BAND: Biomedical Alert News Dataset

Zihao Fu[1], Meiru Zhang[1], Zaiqiao Meng[2,1], Yannan Shen[3], David Buckeridge[3], Nigel Collier[1]

[1]Language Technology Lab, University of Cambridge
[2]School of Computing Science, University of Glasgow
[3]School of Population and Global Health, McGill University
{zf268, mz468, nhc30}@cam.ac.uk, zaiqiao.meng@glasgow.ac.uk
yannan.shen@mail.mcgill.ca, david.buckeridge@mcgill.ca

Abstract

Infectious disease outbreaks continue to pose a significant threat to human health and well-being. To improve disease surveillance and understanding of disease spread, several surveillance systems have been developed to monitor daily news alerts and social media. However, existing systems lack thorough epidemiological analysis in relation to corresponding alerts or news, largely due to the scarcity of well-annotated reports data. To address this gap, we introduce the Biomedical Alert News Dataset (BAND), which includes 1,508 samples from existing reported news articles, open emails, and alerts, as well as 30 epidemiology-related questions. These questions necessitate the model's expert reasoning abilities, thereby offering valuable insights into the outbreak of the disease. The BAND dataset brings new challenges to the NLP world, requiring better inference capability of the content and the ability to infer important information. We provide several benchmark tasks, including Named Entity Recognition (NER), Question Answering (QA), and Event Extraction (EE), to demonstrate existing models' capabilities and limitations in handling epidemiology-specific tasks. It is worth noting that some models may lack the human-like inference capability required to fully utilize the corpus. To the best of our knowledge, the BAND corpus is the largest corpus of well-annotated biomedical outbreak alert news with elaborately designed questions, making it a valuable resource for epidemiologists and NLP researchers alike.

Introduction

In spite of advancements in healthcare, infectious disease outbreaks continue to pose a substantial threat to human health and well-being. To enhance disease surveillance and deepen our understanding of disease transmission, several surveillance systems have been established, including BioCaster (Meng et al. 2022), GPHIN (Mawudeku et al. 2013), ProMED-mail (Yu and Madoff 2004), HealthMap (Freifeld et al. 2008) and EIOS. These systems perform real-time surveillance and analysis of disease outbreaks by monitoring daily news alerts and social media platforms.

Despite the notable achievements of current surveillance systems, most of them concentrate on the detection of outbreak events using social media and news sources. However, there is a dearth of systems that offer automatic and comprehensive epidemiological analysis of corresponding alerts or news. For instance, epidemiologists require such systems that can identify cases where the disease is deliberately released or affects vulnerable populations, such as the elderly or children. The automatic identification of these cases can facilitate preventive measures and prompt rescue efforts. The limited capacity of existing systems can partly be attributed to the scarcity of well-annotated report data, which is critical for training machine learning systems for domain experts. Although several existing datasets (Torres Munguía et al. 2022; Carlson et al. 2023) have been annotated to extract outbreak events, they mostly focus on the statistics (e.g. location, disease names, etc.) of the outbreak event rather than providing a thorough epidemiological analysis for further investigation.

To enrich the capabilities of existing surveillance systems, we present a newly annotated dataset, namely the Biomedical Alert News Dataset (BAND)[1]. This dataset comprises 1,508 samples extracted from recently reported news articles, open emails, and alerts, accompanied by 30 epidemiology-related questions. These questions cover most

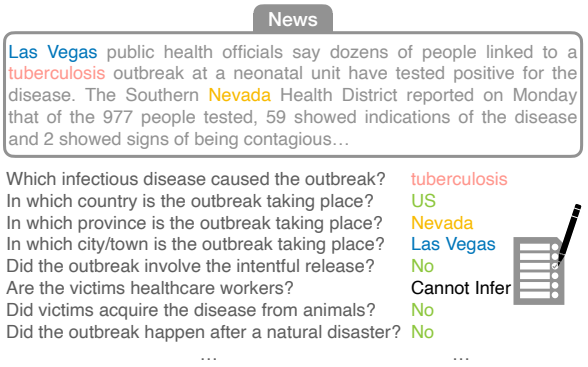

Figure 1: The BAND dataset consists of 30 disease outbreak-related questions designed by epidemiologists. Annotators are required to infer answers based on their understanding of the context when the answer is not explicitly provided in the context (e.g. the answers in green color).

[1] Our dataset and code are available at https://github.com/fuzihaofzh/BAND

of the event-related queries raised in Torres Munguía et al. (2022); Carlson et al. (2023) as well as more detailed inquiries regarding the outbreak event. For instance, we annotate whether an outbreak was an intentional release or involved a pregnant woman (refer to Table 1 for details of all the questions), which are important risk factors considered by human public health analysis. Affirmative responses to these questions serve as indicators for epidemiologists to prioritize and assess the need for further action. The selection of samples and questions is meticulously carried out by domain experts specializing in the fields of epidemiology and NLP. This dataset aims to empower NLP systems to analyze and address several critical questions, which aids the current surveillance systems in identifying significant trends and providing insights on how to improve disease surveillance and management.

This dataset presents new challenges for the NLP community, particularly in the area of common sense reasoning. For example, as illustrated in Figure 1, the system must automatically extract the outbreak country from the given news, even when it is not explicitly stated. This requires the model to infer the country name from context clues such as city name, state name, and report organization. In addition, the dataset requires better content disambiguation capabilities. For instance, when asked to identify the city of the outbreak, many cities worldwide share the same name, making it necessary to provide a geocode[2] to uniquely identify the location. Our datasets can be used to assess the capabilities of state-of-the-art models across a range of benchmark NLP tasks. In particular, we have performed experiments on three prominent tasks, including Question Answering (QA), Named Entity Recognition (NER), and Event Extraction (EE), to showcase the effectiveness of current models in addressing these tasks on this new dataset.

The contribution of this paper can be summarized as follows:

1) We introduce the BAND corpus, which is the largest corpus of well-annotated biomedical alert news with elaborately designed questions to the best of our knowledge.

2) We provide various model benchmarks for a range of NLP tasks, including Named Entity Recognition (NER), Question Answering (QA), and Event Extraction (EE).

3) We present a complete pipeline for annotating biomedical news data that can be leveraged for annotating similar datasets.

The BAND Corpus

The BAND corpus consists of 1,508 authentic biomedical alert news articles and 30 expert-generated questions aiming at enhancing understanding of disease outbreak events and identifying significant incidents requiring special attention. The alert news articles encompass a wide range of sources, including publicly available news articles, emails, and reports. The annotation process is outlined in Figure 2. Initially, epidemiology and NLP researchers select appropriate questions and samples from the alert news. Experienced annotators then conduct an ethics check to filter out unsuitable

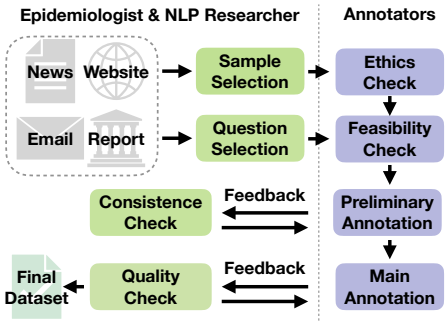

Figure 2: Our annotation workflow.

content and perform a feasibility check to ensure that they can annotate the selected questions. Should any questions arise, annotators consult with the experts for clarification. Subsequently, annotators engage in a preliminary annotation process by answering the selected questions for each sample. Following this, a consistency check is performed to ensure the same understanding among all annotators regarding the questions and samples. Finally, in the main annotation stage, annotators proceed to annotate all remaining samples. Multiple checks and feedback are incorporated throughout the annotation process to ensure a high-quality outcome.

Data Annotation

Question Selection. The question selection process involves the creation of pertinent natural language questions crucial for analyzing new articles within the realm of epidemiology, which were conducted collaboratively with esteemed experts possessing Ph.D. qualifications in the fields of epidemiology and public health. The expertise and knowledge of these professionals contribute to the meticulous design of the questions, which are subsequently categorized into three groups: event questions, epidemiology questions, and ethics questions. Event questions are designed to capture specific information such as the location and disease name, resembling similar disease outbreak events found in Torres Munguía et al. (2022) and Carlson et al. (2023). Epidemiology questions focus on detailed epidemiological information, such as whether the disease was intentionally released, which extends the questions in Conway et al. (2009) and Conway et al. (2010). Ethics questions require the annotators to conduct thorough ethical assessments to prevent potential privacy breaches or inappropriate content. For a detailed list of these questions, please refer to Table 1.

Samples Selection. We obtain our raw news alerts from ProMED-mail[3], a network of medical professionals known for delivering timely information on global disease outbreaks. ProMED-mail covers a wide range of diseases, including infectious diseases, foodborne illnesses, zoonotic diseases, etc. It offers detailed reports on outbreaks containing crucial information such as the number of cases, outbreak locations, associated symptoms, and etc. This information is invaluable for the development of effective strate-

[2] https://www.geonames.org/

[3] https://promedmail.org/

Questions	Short name	Category	Options	Sparse
1) Which infectious disease caused the outbreak?	Disease	Event	-	-
2) In which country is the outbreak taking place?	Country	Event	-	-
3) In which province is the outbreak taking place?	Province	Event	-	-
4) In which city/town is the outbreak taking place?	City	Event	-	-
5) Check and fill country Geo Code (e.g. 1794299):	Countrycode	Event	-	-
6) Check and fill province Geo Code (e.g. 1794299):	Provincecode	Event	-	-
7) Check and fill city Geo Code (e.g. 1815286):	Citycode	Event	-	-
8) Which virus or bacteria caused the outbreak?	Virus	Event	-	-
9) What symptoms were experienced by the infected victims?	Symptoms	Epidemiology	-	-
10) Which institution reported this outbreak?	Reporter	Epidemiology	-	-
11) What is the type of victims?	Victimtype	Epidemiology	Human/Animal/Plant	-
12) How many new infected cases are reported in the specific event in the report? (please input digits like 1, 34, etc.)	Casesnum	Epidemiology	-	-
13) Has the victim of the disease travelled across international borders?	Borders	Epidemiology	YES/NO/Cannot Infer	YES
14) Does the outbreak involve the intentful release?	Intentful	Epidemiology	YES/NO/Cannot Infer	YES
15) Did human victims acquire the infectious disease from an animal?	Fromanimal	Epidemiology	YES/NO/Cannot Infer/ Not Applicable	-
16) Did the victim fail to respond to a drug?	Faildrug	Epidemiology	YES/NO/Cannot Infer/ Not Applicable	-
17) Are healthcare workers included in the infected victims?	Healthcareworkers	Epidemiology	YES/NO/Cannot Infer	YES
18) Are animal workers included in the infected victims?	Animalworkers	Epidemiology	YES/NO/Cannot Infer	YES
19) Is the victim of the disease a military worker?	Militaryworkers	Epidemiology	YES/NO/Cannot Infer	YES
20) Did the outbreak involve a suspected contaminated blood product or vaccine?	Vaccine	Epidemiology	YES/NO/Cannot Infer	YES
21) Are the victims in a group in time and place?	Group	Epidemiology	YES/NO/Cannot Infer/ Not Applicable	-
22) Did the victim catch the disease during a hospital stay?	Hospitalstay	Epidemiology	YES/NO/Cannot Infer	YES
23) Is the victim of the disease a child?	Child	Epidemiology	YES/NO/Cannot Infer	-
24) Is the victim of the disease an elderly person?	Elderly	Epidemiology	YES/NO/Cannot Infer	-
25) Is the victim of the disease a pregnant woman?	Pregnant	Epidemiology	YES/NO/Cannot Infer	YES
26) Has the victim of the disease been in quarantine?	Quarantine	Epidemiology	YES/NO/Cannot Infer	YES
27) Did the outbreak take place during a major sporting or cultural event?	Event	Epidemiology	YES/NO/Cannot Infer	YES
28) Did the outbreak take place after a natural disaster?	Disaster	Epidemiology	YES/NO/Cannot Infer	YES
29) When did the outbreak happen? (Relative to article completion time)	Tense	Epidemiology	Past/Now/Not Yet	-
30) Does the text contain information that can uniquely identify individual people? e.g. names, email, phone, and credit card numbers, addresses, user names.	Sensitive	Ethics	YES/NO	-

Table 1: Epidemiology questions are given by experts in epidemiology.

gies aimed at controlling and preventing the spread of diseases. To conduct our research, we initially collect 36,788 raw alerts available on ProMED-mail, spanning from December 2009 to December 2021. Then, we engage experts with Ph.D. degrees in epidemiology and public health to generate a list of questions and filter samples for further annotation. Specifically, we carefully select 2,458 samples and request the experts to assign scores ranging from 1 to 5 to each sample. The distribution of scores is illustrated in Figure 4. Samples with scores exceeding 4 are chosen as candidate samples for further analysis. Additionally, it has been observed that certain questions, such as Question 14 (Does the outbreak involve intentful release?), have a sparse distribution of answers, as most diseases are not intentionally released. These questions are called "sparse questions" and are listed in Table 1. To ensure an adequate number of data points for these types of questions, the candidate sample set is ranked based on both the expert scoring and the keyword hits[4]. For instance, if a sample contains keywords like "intentful release", the sample will be given one extra point. In this way, samples with more keyword hits are prioritized. However, these kinds of samples are still not enough and a manual search is conducted on ProMED, Wikipedia, and media news platforms to identify relevant articles containing positive answers to these questions.

Annotation. To facilitate the annotation process, we develop a new annotation interface using LabelStudio[5] (see Fu et al. (2023) Appendix Figure 5 for the annotation interface). Subsequently, we employ a professional annotation company to undertake the annotation of the detailed questions. The annotation process is divided into four batches,

[4] A detailed list of keywords can be found in Fu et al. (2023) Appendix Table 7. [5] https://labelstud.io/

comprising 40, 710, 110, and 660 samples, respectively. After each stage, we manually review the annotations and provide feedback to address any systematic annotation issues that may arise.

Consistency Check. To ensure the annotation team delivers high-quality annotations, we conducted a quality check during the preliminary annotation. All five annotators were assigned to annotate the same set of 40 samples, and we manually reviewed the answers to identify any obvious mistakes. Additionally, we assessed the consistency of annotations by comparing the responses from all annotators[6].

Quality Check. To maintain the quality of annotations, we implement a quality check after the completion of each batch by the annotators. First, the annotators conduct a manual review of their annotation results to identify and rectify any typos or erroneous annotations. Subsequently, they submit the annotated batch to the experts, who provide feedback to address any misunderstandings that occur. This iterative feedback loop between the annotators and experts ensures ongoing refinement and enhancement of the quality.

Ethics Check. To ensure compliance with ethical requirements, we initiate a research ethics review and obtain permission from the faculty's research ethics committee prior to conducting the annotation process. During the annotation phase, we instruct the annotators to carefully assess whether the samples violate any ethical rules. Any samples found to be in violation were promptly removed from the corpus without further annotation. This proactive approach ensures that the annotation process adheres to ethical guidelines and maintains the integrity of the research.

Statistics

To gain a comprehensive understanding of the BAND dataset, we present various statistics that provide insights into the data's coverage and highlight its significant contributions to the field of NLP and epidemiology. These statistics effectively demonstrate the breadth and depth of the dataset, showcasing its value and potential impact.

Disease Distribution. The distribution of diseases in the BAND dataset is depicted through a histogram shown in Figure 3 (a). This histogram reveals that our dataset covers a wide range of popular infectious diseases, such as Anthrax, Cholera, and others. This extensive coverage underscores the dataset's potential for training models to effectively monitor and surveil various disease outbreaks.

Location Distribution. In addition, we have generated visualizations of the location distribution in the BAND dataset, highlighting the coverage of various countries (Figure 3 (b)), provinces (Figure 3 (c)), and cities (Figure 3 (d)). These visualizations demonstrate that our dataset encompasses a wide range of locations, affirming its potential for training a model capable of handling daily news reports from around the world. This global coverage further enhances the dataset's applicability in addressing diverse scenarios.

[6] The comparison results are presented in Fu et al. (2023) Appendix Table 8, which demonstrates high level of consistency among the annotators, and validates the qualification for them.

Pathogen Distribution. Our dataset exhibits a comprehensive coverage of various pathogens including bacteria, fungi, protozoa, viruses, and etc. The distribution is shown in Figure 3 (e). It is evident from the statistics that the BAND dataset encompasses mentions of numerous popular infectious pathogens, including bacillus anthracis, rabies virus, vibrio cholera, and many others. This extensive coverage of prominent pathogens enhances the dataset's relevance and suitability for training models to effectively analyze and respond to a wide range of infectious disease scenarios.

Victim Distribution. Within the BAND dataset, the term "victim" refers to the infected host type, which includes humans, animals, and plants. As depicted in Figure 3 (h), our primary focus is on human and animal diseases. However, we have also included a portion of the data (approximately 6%) that describes plant diseases, thereby extending the application domains of the dataset.

Symptoms Distribution. The BAND dataset includes annotations for a diverse range of symptoms, as illustrated in Figure 3 (f). Symptoms such as fever, vomiting, and others are comprehensively covered within the dataset. This broad coverage of symptoms highlights the potential to train a model capable of handling different types of symptoms.

Data Split

We provide two different sampled splits, namely the Rand Split and the Stratified Split, as shown in Table 2.

Rand Split. This split randomly partitions the corpus into train/dev/test sets, without considering any other factors.

Stratified Split. In order to assess the model's ability to accurately answer sparse questions with limited positive answers, it is crucial to focus on these specific samples in upcoming research. To accomplish this, we employ a split strategy that prioritizes samples with positive answers for sparse questions. These samples are divided in a ratio of 5:1:4 for the train/dev/test sets respectively. This ensures that these important samples are adequately trained and evaluated. Then, we randomly sample other instances to complement the dataset. This approach allows for a thorough assessment of the model's performance in addressing the challenges posed by sparse questions.

Experiments

To evaluate the performance of existing NLP models on our newly annotated dataset, we conduct experiments on three widely used NLP tasks: Named Entity Recognition (NER), Question Answering (QA), and Event Extraction (EE). By assessing the performance of various models on these tasks, we can gain insights into their strengths and limitations in handling this dataset.

Experimental Setup

NER Task The NER task aims at extracting named entities belonging to specific categories. In this study, we aim to demonstrate how our annotated biomedical dataset can help advance research in the NER task for specific terms. We focus on extracting entities related to disease names, outbreak

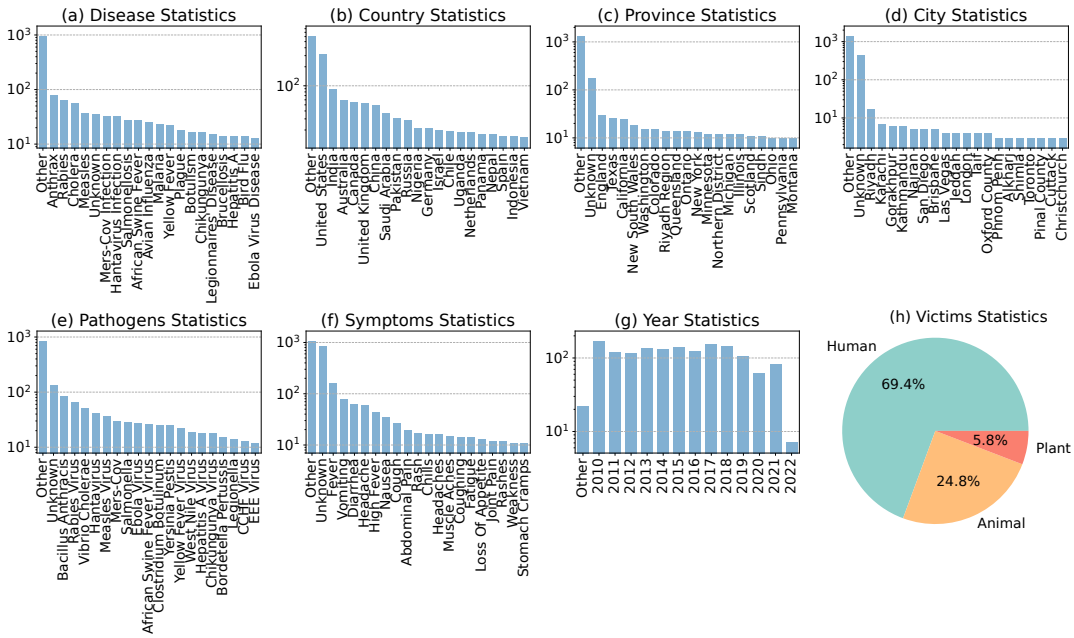

Figure 3: Statistics for BAND corpus.

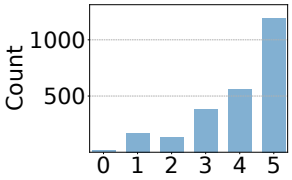

Figure 4: Experts' scores.

	Rand	Stratified
train	1,208	1,126
dev	150	149
test	150	233

Table 2: Data split.

locations (country/province/city), pathogens (viruses/bacteria), and symptoms. To ensure compliance with the requirements of the NER task, we limit our analysis to only those entities that are explicitly mentioned in the original text, without considering any additional inference by annotators. We compare the performance of various models, including CRFBased, TokenBased, SpanBased, and ChatGPT models in extracting. The model details are as follows:

CRFBased model (Lafferty, McCallum, and Pereira 2001; McCallum and Li 2003; Manning et al. 2014) incorporates the contextual information of nearby words to recognize and classify named entities in the provided text. It utilizes a Conditional Random Field (CRF) layer to predict the BIO tags for each input word.

TokenBased model (Kenton and Toutanova 2019; Lee et al. 2020) fine-tunes a pre-trained model on the annotated NER data and uses it to directly predict the BIO tags for each word in the given sequence.

SpanBased model (Lee et al. 2017; Luan et al. 2018, 2019; Wadden et al. 2019; Zhong and Chen 2021) partitions the input text into spans of varying lengths and then directly assigns labels to the spans that correspond to named entities. This technique has been demonstrated to improve the performance of previous NER models. We employ the implementation provided by Zhong and Chen (2021).

ChatGPT model (Ouyang et al. 2022) has shown great potential in performing this task without requiring additional training or fine-tuning. We elicit named entities with corresponding categories directly from the API. We have attempted various prompts to instruct ChatGPT to infer as much information as possible and adhere to the terminology mentioned in the original text.

QA Task The task of question answering involves providing answers to questions based on a given corpus. This task can be categorized as either extractive QA or abstractive QA. In extractive QA, the model selects the relevant answer span from the input text, while in abstractive QA, the model generates an answer based on the input text, which may not be an exact span from the given text. In our task, as the answer to many questions may not exist in the original text, we focus on the abstractive QA setting. We demonstrate the performance of several models to showcase their potential on our dataset. To evaluate the models' performance, we utilize the widely-used accuracy metric. Prior to comparing the models' results to the gold standard label, we normalize all occurrences of "N/A", "Unknown", "na", and "nan" as "Cannot Infer". We use exact match accuracy to evaluate the results. We conduct experiments with following models:

T5 (Raffel et al. 2020) model is built on the Transformer architecture and is pre-trained on large volumes of text data using a diverse range of language modeling tasks. We fine-tune the T5 model on the training set by concatenating the text and the question as input, with the output being the answer to the corresponding question.

Bart (Lewis et al. 2019) model is similar to the T5 model, as it also employs an encoder-decoder architecture. We use

	Random			Stratified		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
CRFBased	0.582	0.674	0.625	0.600	0.663	0.630
TokenBased	0.631	0.691	0.660	0.701	0.730	0.715
SpanBased	0.598	0.694	0.642	0.676	0.759	0.715
ChatGPT	0.326	0.353	0.339	0.424	0.318	0.363

Table 3: Named entity recognition results.

	Precision	**Recall**	**F1-score**
City	0.326	0.500	0.395
Country	0.710	0.760	0.734
Disease	0.583	0.758	0.659
Province	0.616	0.517	0.562
Virus	0.696	0.823	0.754

Table 4: NER results for each domain.

the Bart model as the backbone model and fine-tune it on our annotated dataset using the same setting as the T5 model.

GPT2 (Li and Liang 2021) is a decoder-only language model that concatenates all context sequences, questions, and answers into a single sequence, which is then used to fine-tune the GPT2 model.

GPTNEO (Black et al. 2022) is a transformer-based language model developed by EleutherAI, designed to be an open-source model similar to GPT-3. It was trained on the Pile dataset, which comprises a diverse corpus of text data, including books, websites, and academic papers.

OPT (Zhang et al. 2022) is a decoder-only language model developed by Meta AI, with the aim of providing an open-source model comparable to GPT-3. OPT offers models with parameters ranging from 125M to 175B. In our experiment, we adopt the model with 350M parameters.

Galactica (Taylor et al. 2022) is a decoder-only language model trained on a large-scale scientific corpus. It is designed to handle scientific tasks, including scientific QA, mathematical reasoning, summarization, and document generation. The model may have been trained with corresponding disease and country names, making it more likely to understand the news text in our dataset. Galactica comes in a range of model sizes, from 125M to 120B parameters, and we test the 125M model in our experiment as larger models tend to explain the answer with their own words instead of our pre-defined format, leading to a degenerated output.

BLOOM (Scao et al. 2022) is an autoregressive large language model that outputs coherent text in 46 languages and 13 programming languages. Additionally, it can complete diverse text tasks, even those it was not directly trained for, by framing them as text generation tasks.

ChatGPT model (Ouyang et al. 2022) is a zero-shot model unsuitable for fine-tuning with our data. It is used via its API as in the NER task. We prompt it to read a paragraph and answer questions sequentially, with instructions detailed in Fu et al. (2023) Appendix Figure 6.

Model	**Rand**	**Stratified**	**Size**	**Mode**
T5	0.674	0.591	220M (base)	Finetune
Bart	0.666	0.510	140M (base)	Finetune
GPT2	0.663	0.647	124M	Finetune
OPT	0.699	0.687	125M	Finetune
GPTNEO	0.695	0.695	125M	Finetune
Galactica	0.717	0.710	125M	Finetune
BLOOM	0.735	0.751	560M	Finetune
ChatGPT	0.497	0.413	-	Zero-Shot

Table 5: Question answering results.

EE Task The event extraction task focuses on identifying and extracting event-specific information from unstructured text. Detecting disease outbreak events is more challenging than general event extraction because of the diverse terminology used. Although triggers like "outbreak", "epidemic", or "pandemic" may be utilized, their absence can limit the effectiveness of traditional keyword-based approaches. In response, we defined a set of attributes for outbreak events: disease name, location, pathogens involved, victim type, and associated symptoms. Our task identifies and classifies entities linked to these attributes in a document, using autoregressive models similar to those in our QA baselines.

Experimental Results

NER Task. The results for NER task are shown in Table 3. The results indicate that: 1) The existing supervised models (CRFBased, TokenBased, SpanBased) achieve good performance than the zero-shot model (ChatGPT), which suggests that training the model with the BAND corpus can aid in identifying commonly used named entities in disease outbreak news. 2) ChatGPT does not perform as well as the supervised models. This could be due to several factors: firstly, our data is newly annotated and belongs to a highly specialized domain that ChatGPT may not have been extensively trained on. Additionally, ChatGPT prefers to use its own words to provide the name (which is usually more formal), leading to lower scores. We have attempted to utilize multiple instructions to encourage it to use the original text (as shown in Figure 6), but it remains unresponsive.

We also show the NER results for each domain in Table 4. The following observations can be made: 1) The NER model performs well in the country, disease, and virus domains. This is likely because the named entities in the testing set are also present in the training set, and the model has learned to recognize these types of entities. 2) In the province and city domains, F1 scores drop significantly due to unmentioned names in the training set, requiring models with improved few-shot/zero-shot abilities and introducing new challenges.

The QA Task. The results of the QA task are shown in Table 5. It can be observed from the results that 1) The decoder-only models, like GPT2 and Galactica, tend to outperform encoder-decoder ones, such as T5 and Bart, possibly due to the former's pre-training on more extensive text data. 2) The BLOOM model outperforms other generative models, which may be due to its training on corpora more relevant to our

Model	Overall F1	Individual F1									
		Disease	Country	Province	City	Country code	Province code	City code	Pathogen	Symptoms	Victim
T5	0.609	0.764	0.879	0.564	0.580	0.686	0.151	0.023	0.662	0.768	0.973
Bart	0.609	0.683	0.882	0.495	0.533	0.852	0.328	0.066	0.536	0.615	0.987
GPT2	0.548	0.719	0.807	0.405	0.427	0.757	0.113	0.019	0.570	0.532	0.932
OPT	0.589	0.724	0.826	0.466	0.435	0.820	0.324	0.073	0.619	0.522	0.966
GPTNEO	0.433	0.601	0.651	0.272	0.328	0.559	0.084	0.018	0.471	0.317	0.891
Galactica	0.560	0.722	0.824	0.458	0.477	0.807	0.266	0.044	0.652	0.523	0.740
Bloom	0.586	0.689	0.839	0.424	0.418	0.826	0.351	0.061	0.586	0.579	0.929
ChatGPT	0.477	0.562	0.792	0.473	0.462	0.516	0.073	0.044	0.280	0.450	0.835

Table 6: Event extraction results on random split.

domain. Additionally, finetuning a larger BLOOM model can force it to use our desired output style, while other larger backbone models tend to use their own words to answer. 3) We also attempt to utilize ChatGPT, but its performance was not as good as fine-tuned models. This might be due to ChatGPT being a zero-shot model with limited training on our new dataset, and it showed no inference capabilities despite various prompting attempts. In Appendix Error Analysis, we conduct detailed experiments to discuss these issues.

The EE Task. In Table 6, overall and individual F1 scores offer performance insights into various models on event extraction task. All models exhibit challenges with "province code" and "city code" extraction, highlighting the intricacy of context-dependent information extraction. In contrast to the QA task, encoder-decoder models like T5 and Bart excel over decoder-only models such as GPT2 and GPTNEO in tasks, particularly in extracting context attributes, possibly due to the absence of YES/NO questions in the EE task. Interestingly, OPT performs better in geocode prediction despite a lower F1 score in location prediction. This may be due to geocode-related documents in its pretraining data, showing how these models can use latent knowledge from the pretraining phase. ChatGPT, though less effective overall, achieves a notable score in the "city" category. This, coupled with its weaker zero-shot performance, indicates its robustness in specific context-dependent tasks.

Related Works

Numerous epidemiology disease surveillance systems have been developed to monitor disease outbreak events. Among these, the BioCaster system (Meng et al. 2022) automatically gathers news and alerts from social media, while GPHIN (Mawudeku et al. 2013) uses global surveillance and data analysis to detect potential public health threats. ProMED-mail (Yu and Madoff 2004) relies on a network of experts to provide real-time news alerts and expert notifications on emerging diseases and outbreaks. HealthMap (Freifeld et al. 2008) aggregates disease data from various sources to provide real-time disease outbreak monitoring and visualization. FluTrackers[7] offers real-time monitoring and analysis of influenza. The ECDPC[8] is utilized for real-time monitoring, risk assessment, and outbreak investigation of infectious diseases in Europe. However, these systems have only made use of simple NLP tools such as extraction tools, and further data is needed to train models with a deeper understanding of news articles and reports.

In order to enhance the performance of disease surveillance systems, several pandemic and epidemic datasets have been annotated. Conway et al. (2010) have annotated a disease outbreak dataset comprising 200 samples. Meanwhile, Torres Munguía et al. (2022) have provided a dataset that includes 2,227 samples; however, this dataset mainly concentrates on the outbreak event itself and does not contain report text annotations. Chan et al. (2010); Carlson et al. (2023) have utilized WHO's data to examine the outbreak event. Mutuvi et al. (2020b,a) have emphasized multilingual outbreak detection, whereas Balashankar et al. (2019) and Lamsal (2021) have centered on outbreak news for MERS and COVID-19, respectively. Nevertheless, these datasets are either relatively small or emphasize outbreak statistics ignoring other important information for epidemiologists.

Conclusions

In this paper, we contribute a new Biomedical Alert News Dataset (BAND) designed to provide a more comprehensive understanding of disease spread and epidemiology-related questions by enabling NLP systems to analyze and answer several important questions. Our dataset contains 1,508 samples from recent news articles, open emails, and alerts, as well as 30 event and epidemiology-related questions. The questions and samples are carefully selected by domain experts in the fields of epidemiology and NLP and require the common sense reasoning capability of NLP models. BAND is the largest corpus of well-annotated biomedical alert news with elaborately designed questions, and we provide a variety of model benchmarks for NER, QA and EE tasks in the epidemiology domain. The experimental results show the new dataset can help train NLP models to better understand outbreak and answer important epidemiology questions.

Acknowledgments

The authors gratefully acknowledge the support of the funding from UKRI under project code ES/T012277/1. We would also like to express our sincere gratitude to Anya Okhmatovskaia and Nicholas King for their invaluable assistance and insightful discussions.

[7] https://flutrackers.com/forum/

[8] https://www.ecdc.europa.eu/en

References

Balashankar, A.; Dugar, A.; Subramanian, L.; and Fraiberger, S. 2019. Reconstructing the MERS disease outbreak from news. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, 272–280.

Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Carlson, C. J.; Boyce, M. R.; Dunne, M.; Graeden, E.; Lin, J.; Abdellatif, Y. O.; Palys, M. A.; Pavez, M.; Phelan, A. L.; and Katz, R. 2023. The World Health Organization's Disease Outbreak News: A retrospective database. *PLOS Global Public Health*, 3(1): e0001083.

Chan, E. H.; Brewer, T. F.; Madoff, L. C.; Pollack, M. P.; Sonricker, A. L.; Keller, M.; Freifeld, C. C.; Blench, M.; Mawudeku, A.; and Brownstein, J. S. 2010. Global capacity for emerging infectious disease detection. *Proceedings of the National Academy of Sciences*, 107(50): 21701–21706.

Conway, M.; Doan, S.; Kawazoe, A.; and Collier, N. 2009. Classifying disease outbreak reports using n-grams and semantic features. *International journal of medical informatics*, 78(12): e47–e58.

Conway, M.; Kawazoe, A.; Chanlekha, H.; Collier, N.; et al. 2010. Developing a disease outbreak event corpus. *Journal of medical Internet research*, 12(3): e1323.

Freifeld, C. C.; Mandl, K. D.; Reis, B. Y.; and Brownstein, J. S. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association*, 15(2): 150–157.

Fu, Z.; Zhang, M.; Meng, Z.; Shen, Y.; Okhmatovskaia, A.; Buckeridge, D.; and Collier, N. 2023. BAND: Biomedical Alert News Dataset (Full Version with Appendix). *arXiv preprint arXiv:2305.14480*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Lafferty, J.; McCallum, A.; and Pereira, F. C. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lamsal, R. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *applied intelligence*, 51: 2790–2804.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234.

Lee, K.; He, L.; Lewis, M.; and Zettlemoyer, L. 2017. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 188–197.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597.

Luan, Y.; He, L.; Ostendorf, M.; and Hajishirzi, H. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3219–3232.

Luan, Y.; Wadden, D.; He, L.; Shah, A.; Ostendorf, M.; and Hajishirzi, H. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3036–3046.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Mawudeku, A.; Blench, M.; Boily, L.; St. John, R.; Andraghetti, R.; and Ruben, M. 2013. The global public health intelligence network. *Infectious disease surveillance*, 457–469.

McCallum, A.; and Li, W. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 188–191.

Meng, Z.; Okhmatovskaia, A.; Polleri, M.; Shen, Y.; Powell, G.; Fu, Z.; Ganser, I.; Zhang, M.; King, N. B.; Buckeridge, D.; et al. 2022. BioCaster in 2021: automatic disease outbreaks detection from global news media. *Bioinformatics*, 38(18): 4446–4448.

Mutuvi, S.; Boros, E.; Doucet, A.; Lejeune, G.; Jatowt, A.; and Odeo, M. 2020a. Multilingual epidemiological text classification: a comparative study. In *COLING, International Conference on Computational Linguistics*, 6172–6183.

Mutuvi, S.; Doucet, A.; Lejeune, G.; and Odeo, M. 2020b. A dataset for multi-lingual epidemiological event extraction. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 4139–4144.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text

Transformer. *Journal of Machine Learning Research*, 21: 1–67.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Torres Munguía, J. A.; Badarau, F. C.; Díaz Pavez, L. R.; Martínez-Zarzoso, I.; and Wacker, K. M. 2022. A global dataset of pandemic-and epidemic-prone disease outbreaks. *Scientific data*, 9(1): 683.

Wadden, D.; Wennberg, U.; Luan, Y.; and Hajishirzi, H. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5784–5789.

Yu, V. L.; and Madoff, L. C. 2004. ProMED-mail: an early warning system for emerging diseases. *Clinical infectious diseases*, 39(2): 227–232.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhong, Z.; and Chen, D. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 50–61.