Original Paper

Automated Extraction of Mortality Information From Publicly Available Sources Using Large Language Models: Development and Evaluation Study

Mohammed Al-Garadi[1], PhD; Michele LeNoue-Newton[1], PhD; Michael E Matheny[1], MD, MS, MPH; Melissa McPheeters[2], MPH, PhD; Jill M Whitaker[1], MSN; Jessica A Deere[1], MPH; Michael F McLemore[1], BSN, RN; Dax Westerman[1], MS; Mirza S Khan[1], MS; José J Hernández-Muñoz[3], PhD; Xi Wang[3], PhD; Aida Kuzucan[3], PharmD; Rishi J Desai[4], MS, PhD; Ruth Reeves[1], PhD

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[2]Research Triangle Park, NC, United States

[3]United States Food and Drug Administration, Silver Spring, MD, United States

[4]Harvard University, Cambridge, MA, United States

Corresponding Author:

Mohammed Al-Garadi, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Avenue
Nashville, TN 37203
United States
Phone: 1 2139151696
Email: mohammed.a.al-garadi@vumc.org

Abstract

Background: Mortality is a critical variable in health care research, especially for evaluating medical product safety and effectiveness. However, inconsistencies in the availability and timeliness of death date and cause of death (CoD) information present significant challenges. Conventional sources such as the National Death Index and electronic health records often experience data lags, missing fields, or incomplete coverage, limiting their utility in time-sensitive or large-scale studies. With the growing use of social media, crowdfunding platforms, and web-based memorials, publicly available digital content has emerged as a potential supplementary source for mortality surveillance. Despite this potential, accurate tools for extracting mortality information from such unstructured data sources remain underdeveloped.

Objective: The aim of the study is to develop scalable approaches using natural language processing (NLP) and large language models (LLMs) for the extraction of mortality information from publicly available web-based data sources, including social media platforms, crowdfunding websites, and web-based obituaries, and to evaluate their performance across various sources.

Methods: Data were collected from public posts on X (formerly known as Twitter), GoFundMe campaigns, memorial websites (EverLoved and TributeArchive), and web-based obituaries from 2015 to 2022, focusing on US-based content relevant to mortality. We developed an NLP pipeline using transformer-based models to extract key mortality information such as decedent names, dates of birth, and dates of death. We then used a few-shot learning (FSL) approach with LLMs to identify primary and secondary CoDs. Model performance was assessed using precision, recall, F_1-score, and accuracy metrics, with human-annotated labels serving as the reference standard for the transformer-based model and a human adjudicator blinded to the labeling source for the FSL model reference standard.

Results: The best-performing model obtained a microaveraged F_1-score of 0.88 (95% CI 0.86-0.90) in extracting mortality information. The FSL-LLM approach demonstrated high accuracy in identifying primary CoD across various web-based sources. For GoFundMe, the FSL-LLM achieved 95.9% accuracy for primary cause identification compared to 97.9% for human annotators. In obituaries, FSL-LLM accuracy was 96.5% for primary causes, while human accuracy was 99%. For memorial websites, FSL-LLM achieved 98% accuracy for primary causes, with human accuracy at 99.5%.

Conclusions: This study demonstrates the feasibility of using advanced NLP and LLM techniques to extract mortality data from publicly available web-based sources. These methods can significantly enhance the timeliness, completeness, and granularity of mortality surveillance, offering a valuable complement to traditional data systems. By enabling earlier detection of mortality signals and improving CoD classification across large populations, this approach may support more responsive public health monitoring and medical product safety assessments. Further work is needed to validate these findings in real-world health care settings and facilitate the integration of digital data sources into national public health surveillance systems.

J Med Internet Res 2025;27:e71113; doi: 10.2196/71113

Keywords: natural language processing; NLP; large language model; LLM; mortality data; few-shot learning; public health informatics; social media analysis; obituaries; automated surveillance; health care analytics

Introduction

Mortality is a critical variable in health care research, and all-cause mortality is one of the most studied end points [1-4]. Accurate identification of the fact, timing, and cause of death (CoD) is essential for various types of medical research, including clinical trials, observational studies, and postmarketing surveillance programs such as the US Food and Drug Administration (FDA) Sentinel System [5-8].

A recent report identified limitations in the availability of date and CoD information as a major cause for study insufficiency when considering the use of the Sentinel Active Risk Identification and Analysis system to address regulatory questions [9]. Failing to identify deaths may result in substantial underestimation of mortality outcomes related to medical products, so efforts to identify additional data sources to supplement current systems have far-reaching consequences. Vital statistics data, collected in the United States through death certificates and submitted at the state level, serve as the “reference standard” for mortality information. Depending on US state laws (and sometimes the manner of death), death certificates may be completed by coroners, medical examiners, or physicians within the health care system. Once submitted to state systems, death certificate data are forwarded to the Centers for Disease Control and Prevention, which codes the underlying CoD and adds it to national records. However, this process is slow—vital statistics data are typically delayed by at least 9 months, and the National Death Index often lags by up to 2 years. There are other data sources for death information, including claims databases and medical records, but each of these sources has limitations [10,11]. Claims databases may underrepresent uninsured populations, while medical records often lack standardization between health care providers, complicating data aggregation and comparison [12]. In most claims databases, death-related information, including occurrence and CoD, is often incomplete or not directly recorded. Similarly, health care system–based data sources, such as electronic health records (EHRs), frequently lack comprehensive mortality data, particularly when patients are not under the care of the health care system at the time of death. This poses significant challenges for researchers and clinicians relying on these data sources for epidemiological studies, outcomes research, and health care quality assessments.

The rise in the use of social media has introduced potential sources of mortality-related information, including web-based obituaries and the sharing of death information in social networks through Twitter (subsequently rebranded X) and other channels. There is growing precedent for the use of social media in public health and other health-related research, and user posts have been used to track illnesses [13-17], measure behavioral risk factors [18-22], localize diseases geographically [21,23,24], and analyze symptoms and medication use [25-30]. Nonetheless, a key challenge inherent in social media data for mortality information is the capacity to extract the data and CoD at scale and with replicable methods. These social media sources offer potential advantages in timeliness, context, and coverage compared to traditional mortality data sources.

In this study, we sought to develop a set of NLP tools to extract both the fact and CoD from publicly available records and to assess the relative information density of illness and death information within these records. These types of data, when combined with other sources, could improve ascertainment in downstream studies that require the use of the facts and causes of mortality among EHR and claims data analyses.

The innovative approach leverages publicly available data to provide timely insights into population health trends, potentially enabling faster responses to emerging health threats. By linking social media and obituary data with patient records, the system could offer a more comprehensive view of health outcomes and risk factors as well as system evaluation.

This study highlights the transformative potential of LLMs in comparison to traditional NLP approaches. While earlier methods rely heavily on predefined rules or extensive labeled datasets, LLMs offer greater flexibility through few-shot learning (FSL) and contextual understanding of complex narratives. This capability is especially critical for extracting nuanced mortality information from diverse and informal web-based text, where structure and terminology often vary widely across sources.

In this study, we developed and evaluated a pipeline combining transformer-based NLP models and FSL with LLMs to extract mortality information, including fact and CoD, from publicly available web-based sources. Our goal was to assess the feasibility, accuracy, and utility of this approach for supplementing traditional mortality data in health care research and surveillance.

Methods

Overview

We developed and evaluated natural language processing (NLP) techniques to extract mortality information from publicly available web-based sources, focusing on US-related data. The research methodology included data collection, NLP model development, and performance assessment.

Data Sources and Study Cohort

Data were collected from X (formerly known as Twitter), GoFundMe, web-based obituaries (Obituaries), and memorial websites (EverLoved and TributeArchive) between the years 2015 and 2022 in the United States, which are publicly available and aggregated for research purposes in accordance with fair use. The web-based obituary sources provide more robust metadata for determining inclusion criteria than records obtained from Twitter or GoFundMe. Our collection methods, therefore, differed by source.

Our search on X used around 50 derived keywords (the list is in Multimedia Appendix 1) for English-language posts while excluding non-English content. Keywords included terms like “death,” “expired,” and “deceased.” Using Twitter’s official research application programming interface, this approach yielded approximately 40 million tweets. Using similar keywords (provided in Multimedia Appendix 1), we identified and retrieved posts from GoFundMe and memorial websites (EverLoved and TributeArchive) containing mortality-related information. For obituaries, we acquired reports from 2015 to 2022, which contained millions of records. For the obituary data sources, we collected structured metadata (eg, first name, last name, date of death, date of birth, and location) and extracted accompanying textual information. NLP techniques were subsequently used on this textual content to supplement or complete missing or incomplete metadata fields. This approach allowed us to maximize the information extracted from each obituary, enhancing the overall quality and completeness of our dataset.

Reference Standard

To construct a human-based reference dataset for training and testing our models, we developed an annotation process that captured the deceased’s name, names of related individuals, key dates (including death, birth, and other relevant dates), and CoD. First, annotators were instructed to accurately classify names with postnominals, avoid names in Twitter handles, and use specific relationship attributes for related persons (eg, spouse, sibling, and child). Second, annotated dates included exact, partial, or relative expressions, with clear distinctions for death and birth dates. Third, the CoDs were annotated with attributes indicating assertion (positive, negative, and uncertain) and patient versus nonpatient (reference to the deceased or to someone else). Finally, if no relevant data were found in a document, annotators classified the document as “No data.”

A corpus of 4200 notes, 1050 from each of the data sources, was randomly sampled. We split the 4200 annotated posts from all data sources into training (70%), testing (20%), and validation (10%) datasets. The training data contained 81,082 tokens (words), and the test data contained 27,834 tokens (words).

Annotation

The data were annotated by 3 trained nurse annotators who closely followed a detailed annotation guideline, categorizing each post into first and last names, dates of birth, dates of death, and CoD. The training was initiated using records from Twitter, GoFundMe, the memorial website (EverLoved or TributeArchive), and Obituaries, with all 3 annotators independently labeling the same documents in rounds of 15 documents from each source (n=45).

After each training round of annotation was completed by all 3 members, agreement rates were computed between pairs of annotation sets. The overall interannotator agreement (IAA) was evaluated using Cohen $\varkappa$ [31], and annotators were required to achieve an overall IAA threshold of 0.80 on the training set before proceeding with the full annotation process. When the targeted threshold was not met, the annotation team performed a consensus annotation over each document in a given annotation round, discussed their differences, and updated or clarified the annotation guidelines. Once trained, each annotator independently labeled a subset of a corpus totaling 4200 documents (1050 per source). To assess reliability, 100 additional documents (25 per source) were randomly assigned to all 3 annotators, and an independent IAA was conducted. The eHOST annotation tool was used to annotate the documents [32].

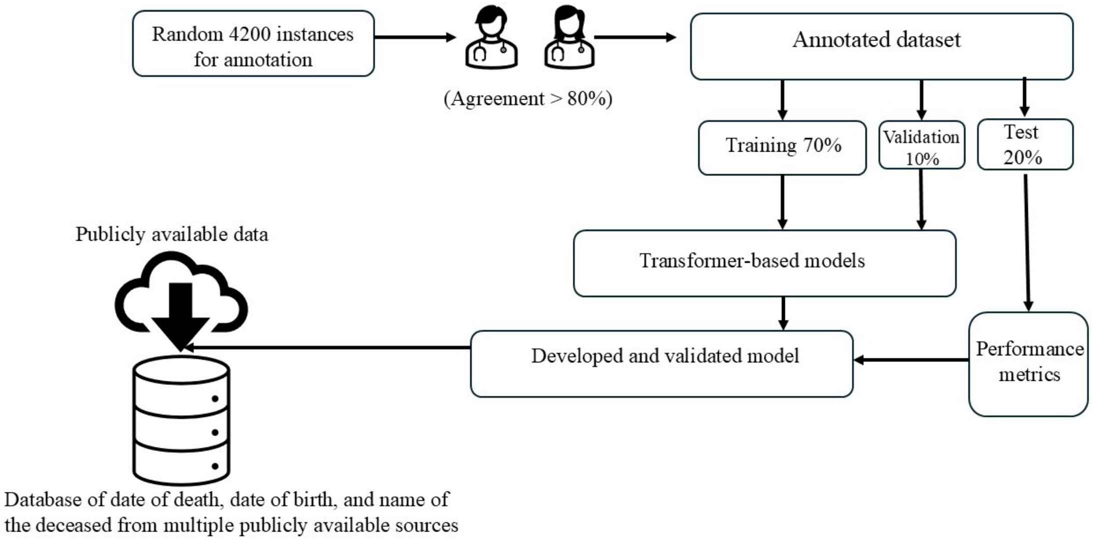
Information Density Assessment

Annotations completed by nurse annotators were used to assess the information density of web-based sources, such as social media platforms like Twitter, to determine if they contained sufficient details for reliable patient linkage and augmentation of date of death in health care systems. Sources with inadequate information were excluded from further analysis. Assessment of CoD availability was completed using the 600 document annotations used in the FSL validation with verification by the nurse adjudicator of CoDs mentioned within the post.

NLP Development and Implementation

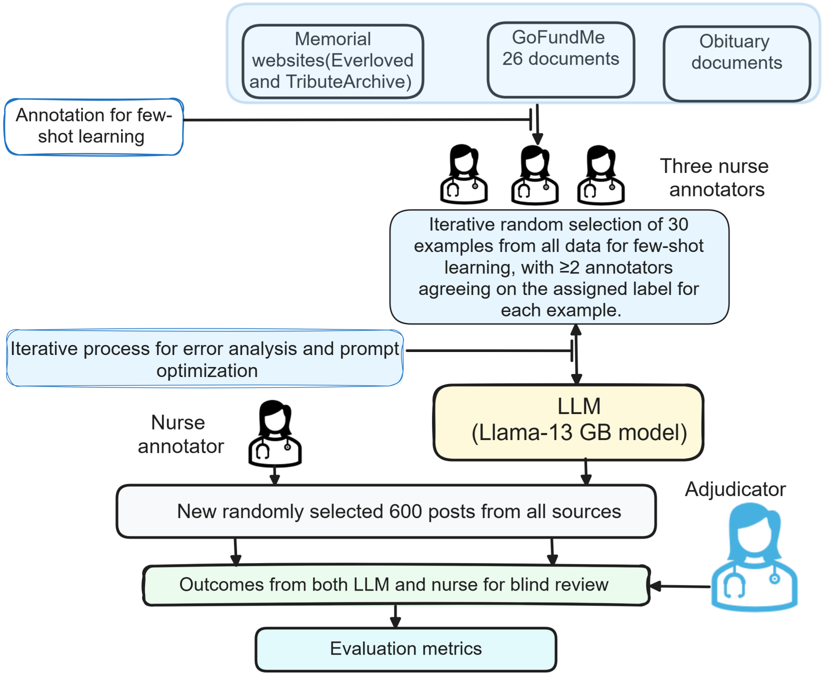
We developed in parallel 2 NLP tools for information extraction from the previously described social media sources. First, we adapted 4 deep learning transformer-based methods including Bidirectional Encoder Representations from Transformers (BERT) [33], Robustly Optimized BERT Pretraining Approach (RoBERTa) [34], A Lite BERT (ALBERT) [35], and BERTweet [36] to extract the decedent’s name, date of birth, and date of death and to exclude any irrelevant dates. The technical pipeline overview for the transformer-based model is illustrated in Figure 1.

Figure 1. Workflow of the natural language processing pipeline development and evaluation process.


Random 4200 instances for annotation
(Agreement > 80%)
Annotated dataset
Training 70%
Validation 10%
Test 20%
Transformer-based models
Publicly available data
Developed and validated model
Performance metrics
Database of date of death, date of birth, and name of the deceased from multiple publicly available sources

To identify CoD, we used an FSL approach to leverage an open-source large language model (LLM; Figure 2). The decision to forgo transformer models for this phase of information extraction was based on the need for a nuanced understanding of both the extracted cause and its contextual relevance in predicting CoD. Instead, we used an iterative prompting strategy incorporating annotated examples and structured guidelines to delineate primary and secondary CoD. High-quality annotation labels, determined by consensus between at least 2 annotators, ensured the reliability of the prompts.

Figure 2. Workflow for few-shot learning and evaluation. LLM: large language model.


Memorial websites(Everloved and TributeArchive)
GoFundMe 26 documents
Obituary documents
Annotation for few-shot learning
Three nurse annotators
Iterative random selection of 30 examples from all data for few-shot learning, with ≥2 annotators agreeing on the assigned label for each example.
Iterative process for error analysis and prompt optimization
LLM (Llama-13 GB model)
Nurse annotator
Adjudicator
New randomly selected 600 posts from all sources
Outcomes from both LLM and nurse for blind review
Evaluation metrics

For example, in the post, “Jane Smith died from a severe infection following surgery. She also had diabetes and hypertension, which contributed to her deteriorating health,” the main cause would be noted as “severe infection following surgery,” and the secondary causes as “diabetes” and “hypertension.” The initial prompt engineering stage ensures that the LLM properly formulates the type of information to extract or predict. We used the LLaMA model, a 13 GB language model developed by Facebook AI Research, for processing the data [37]. LLaMA, which stands for “Large Language Model Meta AI,” is a foundational language model that exhibits remarkable performance across various NLP tasks [37]. A smaller version, such as the 13 GB variant, of the LLaMA model can be run locally on a machine with sufficient computational resources, making it more accessible and efficient for certain applications. We started with 30 randomly selected examples from the manually annotated data (training split) for prompting (the LLM-prompt example is available in Multimedia Appendix 1) and 30 for assessing the model’s performance, where at least 2 annotators agreed on the annotated instance. The prompts and assessment examples went through several iterations of LLM refinement, totaling 4 iterations, until the identified CoD was correct in most cases across the various assessment sets. The accuracy during the prompting process was evaluated qualitatively to understand where the model performed correctly and where it made errors.

During the testing phase, we evaluated our final prompting design on a new set of 600 examples. The evaluation process involved 3 steps:

- A nurse annotator identified the CoD in these examples following the provided guidelines.
- Simultaneously, our refined language model (LLM) automatically extracted the CoD from the same 600 examples.
- A second trained nurse, acting as an adjudicator, independently reviewed both sets of results. This review ensured that the annotations adhered to the guidelines and that the primary CoD was accurately identified in each case.

Following the evaluation, we analyzed the results by determining the accuracy of primary CoD and additional identified causes from both the human annotator and the LLM per the adjudication. We then determined true positives, true negatives, false positives, and false negatives to compute relevant statistical metrics, allowing us to assess the accuracy and effectiveness of both human and automated CoD identification methods.

Statistical Metrics for Model Evaluation

For the transformer-based model evaluation, we calculated sensitivity, positive predictive value, and the F_1-score to evaluate model performance, and we computed microaverages for each to compute the average metric for a global measure of performance (all metric definitions are provided in Multimedia Appendix 1). We used bootstrap to calculate the CI by resampling the test set, calculating the required metrics for each resample, and using percentiles of these metrics to form the CI. We also assessed the information density of web-based posts from each data source to determine their adequacy for reliable patient linkage and mortality information augmentation in health care systems.

For the LLM CoD information extraction module, we calculated the F_1-score, accuracy, precision, and recall for the primary CoD. However, for all potential CoD, due to the variation in the number of causes and the challenge of measuring performance using traditional NLP metrics, we asked the adjudicator to qualitatively assess the number of cases where the LLM correctly identified all the contributing CoD mentioned in the posts and to determine if the LLM and human annotators correctly identified the primary CoD. As such, we addressed the ambiguity of using a static output for each social media post. The adjudicator focused on whether the predicted CoD was accurate, regardless of whether it was explicitly mentioned in the post or inferred from the overall understanding of the post.

Phrases classified as “No CoD” indicated no specific medical CoD. These included “brief or sudden or extended or chronic illness,” “unexpected” or “sudden death or passing,” “natural causes,” “no mention” of cause, “none,” and “unknown or unspecified reasons or cause.” Posts containing only such phrases were categorized as “No CoD.” Correct identification of these cases by the language model counted as true negatives in the CoD identification process. This approach ensured that vague or nonmedical descriptions were not misclassified as specific CoDs.

Application of NLP and Final Data Collection

The final phase of our study involved compiling the extracted data into a comprehensive dataset ready for analysis. We applied a series of cleaning filters and NLP techniques to ensure that only documents with reliable mortality-related information were included. This thorough process resulted in a dataset. This dataset, enriched with mortality information from various sources, is poised to serve as a valuable resource for public health surveillance and future research efforts.

Ethical Considerations

Given its focus on public health surveillance using open-source information, this research qualifies for exemption from FDA and Vanderbilt University Medical Center Institutional Review Board oversight. This study used publicly available web-based data and did not involve any interaction with human participants. Data collection and analysis were conducted in accordance with ethical guidelines and fair use principles for research purposes.

Results

Annotation IAA

Overall IAA with respect to GoFundMe achieved a 92.5% agreement rate in the final iteration, while the IAA within Twitter data maintained an 85.7% agreement rate after 3 rounds of assessment. IAA achieved within data sourced from the obituary websites demonstrated strong overall agreement, with a 91.5% agreement rate after the third round of assessment.

Information Density in Social Media

Analysis of information density in web-based posts revealed varying levels of utility for patient linkage and mortality information augmentation in health care systems. Among the examined sources, 3 demonstrated high information density for annotated names, ranging from 87.81% to 97.81% (Table 1). These sources provided sufficient detail for reliable patient identification and mortality data enhancement, whereas X had low information density for patient identification and was excluded from subsequent analysis.

Table 1. Information density of annotated names across different web-based public sources.

Source type	Names annotated, n (%)
Twitter	13.71
Obituary	96.48
GoFundMe	87.81
Memorial Website	97.81

Extracting Mortality Information Results

Evaluated on the manually annotated test data, the RoBERTa model achieved the highest overall performance for extracting the targeted information, with a microaveraged F_1-score of 0.88 (95% CI 0.86-0.90; Table 2). Confusion matrices are provided in Multimedia Appendix 1. This model outperformed others in all 3 tasks, achieving an F_1-score of 0.85 (95% CI 0.84-0.86) for decedent name, 0.89 (95% CI 0.88-0.90) for date of death, and 0.94 (95% CI 0.92-0.94) for date of birth. The ALBERT model attained an F_1-score of 0.87 (95% CI 0.86-0.89) for date of death, 0.83 (95% CI 0.82-0.86) for decedent name, and 0.91 (95% CI 0.90-0.93) for date of birth. BERTweet achieved an F_1-score of 0.90 (95% CI 0.89-0.91) for date of birth, 0.82 (95% CI 0.81-0.83) for decedent name, and 0.85 (95% CI 0.84-0.86) for date of death. BERT's performance was marginally lower, with F_1-scores of 0.81 (95% CI 0.80-0.83), 0.84 (95% CI 0.82-0.86), and 0.89 (95% CI 0.88-0.90) for decedent name, date of death, and date of birth, respectively.

Table 2. Performance comparison of fine-tuned transformer models (on named entity recognition tasks: decedent name, date of death, and date of birth).

	RoBERTa[a]			BERT[b]			ALBERT[c]			BERTweet		
	Precision (PPV[d])	Recall (sensitivity)	F_1-score (95% CI)	Precision (PPV)	Recall (sensitivity)	F_1-score (95% CI)	Precision (PPV)	Recall (sensitivity)	F_1-score (95% CI)	Precision (PPV)	Recall (sensitivity)	F_1-score (95% CI)
Decedent name	0.86	0.84	0.85 (0.84-0.86)	0.81	0.80	0.81 (0.80-0.83)	0.84	0.82	0.83 (0.82-0.86)	0.83	0.81	0.82 (0.81-0.83)
Date of death	0.87	0.91	0.89 (0.88-0.90)	0.82	0.87	00.84 (0.82-0.86)	0.86	0.88	0.87 (0.86-0.89)	0.86	0.84	0.85 (0.84-0.86)
Date of birth	0.95	0.93	0.94 (0.92-0.94)	0.90	0.89	0.89 (0.88-0.90)	0.92	0.91	0.91 (0.90-0.93)	0.91	0.89	0.90 (0.89-0.91)
Microaverage	0.88	0.88	0.88 (0.86-0.90)	0.83	0.85	0.84 (0.82-0.86)	0.85	0.87	0.86 (0.84-0.86)	0.84	0.85	0.84 (0.82-0.86)

[a] ROBERTa: Robustly Optimized BERT Pretraining Approach.
[b] BERT: Bidirectional Encoder Representations from Transformers.
[c] ALBERT: A Lite BERT.
[d] PPV: positive predictive value.

The accuracy of the primary CoD identification and all CoD identification for both FSL-LLM and human identification is as follows: for GoFundMe, FSL-LLM achieved an accuracy of 95.9% for primary cause and 56.4% for all causes, while human accuracy was 97.9% for primary cause and 93.3% for all causes. For Obituary, FSL-LLM accuracy was 96.5% for primary and 96% for all causes, with human accuracy at 99% for primary causes and 98.5% for all causes. For memorial websites, FSL-LLM accuracy was 98% for primary causes and 93.5% for all causes, whereas human accuracy was 99.5% for primary causes and 99% for all causes (Table 3).

Table 3. Accuracy of cause of death (CoD) identification (few-shot learning [FSL]-large language model [LLM] vs human).

Source	FSL-LLM: primary CoD identification accuracy (%)	Human: primary CoD identification accuracy (%)	LLM: all CoD identification accuracy (%)	Human: all CoD identification accuracy (%)
GoFundMe	95.9	97.9	56.4	93.3
Obituary	96.5	99	96	98.5
Memorial websites	98	99.5	93.5	99

The precision, recall, and F_1-score for the LLM's versus human detection of the primary CoD were computed for each source. The metrics are presented in Table 4.

Table 4. Precision, recall, and F_1-score for few-shot learning (FSL)-large language model (LLM) versus human (primary cause of death).

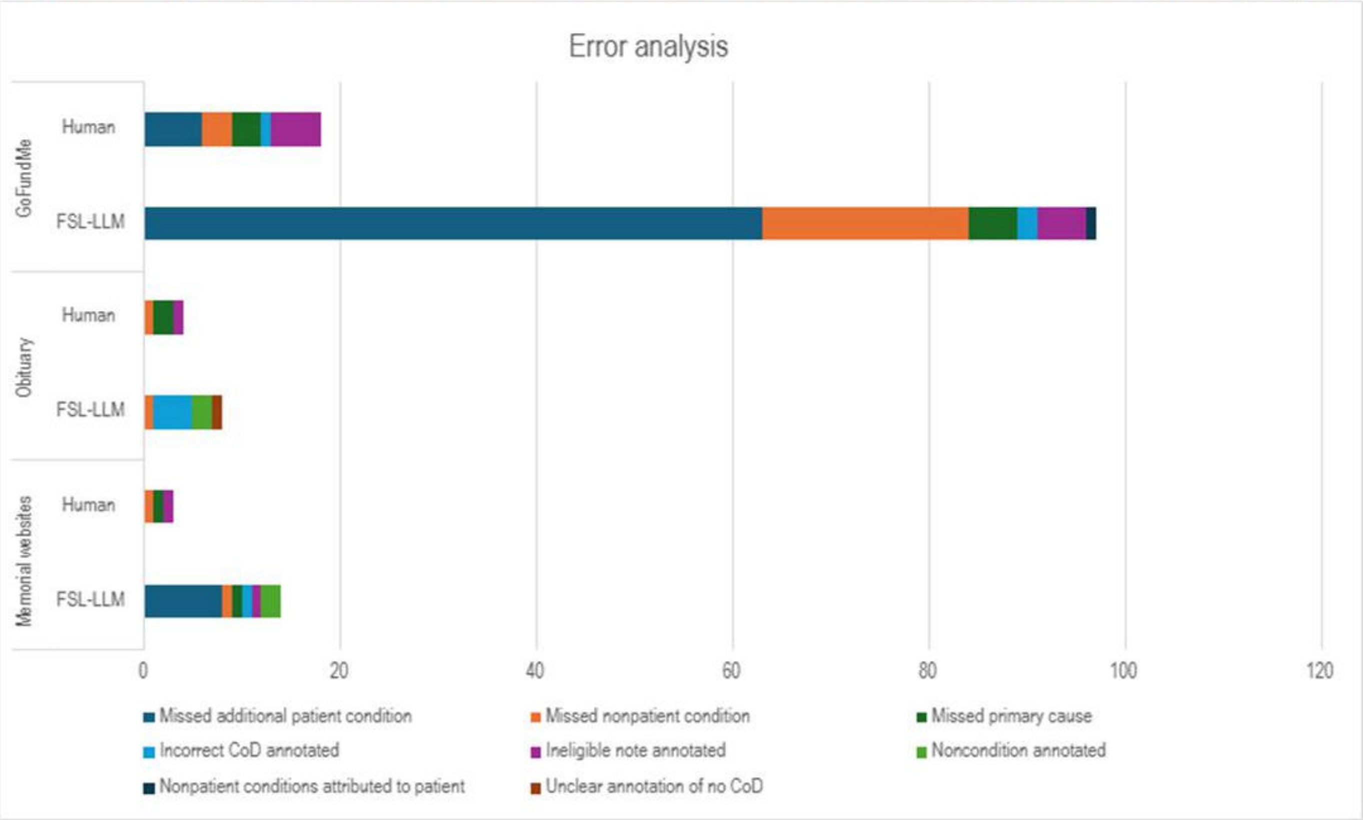
Sources	FSL-LLM			Human		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
GoFundMe	0.97	0.95	0.96	1.00	0.98	0.99
Obituary	0.61	1.00	0.76	1.00	0.82	0.90
Memorial websites	0.94	0.98	0.96	1.00	0.98	0.99

Assessment of CoD Availability and Classification Error Analysis Across Social Media Sources

As shown in Multimedia Appendix 2, social media sources varied significantly in the availability of CoD information. Obituaries had a very low density of CoD mentions, with only 6% (n=63) of 1,050 annotated posts containing any CoD. EverLoved posts primarily contained a single potential CoD, with 89% (n=934) of 1,050 posts including one cause. GoFundMe was the richest source, with 43% (n=451) of 1,050 posts containing a single CoD and 50% (n=525) containing multiple potential CoD mentions. However, not all mentioned CoDs or conditions pertained to the deceased individual referenced in the social media post.

The distribution and comparison of errors made by LLM and human annotators across the test dataset are illustrated in Figure 3. Each post may have multiple errors or error types. The analysis focuses on discrepancies in both primary and additional CoD annotations, providing a detailed breakdown of error types and frequencies. The errors include missed additional patient conditions, missed nonpatient conditions, missed primary causes, incorrect CoD annotated, ineligible notes annotated, nonpatient conditions attributed to the patient, and unclear annotation of no CoD.

Figure 3. Types of errors between LLM versus human annotation on the test dataset (errors per post). CoD: cause of death; FSL: few-shot learning; LLM: large language model.



The disparate information density across the data sources (Figure 3) influenced the types of errors found within the annotations, though the human annotator consistently had higher rates of agreement with the adjudicator than the computer annotations. Obituaries had a low density of CoD information and very low error rates. The most common error made by the FSL algorithm was in the annotation of a CoD that was not mentioned in the post (3.5%), whereas the human annotator missed a mentioned CoD in 1% of posts. For memorial websites, both human and FSL-LLM annotations exhibited a small number of errors. FSL-LLM annotations missed mentions of medical conditions in 4.5% of posts and attributed primary causes incorrectly in only 2% of posts, whereas human annotators had an error rate of less than 1% for any category. For GoFundMe, which regularly mentions multiple patient and nonpatient conditions, the FSL-LLM model has similar error rates to human annotation except for “missed nonpatient condition” (10.5%) and “missed additional patient conditions” (31.5%) categories, indicating a performance gap compared to human annotations (1.5% and 3%, respectively) in identification of all potential medical conditions within the post though very low error rate in identification of the primary CoD.

Final Collected Records

After applying the cleaning filters and NLP techniques, we successfully identified and extracted mortality-related information from a substantial number of documents across various sources. Table 5 provides a summary of the total documents retained from each source.

Table 5. Number of documents with mortality-related information identified from each source.

Source	Total documents, n
GoFundMe	23,615
Memorial website (TributeArchive and EverLoved)	733,754
Obituaries	7,375,229
Total	8,132,598

Discussion

Principal Findings

We used a novel approach to extract mortality data from web-based sources using transformer-based NLP models and FSL with LLMs. Our analysis demonstrated the effectiveness of fine-tuned transformer-based NLP models in extracting mortality data from diverse web-based sources, showcasing their potential to enhance traditional data collection methods. We also developed an FSL approach with LLMs to effectively identify primary CoD from web-based unstructured text data, achieving high agreement with human annotators. By leveraging publicly available web-based data, our approach has the potential to supplement conventional mortality databases, facilitating a more timely, comprehensive, and granular understanding of population-level mortality trends and risk factors.

Our study is consistent with other published papers that use social media data generally and obituary data specifically to improve the ability of health and health care research to accurately measure outcomes at the population level. For example, some studies have successfully used data from the Twitter platform to predict opioid overdose [38] and heart disease mortality [39], outperforming traditional demographic and health risk factors in predicting mortality. Additional studies have used GoFundMe data to identify disease categories in 89,645 medical crowdfunding campaigns [40] and to identify factors associated with cancer fundraising success [41]. An additional set of studies has used a range of techniques in web-based obituaries specifically, including automated surveillance of cancer mortality trends [42], extraction of kinship data for genetic research [43], and reporting of drug overdose [44].

Our study extends the existing literature by using transformer-based NLP models, which enhanced the extraction of key components of mortality data across public sources. Models such as RoBERTa, ALBERT, BERTweet, and BERT showed strong performance in handling unstructured data to extract decedent names (first and last), dates of birth, and dates of death, with RoBERTa achieving the highest microaveraged F_1-score of 0.88 (95% CI 0.86-0.90).

For primary CoD identification, our FSL-LLM approach demonstrated high accuracy across all sources (GoFundMe: 95.9%, obituaries: 96.5%, memorial websites: 98%), approximating human annotator performance (97.9%, 99%, and 99.5% respectively). Detailed performance metrics revealed robust results for GoFundMe (precision=0.97; recall=0.95; F_1=0.96) and memorial websites (precision=0.94; recall=0.98; F_1=0.96). Obituaries achieved high accuracy, though the precision-recall pattern (precision=0.61; recall=1.0; F_1=0.76) suggests potential for optimization in processing such data format. These findings demonstrate the model’s effectiveness while highlighting opportunities for source-specific improvements.

FSL-LLM demonstrated equivalent performance to human annotations for CoD identification across all sources; there remains room for further enhancement to identify potential contributing CoDs. The error analysis indicates that FSL-LLM exhibited higher error in categories such as “missed nonpatient condition” and “missed additional patient condition,” whereas it exhibited very low rates of error in identifying primary CoD or appropriately classifying a note as having no specific CoD noted. This was primarily noted in GoFundMe data, as it was the only data source with significant posts containing more than 1 medical condition. Targeted improvements in the model’s ability to identify nonpatient conditions and additional potential contributing causes are necessary to reduce these errors. The observed variation in error rates underscores the need for data-specific tuning to optimize model accuracy across different sources. To further enhance the FSL-LLM’s performance, focused fine-tuning on the identified error types and the integration of more diverse training datasets are recommended.

An additional finding was the low information density observed in the data from the X platform relative to the other data sources allowing linkage to specific persons. The absence of reliable person identification in the data hinders reliable patient linkage, an essential element in the augmentation of mortality information and subsequent integration into the health care system. We therefore excluded Twitter data from the analysis after the annotation phase.

Automated extraction of key mortality information from web-based sources has the potential to significantly improve traditional mortality databases, which often experience delays and incomplete data. This approach enables the timely collection of crucial details surrounding mortality, such as decedent names, dates of birth, and dates of death, which could enable linkage to other health care data sources such as EHRs to facilitate clinical research. For instance, in studies monitoring medical product safety and effectiveness using insurance claims and EHR data, such as in the FDA Sentinel system, mortality information from publicly available sources using approaches described here could allow investigators to study inferential questions regarding the impact of medical products on overall and cause-specific mortality. Integrating these methods with health care systems can improve reporting timeliness and completeness [45].

Limitations

Despite promising results, this study has several limitations. First, social media data may not fully represent all population segments due to use and sharing biases. Second, although the NLP pipeline achieved high accuracy, the inherent ambiguity and scarcity of specific CoD mentions in the source data resulted in the underdetermination of some portions of the targeted information, as indicated by the human reference standard reviewers. Consequently, the NLP system may still misclassify some data points. Additionally, our reliance on keyword-based searches may miss relevant posts because of variations in language, slang, or indirect references to mortality, and the exclusion of non-English posts may limit generalizability. Moreover, social media posts may underreport stigmatized conditions such as HIV, suicide, or opioid-related deaths due to reporting bias, and the exclusion of nonusers may further affect data representativeness. Finally, CoD identification from text remains challenging, often requiring an understanding of context and relationships between mentioned conditions. While the FSL with the LLM algorithm performed well in identifying primary CoD, further work is needed to improve its ability to extract multiple contributing causes from individual posts. The use of a nurse adjudicator may also introduce correlated errors with the human-based reference, potentially providing a conservative estimate of model performance.

Future Directions

At the population level, future research could focus on comparing CoD derived from web-based public data with those reported by official agencies. This comparison could help validate the accuracy and timeliness of web-based sourced mortality information. If validated, such data could potentially provide near real-time insights into emerging mortality trends, particularly for rapidly spreading causes such as infectious diseases or environmental exposures.

The integration of web-based sourced mortality data into existing surveillance systems would require careful validation against official records to ensure accuracy and reliability. This process would likely involve collaboration between researchers and public health agencies. Such collaborations could help develop protocols for effectively incorporating web-based data into public health surveillance and decision-making processes, potentially enhancing the speed and breadth of public health responses. Future research also should assess the plausibility of the CoD distribution from web-based sources by comparing it to sex- and age-adjusted national mortality statistics and investigating the potential underreporting of specific causes.

Conclusions

We have demonstrated a promising application of advanced NLP techniques, including transformer-based models and FSL with LLMs, to extract critical mortality information and identify CoDs from diverse web-based public data sources. The successful development of an NLP pipeline and the strong performance of the FSL algorithm highlight the potential of these approaches to address limitations in traditional mortality databases and improve the timeliness, comprehensiveness, and granularity of mortality monitoring. However, the study acknowledges several limitations, such as potential biases in web-based data representation and challenges in extracting multiple contributing CoDs. Future research should focus on validating the usefulness of these methods in real-world settings, studying the correlation between digital-derived CoDs and official records, and improving the integration of web-based data into public health surveillance systems. Addressing these challenges and opportunities will strengthen the application of advanced NLP techniques to web-based public data for enhancing mortality surveillance.

Acknowledgments

This project was supported by Task Order 75F40123F19010 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA). FDA coauthors reviewed the study protocol, statistical analysis plan, and the manuscript for scientific accuracy and clarity of presentation. Representatives of the FDA reviewed a draft of the manuscript for the presence of confidential information and accuracy regarding the statement of any FDA policy. The views expressed are those of the authors and not necessarily those of the US FDA.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the inclusion of sensitive health information and participant confidentiality constraints but are available from the corresponding author upon reasonable request. The code used for analysis is available at reference [46].

Authors' Contributions

MA-G, RR, MEM, RJD, and JJH-M conceptualized the study. MA-G, ML-N, RJD, and MEM developed the methodology. MA-G, DW, ML-N, MEM, and MSK developed the software. JAD, JMW, ML-N, RR, and MFM performed validation. MA-G, XW, and AK conducted the formal analysis. MA-G, ML-N, and JMW led the investigation. MEM, MM, and RJD provided key resources. MFM, JAD, JMW, ML-N, and RR curated the data. MA-G, RJD, MEM, ML-N, and RR drafted the manuscript. RR, RJD, MM, JJH-M, XW, and AK reviewed and edited the manuscript. MA-G, DW, and ML-N created the visualizations. RR, MEM, RJD, and JJH-M supervised the study. ML-N led project administration. MEM, RJD, JJH-M, XW, and AK acquired funding. All authors had full access to the code, contributed to the manuscript, and approved the final version for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Annotation guidelines, metric definitions, confusion matrices, and prompt examples used for extracting mortality information and causes of death from public web-based posts.
[DOCX File (Microsoft Word File), 380 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Availability of cause of death within social media posts. Social media posts were annotated for potential causes of death using the annotation guidelines outlined in Multimedia Appendix 1. Social media sources were assessed for information density related to cause of death information by categorizing a post as having either no medical conditions mentioned, a single medical condition mentioned, or having multiple unique medical conditions within the post.
[PNG File (Portable Network Graphics File), 108 KB-Multimedia Appendix 2]

References

1. Weiss NS. All-cause mortality as an outcome in epidemiologic studies: proceed with caution. Eur J Epidemiol. Mar 2014;29(3):147-149. [doi: 10.1007/s10654-014-9899-y] [Medline: 24729152]
2. Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. JAMA. Jan 2, 2013;309(1):71-82. [doi: 10.1001/jama.2012.113905] [Medline: 23280227]
3. Berrington de Gonzalez A, Hartge P, Cerhan JR, et al. Body-mass index and mortality among 1.46 million White adults. N Engl J Med. Dec 2, 2010;363(23):2211-2219. [doi: 10.1056/NEJMoa1000367] [Medline: 21121834]
4. Starfield B. Is US health really the best in the world? JAMA. Jul 26, 2000;284(4):483-485. [doi: 10.1001/jama.284.4.483] [Medline: 10904513]
5. Mieno MN, Tanaka N, Arai T, et al. Accuracy of death certificates and assessment of factors for misclassification of underlying cause of death. J Epidemiol. 2016;26(4):191-198. [doi: 10.2188/jea.JE20150010] [Medline: 26639750]
6. Lauer MS, Blackstone EH, Young JB, Topol EJ. Cause of death in clinical research: time for a reassessment? J Am Coll Cardiol. Sep 1999;34(3):618-620. [doi: 10.1016/s0735-1097(99)00250-8] [Medline: 10483939]
7. Hart JD, Sorchik R, Bo KS, et al. Improving medical certification of cause of death: effective strategies and approaches based on experiences from the Data for Health Initiative. BMC Med. Mar 9, 2020;18(1):74. [doi: 10.1186/s12916-020-01519-8] [Medline: 32146900]
8. Ter-Minassian M, Basra SS, Watson ES, Derus AJ, Horberg MA. Validation of US CDC National Death Index mortality data, focusing on differences in race and ethnicity. BMJ Health Care Inform. Jul 2023;30(1):e100737. [doi: 10.1136/bmjhci-2023-100737]
9. Antolini L, DiFrancesco JC, Zedde M, et al. Spontaneous ARIA-like events in cerebral amyloid angiopathy-related inflammation: a multicenter prospective longitudinal cohort study. Neurology (ECronicon). Nov 2, 2021;97(18):e1809-e1822. [doi: 10.1212/WNL.0000000000012778] [Medline: 34531298]
10. Riley GF. Administrative and claims records as sources of health care cost data. Med Care. Jul 2009;47(7 Suppl 1):S51-5. [doi: 10.1097/MLR.0b013e31819c95aa] [Medline: 19536019]
11. Haut ER, Pronovost PJ, Schneider EB. Limitations of administrative databases. JAMA. Jun 27, 2012;307(24):2589; [doi: 10.1001/jama.2012.6626] [Medline: 22735421]

12. Taksler GB, Dalton JE, Perzynski AT, et al. Opportunities, pitfalls, and alternatives in adapting electronic health records for health services research. Med Decis Making. Feb 2021;41(2):133-142. [doi: 10.1177/0272989X20954403] [Medline: 32969760]
13. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health. Apr 2, 2020;41:101-118. [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
14. Abdullah S, Choudhury T. Sensing technologies for monitoring serious mental illnesses. IEEE MultiMedia. 2018;25(1):61-75. [doi: 10.1109/MMUL.2018.011921236]
15. Charles-Smith LE, Reynolds TL, Cameron MA, et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. PLoS ONE. 2015;10(10):e0139701. [doi: 10.1371/journal.pone.0139701] [Medline: 26437454]
16. Dey V, Krasniak P, Nguyen M, Lee C, Ning X. A pipeline to understand emerging illness via social media data analysis: case study on breast implant illness. JMIR Med Inform. Nov 29, 2021;9(11):e29768. [doi: 10.2196/29768] [Medline: 34847064]
17. Chapman B, Raymond B, Powell D. Potential of social media as a tool to combat foodborne illness. Perspect Public Health. Jul 2014;134(4):225-230. [doi: 10.1177/1757913914538015] [Medline: 24990140]
18. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proc Int AAAI Conf Weblogs Soc Media. 2013;7. [doi: 10.1609/icwsm.v7i1.14432]
19. Centola D. Social media and the science of health behavior. Circulation. May 28, 2013;127(21):2135-2144. [doi: 10.1161/CIRCULATIONAHA.112.101816] [Medline: 23716382]
20. De Choudhury M, Counts S, Horvitz E. Predicting postpartum changes in emotion and behavior via social media. Presented at: CHI ’13: CHI Conference on Human Factors in Computing Systems; Apr 27 to May 2, 2013; Paris, France. [doi: 10.1145/2470654.2466447]
21. Al-Garadi MA, Khan MS, Varathan KD, Mujtaba G, Al-Kabsi AM. Using online social networks to track a pandemic: a systematic review. J Biomed Inform. Aug 2016;62:1-11. [doi: 10.1016/j.jbi.2016.05.005] [Medline: 27224846]
22. Naslund JA, Bondre A, Torous J, Aschbrenner KA. Social media and mental health: benefits, risks, and opportunities for research and practice. J Technol Behav Sci. Sep 2020;5(3):245-257. [doi: 10.1007/s41347-020-00134-x] [Medline: 33415185]
23. Stefanidis A, Crooks A, Radzikowski J. Harvesting ambient geospatial information from social media feeds. GeoJournal. Apr 2013;78(2):319-338. [doi: 10.1007/s10708-011-9438-2]
24. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One. 2013;8(12):e83672. [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]
25. Sarker A, Ginn R, Nikfarjam A, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform. Apr 2015;54:202-212. [doi: 10.1016/j.jbi.2015.02.004] [Medline: 25720841]
26. Al-Garadi MA, Yang YC, Cai H, et al. Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC Med Inform Decis Mak. Jan 26, 2021;21(1):27. [doi: 10.1186/s12911-021-01394-0] [Medline: 33499852]
27. Thackeray R, Neiger BL, Smith AK, Van Wagenen SB. Adoption and use of social media among public health departments. BMC Public Health. Mar 26, 2012;12:1-6. [doi: 10.1186/1471-2458-12-242] [Medline: 22449137]
28. Vereen RN, Kurtzman R, Noar SM. Are social media interventions for health behavior change efficacious among populations with health disparities?: A meta-analytic review. Health Commun. Jan 2023;38(1):133-140. [doi: 10.1080/10410236.2021.1937830] [Medline: 34148445]
29. Gough A, Hunter RF, Ajao O, et al. Tweet for behavior change: using social media for the dissemination of public health messages. JMIR Public Health Surveill. Mar 23, 2017;3(1):e14. [doi: 10.2196/publichealth.6313] [Medline: 28336503]
30. Fung ICH, Tse ZTH, Fu KW. The use of social media in public health surveillance. Western Pac Surveill Response J. 2015;6(2):3-6. [doi: 10.5365/WPSAR.2015.6.1.019] [Medline: 26306208]
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282. [Medline: 23092060]
32. South B, Shen S, Leng J, Forbush TB, DuVall SL, Chapman WW. A prototype tool set to support machine-assisted annotation. Presented at: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; Jun 8, 2012; Montreal, Canada.
33. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019. [doi: 10.48550/arXiv.1810.04805]
34. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]

35. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for self-supervised learning of language representations. arXiv. Preprint posted online on Feb 9, 2020. [doi: 10.48550/arXiv.1909.11942]
36. Nguyen DQ, Vu T, Nguyen AT. BERTweet: a pre-trained language model for english tweets. arXiv. Preprint posted online on Oct 5, 2020. [doi: 10.48550/arXiv.2005.10200]
37. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
38. Giorgi S, Yaden DB, Eichstaedt JC, et al. Predicting U.S. county opioid poisoning mortality from multi-modal social media and psychological self-report data. Sci Rep. Jun 3, 2023;13(1):9027. [doi: 10.1038/s41598-023-34468-2] [Medline: 37270657]
39. Eichstaedt JC, Schwartz HA, Kern ML, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci. Feb 2015;26(2):159-169. [doi: 10.1177/0956797614557867] [Medline: 25605707]
40. Doerstling SS, Akrobetu D, Engelhard MM, Chen F, Ubel PA. A disease identification algorithm for medical crowdfunding campaigns: validation study. J Med Internet Res. Jun 21, 2022;24(6):e32867. [doi: 10.2196/32867] [Medline: 35727610]
41. Zhang X, Tao X, Ji B, Wang R, Sörensen S. The success of cancer crowdfunding campaigns: project and text analysis. J Med Internet Res. Mar 3, 2023;25:e44197. [doi: 10.2196/44197] [Medline: 36692283]
42. Tourassi G, Yoon HJ, Xu S. A novel web informatics approach for automated surveillance of cancer mortality trends. J Biomed Inform. Jun 2016;61:110-118. [doi: 10.1016/j.jbi.2016.03.027] [Medline: 27044930]
43. He K, Yao L, Zhang J, Li Y, Li C. Construction of genealogical knowledge graphs from obituaries: multitask neural network extraction system. J Med Internet Res. Aug 4, 2021;23(8):e25670. [doi: 10.2196/25670] [Medline: 34346903]
44. Warren K. From Death Notice to the Cyber Obit: The History of the Overdose Obituary (Unpublished manuscript). URL: https://projects.iq.harvard.edu/files/historyopioidepidemic/files/katherine_warren_paper.pdf?utm_source=chatgpt.com [Accessed 2025-07-16]
45. LeNoue-Newton M, al-Garadi M, Ngan K, et al. Augmenting fact and date of death in electronic health records using internet media sources: a validation study from two large healthcare systems. medRxiv. Preprint posted online on Jan 27, 2025. [doi: 10.1101/2025.01.24.25321042]
46. Al-Garadi M, LeNoue-Newton M, Matheny ME, et al. Automated extraction of mortality information from publicly available sources using language models. medRxiv. Preprint posted online on Nov 1, 2024. [doi: 10.1101/2024.10.28.24316027]

Abbreviations

ALBERT: A Lite BERT
BERT: Bidirectional Encoder Representations From Transformers
CoD: cause of death
EHR: electronic health record
FDA: Food and Drug Administration
FSL: few-shot learning
IAA: interannotator agreement
LLM: large language model
NLP: natural language processing
RoBERTa: Robustly Optimized BERT Pretraining Approach

Edited by Javad Sarvestan; peer-reviewed by J C Heemelaar, Peter W Young; submitted 10.01.2025; final revised version received 15.04.2025; accepted 16.04.2025; published 18.08.2025

Please cite as:
Al-Garadi M, LeNoue-Newton M, Matheny ME, McPheeters M, Whitaker JM, Deere JA, McLemore MF, Westerman D, Khan MS, Hernández-Muñoz JJ, Wang X, Kuzucan A, Desai RJ, Reeves R
Automated Extraction of Mortality Information From Publicly Available Sources Using Large Language Models: Development and Evaluation Study
J Med Internet Res 2025;27:e71113
URL: https://www.jmir.org/2025/1/e71113
doi: 10.2196/71113