This count for topic t in month m is denoted as $C_{t,m}$.

[2] https://shabellemedia.com/

[3] https://cloud.google.com/translate

[4] A careful reader might see that certain topics seem similar due to shared keywords. However, this does not imply the articles within those topics are alike. They might differ in aspects like sentiment. Nevertheless, it is important to note that LLMs are not flawless in detecting underlying semantic patterns.

3. **Calculation of relative topic frequency**: First, we compute the total number of articles published per month, denoted as N_m. Then, for each topic in each month, the relative frequency is calculated as $F_{t,m} = \frac{C_{t,m}}{N_m}$, indicating the proportion of articles about topic t in month m.
4. **Formation of topic frequency matrix**: The relative frequencies $F_{t,m}$ are put into a $100 \times n$ matrix M, with n being the number of months in the data set, to represent the topic frequencies over time. To smoothen these frequencies, we apply an averaging rolling window of 3 months.
5. **Dimensionality reduction via PCA**: We apply PCA with $P = 5$ components to M. This step reduces the dimensionality of our data, grouping topics that move similarly over time to reduce multicollinearity and to prevent overfitting. This step results in a $(5, n)$ matrix denoted D.
6. **Integration with classic food insecurity features**: The features D derived from PCA are joined with classic food insecurity features Z. The classic food insecurity features are based on data collected by Somalia's Food Security and Nutrition Analysis Unit (FSNAU). These features are:
   - *Rainfall*, estimated from remote sensing
   - *NDVI*, a remote sensing proxy for vegetation
   - *CMB*, estimated minimal cost of a food basket covering all nutrients
   - *Total alarms*, the number of factors exceeding expert-established thresholds, covering a diverse range of factors such as conflict fatalities, food prices, disease prevalence, and displacement.[5]

   These features are monitored on a monthly basis at the district level by the FSNAU. To create regional variants of these variables, an average is taken across all districts within a region. Additionally, we add to the model the previous IPC variable $IPC^{w}_{m-w,r}$; this is the most recent IPC variable available at the current time minus the forecasting window.
7. **Prediction of IPC variable**: Using these features, we predict the IPC variable out-of-sample for future time points at a lead time w of 1 to 3 months:

   For the predictions, we use ordinary least squares with the following formulation:

$$IPC^{w}_{m,r} = \beta_0 + \alpha \cdot Z_{m-w,r} + \beta D_{m-w} + \gamma IPC^{w}_{m-w,r} + \epsilon_{m,r},$$

   with m equal to the month, r equal to the region, and w equal to the lead time. Notice that we run a separate regression for each lead time w. In all instances, the predictions are rounded to the nearest integer in the range 1 to 5.

For the evaluation of our method, we compare the results of our model to a baseline model, which contains all variables considered except for the news features $D_{m-w,r}$, as well as a human baseline using FEWS NET projections of the most likely future IPC phase. In addition, we test a model that includes $IPC^{w}_{m-w,r}$ and D_{m-w}, but foregoes $Z_{m-w,r}$. We test this model in order to see how our method could replace contextual data when such data are hard to obtain. There are two types of projections: short-term projections that consider the near future (1–4 months) and medium-term projections that consider a medium horizon (6–8 months). Before 2020, projections were only released three times a year; as such, we only evaluate short-term projections after 2020, as we do not have the necessary ground truth to test the projections before 2020.

To prevent data leakage by the incorporation of future information (for instance, using data from 2018 to make predictions for 2017), we partition our data set into distinct training and test sets based on time. We employ a temporal cross-validation method, depicted in Figure 3, as it provides a more comprehensive evaluation of our approach's performance. It is important to maintain a clear division not only between the training and test data sets in the linear model but also during the topic modeling phases. Consequently, topic generation exclusively uses data available up until the end of the training phase. For instance, in the testing phase covering 2019–2020, no information about COVID-19 is included, leading to articles on COVID-19 being associated with the most relevant existing topic, such as one related to diseases. In each split, we iteratively recalibrate the BERTopic model to ensure it integrates the most up-to-date information without incorporating future data. Moreover, we abstain from interpolating any variables to avoid data leakage, opting to forward-fill all missing values instead. The model's training and evaluation are confined to the periods when the IPC is published.

In Online Appendix C, we test the robustness of various choices, such as the number of topics K and the number of principal components P, as well as test ablations of our method, such as replacing the topic modeling step by a simpler keyword counting approach. We find that our method is relatively robust to perturbations, and performs favorably to ablations. In addition, we test an approach where we incorporate region-dependent topic frequencies $F_{t,m,r}$ based on the relative number of topic mentions in articles tagged to region r. We do not find that this improves prediction performance, so it is left out in the main part of the article and can be viewed in Online Appendix D.

Our choice of PCA and K-means compared to the more complex default algorithms (HDBSCAN and, respectively, UMAP) is based on the sheer number of articles considered. With over four million paragraphs, the favorable runtime of PCA and K-means readily allowed us to iterate and experiment.

After having presented our method, we spend the next two sections evaluating it. We do this in two ways. First, we want to establish to the best of our ability how grounded the relative topic frequencies $F_{t,m}$ are to Somalia's developments over time. For this, we use data from the FSNAU to see if topics with the highest absolute correlation to food insecurity features are indeed logically related to those features in a way

[5] For the threshold definition per variable, please visit: https://dashboard.fsnau.org/application/cache/images/EWEA_Dashboard_Indictor_Thresholds.pdf

one would expect. Next, we evaluate whether the final result of our method, the predictions of the IPC variable, forms an improvement over currently used methods. We do this by comparing the predictions to a similar baseline model that does not contain the matrix D, in addition to IPC projections made by FEWS NET experts.

7 | RELEVANCE OF THE TOPIC FEATURES

To determine the extent to which our topic features are grounded in existing socioeconomic data, we take a look at the Pearson correlation over time between a classic food insecurity feature $A_{f,m}$ and a topic feature $F_{t,m}$ with f representing the type of classic feature and m representing the month, and t representing the topic ($t \in \{1, 2, \ldots, 100\}$). We then report per factor f, all topics t with a statistically significant correlation at the 0.1% confidence level. In addition, we report whether the correlation is positive or negative, as well as its magnitude. Finally, for each significant correlation, we report whether there is a plausible mechanism behind the effect. The results are reported in Table 3.

As can be seen from Table 3, there is correspondence between news and classic features. For some variables, such as ACLED fatalities, this correspondence is very clear, with all but one topic showing a clear relationship to conflict fatalities. In addition, we find that for acute malnutrition cases, half the topics are directly related to food insecurity or hospitalization. As for acute watery diarrhea (AWD) and cholera cases, there is an established link between drought and cholera (Charnley et al., 2021). Tracking food prices using topic features is challenging. Although many topics strongly correlate with sorghum prices—a major grain in Somalia—they do not directly relate to food insecurity. A logical explanation for this is that food price changes are not only related to food insecurity but also to a host of other conditions, such as violence and political stability. Another explanation for this is that in dire times, certain topics, such as piracy, may be underreported. This hinders the direct interpretability of our method and positions our method as providing a host of signals on latent socioeconomic phenomena, rather than a precise identification of food insecurity factors.

A notable example of a particularly strong correspondence between news features and classic features is the correlation between topic features and conflict fatalities reported by ACLED. Specifically, the topic feature exhibiting the strongest correlation with conflict fatalities involves topics related to killings, as detailed in Table 3. Additionally, significant correlations are observed with topics predominantly concerning attacks by the Islamist militant group Al-Shabaab, including events like bombings.

This marked correspondence between topic features and ACLED conflict fatalities is likely attributable to the common source of data: news reports. A major contributor to ACLED data in Somalia, the Shabelle Media Network, is also significantly represented in our corpus. Topics displaying the most substantial correlation with ACLED fatalities primarily involve Al-Shabaab's attacks. In contrast, topics showing the largest negative correlation are generally associated with peace talks and humanitarian efforts.

While this correlation is not a direct validation of our news features' groundedness, given that ACLED relies on second-hand evidence, it is noteworthy. The rigorous evidence-checking standards followed by ACLED analysts in conflict reporting suggest the validity of our method, as we capture similar signals as done by thorough and industry-standard human analysis. It is notable to state that the ACLED database is indeed used by FEWS NET experts to make IPC forecasts and projections (FEWS NET, 2024).

8 | RESULTS

In this section, we evaluate our method as described in Section 6. As seen in Table 4, we generally observe improvements over the baseline in near-term forecasts, that is, 1, 2, and 3 months ahead, in all test periods. We find that experts perform significantly better at a 1- and 2-month lead but fall short of the news-based model at a lead time of 3 months. Based on the consistency of our results across 6 years, we conclude that the method can identify predictive signals relevant to food insecurity in the short term. Nevertheless, the gap between the news-based model and the baseline is not large. Interestingly, we see that the model that does not use contextual information, the "news-only" model, performs similarly to the baseline model.

In Table 12 of Online Appendix C, we show how all methods perform at lead times up to 12 months. At lead times longer than 3 months, the performance of our news model is less consistent and, on average, does not improve the baseline model's performance. However, we observe that the news-only model consistently outperforms both experts and the models including contextual information at forecasting windows over 5 months. Thus, it appears that contextual information from over 5 months ago harms the forecasting performance, rather than improving it. This makes sense, as Somalia experiences two rainy seasons per year, so climatological data from over 6 months ago will be outdated. In addition, food prices can change fast. The news, however, might provide signals on more structural issues, such as emerging political instability. As these issues are more long-lasting in nature, they might offer better predictability for long-term food insecurity.

The mean absolute error (MAE), however, provides only a limited overview of forecasting errors. Overpredictions of IPC may be evaluated differently than underpredictions. For example, in scenarios where the humanitarian community has fewer resources available than necessary, overpredictions of IPC might be more costly, as funds allocated due to overprediction could be better spent in other regions. Conversely, in times of resource abundance, the opposite is true. Without making any value judgments on over- or

TABLE 3 Topics with the highest degree of linear correlation to various metrics.

Considered factor	Topic keywords	Pearson's *r*	Sign	Reasonable relation
ACLED fatalities	killed, injured, explosion, attack	0.53	+	✓
	injured, died, people, hospital	0.48	+	✓
	people, war, attacks, violence	0.43	+	✓
	security, peace, military, forces	0.40	+	✓
	security, forces, alshabaab, government	0.39	+	✓
	alshabaab, mogadishu, forces, region	0.36	+	✓
	school, family, years, mother	0.33	+	
Global acute malnutrition cases	development, food, economic, countries	0.47	+	✓
	delegation, mogadishu, somali, capital	0.45	−	
	000, million, food, water	0.44	+	✓
	injured, died, people, hospital	0.41	+	✓
	attack, forces, attacks, security	0.41	−	
	security, peace, military, forces	0.37	+	
Acute watery diarrhea (AWD)/cholera cases	refugees, somalis, 000, drought	0.36	+	✓
Sorghum prices	pirates, people, police, arrested	0.48	−	
	committee, meeting, election, members	0.42	+	
	president, meeting, chairman, mohamed	0.41	+	
	somali, somalis, somalia, killed	0.40	−	
	somali, journalists, somalia, somalis	0.39	−	
	security, forces, meeting, police	0.37	+	
	somalia, peace, government, somaliland	0.37	−	
	troops, somalia, african, mission	0.35	−	
	people, said, children, women	0.33	−	
	somalia, war, attacks, troops	0.32	−	
	000, million, food, water	0.32	+	✓
	president, minister, party, mr	0.32	+	
	food, million, refugees, water	0.32	+	✓
	trump, investigation, said, security	0.31	+	
	president, hassan, sheikh, minister	0.31	+	
	saudi, war, terrorism, arabia	0.30	−	
	somali, somalis, refugee, somalia	0.30	−	
	mohamed, sheikh, ahmed, ali	0.30	+	
	development, food, economic, countries	0.30	+	✓
	attacks, war, forces, groups	0.29	−	
	africa, million, food, ethiopia	0.28	+	✓
	kenya, refugees, africa, ethiopia	0.28	−	
	troops, forces, somalia, amisom	0.26	−	

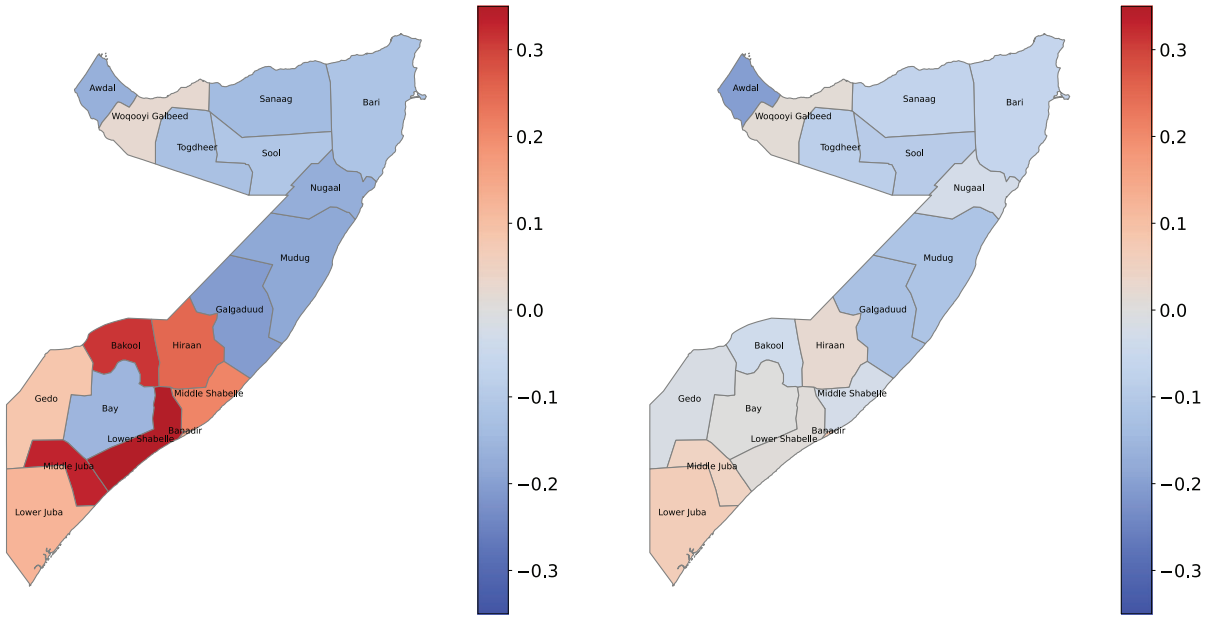
underpredictions, it is still useful to separate our errors into positive (overpredictions) and negative (underpredictions).

In Figure 4, we plot the prediction error on a map of Somalia. It should be noted that the northern regions of Somalia, consisting of the self-proclaimed autonomous regions Somaliland and Puntland, are generally more stable than the south and, on average, experience lower IPC levels. The baseline model generally underestimates IPC in the north and overestimates it in the south. The news-based model slightly increases this bias.

This increase in bias observed in the news-based model can be explained by its decreased reliance on the previous IPC value in making forecasts. Table 5 shows the average coefficients of the classical features, as well as the previous IPC value, multiplied by the data's standard deviation to capture the effect size of each variable for both models. The effect size of the IPC value is significantly smaller in the news-based model compared to the baseline. This indicates that the news-based model is more likely to deviate from the previous IPC value when making a prediction. This has both

TABLE 4 Performance of near-term forecasts in terms of mean absolute error (MAE). Bold terms represent the best performing method for a given setup.

Lead time (months)	1				2				3			
Cutoff year	**Baseline**	**Experts**	**News-based**	**News-only**	**Baseline**	**Experts**	**News-based**	**News-only**	**Baseline**	**Experts**	**News-based**	**News-only**
2017	0.39	–	**0.37**	**0.37**	0.57	–	0.54	**0.50**	0.59	–	**0.56**	**0.56**
2018	0.36	–	0.36	**0.32**	**0.14**	–	0.15	**0.14**	0.18	–	**0.17**	0.18
2019	**0.44**	–	**0.44**	**0.44**	0.57	–	**0.46**	0.57	0.56	–	**0.43**	0.50
2020	0.37	**0.15**	0.37	0.37	0.31	0.28	**0.20**	0.31	0.31	0.26	**0.22**	0.31
2021	0.48	**0.20**	0.46	0.48	0.61	**0.46**	0.63	0.61	0.61	0.59	**0.56**	0.61
2022	0.63	**0.52**	0.63	0.69	0.61	**0.33**	0.64	0.69	0.61	0.69	**0.58**	0.61
Average:	0.45	**0.29**	0.44	0.43	0.47	**0.36**	0.44	0.47	0.48	0.52	**0.42**	0.46



(a) $\hat{y}_{\text{baseline}} - y_{\text{true}}$ (b) $\hat{y}_{\text{news-based}} - \hat{y}_{\text{baseline}}$

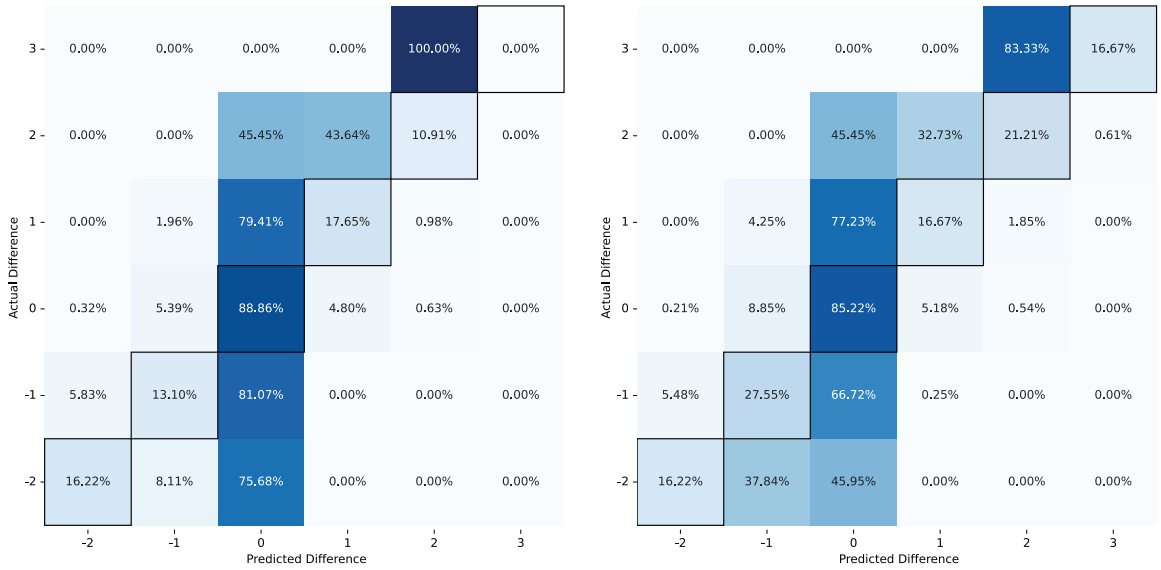
FIGURE 4 Average baseline model error, and the difference in average error between the news-based and the baseline model for forecasts up to a 3 months lead time.

TABLE 5 Average coefficients of the classical features and the lagged Integrated Food Security Phase Classification (IPC) value, standardized by multiplying with the data's standard deviations for forecasting windows of up to 3 months; 95% confidence windows are shown in parentheses.

Model	Constant	Rainfall	NDVI	Cost of minimum basket (CMB)	Total alarms	Previous IPC
Baseline	0.757 (0.671, 0.844)	−0.002 (−0.098, 0.093)	−0.082 (−0.125, −0.039)	0.004 (−0.030, 0.039)	0.006 (−0.020, 0.033)	0.586 (0.582, 0.590)
News-based	0.897 (0.837, 0.957)	−0.004 (−0.103, 0.095)	−0.088 (−0.130, −0.047)	0.018 (−0.005, 0.042)	0.100 (0.071, 0.128)	0.520 (0.515, 0.525)

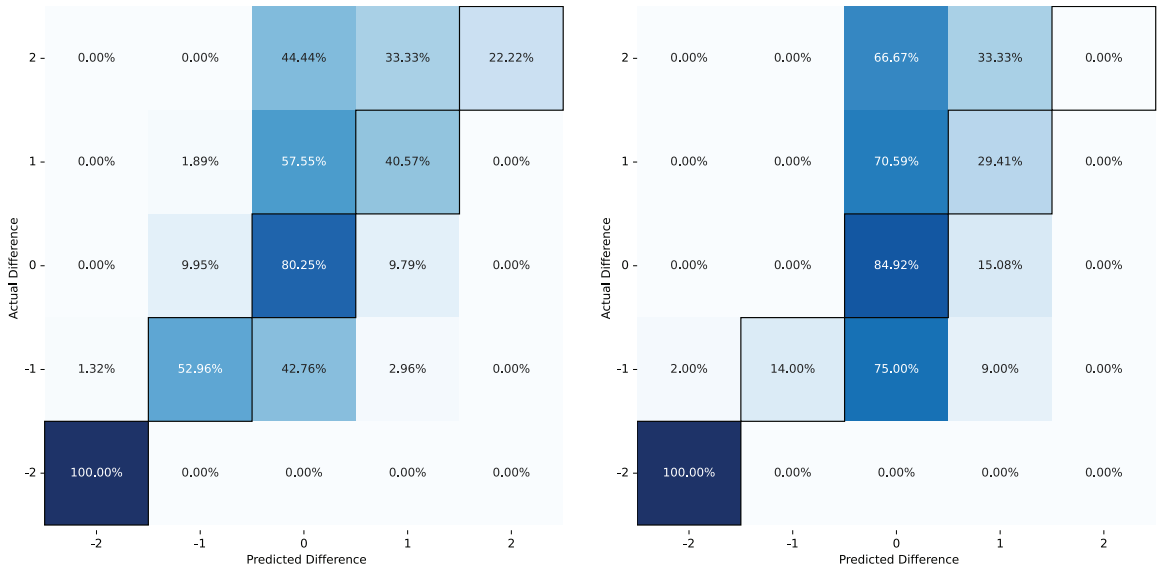
advantages and disadvantages. The upside is that the model is more likely to forecast changes in IPC, rather than making constant predictions. The downside is an increase in model bias, as shown in Figure 4.

In Figure 5, we confirm that the news-based model is generally less conservative than the baseline model and more often correctly identifies changes in IPC, rather than making constant predictions. Figure 6 shows that experts are generally good at accurately predicting changes in the very short term, but for forecasting intervals of 3 months, experts tend to be highly conservative as well. Thus, one advantage of using news as an additional data source might be its ability to more frequently identify changes while maintaining similar overall accuracy to the baseline model.



(a) Baseline model (b) News-based model

FIGURE 5 Confusion matrices of the baseline model vs. the news-based model for forecasts up to a 3 months lead time.



(a) Experts at up to a 3 months lead time. (b) Experts at strictly a 3 months lead time.

FIGURE 6 Confusion matrices of expert's predictions.

The results of this analysis are perhaps unsurprising. The starting point that led us to develop this method was the observation that the commonly used indicators, such as the NDVI (Normalized Difference Vegetation Index) and food prices only have weak predictive power on IPC, compared to a large autocorrelation factor in the IPC itself. We see that the news model relies less on autocorrelation of the IPC, and gains performance in instances where the IPC deviates from its past value. While far from perfect, our method can offer small corrective signals in IPC prediction as the news features correlate with IPC to a higher degree than the commonly used indicators.

We speculate that our news features may give a more nuanced view of past food insecurity than past IPC values alone, for example, a transient food crisis may be reported differently than one that is expected to be longer term. In addition, the IPC is constrained to five values, resulting in different underlying situations getting the same IPC rating, in this case, the news features might provide a more detailed figure. Further filtering might improve current results, or alternative avenues might be explored to further refine topic modeling by having an encoding strategy that also separates positive and negative sentiment.

TABLE 6 The topics most influential to the formation of each principal component, split by topic term, and positive versus negative loading weight.

PCA component	Negative weight	Positive weight
1	president, mohamed, sheikh, online	food, refugees, million, water, 000, children, somalis, somalia, people, migrants, development, displaced, training, emergency, teams, somaliland, aid, kenya
2	somalia, peace, somaliland, government, development, somali, war, security, somalis, kenya, constitution, political, elections, meeting, agreement	000, million, water, children, explosion, attack, killed, injured
3	said, trump, disclaimer, person, president, did, meeting, committee, parliament, election, money, education, express	somalia, war, security, somali, alshabaab, al, forces, troops, region, killed, soldiers, attack, journalists, somalis
4	somalia, war, security, somali, terrorism, saudi	meeting, agreement, government, peace, somalia, development, somaliland, federal, 000, million, water, children, training, emergency, teams, minister, president, mogadishu, somali, capital, mohamed, sheikh, online
5	refugees, somalis, 000, somalia, displaced, migrants, president, sheikh, hassan, ahmed, security, government, said, conflict, constitution, trump	somalia, government, development, somaliland, ministry, puntland, project, aid, peace, somali

To interpret our method, we take a look at the loadings of the PCA. That is to say, for each principal component, we take a look at the 10 topics most influential to that principal component. We then split the topics by positive and negative influence on the component. We opt to use the cutoff year 2020 for this analysis as it is in the middle of our cutoff year range (see Table 6). The results are broadly similar for the other years. We find that the first principal component is focused on food crises and displacement; this component is intuitively most directly tied to the IPC. However, in terms of magnitude and statistical significance of the regression coordinates, we find that components 3 and 4 perform favorably (with p-values of 0.018 and $2.6 \cdot 10^{-5}$, respectively). Component 3 seems to be focused on al-Shabaab attacks and the civil war, and component 4 seems to be focused on peace agreements and federal governance, while also having a focus on food insecurity and emergencies. Both components display a positive relationship to IPC, with coefficients of, respectively, 5.3 and 23.4. The other components are not statistically significant predictors of IPC at the 5% level. Based on these results, it seems likely the model's improvements on the baseline can be attributed to better capturing food insecurity and conflict dynamics. In particular, the link between conflict and acute food insecurity has been well established in qualitative studies (e.g., Hendrix & Brinkman, 2013), which our results are in accordance with.

9 | CONCLUSION

Our approach to enhancing food insecurity prediction introduces an unsupervised system that uses news signals. By doing so, we are able to take into account the context of the news and avoid reliance on pre-existing ideas about (causes of) food insecurity. This allows for predictions to take into account a more comprehensive understanding of the complex issue of food insecurity, besides a fixed set of factors. We argue that this broader perspective is essential, as current data analytics methods are not able to fully explain the causes and factors involved in food insecurity. Additionally, because our approach does not require expert curation, it can be easily adapted and applied to other countries.

We have shown that local news sources can easily be incorporated by using translation as a preprocessing step. We consider the use of local news sources to be preferable, as they are aligned to local context, and provide information closer to the source of the events. The system, based on local news sources, like Hiiraan Online and Caasimada, shows a high degree of correspondence with existing sources and is capable of providing predictions competitive with current food insecurity predictions without using additional data.

To our surprise, incorporating news features specific to each of Somalia's regions did not improve performance. We establish that this phenomenon is not caused by a poor article localization method. While this may be related to a lack of general data when looking at spatially desegregated features, it warrants further research. It may be that shocks on a smaller, regional scale do not influence the rather coarse IPC measure as much. Alternatively, shocks may influence the IPC in a highly nonlinear manner, unfortunately, we lack sufficient data to properly test this hypothesis.

It is important to consider that there may be significant overlap in the information used to determine an IPC phase and that used to perform near-term forecasts, leading to confirmation bias. A similar concern can be made about news-based forecasts, for example, when news sentiment moves according to IPC reports and when IPC reports move according to news sentiment. Such concerns highlight the need for a food security index based on a fixed set of measurable factors. An example of such an index is the Global Food Security Index,[6] however, it is only updated annually, at a country-level granularity.

Further, it must be noted that the performance difference between a model relying on both contextual data, as well as news features and a baseline relying only on contextual data, is minor and limited to forecasting windows below 4 months. Surprisingly, we find that a model relying merely on news features and the previous IPC value can offer consistent

[6] https://impact.economist.com/sustainability/project/food-security-index

competitive performance with models using contextual information. We find that this model slightly improves the results over all tested forecasts, including human expert ones for forecasts six or more months in advance. We thus recommend that this model be considered to augment expert forecasts in the medium to long term.

While this method's performance surpasses human expert projections in certain areas, we advise against using the model output directly in sensitive decision making, such as IPC classification. This caution arises from the relatively high margin of error, with MAEs approximately 50% of the IPC standard deviation ($\sigma_{IPC} = 0.81$) in our data set. Typically, IPC changes by only one phase at a time, making such changes significant for humanitarian actions. Although the model's average performance suggests it is generally accurate, it can still be one phase off in many instances. Therefore, we recommend integrating these models into the existing consensus-building process for IPC as an additional perspective alongside stakeholders' input. For example, if the model predicts IPC phase 3 and experts suggest IPC phase 2, or vice versa, the model's output should prompt further investigation, such as phone surveys. This approach allows the model to add value by informing investigative efforts, rather than directly influencing humanitarian planning, which could be costly if incorrect.

Relatively modest performance is also observed in our literature analysis on the topic. This is perhaps unsurprising, as the IPC is not originally made for statistical analysis, but rather to inform decision-makers in a concise manner (IPC Global Partners, 2021). One cause of this may be that the evidence on which the IPC is decided is not consistent over time; this is a reality of the turbulent environment of food-insecure nations. Thus, one general recommendation unrelated to the specifics of our method is to invest in a system besides the IPC that trades conciseness and ease of use for a larger degree of transparency and consistency. It would be easier to train and verify models on such a measure whose results could then be used to inform IPC decision making.

Additionally, it can be noted that a lack of data has led us to make design decisions in our model, which may be suboptimal if more IPC data were available. For example, the use of PCA compared to incorporating all news features separately, allows us to reduce overfitting in the linear model but also hampers interpretability. Thus, in cases where more data are available, these decisions might need to be rethought. Additionally, given the significant presence of government censorship, Somalia might be a relatively challenging context for news-based forecasting. It would be interesting to test our method in countries with similar structural food insecurity problems, but a lower degree of government censorship. In these countries, humanitarian access will also be generally higher. Thus, one could additionally test whether news analysis can provide additional value when data availability is already generally high.

ORCID


Cascha van Wanrooij https://orcid.org/0009-0002-5394-4489

REFERENCES

Ahn, Y., Yan, M., Lin, Y.-R. & Wang, Z. (2023) Hungergist: An interpretable predictive model for food insecurity. In: He, J., Palpanas, T., Hu, X., Cuzzocrea, A., Dou, D., Slezak, D., Wang, W., Gruca, A., Lin, J. C.-W. & Agrawal, R. (Eds.) *2023 IEEE International Conference on Big Data (BigData)*, Sorrento, Italy: IEEE, pp. 1591–1600.

Andree, B., Chamorro, A., Kraay, A., Spencer, P. & Wang, D. (2020) *Predicting food crises*. World Bank Working Paper, 9412.

Armstrong, J., Gabrielle, T., Halma, A., Korpi, K., Mosconi, F., Rabier, M. et al. (2021) *IPC ATARI report #2 pilots for increased coverage ABD frequency of IPC classifications*. Available at: https://www.ipcinfo.org/fileadmin/user_upload/ipcinfo/docs/IPC-ATARI-Report-_2.pdf. [Accessed 07 January 2024].

Artetxe, M. & Schwenk, H. (2019) Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597–610.

Balashankar, A., Subramanian, L. & Fraiberger, S.P. (2023) Predicting food crises using news streams. *Science Advances*, 9(9), eabm3449.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(null), 993–1022.

Charnley, G.E.C., Kelman, I. & Murray, K.A. (2021) Drought-related cholera outbreaks in Africa and the implications for climate change: a narrative review. *Pathogens and Global Health*, 116(1), 3–12.

Churchill, R. & Singh, L. (2022) The evolution of topic modeling. *ACM Computing Surveys*, 54(10s), 1–35.

Das, D., Chen, D., Martins, A.F.T., Schneider, N. & Smith, N.A. (2014) Frame-semantic parsing. *Computational Linguistics*, 40(1), 9–56.

FAO, IFAD, U.W.W. (2022) The state of food security and nutrition in the world 2022: Repurposing food and agricultural policies to make healthy diets more affordable. Rome, Italy: FAO, IFAD, UNICEF, WFP, WHO. 260.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N. & Wang, W. (2022) Language-agnostic BERT sentence embedding. In: Muresan, S., Nakov, P. & Villavicencio, A. (Eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland. Association for Computational Linguistics. pp. 878–891.

FEWS NET (2023) FEWS NET in brief. USAID-funded activity. https://fews.net/sites/default/files/2023-03/White%20Paper%20-%20SD_0.pdf [Accessed 21st December 2023].

FEWS NET (2024) *Conflict* Available at: https://fews.net/topics/conflict. [Accessed 9th January 2024].

Grootendorst, M. (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 1–10.

Hendrix, C. & Brinkman, H.-J. (2013) Food insecurity and conflict dynamics: Causal linkages and complex feedbacks. *Stability: International Journal of Security and Development*, 2(2), 26.

IPC Global Partners (2021) Integrated food security phase classification technical manual version 3.1. Evidence and standards for better food security and nutrition decisions. Rome. https://www.ipcinfo.org/fileadmin/user_upload/ipcinfo/manual/IPC_Technical_Manual_3_Final.pdf

Kherwa, P. & Bansal, P. (2019) Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), 1–16.

Lentz, E., Michelson, H., Baylis, K. & Zhou, Y. (2019) A data-driven approach improves food insecurity crisis prediction. *World Development*, 122, 399–409.

Martini, G., Bracci, A., Riches, L., Jaiswal, S., Corea, M., Rivers, J. et al. (2022) Machine learning can guide food security efforts when primary data are not available. *Nature Food*, 3(9), 716–728.

Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), Article 30, 1–40. https://doi.org/10.1145/3605943

Mwebaze, E., Okori, W. & Quinn, J.A. (2010) Causal structure learning for famine prediction. In *AAAI Spring Symposium: Artificial Intelligence for Development*. Palo Alto, California: Association for the Advancement of Artificial Intelligence (AAAI), pp. 61–66.

Okori, W. & Obua, J. (2011) Machine learning classification technique for famine prediction. In: Ao, S.I., Gelman, L., Hukins, D.W.L., Hunter, A. & Korsunsky, A.M. (Eds.) *Proceedings of the world congress on engineering*. 2011, Vol II, 991–996.

Phuong, M. & Hutter, M. (2022) Formal algorithms for transformers. *arXiv preprint arXiv:2207.09238*.

Raleigh, C., Linke, A., Hegre, H. & Karlsen, J. (2010) Introducing ACLED: An Armed Conflict Location and Event Dataset: Special data feature. *Journal of Peace Research*, 47(5), 651–660.

Vayansky, I. & Kumar, S.A. (2020) A review of topic modeling methods. *Information Systems*, 94, 101582.

Wang, D., Andree, B., Chamorro, A. & Spencer, P. (2020) *Stochastic modeling of food insecurity*. World Bank Working Paper, 9413.

Westerveld, J.J., van den Homberg, M.J., Nobre, G.G., van den Berg, D.L., Teklesadik, A.D. & Stuit, S.M. (2021) Forecasting transitions in the state of food security with machine learning using transferable features. *Science of the Total Environment*, 786, 147366.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: van Wanrooij, C., Cruijssen, F. & Olier, J.S. (2024) Unsupervised news analysis for enhanced high-frequency food insecurity assessment. *Decision Sciences*, 1–15. https://doi.org/10.1111/deci.12653

AUTHOR BIOGRAPHIES

Cascha van Wanrooij At the time of this research, Cascha was a junior researcher at the Zero Hunger Lab, with his position sponsored by Kickstart AI. His work focused on using AI to improve food security classifications and enhance decision making on humanitarian aid by applying machine learning methods to large public data sources, such as news and satellite images.

Dr. Frans Cruijssen has a background in both the corporate world and academia, having spent time at financial and logistics firms before starting a commercial venture based on academic research. Frans now works as a senior researcher at the Zero Hunger Lab of Tilburg University, where he conducts individual research, supervises PhDs, and conducts applied research with NGOs.

Dr. Juan Sebastian Olier completed his PhD in Machine Learning and Interactive Environments at Eindhoven University of Technology. He is currently an assistant professor at Tilburg University. He continues to specialize in deep learning and its applications, with a growing interest in using AI to analyze societal phenomena.